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

Vincent Spinelli

Stage de fin d'études BIOCOMP / ENSAT
Encadrantes : Nathalie Vialaneix, Andrea Rau
Structure d'accueil : MIAT – INRAE Toulouse
février - juillet 2025

Table des matières

Chapitre 1

Introduction

1.1 Problématique

Les espèces domestiques d'élevage ont fait l'objet de fortes pressions de sélection visant à optimiser des caractères économiques, sanitaires ou de production. Si ces efforts ont permis des gains significatifs en performance, ils ont également entraîné une diminution de la diversité génétique exploitée, réduisant les leviers génétiques disponibles pour faire face aux pathogènes et aux stress environnementaux. Dans ce contexte, et face à l'émergence de nouveaux enjeux sociétaux liés au changement climatique et à l'accroissement de la population mondiale, il apparaît pertinent de caractériser et d'explorer cette diversité négligée, potentiellement porteuse de traits d'intérêt pouvant répondre à ces enjeux.

Depuis plusieurs dizaines d'années, un renouveau dans l'exploration du vivant est rendu possible par l'essor des approches multi-omiques et des technologies de séquençage à haut débit, qui permettent aujourd'hui d'accéder à des niveaux d'informations riches et variées (génotype, expression génique, régulation épigénétique, interactions métaboliques), le tout à moindre coût. Cette richesse ouvre la voie à une exploration et à une exploitation de la variabilité génétique et moléculaire. Toutefois, intégrer ces données massives dans un cadre analytique cohérent, comme les **études d'association pangénomiques de l'expression des gènes** (**eGWAS**), demeure un défi méthodologique et computationnel de taille. L'identification de **variants causaux** (**eQTLs**) via les études d'association est limitée par plusieurs contraintes, à la fois méthodologiques et biologiques.

Sur le plan statistique, le volume important de données générées à partir d'effectifs souvent modestes implique la réalisation d'un très grand nombre de tests, ce qui réduit la puissance des analyses et augmente le risque de faux positifs. Par ailleurs, le **déséquilibre de liaison** (en anglais, *linkage desequilibrium*, **LD**) complique la localisation précise des variants causaux, ceux-ci étant fréquemment confondus avec des variants proches et corrélés. D'autre part, la **structure génétique des populations**, ici assimilées à des races animales, introduit des effets intra et inter-races susceptibles de biaiser les signaux d'association.

D'un point de vue biologique, une étude préliminaire menée par Ko *et al.* [1] a mis en évidence l'existence de variants complexes dont l'effet associé varie selon les sous-populations. Ces variants, non modélisés explicitement par les approches classiques, pourraient conduire à des résultats incomplets, voire erronés.

Comment identifier de manière robuste les associations complexes entre variants ponctuels et expression génétique dans un contexte multi-populations, afin de mieux caractériser la diversité génétique des espèces agricoles ?

1.2 Objectifs du stage

L'objectif de ce stage est d'évaluer dans quelle mesure les modèles actuels peuvent identifier des associations complexes entre variants génétiques et expression génique dans un contexte multi-populations structuré. Pour cela, trois objectifs principaux ont été définis afin de répondre à cette problématique :

1. **Réaliser un état de l'art** des approches méthodologiques existantes pour les analyses eGWAS en contexte multi-populations. Cela inclut l'étude de **modèles linéaires mixtes global** et **intra-populations**, ainsi que des méthodes de **méta-analyse** permettant de combiner les résultats obtenus pour chaque population.
2. **Mettre en place un cadre de simulation réaliste**, à partir de données génomiques réelles issues de deux races porcines, afin de simuler des associations réalistes entre variants et gènes selon différents scénarios. Ces scénarios dépendent notamment du type d'effet inter-populations choisi (**effets identiques, différents, opposés**) ainsi que de la répartition des variants au sein des deux races.
3. **Évaluer la capacité des modèles à détecter des associations**, d'abord dans le cas de référence avec des effets identiques, puis dans le cas plus complexe des effets différents, voire opposés, entre races. Ces analyses devront permettre d'élucider les forces et faiblesses des modèles testés dans ce contexte génétiquement structuré.

1.3 Organisation du rapport

Ce rapport est structuré de manière à accompagner progressivement le lecteur, depuis le contexte général du stage à la mise en place de l'**étude de simulation**, jusqu'à l'analyse des résultats et les perspectives de recherche qui en découlent.

Le **Chapitre 1** introduit le contexte de l'étude, en présentant différentes problématiques liées aux études d'association, puis les objectifs du stage.

Le **Chapitre 2** décrit le contexte institutionnel du stage, avec une présentation de la structure d'accueil **MIAT**, du partenariat avec l'unité **GABI**, et du positionnement du stage dans le cadre du projet **AgroDiv**.

Le **Chapitre 3** définit le cadre théorique de l'étude. Il introduit les notions de génétique nécessaires à la bonne compréhension du rapport puis présente le principe des analyses d'association pangénomiques et des études de simulation.

Le **Chapitre 4** détaille la méthodologie employée. Il décrit les données utilisées, les étapes de pré-traitement réalisées, les outils mobilisés, la stratégie de simulation mise en œuvre ainsi que les critères retenus pour comparer les performances des méthodes.

Le **Chapitre 5** présente les résultats obtenus lors des simulations pour les différents modèles. Il présente d'abord une analyse préliminaire des principaux paramètres du système, avant de comparer les performances relatives des modèles dans le cadre de base, puis en fonction des effets complexes.

Le **Chapitre 6** présente les limites méthodologiques identifiées au cours du stage, et propose une ouverture vers une approche statistique complémentaire, non explorée dans l'étude mais prometteuse pour répondre aux problématiques évoqués.

Le **Chapitre 7** conclut le rapport en dressant une synthèse générale du travail effectué et des résultats obtenus, en exposant les perspectives qui en découlent.

Enfin, la partie **Annexes** contient les bilans personnels rédigés à l'issue du stage ainsi que les références bibliographiques.

Chapitre 2

Contexte et structure d'accueil

2.1 Présentation d'INRAE, de l'UR MIAT et de l'UMR GABI

INRAE, ou **Institut national de recherche pour l'agriculture, l'alimentation et l'environnement**, est un établissement public à caractère scientifique, placé sous la tutelle conjointe du ministère de l'Enseignement supérieur et de la Recherche, ainsi que du ministère de l'Agriculture et de la Souveraineté alimentaire. Il est né en 2020 de la fusion de deux organismes : **l'Institut national de la recherche agronomique** (**INRA**) et **l'Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture** (**IRSTEA**).

INRAE a pour mission de produire et de diffuser des connaissances scientifiques afin de répondre aux grands enjeux de société dans les domaines de l'agriculture, de l'alimentation, de l'environnement et de la gestion durable des ressources naturelles. Avec ses 14 départements scientifiques répartis dans 18 centres régionaux, il constitue aujourd'hui le premier organisme de recherche français dans ces domaines.

Ce stage de fin d'études a été réalisé au sein de l'unité de recherche (UR) **Mathématiques et Informatique Appliquées de Toulouse (MIAT)**, située au centre INRAE Occitanie-Toulouse. L'unité MIAT développe des méthodes statistiques, bio-informatiques et mathématiques appliquées à la biologie et à l'agronomie. Mon environnement de travail au MIAT m'a permis de côtoyer des ingénieurs de recherche, des doctorants, des post-doctorants, des contractuels ainsi que de nombreux stagiaires. Cette immersion m'a permis de découvrir le fonctionnement quotidien d'un laboratoire de recherche publique, tant sur le plan scientifique qu'humain.

Ce stage s'est également déroulé en collaboration étroite avec l'unité mixte de recherche (UMR) **Génétique Animale et Biologie Intégrative** (**GABI**), basée au centre INRAE de Jouy-en-Josas, spécialisée dans la génétique animale, la génomique fonctionnelle et la biologie des systèmes.

Le stage a été co-encadré par Nathalie Vialaneix, directrice de recherche à MIAT, et Andrea Rau, directrice de recherche à GABI.

2.2 Le PEPR AgroEcoNum et le projet AgroDiv

Ce stage s'inscrit dans le projet national **AgroDiv**, porté par le **Programme et Équipements Prioritaires de Recherche** (PEPR) **Agroécologie et Numérique** (**AgroEcoNum**), une initiative lancée en 2023 dans le cadre de **France 2030**. Doté d'un budget de 65 millions d'euros sur huit ans et copiloté par INRAE et l'Inria, ce programme national vise à accélérer la transition agroécologique par le biais des technologies numériques, en réponse aux enjeux liés à la sécurité alimentaire, au changement climatique et à la durabilité des systèmes agricoles.

Le PEPR s'organise autour de plusieurs axes de recherche stratégiques : étude de l'impact des politiques publiques et des technologies sur les pratiques agricoles, **caractérisation des ressources génétiques au service de l'agroécologie**, développement d'agro-équipements innovants (robotique, agriculture de précision), et conception d'outils numériques pour l'analyse et l'aide à la décision. En soutenant à la fois des travaux théoriques et appliqués, il vise à positionner la France comme un leader de l'agriculture numérique durable.

Dans ce cadre, le projet AgroDiv a pour objectif de mieux caractériser et valoriser la diversité génétique et fonctionnelle des principales espèces agricoles. Il s'agit notamment de renforcer l'adaptation des plantes et animaux domestiques aux contraintes environnementales croissantes, en soutenant des pratiques de sélection plus efficaces et compatibles avec les principes de l'agroécologie.

Le stage s'insère dans le volet **WP3** du projet, dédié au développement de méthodes statistiques et bio-informatiques avancées pour caractériser la diversité inter-espèces. Identifier des associations complexes entre variations génétiques et expression génique contribue en effet à mieux caractériser et valoriser la diversité génétique dans les programmes de sélection, et de fait, à potentiellement renforcer la résilience des systèmes d'élevage face aux changements environnementaux.

Chapitre 3

Cadre théorique et état de l'art

3.1 Rappel de génétique

L'ensemble du vivant repose sur une information codée au sein de la molécule d'**ADN** (**acide désoxyribonucléique**), support de l'hérédité chez la grande majorité des organismes. L'ADN est composé de longues chaînes de quatre **bases azotées** (**adénine, thymine, cytosine et guanine**), dont l'ordre au sein des deux hélices détermine l'information génétique (figure 3.1). Les **gènes** sont des segments d'ADN jouant un rôle fonctionnel essentiel. Ils se composent d'une séquence codante, contenant l'information nécessaire à la production des molécules biologiques (surtout des **protéines**), ainsi que de régions régulatrices situées en amont et en aval qui contrôlent l'expression de cette information. La molécule d'ADN est capable de se **répliquer** à l'identique à chaque division cellulaire, assurant ainsi la transmission de l'information génétique de la cellule mère aux cellules filles.

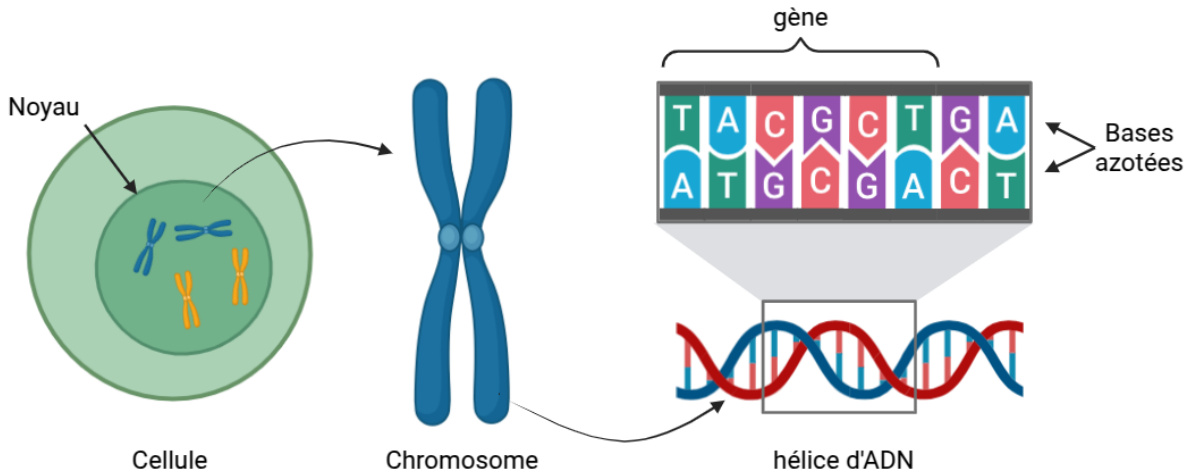


FIGURE 3.1 – **Structure de la molécule d'ADN.**

L'ensemble du matériel génétique d'un individu forme son **génome**, composé des régions codantes (gènes) et non codantes (régions régulatrices, introns, séquences répétées, etc.). La lecture et la régulation de ce génome déterminent le développement, le fonctionnement et l'identité des organismes.

L'**expression génique** désigne l'ensemble des mécanismes par lesquels l'information contenue dans un gène est utilisée pour produire une molécule fonctionnelle. Ce processus comprend deux étapes principales (figure 3.2). La première est la **transcription**, au cours de laquelle la séquence du gène est copiée en une molécule d'**ARNm** (**Acide ribonucléique messager**). Cette molécule, complémentaire à l'ADN, transporte l'information génétique depuis le noyau vers le cytoplasme. La seconde étape de l'expression génique est la **traduction** : l'ARNm est alors lu par un ribosome, un élément cellulaire qui associe à chaque triplet de bases azotés (appelé codon), un **acide aminé** correspondant. Ces acides aminés sont ensuite assemblés pour former une chaîne, qui se repliera en une protéine fonctionnelle.

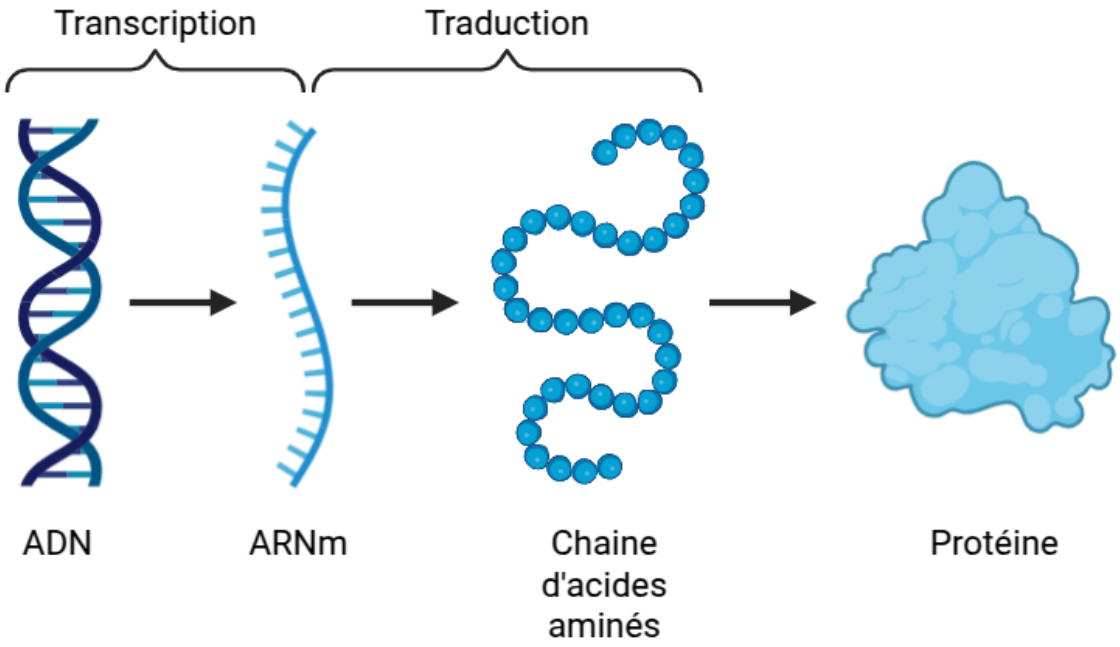


FIGURE 3.2 – **Étapes principales de l'expression génique**

Des modifications ponctuelles ou structurales, appelées **mutations** ou **variants**, peuvent survenir au sein d'un brin d'ADN. Elles résultent d'erreurs lors de la réplication de l'ADN, de mécanismes de recombinaison ou encore de l'exposition à des agents mutagènes. Selon leur nature et leur localisation dans le génome, ces mutations peuvent être neutres, délétères ou bénéfiques. Chez l'humain, on estime qu'un individu porte en moyenne 4 à 5 millions de variants répartis dans son génome [2].

La majorité des études scientifiques sur les variants se concentrent sur les **polymorphismes génétiques ponctuels** (en anglais, *single nucleotide polymorphisms*, **SNP**), qui correspondent à la substitution d'une base azotée par une autre à une position donnée du génome (figure 3.3). Un second type de mutations très étudié, les indels, correspond à l'insertion ou à la délétion de courtes séquences d'ADN.

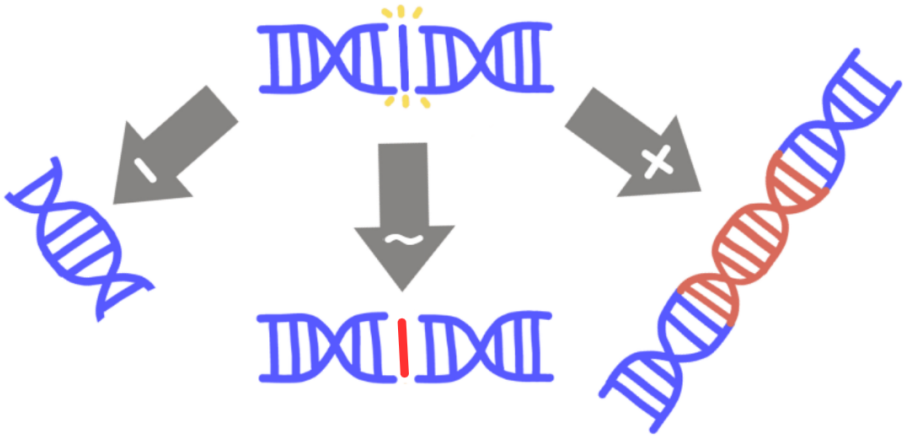


FIGURE 3.3 – **Principaux types de mutations génétiques**. De gauche à droite : délétion, substitution et insertion.

Les différents variants associés à une position donnée sont appelés **allèles**, et sont généralement caractérisés par leur **fréquence allélique** dans la population étudiée (en anglais, *Variant Allele Frequency*, **VAF**).

3.2 Études GWAS et eGWAS

Les études pangénomiques (en anglais, *Genome-Wide Association Studies*, **GWAS**) sont des études d'association visant à identifier des variants génétiques liés à des **caractères phénotypiques** qualitatifs ou quantitatifs, comme la résistance à une maladie ou la qualité de la viande. La méthode repose sur l'analyse statistique de la corrélation entre les variants génétiques et le phénotype étudié, issus d'une population d'individus génotypés (figure 3.4). Pour chaque variant, un test d'association est réalisé (souvent un test de régression linéaire), produisant une p-valeur qui mesure la probabilité d'observer une association au moins aussi extrême sous l'hypothèse nulle d'absence de lien. Afin de limiter les faux positifs produits par le très grand nombre de tests effectués, des corrections pour tests multiples sont appliquées, comme le seuil de **Bonferroni**.

Les GWAS permettent ainsi d'identifier des variants, ou plus précisément des régions du génome (cf. sous-section 6.2) associées à des traits complexes, sans toutefois en expliciter directement les mécanismes biologiques sous-jacents.

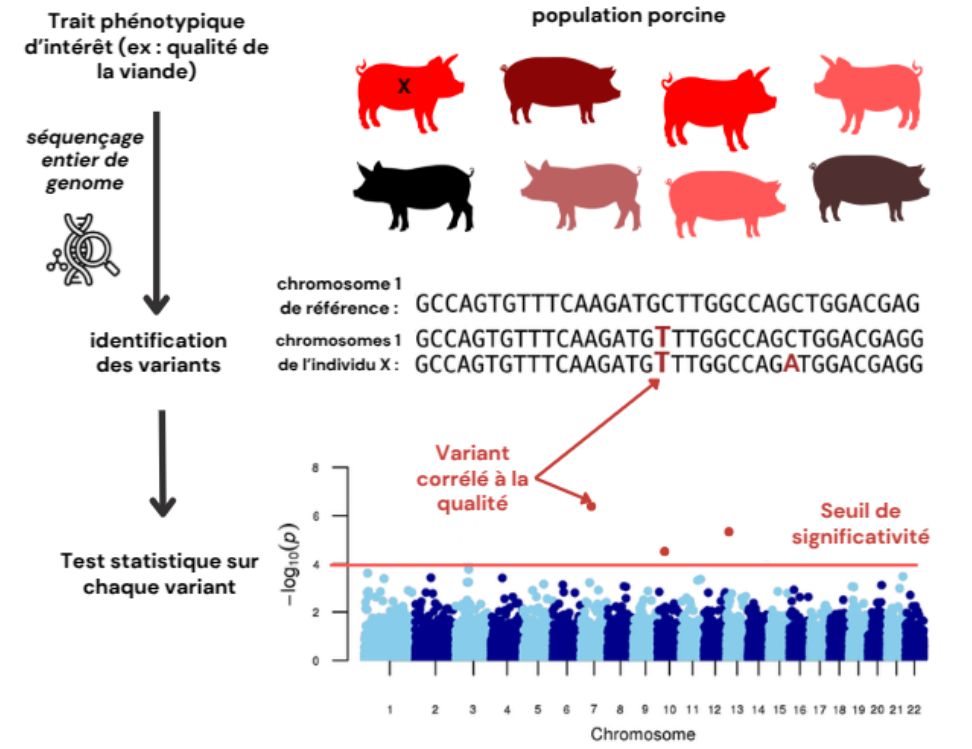


FIGURE 3.4 – **Schéma explicatif du principe de GWAS**, illustré en bas de figure par un *Manhattan plot* représentant l'association génétique entre les variants et le caractère d'intérêt. Chaque point correspond à un variant, situé en fonction de sa position génomique (axe des abscisses) et du degré de significativité de son association (axe des ordonnées, $-\log_{10}(p\text{-value})$). La ligne rouge indique le seuil de significativité statistique après correction pour tests multiples (seuil de Bonferroni). Les variants au-dessus de cette ligne sont considérés comme significativement associés au phénotype étudié.

Les analyses eGWAS (*expression GWAS*), aussi appelées études eQTLs, pour *expression Quantitative Trait Loci*, appliquent cette méthodologie à des phénotypes moléculaires, en prenant comme variable d'intérêt le niveau d'expression de l'ARNm de plusieurs milliers de gènes. Elles visent à détecter des variants ou zones du génome régulant l'expression de gènes d'intérêt. On distingue les **cis-eQTLs**, localisés à proximité du gène cible, et les **trans-eQTLs**, situés au-delà d'un seuil voire sur d'autres chromosomes. L'eGWAS se démarque des GWAS traditionnels en apportant une dimension fonctionnelle : en intégrant l'expression génique, elle permet d'explorer comment certains variants influencent directement l'activité des gènes en fonction des zones du génome impactées. Cette approche s'inscrit dans la stratégie d'intégration multi-omique, combinant différentes couches d'information (génomique et transcriptomique) pour mieux décrypter les mécanismes moléculaires qui régulent les traits complexes.

Toutefois, cette méthodologie implique un nombre de tests beaucoup plus élevé que dans les GWAS classiques, car chaque variant est testé pour son association avec l'expression des milliers de gènes. Cela augmente fortement le risque de faux positifs, et nécessite donc des corrections plus strictes pour le contrôle du taux d'erreurs. En conséquence, la puissance statistique peut être réduite, notamment pour les effets de faible amplitude qui sont généralement plus difficiles à détecter. La qualité des résultats dépend de la densité du génotypage, de la taille de l'échantillon et de la prise en compte de la structure de population.

3.3 Études de simulation appliquées aux eGWAS

Les **études de simulation** jouent un rôle central dans le développement, l'évaluation et la validation des modèles statistiques. Elles consistent à générer des jeux de données artificiels mais réalistes, en contrôlant explicitement la structure des données et en générant artificiellement les éléments que les modèles doivent retrouver. Ces jeux de données dans lesquels la "vérité" est connue, permettent de tester précisément la performance des modèles.

Dans le cadre d'une étude de simulation eGWAS, la première étape consiste à générer les données génotypiques des variants en s'appuyant sur deux approches principales. La première repose sur l'utilisation de génotypes réels, issus par exemple de données de séquençage, permettant de conserver des **structures de dépendance** réalistes entre individus (liées à la parenté et à la stratification génétique), ainsi qu'entre variants (déséquilibre de liaison). La seconde option consiste à générer des données génomiques entièrement simulées, en cherchant à reproduire au mieux les caractéristiques de données réelles.

Des effets génétiques sont ensuite attribués à un sous-ensemble de variants, les **eQTLs**, que les méthodes devront identifier. Ces effets peuvent être définis selon plusieurs paramètres : effet constant ou variable entre sous-populations, distribution des effets (gaussienne, bimodale, etc.), ou encore fréquence allélique spécifique (variants rares, fréquents, ou spécifiques à une race).

Une fois les effets assignés, les niveaux d'expression génique sont simulés pour chaque individu à partir d'un modèle reliant les variants à leurs génotypes (qui peut être additif, multiplicatif ou intégrer des interactions). L'ajout d'un **bruit aléatoire**, généralement gaussien, modélise les sources de variation résiduelles (techniques, biologiques, environnementales) que la part génétique n'explique pas.

En disposant de la vérité biologique (la position des variants causaux, la nature de leurs effets simulés, ainsi que les valeurs exactes des effets), il est possible d'évaluer finement les performances des méthodes d'analyse selon les différents scénarios établis. L'évaluation passe par le calcul de diverses métriques évoquée en sous-section 4.4.8.

Chapitre 4

Méthodologie

4.1 Environnement de travail

Les scripts développés dans le cadre de ce projet sont rédigés en langage **R** (version 4.4.3), sous forme de fichiers `R Markdown` et de scripts `R` classiques. L'environnement de travail est géré avec la librairie `renv`, assurant la reproductibilité des analyses en figeant les versions des librairies utilisées.

Les scripts de l'étude de simulation ont été conçus de manière à ce que toute personne disposant des données de départ puisse exécuter l'ensemble du processus en quelques minutes, sur n'importe quelle machine. La gestion et le suivi du projet sont facilités par l'utilisation de `Git`, un système de gestion de versions permettant un suivi rigoureux des modifications, ainsi qu'un partage structuré des fichiers entre les membres du projet.

L'ensemble des analyses a été réalisé en local, sur un poste `Linux` mis à disposition par INRAE. Les simulations et calculs effectués sont restés modérés en termes de ressources. Le poste INRAE, équipé d'un SSD, de 32 Go de RAM et d'un processeur `Intel Core i7-12700` (12 cœurs physiques, 20 threads) a suffi pour l'ensemble des traitements, qui ne dépassaient rarement plus de quelques minutes de calcul.

4.2 Données utilisées (GENE-SWitCH)

Le projet **GENE-SWitCH** (*The regulatory GENomE of SWine and CHicken*) est un projet financé par l'union européenne visant à annoter fonctionnellement les génomes du porc et du poulet. Il s'inscrit plus généralement dans l'initiative **FAANG** (*Functional Annotation of Animal Genomes*) qui a pour but de caractériser les éléments fonctionnels (dont les éléments régulateurs) des génomes, dans une perspective de production animale durable.

Dans ce cadre, une étude de Crespo-Piazuelo *et al.* [3] s'est appuyée sur un panel de 300 porcs issus de trois races commerciales (*Duroc*, *Landrace* et *Large White*), en combinant des données de séquençage du génome entier (en anglais, *Whole Genome Sequencing* **WGS**) et de l'ARN (en anglais, **RNA-seq**) dans trois tissus cibles : foie, duodénum et muscle squelettique. L'étude a permis de quantifier l'expression de plusieurs milliers de gènes et d'identifier plus de 44 millions de variants génétiques. Les données de séquençage WGS ont constitué la base des jeux de génotypes utilisés dans notre étude de simulation, permettant une représentation réaliste de la variabilité génétique d'un contexte porcin multi-populations.

Des analyses eGWAS réalisées sur l'ensemble des 3 races ont mis en évidence plus de 14 millions d'associations significatives, révélant des signaux *cis* et *trans*, ainsi que des **régions de régulation majeures** (en anglais, *hotspots*) impactant plus d'une dizaine d'expressions de gènes différentes. Il est important de noter que ces associations correspondent à celles partagées par les trois races. Cette étude n'a en effet pas pris en compte les scénarios plus complexes évoqués dans l'introduction, tels que des associations spécifiques à une seule race ou des effets différents entre les races.

4.2.1 Pré-traitement des données

Le travail a débuté par une phase d'exploration et de familiarisation avec les données issues de Crespo-Piazuelo *et al.*. Deux types de données ont été mobilisées :

1. les génotypes des variants génétiques, obtenus à partir des données WGS. Le jeu de données contient plusieurs dizaines de millions de variants, filtrés selon deux critères : une fréquence allélique de variant (**VAF**) supérieure à 5 % et un taux de génotypes manquants sur l'ensemble des individus inférieur à 10 % ;
2. les **méta-données** associées aux porcs, incluant notamment la race, le sexe, et l'identifiant de chaque individu.

Les données de génotypage et les méta-données ont été fournies au format `PLINK`, un standard en génétique des populations. Ce format repose sur trois fichiers complémentaires :

1. `.bed` : un fichier binaire contenant les génotypes des variants pour chaque individu, codés sous forme numérique : 0 pour un homozygote référence (deux allèles identiques à la référence), 1 pour un hétérozygote (un allèle de référence et un allèle alternatif), et 2 pour un homozygote alternatif (deux allèles alternatifs) ;
2. `.bim` : un fichier texte décrivant les caractéristiques des variants, incluant le chromosome associé, la position physique et les allèles ;
3. `.fam` : un fichier texte regroupant les informations relatives aux individus (identifiant, sexe, statut phénotypique).

Les variants ont été convertis au format `VCF` (en anglais, *Variant Call Format*) à l'aide de `PLINK`, puis une matrice binaire représentant les génotypes des variants de chaque individu a été extraite dans `R` à partir du fichier `VCF` pour la suite des analyses.

Une première étape de validation a consisté à vérifier la concordance des identifiants entre les fichiers génotypiques et les métadonnées, à identifier d'éventuelles valeurs manquantes ou aberrantes, et à uniformiser les noms des différents éléments afin de faciliter leur traitement avec `R`.

4.2.2 Exploration des données

Des analyses descriptives des données de génotypage ont ensuite été menées pour caractériser la diversité génétique entre races, notamment via l'analyse des fréquences alléliques, et la réalisation d'**analyses en composantes principales** (**ACP**).

L'analyse des fréquences alléliques, présentée sur la figure 4.1, a mis en évidence une forte prédominance des variants à faible fréquence : plus de la moitié d'entre eux présentent une VAF inférieure à 25%. Cela signifie qu'en moyenne, un variant donné est porté par un quart de la population ou moins. L'absence de variants dont la fréquence est inférieure à 5% (VAF < 0.05) est due à la phase de filtration effectuée en amont.

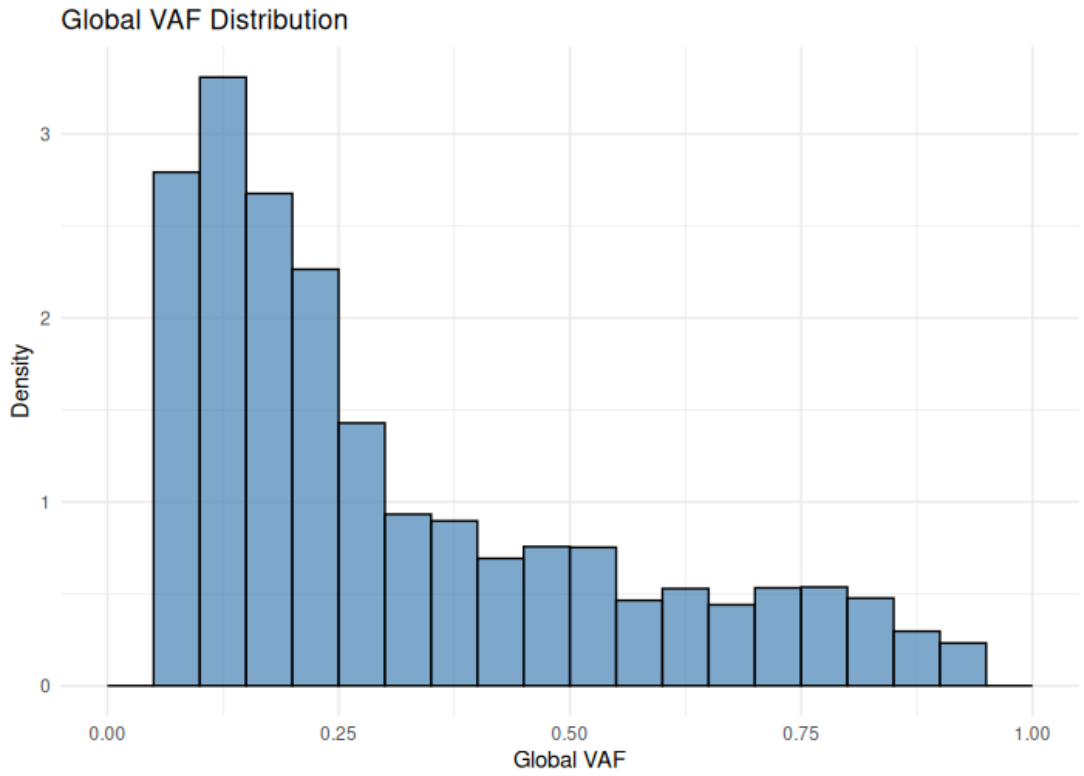


FIGURE 4.1 – **Répartition des fréquences alléliques des variants (VAF) dans l'ensemble des individus.** L'axe des abscisses représente la fréquence de l'allèle alternatif, tandis que l'axe des ordonnées correspond à la densité des variants observés à chaque fréquence.

L'analyse des distributions spécifiques par race a montré des profils globalement similaires entre les deux races, avec la race *Landrace* présentant un pic plus marqué de variants à faible VAF que les deux autres.

L'analyse en composantes principales (**ACP**) est une méthode de réduction de dimension appliquée ici aux génotypes de variants. Elle permet de résumer la variabilité entre individus à l'aide de quelques axes principaux, appelés **composantes principales** (**CPs**), qui capturent les tendances majeures de différenciation génétique. Les CPs sont des combinaisons linéaires des variables (ici le codage des variants) qui maximisent la variance entre les individus. La projection des individus sur les 2 premiers axes de l'ACP (figure 4.2) met en évidence une répartition nette des individus en fonction de leur race. On observe que les individus *Large White* et *Duroc* présentent une plus grande homogénéité génétique, tandis que les individus *Landrace* montrent une variabilité plus marquée dans leurs profils génétiques.

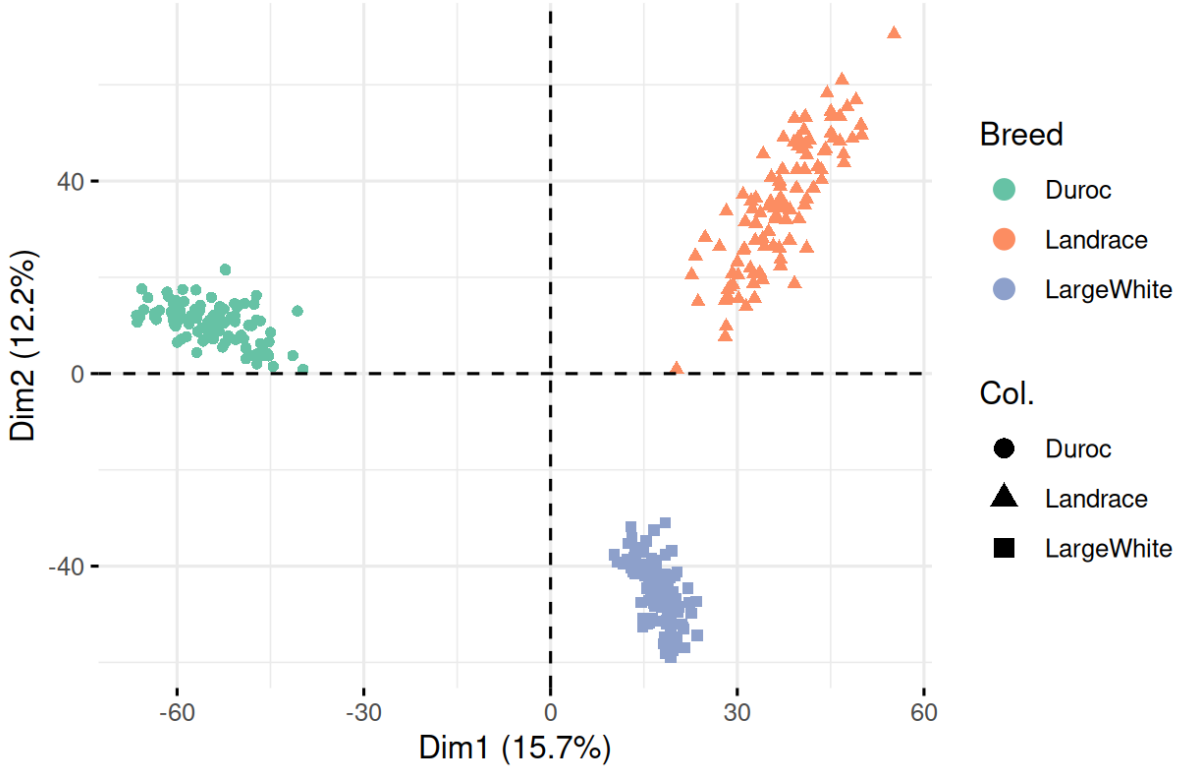


FIGURE 4.2 – **Projection des individus sur les deux premiers axes de l'analyse en composantes principales des génotypes des individus des 3 races.** Chaque point représente un individu.

Pour simplifier l'analyse inter-races, nous avons choisi de nous concentrer uniquement sur deux races : les *Landrace* et les *Large White*, en raison de leur proximité génétique plus importante (figure 4.2). L'étude de simulation a cependant été structurée de manière à pouvoir changer de combinaison de races en fonction de l'objectif recherché.

4.3 Modèles eGWAS utilisés en analyse multi-population

Avant de présenter en détail la stratégie de simulation, il est important de décrire les modèles statistiques couramment utilisés en eGWAS et testés dans notre étude afin de mieux comprendre les choix de scénarios retenus. Pour faciliter la compréhension des différents modèles, les formulations mathématiques présentées dans cette partie se limiteront à l'étude de l'expression d'un seul gène.

4.3.1 Modèle linéaire global

Bien qu'il soit généralement peu performant dans les contextes de populations structurées, le **modèle linéaire simple** mérite d'être introduit car il constitue la base de modèles plus complexes utilisés. Ce modèle est dit *global* ou *inter-races* car il considère l'ensemble des individus, sans distinction de race. Sa formulation mathématique s'appuie sur la régression linéaire classique, ici sans ordonnée à l'origine μ car chaque phénotype Y et chaque génotype G_j est centré :

$$Y_i = \beta_j G_{ij} + \varepsilon_i$$

où :

- Y_i est le niveau d'expression du gène pour l'individu i ;
- G_{ij} représente le génotype de l'individu i pour le variant j, codé par 0, 1 ou 2 selon le nombre d'allèles alternatifs présents chez l'individu. Ces modèles sont dits **additifs** et reposent sur l'hypothèse qu'un individu homozygote pour l'allèle alternatif ($G_{ij} = 2$) exprimera un effet β_j doublé par rapport à un individu hétérozygote ($G_{ij} = 1$).
- β_j représente l'effet du variant j sur l'expression du gène. Un β_j positif se traduit par une augmentation de l'expression en présence du variant, tandis qu'un β_j négatif indique une diminution.
- ε_i est un terme d'erreur aléatoire, supposé suivre une loi normale $\mathcal{N}(0, \sigma^2)$.

Ce modèle est ajusté indépendamment pour chaque couple *(variant, gène)* à l'aide de la méthode des moindres carrés. Il permet de tester l'existence d'une relation linéaire significative entre le génotype d'un variant et l'expression d'un gène donné, via l'hypothèse nulle $H_0 : \beta_j = 0$.

4.3.2 Structuration des populations

Dans les études d'association génétique, il est courant que la population soit structurée en sous-populations génétiquement différenciées, telles que des races animales ou des écotypes végétaux. Cette structuration peut également refléter les possibles liens de parenté entre individus. Elle est particulièrement marquée chez les espèces domestiques, sélectionnées pour des phénotypes spécifiques et issus de parents proches. Cette sélection a conduit à réduire la diversité génétique au sein des groupes et a isoler les races entre elles. Ces deux phénomènes sont illustrés sur la figure 4.3, où la distance entre deux individus reflète le degré de similarité entre leurs génomes.

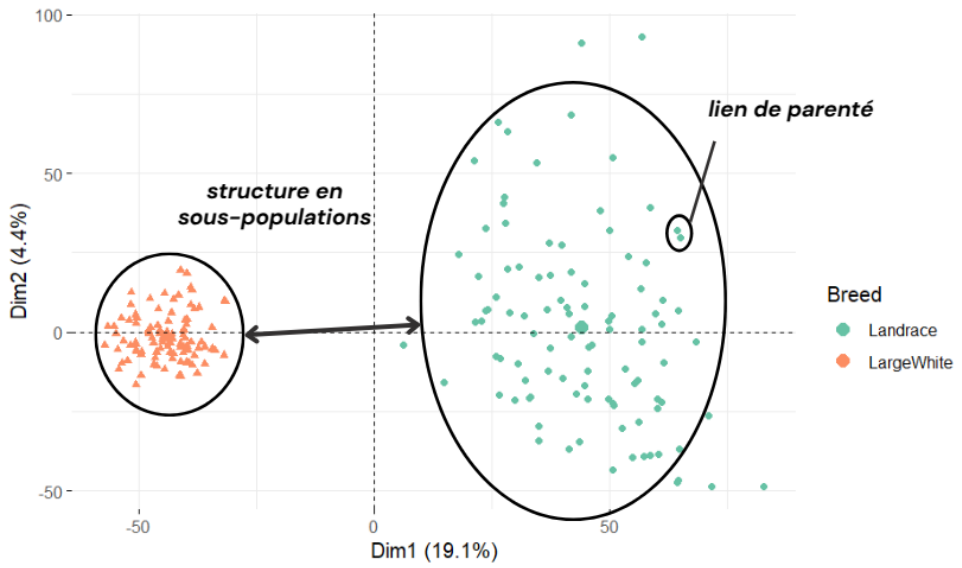


FIGURE 4.3 – **ACP des génotypes des races Landrace et Large White.**

Dans ce contexte, les individus ne sont pas **indépendants et identiquement distribués**. Cette structuration génétique, lorsqu'elle n'est pas correctement prise en compte, peut largement biaiser les résultats de l'eGWAS. Le choix du modèle statistique devient alors déterminant : ignorer ces relations peut conduire à une inflation du taux de faux positifs ou, à l'inverse, entraîner une perte de puissance dans la détection d'associations réelles. Par exemple, une fausse association peut apparaître simplement car un allèle est plus fréquent dans une race que dans une autre, sans lien causal avec le phénotype étudié.

Cette problématique a été illustrée dans l'étude préliminaire de Ko *et al.* [1] sur les données porcines de Crespo-Piazuelo *et al.* [3] considérées dans ce rapport, où l'utilisation d'un modèle linéaire simple a généré un nombre très élevé de faux positifs.

4.3.3 Modèle linéaire mixte global (approximation)

Comme le modèle linéaire global, le **modèle linéaire mixte global** considère l'ensemble des individus toutes races confondues. Afin de prendre en compte la structure génétique sous-jacente présentée en section 4.3.2, ce modèle intègre un effet aléatoire basé sur une **matrice d'apparentement K** qui quantifie la similarité génétique entre les individus à partir de leurs génotypes. On parle alors de **modèle linéaire mixte** (en anglais, *Linear Mixed Models*, **LMM**) car le modèle combine des effets fixes (par exemple, l'effet β du variant) et des effets aléatoires (ici, l'effet structuré selon la matrice d'apparentement). Le modèle s'écrit :

$$Y_i = \beta_j G_{ij} + u_i + \varepsilon_i$$

où u_i désigne l'effet aléatoire associé à l'individu i, issu du vecteur $u \sim \mathcal{N}(0, \sigma^2 \mathbf{K})$, de dimension n, modélisant les effets aléatoires individuels. Ce vecteur permet de capturer la structure génétique entre individus, $\mathbf{K}$ représentant la matrice d'apparentement de taille $n \times n$. Chaque coefficient $\mathbf{K}_{ij}$ reflète le degré de similarité génétique entre les individus i et j : plus il est élevé, plus les individus partagent des variants en commun. Dans notre cas, $\mathbf{K}$ est calculé comme suit :

$$\mathbf{K} = \frac{1}{p}\mathbf{Z}\mathbf{Z}^T$$

avec $\mathbf{Z}$ la matrice des génotypes centrés réduits, et p le nombre de SNPs.

Cette approche est particulièrement pertinente pour identifier des associations communes à plusieurs races, tout en corrigeant les effets de population. Cependant, elle peut être coûteuse en ressources de calcul, notamment lors du calcul de la matrice $\mathbf{K}$. Pour pallier ces limitations, nous avons opté dans notre étude pour une alternative plus légère, permettant d'approximer un modèle linéaire mixte à un faible coût computationnel. Cette approche, largement répandue dans les études d'association génétique [4], consiste à intégrer au modèle linéaire les premières composantes principales issues de l'ACP des génotypes standardisés en tant que **covariables**. Une covariable est une variable explicative ajoutée au modèle pour tenir compte d'un facteur externe susceptible d'influencer la variable d'intérêt, comme le sexe, la race, ou ici la structure génétique.

Le modèle devient alors :

$$Y_i = \beta_j G_{ij} + \sum_{k=1}^{n} \gamma_k \mathrm{CP}_{ik} + \varepsilon_i$$

où CP_{ik} représente la coordonnée de l'individu i pour la $k^{\text{ième}}$ composante principale, et γ_k l'effet fixe associé. Dans notre étude, nous retenons les 10 premières composantes principales, un choix classique permettant de capturer l'essentiel de la structure de population, tout en maintenant un coût computationnel raisonnable.

Parmi les outils permettant l'ajustement rapide de modèles linéaires à grande échelle, `MatrixEQTL` est particulièrement adapté à l'analyse eGWAS [5]. Cet outil, développé en `R`, permet de tester efficacement les associations entre des génotypes et des niveaux d'expression génique. Il repose sur une implémentation optimisée des moindres carrés dans un cadre matriciel, ce qui lui permet de traiter simultanément un très grand nombre de paires *(variant, gène)* tout en incluant des covariables. Dans notre étude de simulation, `MatrixEQTL` a été utilisé pour effectuer les analyses des modèles mixte *global* approximé et mixtes *intra-races* approximés (section 4.3.4).

4.3.4 Modèles linéaires mixtes intra-races (approximation)

Les **modèles intra-races** consistent à réaliser les analyses d'association séparément pour chaque race. Cette approche permet d'identifier des associations spécifiques à une population donnée, sans dilution des effets et bruit causés par la présence d'individus appartenant à d'autres races.

Pour une race r donnée, la formulation d'un **modèle linéaire mixte intra-race** est la suivante :

$$Y_i^{(r)} = \beta_j^{(r)} G_{ij}^{(r)} + u_i^{(r)} + \varepsilon_i^{(r)}$$

où :

- $Y_i^{(r)}$ est le niveau d'expression du gène pour l'individu i dans la race r ;
- $G_{ij}^{(r)}$ est le génotype de l'individu i pour le variant j ;
- $\beta_j^{(r)}$ est l'effet additif estimé du variant j dans la race r ;
- $u_i^{(r)}$ est l'effet aléatoire lié à la structure génétique au sein de la race r ;
- $\varepsilon_i^{(r)}$ est le terme d'erreur aléatoire.

Un avantage notable de cette approche est qu'elle autorise une estimation indépendante de l'effet $\beta_j^{(r)}$ dans chaque race. Cela permet de détecter des effets *différents* entre races, c'est-à-dire des effets dont les valeurs estimées varient d'une race à l'autre, y compris avec des signes opposés. Toutefois, une limite importante de cette méthode est qu'elle ne permet pas de tester formellement si ces différences sont **statistiquement significatives**.

Par ailleurs, cette séparation des analyses par sous-groupe entraîne une diminution de la taille d'échantillon par modèle, ce qui peut réduire la puissance statistique, en particulier pour les variants rares. Enfin, pour tirer pleinement parti des approches intra-races, une étape de synthèse entre les résultats est souvent nécessaire, en particulier quand le nombre de races analysées devient important. Celle-ci peut être réalisée via des méthodes de méta-analyse, décrites dans la section suivante.

4.3.5 Méta-analyse des modèles intra-races

Les approches de **méta-analyse** consistent à combiner les résultats d'analyses d'association effectuées séparément dans chaque race (voir sous-section 4.3.4). Elles permettent d'évaluer la robustesse et la cohérence des effets détectés à travers les différentes sous-populations, et potentiellement de distinguer les effets communs à plusieurs races de ceux qui sont spécifiques à une seule.

Contrairement à un modèle *global* unique, la méta-analyse repose sur une agrégation secondaire des résultats. Cela permet de respecter la structuration de population tout en permettant une synthèse facilitée.

Parmi les outils décrits dans la littérature, **METAL** [6] est couramment utilisé pour réaliser des méta-analyses à grande échelle. Il permet de combiner les résultats d'études d'association en s'appuyant sur les statistiques de test ou sur les effets estimés pondérés par leur erreur standard.

Pour un variant donné, les estimateurs d'effet $\hat{\beta}_j^{(r)}$ dans chaque race r sont combinés en un effet global $\hat{\beta}_{j,\text{meta}}$ selon une moyenne pondérée [6] :

$$\hat{\beta}_{j,\text{meta}} = \frac{\sum_r w_r \hat{\beta}_j^{(r)}}{\sum_r w_r}, \quad \text{où } w_r = \frac{1}{\text{Var}(\hat{\beta}_j^{(r)})}.$$

La variance de l'effet combiné peut alors être estimée par :

$$\text{Var}(\hat{\beta}_{j,\text{meta}}) = \frac{1}{\sum_r w_r},$$

ce qui permet de calculer une statistique de test pour évaluer la significativité de l'association au niveau agrégé, via un score z :

$$z = \frac{\hat{\beta}_{j,\text{meta}}}{\sqrt{\text{Var}(\hat{\beta}_{j,\text{meta}})}} = \hat{\beta}_{j,\text{meta}} \cdot \sqrt{\sum_r w_r}.$$

La fiabilité des résultats dépend fortement de la précision des effets estimés dans chaque sous-population. Lorsque les effectifs sont faibles ou déséquilibrés, la variance des estimateurs augmente, ce qui peut réduire la puissance globale de la méta-analyse et rendre plus difficile l'interprétation des résultats.

Dans le cadre de notre étude de simulation, nous avons fait le choix de ne pas inclure de méta-analyse combinant les résultats de seulement deux races. l'outil `METAL`, bien que largement utilisé dans ce contexte, fonctionne exclusivement en ligne de commande et nécessite des formats de données d'entrée spécifiques, incompatibles en l'état avec ceux générés par `MatrixEQTL`. Cela rend son intégration difficile dans notre pipeline de simulation, entièrement développé sous `R`. Nous nous sommes donc limités à l'analyse simultanée des résultats des deux modèles *intra-races*, sans combinaison secondaire par méta-analyse.

4.3.6 Synthèse des modèles comparés

Le tableau suivant présente une synthèse comparative des modèles étudiés, en mettant en évidence leurs points forts et leurs limites.

TABLE 4.1 – **Comparaison des modèles étudiés pour l'analyse multi-population**

Modèle	**Implémentation**	**Avantages**	**Limites**	**Stratégie adoptée**
Modèle linéaire global	`MatrixEQTL`, `lm()`	Permet de détecter des effets communs à plusieurs races, sert de base à différents modèles linéaires plus complexes.	A un risque élevé de faux positifs car la structuration génétique est ignorée.	Global
Modèle global corrigé	`MatrixEQTL`, `lm()`	Corrige la structure de population, réduit les biais, utilise toute l'information génétique à disposition.	Coût de calcul parfois élevé (calcul de la matrice d'apparentement), ne différencie pas les races.	Global
Modèles intra-race corrigés	`MatrixEQTL`, `lm()`	Permet d'identifier des effets spécifiques à une population, sans dilution.	Faible puissance statistique lorsque les effectifs par race sont faibles.	Intra-races
Méta-analyse des modèles intra	`METAL`	Agrège les résultats par race, permet de distinguer effets partagés et spécifiques.	Moins efficace si les tailles d'échantillons sont faibles ou déséquilibrées.	Hybride

4.4 Étude de simulation réalisée

Dans cette section, nous décrivons la stratégie de simulation adoptée pour générer les jeux de données de génotypes et d'expression génique utilisés pour l'évaluation des modèles. La stratégie de simulation est illustrée sur la figure 4.4.

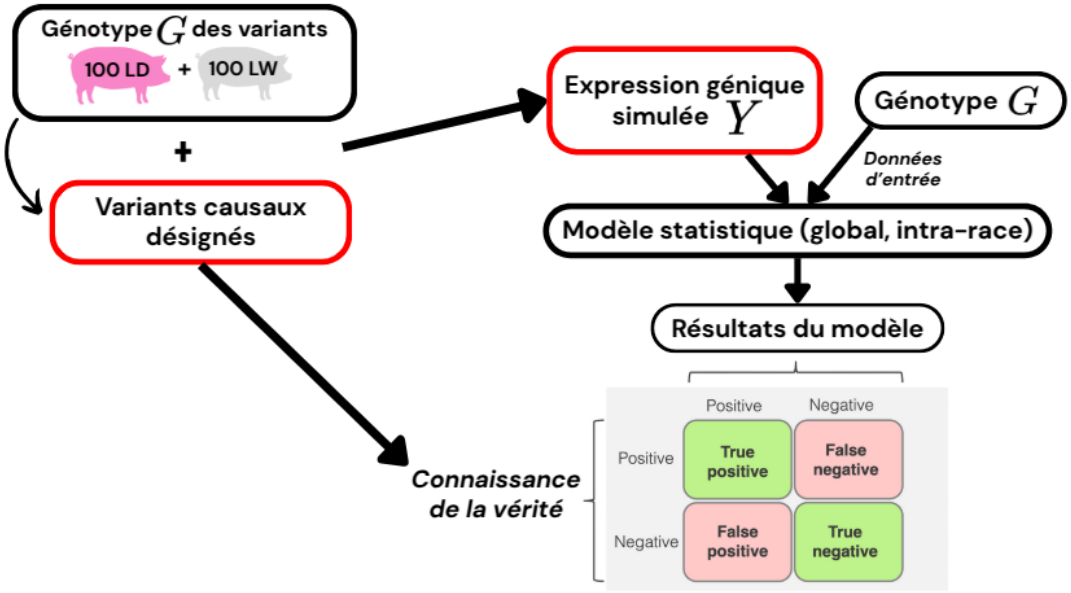


FIGURE 4.4 – **Schéma simplifié de l'étude de simulation réalisée durant le stage.** Grâce à la connaissance en amont des eQTL simulés, chaque association *(gène, variant)* identifiée par les modèles peut être évaluée à l'aide d'une matrice de confusion, distinguant les vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN).

4.4.1 Réduction des données d'entrée

Pour chaque simulation, un échantillon de 5 000 variants a été sélectionné aléatoirement parmi ceux situés sur le chromosome 18. Ce nombre a été retenu car il représentait permettait une dimensionnalité suffisamment élevée pour simuler des scénarios réalistes, sans pour autant entraîner des temps de calculs trop importants. En complément des variants, un total de 100 gènes (notés g), situés sur le chromosome 18 et issus d'une base de données `Ensembldb`, ont été échantillonnés. Pour chacun de ces gènes, 3 cis-eQTLs ont été sélectionnés parmi les 5 000 SNPs, en échantillonant aléatoirement des SNPs contenus dans une fenêtre de 100 000 paires de bases centrée sur la séquence codante du gène. Chaque simulation génère ainsi un ensemble de 300 couples {gène, eQTL} constituant la « **vérité** » à retrouver. Cette valeur de 3 eQTLs par gène a été fixée à la suite de tests montrant qu'elle permettait un bon compromis entre la complexité des signaux simulés et leur détectabilité. En effet, un nombre trop élevé de eQTLs associés à une même expression tend à diluer leur signal individuel, ce qui peut compromettre l'évaluation des performances des modèles testés. Cet effet est détaillé dans la sous-section 5.1.3.

Afin de diversifier les associations générées, des dizaines de simulations ont été réalisées, chacune reposant sur un nouvel échantillonnage aléatoire de variants et de gènes. Cette stratégie a permis de produire des ensembles d'associations distincts entre simulations, renforçant ainsi la robustesse de l'analyse grâce à l'agrégation des résultats obtenus.

4.4.2 Catégorisation des variants

Un des objectifs de l'étude est d'évaluer l'impact de l'hétérogénéité des variants entre les races. Pour cela, nous avons défini une métrique permettant de les catégoriser. Les variants ont d'abord été classés selon leur fréquence allélique globale (VAF) ainsi que leurs fréquences spécifiques dans chacune des deux races considérées : LW_VAF pour la race Large White et LD_VAF pour la race Landrace. En début d'étude, nous avons formulé l'hypothèse selon laquelle les variants présentant une fréquence allélique proche de 0,5 devaient être associés à une variabilité génotypique plus élevée, et étaient donc plus susceptibles d'être détectés. Afin de quantifier cette propriété, nous avons défini une métrique notée VAF_balance, calculée comme suit :

$$\text{VAF\_balance} = 1 - 2 \times |\text{VAF} - 0{,}5|$$

Cette métrique bornée entre 0 et 1 atteint sa valeur maximale lorsque VAF = 0,5, ce qui correspond à une variance génotypique maximale, et s'annule lorsque VAF = 0 ou VAF = 1, c'est-à-dire en l'absence totale de variabilité. De manière analogue, nous avons défini les métriques LD_VAF_balance et LW_VAF_balance pour caractériser la variabilité spécifique à chaque race.

À partir de ces métriques, les variants ont été répartis en trois grandes classes :

- **Variants spécifiques à une race** : la fréquence allélique du variant est strictement nulle dans l'une des deux races `breed_specific`. 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** : la différence absolue entre les métriques LD_VAF_balance et LW_VAF_balance est supérieure à 0,40, traduisant une forte disparité de variabilité entre les races. Deux sous-catégories sont définies : `LD_contrasted` et `LW_contrasted`, selon la race avec la plus grande variabilité ;
- **Variants homogènes** : la différence entre les VAF_balance des deux races est inférieure ou égale à 0,40, indiquant une variabilité allélique relativement équilibrée entre races. Ces variants sont regroupés dans la catégorie `homogeneous`.

4.4.3 Stratégie de simulation des effets

Pour évaluer la performance de différents modèles dans la détection d'effets **eQTLs** variables entre races, plusieurs types d'effets biologiquement plausibles ont été générés :

- des effets **identiques** entre races `"same_effect"` (scénario de référence) ;
- des effets **différents**, de même signe mais d'amplitudes différentes `"different_effect"` ;
- des effets **opposés**, de signes inversés entre races `"opposite_effect"`.

Pour chaque eQTL j désigné, un effet $\beta_j^{(r)}$ a été simulé pour chaque race r. La distribution de ces effets dépend à la fois du type de variant (spécifique, contrasté ou homogène) et du scénario de variation inter-race considéré (identique, différencié ou opposé).

La génération des effets repose sur un modèle de référence commun : une loi normale centrée en 1, avec un écart-type de 0,1, à laquelle est appliqué un signe aléatoire suivant une loi de Bernoulli symétrique. Mathématiquement :

$$\beta_j^{(r)} \sim \text{Ber}(\{-1, 1\}; 0{,}5) \times \mathcal{N}(1, 0{,}1)$$

Ce modèle simule un effet additif modéré, de signe aléatoire, avec une faible variabilité contrôlée par la variance $\sigma_\beta^2 = 0{,}1$, permettant un compromis entre réalisme et simplicité.

4.4.4 Effets des variants spécifiques

Pour les eQTLs **spécifiques** à une race, l'effet $\beta_j^{(r)}$ pour cette race est simulé selon le modèle de référence. L'effet pour l'autre race est fixé à zéro.

4.4.5 Effets des variants contrastés ou homogènes

Pour les eQTLs partagés entre races (variants **contrastés** et **homogènes**), les effets simulés sont répartis en trois catégories selon les proportions suivantes :

— 33 % d'effets **identiques** entre races,
— 33 % d'effets **différents**,
— 33 % d'effets **opposés**.

Effets identiques

Les deux races reçoivent le même effet, généré via le modèle de référence.

$$\beta_j^{(1)} = \beta_j^{(2)}$$

Effets différents

Les deux races reçoivent des effets de même signe mais d'intensités distinctes, tirées dans deux lois normales différentes :

$$\beta_j^{(1)} \sim \mathcal{N}(0.5, 0.1), \quad \beta_j^{(2)} \sim \mathcal{N}(1.5, 0.1)$$

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. Plus précisément :

$$\beta_j^{(1)} \sim \mathcal{N}(1, 0.1), \quad \beta_j^{(2)} \sim \mathcal{N}(-1, 0.1)$$

L'attribution de ces deux effets aux races est également réalisée aléatoirement.

4.4.6 Synthèse des différents scénarios d'effets

La stratégie de simulation repose sur la combinaison de deux dimensions : d'une part, la catégorisation des variants selon leur variabilité génotypique entre races (variants spécifiques, contrastés ou homogènes) ; d'autre part, la nature des effets simulés entre races (identiques, différents ou opposés). Le tableau 4.2 résume les configurations utilisées dans les simulations pour chacune de ces combinaisons. La figure 4.5 illustre les densités des lois utilisées pour générer les trois types d'effets.

TABLE 4.2 – **Scénarios utilisés selon le type d'eQTL et le type d'effet inter-races**

Type de variant	**Type d'effet**	**Effet race 1**	**Effet race 2**
Spécifique	Un seul effet simulé, pour la race ayant une fréquence allélique non nulle.	$\beta_j^{(1)} \sim$ modèle de base	$\beta_j^{(2)} = 0$
Contrasté ou homogène	**Effets identiques** : même effet simulé pour les deux races (modèle de base).	$\beta_j^{(1)} = \beta_j^{(2)}$	$\beta_j^{(1)} = \beta_j^{(2)}$
	Effets différents : même signe, amplitudes différentes ; signe et attribution aléatoires.	$\beta_j^{(1)} \sim \mathcal{N}(0.5, \sigma^2)$	$\beta_j^{(2)} \sim \mathcal{N}(1.5, \sigma^2)$
	Effets opposés : valeurs opposées, attribution aléatoire entre races.	$\beta_j^{(1)} \sim \mathcal{N}(1, \sigma^2)$	$\beta_j^{(2)} \sim \mathcal{N}(-1, \sigma^2)$

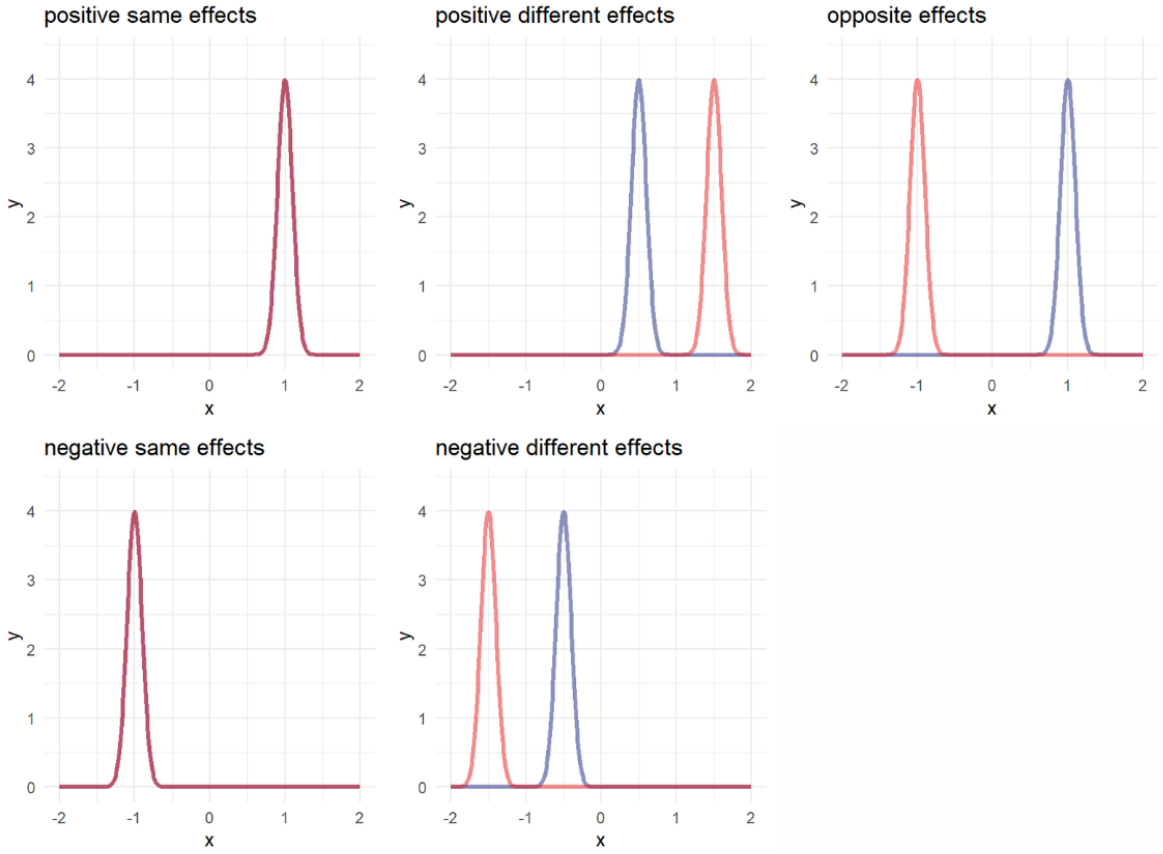


FIGURE 4.5 – **Représentation des lois de densité utilisées pour simuler les coefficients β dans les deux races selon les trois scénarios d'effet**. Dans le cas des effets identiques, la valeur identique ds deux β est tirée d'une seule loi.

4.4.7 Génération des données d'expression

Une fois les 2 effets $\beta_j^{(r)}$ simulés pour chaque eQTL selon les scénarios définis précédemment, nous avons généré les niveaux d'expression génique correspondants pour chaque gène de chaque individu. L'expression simulée d'un gène donné pour un individu i dans une race r est obtenue par la combinaison linéaire des génotypes $G_i^{(r)}$ des eQTLs associés à ce gène, pondérés par leurs effets simulés $\beta_j^{(r)}$, et perturbée par un bruit aléatoire contrôlant l'héritabilité du gène. Mathématiquement, si un gène est associé à m QTLs, alors son expression simulée $Y_i^{(r)}$ est donnée par :

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

Le terme $\varepsilon_i^{(r)}$ modélise la composante résiduelle de l'expression génique, c'est-à-dire la part non expliquée par les effets des eQTLs. Sa variance, notée σ_r^2, est calibrée de manière à atteindre un niveau d'**héritabilité** cible fixé à $h^2 = 0{,}8$. Dans ce contexte, l'héritabilité correspond à la proportion de la variance de l'expression génique attribuable aux effets génétiques simulés, ici les eQTLs.

Statistiquement, cela revient à fixer σ_r^2 de façon à ce que, en moyenne, 90 % de la variance de l'expression simulée soit expliquée par les composantes génétiques. Ce contrôle du rapport signal/bruit est réalisé en deux étapes :

1. Calcul de la variance des composantes génétiques simulées, notée $\mathrm{Var}(G\beta)$, à partir des effets $\beta_j^{(r)}$ appliqués aux génotypes $G_{ij}^{(r)}$;
2. Ajustement de la variance résiduelle selon la relation suivante :

$$\sigma_r^2 = \frac{\mathrm{Var}(G\beta) \times (1 - h^2)}{h^2}$$

Cette formule découle directement de la définition de l'héritabilité :

$$h^2 = \frac{\mathrm{Var}(G\beta)}{\mathrm{Var}(G\beta) + \sigma_r^2}$$

Chaque simulation aboutit ainsi à une matrice d'expression de taille $n \times g$ (individus $\times$ gènes), utilisée comme variable dépendante dans les analyses.

4.4.8 Indicateurs de performance

Dans les études de simulation, plusieurs indicateurs statistiques sont utilisés pour comparer rigoureusement les performances des différents modèles. Le premier critère d'évaluation est la **sensibilité** (ou *recall*), qui mesure la proportion de signaux réellement associés et correctement détectés par l'approche. Elle est définie comme :

$$\text{Sensibilité} = \frac{\text{VP}}{\text{VP} + \text{FN}}.$$

En complément, la **précision** (ou *precision*) quantifie la proportion de signaux détectés par le modèle qui correspondent effectivement à des associations simulées :

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}.$$

Cependant, ces deux métriques ne suffisent pas à résumer l'ensemble des performances, en particulier lorsque les classes positives et négatives sont déséquilibrées. C'est pourquoi le **coefficient de corrélation de Matthews** (en anglais, *Matthews Correlation Coefficient*, **MCC**) est souvent utilisé. Il s'agit d'un indicateur robuste prenant en compte l'ensemble des prédictions :

vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN).

Contrairement à d'autres métriques comme la précision ou le rappel, le MCC reste fiable même en présence d'un fort déséquilibre entre les classes. Il fournit une mesure globale de la qualité de la classification, avec des valeurs comprises entre :

— $+1$: prédictions parfaitement correctes ;
— 0 : performance équivalente au hasard ;
— -1 : prédictions parfaitement inverses.

La formule du MCC est la suivante :

$$\text{MCC} = \frac{\text{VP} \times \text{VN} - \text{FP} \times \text{FN}}{\sqrt{(\text{VP} + \text{FP})(\text{VP} + \text{FN})(\text{VN} + \text{FP})(\text{VN} + \text{FN})}}$$

Bien que le MCC soit une métrique de référence dans l'évaluation de performances, dans le cadre de notre étude, l'enjeu principal est avant tout de détecter les variants causaux (sensibilité). Le taux de faux de positifs étant impactés par des facteurs non contrôlés (comme le déséquilibre de liaison, évoqué dans la partie discussion), la précision et le MCC n'ont pas été gardés comme métriques de performance.

Par ailleurs, l'évaluation des performances ne se limite pas à la simple détection des eQTLs. Une analyse complémentaire, visant à affiner l'interprétation des résultats, consiste à examiner la capacité des modèles à estimer correctement les effets simulés $\beta^{(r)}$. En effet, un modèle peut détecter une association réelle tout en **fournissant une estimation erronée de l'effet**, que ce soit en termes d'amplitude ou de signe. Une telle erreur compromet la qualité de l'association identifiée et peut entraîner des interprétations biologiques inexactes. Cette évaluation complémentaire repose sur la comparaison directe entre les valeurs simulées des effets $\beta^{(r)}$ et celles estimées par les modèles. Elle permet d'apprécier la fidélité des estimations, indépendamment des seules métriques de classification.

Chapitre 5

Résultats

Dans ce chapitre, nous présentons les résultats obtenus au cours de l'étude. L'objectif premier est d'évaluer la qualité des prédictions fournies par **les deux types de modèles basés sur l'approximation du modèle mixte** : d'une part, le modèle *global* ajusté sur l'ensemble des individus races confondues, et d'autre part, deux modèles *intra-races*, ajustés séparément pour les 2 races *Landrace* (*LD_ model*) et *Large White* (*LW_ model*).

Afin d'interpréter correctement les résultats, nous avons dans un premier temps examiné l'impact de certains facteurs techniques propres au système de simulation. Cette étape inclut notamment l'effet des composantes principales (PCs) ajoutées comme covariables, l'influence des variations de fréquence allélique (VAF) entre populations ainsi que l'impact de l'héritabilité et du nombre d'eQTLs par gène.

Cette analyse préliminaire vise à identifier puis à isoler l'effet de ces facteurs de variabilité afin de ne pas biaiser l'interprétation des résultats liés aux scénarios d'effets simulés.

5.1 Analyse du système de simulation

5.1.1 Effet des composantes principales

Pour confirmer l'importance de la prise en compte de la structure de population dans les analyses eQTL, nous avons évalué l'effet de l'ajout des composantes principales comme covariables dans les modèles d'association. Cette évaluation a été réalisée à la fois pour le modèle *global* et pour les modèles *intra-races*.

Chaque modèle a été comparé avec et sans ajustement sur les 10 premières PCs, selon deux critères principaux : la variation du taux de faux positifs et la variation du taux de vrais positifs.

L'ajout des PCs (figure 5.1) a entraîné une **réduction très importante du taux de faux positifs**, allant en moyenne de 60 % pour les modèles *intra-races*, à plus de 95 % pour le modèle *global*. Cet effet particulièrement marqué pour le modèle *global*, s'explique par la présence conjointe de structures de sous-populations et d'apparentement dans cet échantillon.

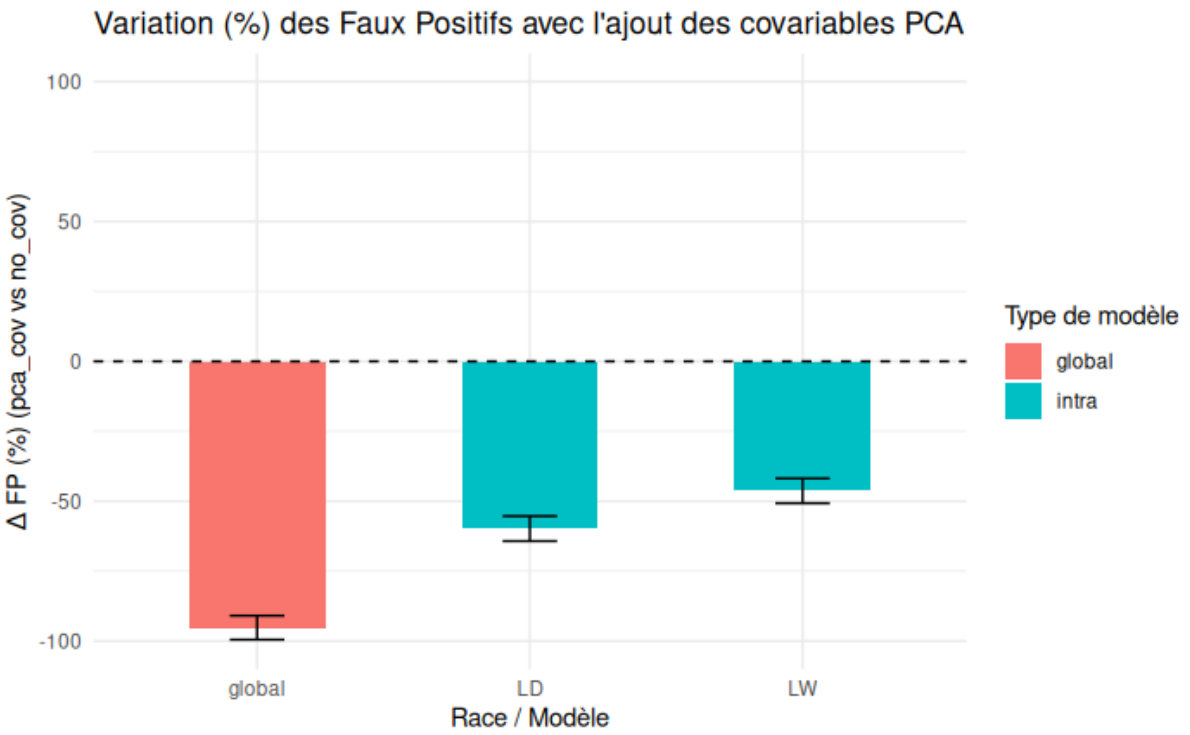


FIGURE 5.1 – **Évolution du taux de faux négatifs avec l'ajout des covariables PCs** pour le modèle global et les deux modèles intras-races.

Il est également essentiel de vérifier que l'ajout des covariables ne diminue pas la capacité à détecter les eQTLs simulés. Après analyse, **les variations du taux de faux négatifs se sont révélées globalement centrées autour de zéro**, ce qui indique que l'ajout des PCs n'affecte pas de manière importante la détection des variants causaux.

Cette forte réduction des faux positifs, combinée à une conservation de la puissance statistique, a justifié le choix méthodologique d'inclure systématiquement les PCs comme covariables dans l'ensemble des analyses suivantes.

5.1.2 Effet de la fréquence allélique des variants

Comme illustré dans la figure 4.1 de la sous-section 4.2.2, les variants simulés présentent une forte dispersion de fréquences alléliques dans la population globale, avec une prépondérance de variants à faible fréquence. Or, la fréquence allélique (VAF) constitue un paramètre clé influençant la puissance de détection des associations génétiques. En effet, plus un variant est représenté de manière équilibrée entre ses différentes formes génotypiques (homozygote référence, hétérozygote, homozygote alternatif), plus la variance génotypique observée risque d'être élevée. Cette variance accrue augmente ainsi sa contribution à la variance totale d'expression du gène cible, améliorant la probabilité de détection par un modèle d'association. À l'inverse, les variants rares, présents chez un nombre limité d'individus, induisent une plus forte incertitude sur l'estimation de leur effet, ce qui limite la puissance statistique du test et complique leur détection.

Dans les modèles linéaires employés, il est généralement admis que les variants à faible ou forte fréquence allélique sont traités de manière symétrique : Théoriquement, pour un coefficient d'effet simulé β donné, un variant dont la fréquence allélique est de 5 % ou de 95 % devrait conduire à une estimation $\hat{\beta}$ identique, avec une variance comparable de l'estimateur. Cette propriété repose sur l'hypothèse implicite d'une distribution génotypique similaire entre ces deux cas extrêmes, ce qui est souvent vérifié en pratique. Toutefois, cette approximation perd en validité pour les variants à fréquence intermédiaire, dont la structure génotypique peut être plus hétérogène. Les limites de cette hypothèse sont discutées plus en détail dans la section 6.1.

Pour évaluer l'impact de la fréquence allélique sur la capacité de détection et vérifier la cohérence de notre méthode, nous avons comparé la distribution des fréquences alléliques entre les variants détectés *VP* et non détectés *FN* par le modèle *global* (figure 5.2).

Afin de prendre en compte la symétrie des effets entre les variants rares et très fréquents (fréquence proche de 0 ou de 1), nous avons projeté sur l'axe des abscisses la fréquence de l'allèle mineur (en anglais, *Minor Allele Frequency*, **MAF**) calculée de la manière suivante :

$$\text{MAF} = \min(\text{VAF}, 1 - \text{VAF})$$

de façon à ce que la fréquence allélique maximale représentée soit 0,5.

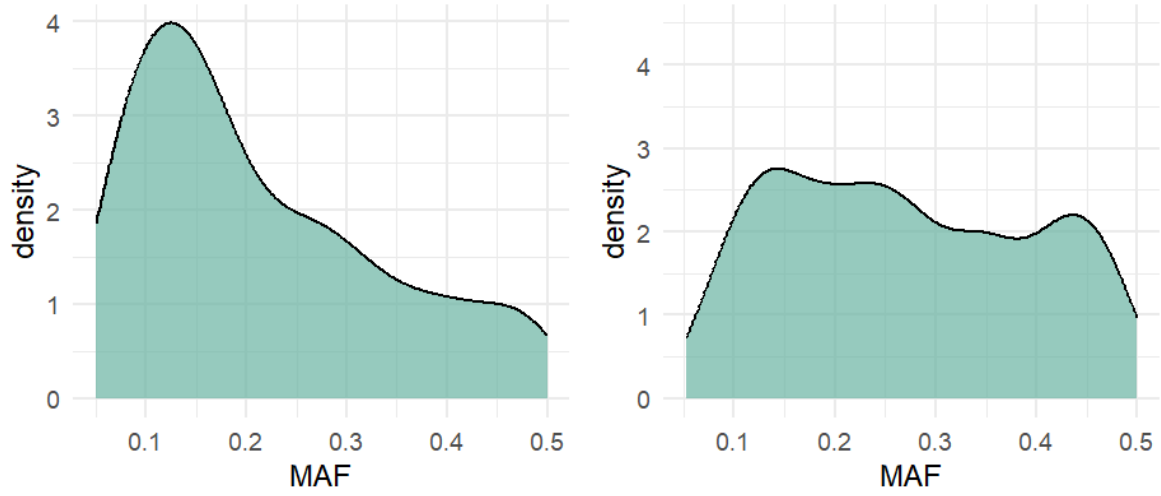


FIGURE 5.2 – **Distribution des faux négatifs (A) et des vrais positifs (B) en fonction de leur fréquence allélique mineure**. L'axe des abscisses représente la $\text{MAF} = \min(\text{VAF}, 1 - \text{VAF})$.

Comme attendu, les variants causaux présentant une faible MAF sont majoritairement associés à **une probabilité plus grande d'être non détectés**. Cela se traduit par une densité plus élevée de faux négatifs dans les classes de faibles fréquences, comparé aux vrais positifs dont la distribution est plus étalée vers des fréquences intermédiaires.

5.1.3 Effet de l'héritabilité et du nombre d'eQTLs

Dans cette sous-section, nous explorons deux facteurs susceptibles d'affecter la qualité de l'estimation des effets dans les analyses eQTL : l'héritabilité de l'expression génique et le nombre d'eQTLs par gène. Afin de mieux comprendre l'impact de ces paramètres, nous avons comparé différentes configurations de simulation, en partant d'un cas de référence.

Le cas de base considéré représente une situation théorique idéale : un unique eQTL par gène et une héritabilité égale à 1, impliquant l'absence de variance résiduelle. Dans ce scénario, les valeurs estimées des effets β par le modèle *global* coïncident parfaitement avec les valeurs simulées. Ce comportement attendu se reflète visuellement par l'alignement des points simulés le long de la bissectrice $y = x$ dans le graphique **A** de la figure 5.3.

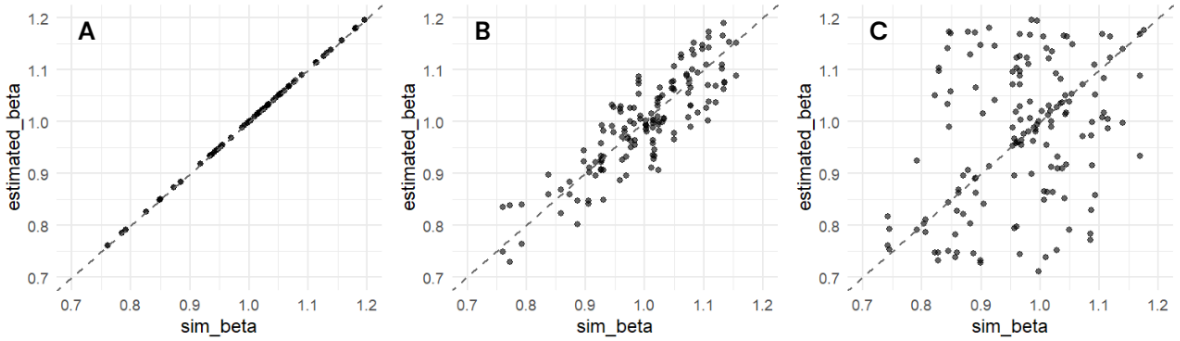


FIGURE 5.3 – **Effets de l'héritabilité et du nombre d'eQTLs par gène sur l'estimation des effets**. **A** : un seul eQTL par gène, héritabilité $h^2 = 1$. **B** : un seul eQTL par gène, héritabilité $h^2 = 0.8$. **C** : deux eQTLs par gène, héritabilité $h^2 = 1$.

Nous avons ensuite étudié deux variantes de ce cas de base :
— un unique eQTL par gène, avec une héritabilité réduite à $h^2 = 0.8$ (graphique **B**),
— deux eQTLs par gène, avec une héritabilité totale de $h^2 = 1$ (graphique **C**).

Les résultats obtenus montrent que chacune de ces modifications **dégrade la qualité de l'estimation des effets**, avec une dégradation plus prononcée pour l'augmentation du nombre de eQTLS.

Une diminution de l'héritabilité introduit une composante de bruit, ce qui réduit la précision des estimations. De même, la présence de plusieurs eQTLs par gène complexifie le signal génétique sous-jacent, pouvant introduire des biais ou des confusions dans l'attribution des effets individuels. Ces observations soulignent l'importance de ces paramètres dans la conception des analyses eQTL et la nécessité d'en tenir compte pour interpréter correctement les résultats.

L'exploration de différents scénarios nous a poussé à fixer le nombre d'eQTLs par gène à trois et l'héritabilité à $h^2 = 0.8$, car cela représentait un bon compromis entre complexité du signal génétique et détectabilité.

5.2 Performances classiques des modèles

Après la phase d'analyses préliminaires ayant amélioré notre compréhension du système, nous avons poursuivi avec **l'évaluation des approches identifiées**. Afin d'établir une base de comparaison solide entre le modèle *global* et les modèles *intra-races*, nous avons commencé par tester leurs performances sur un scénario favorable aux modèles : celui de variants présentant un effet identique entre les deux races (*same_ effect*). Pour chaque test, nous avons effectué 10 simulations et pris la moyenne des résultats obtenus.

5.2.1 Taux de détection des eQTLs

Comme évoqué dans la sous-section sur les métriques de performances 4.4.8, nous souhaitons évaluer la capacité des modèles à retrouver les eQTLs simulés. Pour cela, nous avons réalisé une carte de chaleur (figure 5.4) de la détection des eQTLs *same_ effect* non spécifiques en fonction des différentes catégories de fréquences alléliques définies en section 4.4.2.

Nous observons sur la figure que, dans ce scénario, **le modèle *global* détecte en moyenne davantage d'eQTLs que les modèles *intra-races***, quelle que soit la catégorie. Cette différence s'explique par la perte de puissance statistique dans les modèles intra, liée à la réduction de la taille d'échantillon induite par la séparation des données par race. Ce résultat confirme notre hypothèse initiale : en l'absence d'hétérogénéité d'effets entre sous-populations, un modèle *global* bénéficiant d'un effectif plus important est plus performant.

Les modèles *intra-races* se révèlent plus performants pour détecter les eQTLs contrastés en faveur de leur race associée. ce résultat semble cohérent avec notre hypothèse d'une fréquence allélique proche de 0,5 entrainant une plus forte détectabilité.

Le modèle *global* affiche de meilleures performances pour la détection des variants contrastés envers la race LW, ce qui peut s'expliquer par une variabilité génétique plus importante pour ce type de variants.

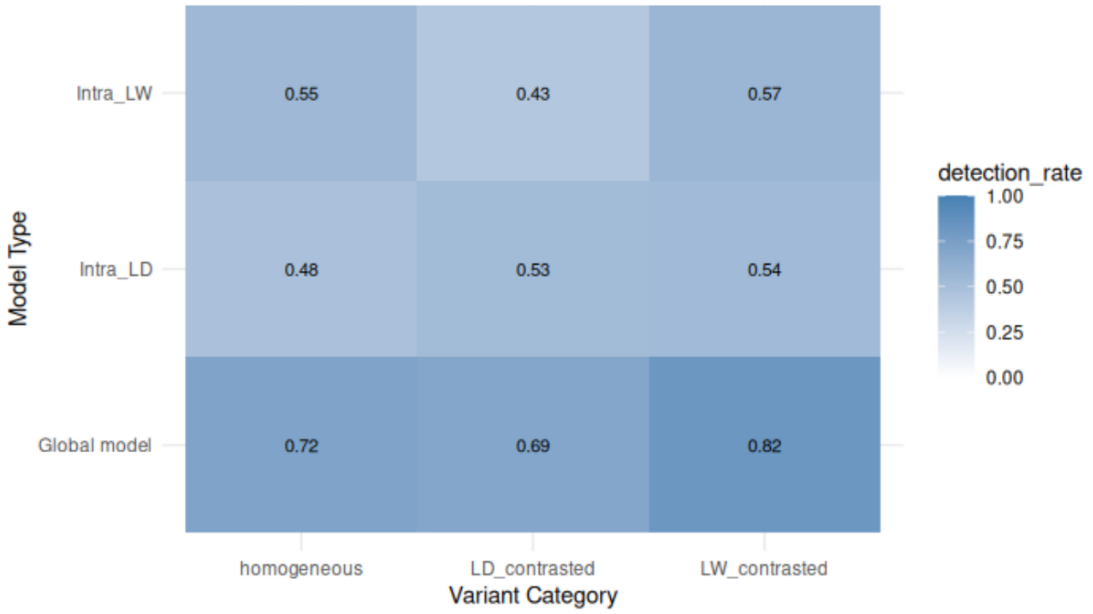


FIGURE 5.4 – **Carte de chaleur de la détection des eQTLs *same_ effect* en fonction des catégories de fréquences.**

Un second résultat en faveur du modèle *global* concerne la détection des eQTLs spécifiques à une seule race (figure 5.5). Les variants *breed-specific* ont été répartis en deux sous-groupes (*Landrace* et *Large White*), puis la performance du modèle *global* sur les deux sous-groupes a été comparée à celle des modèles *intra-races* correspondants. Le modèle *global* s'avère **légèrement plus performant** que les modèles *intra-races*.

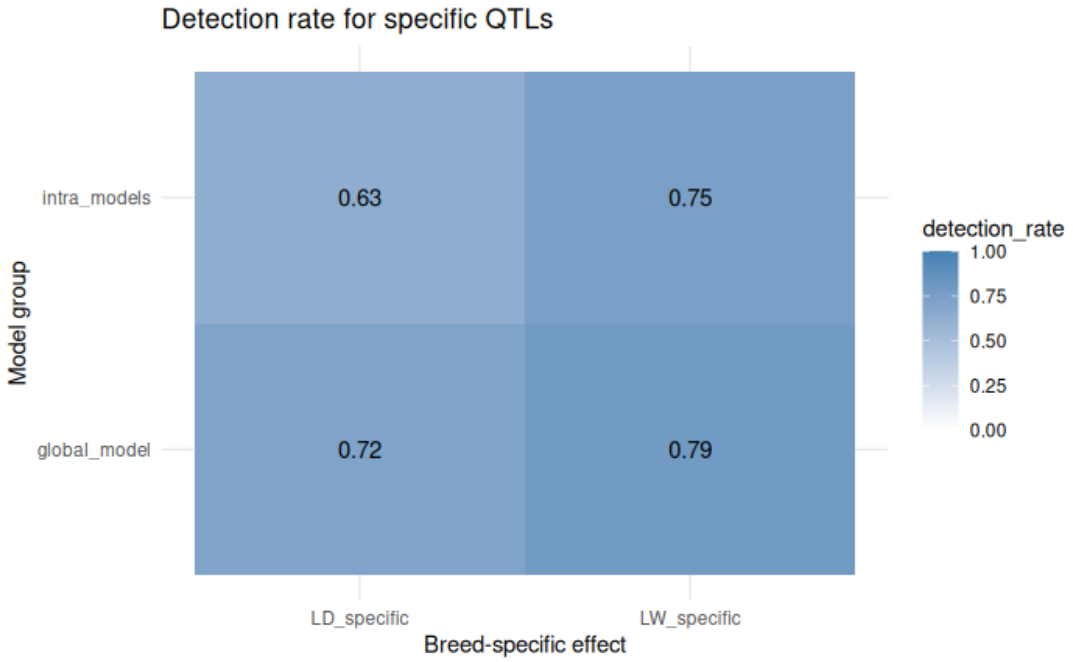


FIGURE 5.5 – **Carte de chaleur de la détection des eQTLs *same_ effect* en fonction des catégories de fréquences.**

Ce résultat peut s'expliquer par le fait que, bien que les individus de l'autre groupe (non porteurs du variant) ne soient pas directement influencés par celui-ci, leur inclusion dans le modèle *global* en tant qu'homozygotes référence permet d'estimer l'effet avec une plus grande précision, car elle augmente la taille de l'échantillon. En effet, l'augmentation du nombre d'individus réduit la variance de l'estimateur, puisque celle-ci est inversement proportionnelle à la racine carrée du nombre d'individus.

5.2.2 Estimation des effets identiques entre races

Pour compléter l'analyse précédente centrée sur la détection des eQTLs, il est crucial d'évaluer la précision des effets estimés par les modèles sur les variants détectés. Cette évaluation est réalisée en comparant les effets estimés aux effets simulés en amont, comme illustré dans la figure 5.6. Par souci de lisibilité, les variants simulés avec un effet négatif ont été inversés pour avoir une apparence similaire aux variants simulés positifs. Ce choix se justifie car les deux effets positifs et négatif présentent une symétrie de résultats.

Globalement, les modèles *global* et *intra-races* montrent des performances similaires, avec des points globalement centrés le long de la bissectrice $y = x$, indiquant une bonne concordance entre les valeurs simulées et estimées.

On observe toutefois une légère dispersion plus importante dans les modèles *intra-races*, traduisant une estimation un peu moins précise. À l'inverse, le modèle *global* présente une répartition des points légèrement plus resserrée autour de la diagonale, suggérant une meilleure stabilité dans l'estimation des effets, probablement due à l'effectif plus important.

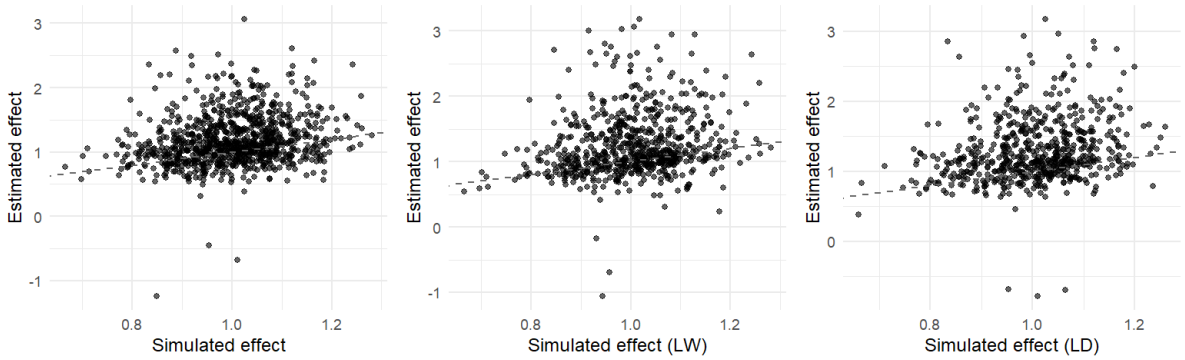


FIGURE 5.6 – **Comparaison des β estimés par les modèles face aux β *same_effect* simulés en amont**. Chaque point représente un eQTL détecté. L'axe des abscisses représente la valeur du β simulé, l'axe des ordonnées représente celle des β estimés par les modèles. La droite en pointillé représente la bissectrice $y = x$

En conclusion, dans le cas d'effets simulés identiques entre races, **le modèle *global* surpasse systématiquement les modèles *intra-races***, tant en termes de détection que en qualité d'estimation des effets.

5.3 Impact des scénarios d'effets

L'analyse suivante se concentre sur l'impact des scénarios d'effets complexes sur la performance des modèles. Il est essentiel de rappeler que seuls les modèles intégrant plusieurs populations, en l'occurrence le modèle *global*, peuvent théoriquement être affectés par la nature des effets variables entre races. En effet, les modèles *intra-races* ne traitent qu'une population homogène ; tous les variants sont donc soumis à un seul et même effet dans ce cadre, quel que soit le scénario d'effet défini.

Par conséquent, nous ne nous attendons pas à observer de différences majeures entre les deux types d'effets simulés (*different*, *opposite*) dans les modèles *intra-races*. À l'inverse, le modèle *global* pourrait subir un impact causé par l'hétérogénéité d'effets entre sous-populations.

5.3.1 Taux de détection des eQTLs

Dans un premier temps, nous avons analysé la matrice de chaleur (figure 5.7) représentant la performance des 2 types de modèles en termes de détection des eQTLs, selon les catégories d'effet définies.

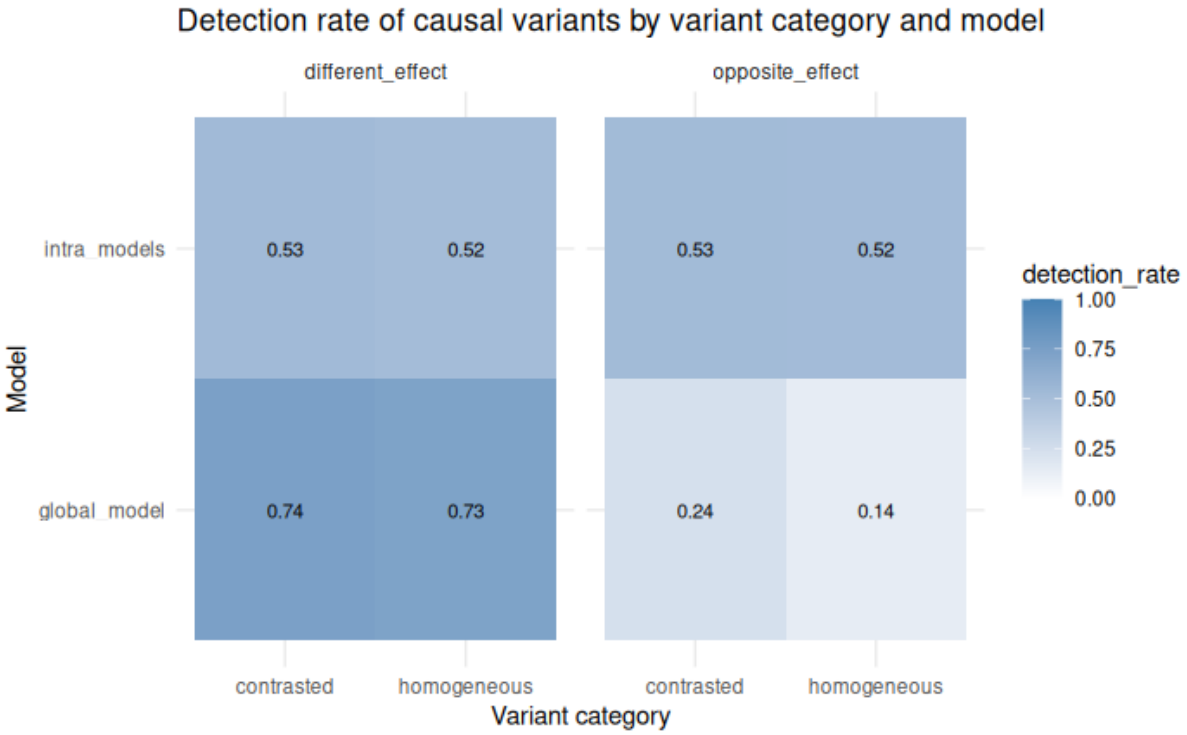


FIGURE 5.7 – **Carte de chaleur de la détection des eQTLs en fonction des types d'effets complexes et des catégories de fréquences.**

Cette visualisation met en évidence plusieurs tendances majeures. Comme prévu, les modèles intras retrouvent des performances similaires à celles obtenues pour les variants *same_ effect*. Le modèle *global* détecte efficacement les variants *different_ effect* avec une performance comparable à celle obtenue pour les variants à *same_ effect*. En revanche, **sa capacité à détecter les variants à effet opposé entre races s'effondre** : en moyenne, environ 2 variants sur 10 sont détectés dans ce scénario, le modèle est donc incapable de détecter 80% des effets eQTLs simulés opposés.

Cette baisse de performance substantielle révèle donc une limite structurelle importante des modèles actuels : lorsqu'un même variant exerce des effets antagonistes entre les populations, les signaux se neutralisent partiellement dans un modèle *global*, rendant leur détection particulièrement difficile.

5.3.2 Estimation des effets différents et opposés

Comme dans la sous-section 5.2.2, nous souhaitons ici analyser la précision d'estimation des modèles en fonction des types d'effets complexes simulés.

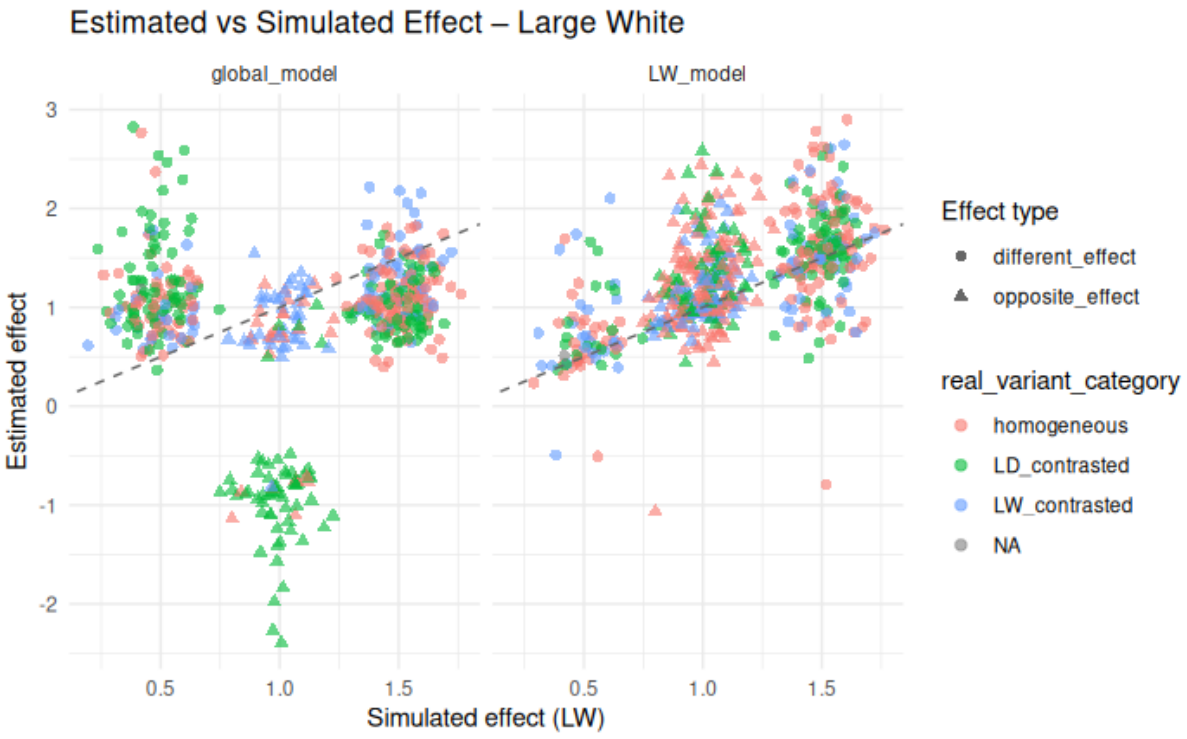


FIGURE 5.8 – **Comparaison des β estimés par les modèles faces aux β complexes simulés en amont**. Chaque point représente un eQTL détecté, catégorisé selon le type d'effet associé. L'axe des abscisses représente la valeur du β simulé, l'axe des ordonnées représente celle des β estimés par les modèles. La droite en pointillés représente la bissectrice $y = x$

Dans la figure 5.8, Pour les variants à effets différents (*different_ effect*), deux groupes sont attendus : l'un centré autour de 0,5 et l'autre autour de 1,5, reflétant les deux distributions distinctes utilisées lors de la simulation. Pour les deux graphiques, nous avons choisi de prendre comme référence les β simulés pour la race *Large White*.

Le graphique du modèle *intra* montre des résultats globalement cohérents avec nos attentes : chaque catégorie d'effet est centrée autour de la bissectrice, ce qui suggère que les effets simulés sont correctement capturés lorsque le modèle est appliqué à une seule race. Il est à noter que la plus faible présence de points détectés en dessous de la bissectrice, notamment autour de la valeur simulée 0,5, s'explique par l'incapacité des modèles à estimer des effets trop proches de zéro.

Par ailleurs, les trois nuages de points présentent une dispersion relativement importante, traduisant une variance notable dans l'estimation des effets.

Le modèle *global* présente des résultats plus contrastés. Pour les variants *different_effect*, **les nuages de points sont décalés par rapport à la bissectrice**, aussi bien pour les groupes centrés autour de 0,5 que de 1,5. Cela reflète le fait que le modèle *global*, contraint à une estimation unique de l'effet, tend à produire une moyenne des effets portés par les deux races, induisant ainsi un biais.

En examinant plus en détail les catégories de variants composant ces groupes, on observe que les effets les moins bien estimés correspondent majoritairement aux variants présentant une forte variabilité dans la race *Landrace*. À l'inverse, les variants les mieux estimés sont principalement les *LW_contrasted*, pour lesquels là variabilité est plus importante au sein de la race *Large White*.

Le cas des variants à effets opposés est encore plus problématique. **Environ la moitié des variants détectés dans ce scénario sont associés à un effet estimé de signe inversé par rapport à l'effet simulé**. De plus, **tous les variants incorrectement estimés sont catégorisés *LD_contrasted* ou *homogeneous***. Cette inversion résulte de l'annulation partielle des effets portés par les deux races dans le modèle *global*, contraint à fournir une estimation unique. Le signe attribué à l'effet dépend alors principalement de la race dans laquelle le variant est le plus exprimé.

Ces observations soulignent une limite importante des modèles classiques utilisés dans cette étude : leur incapacité structurelle à estimer de manière flexible des effets différents entre races. Ce constat prouve la nécessité d'explorer des modèles plus flexibles, capables de modéliser explicitement les effets différenciés ou antagonistes entre sous-populations.

Chapitre 6

Discussion

Ce chapitre de discussion vise à replacer les résultats obtenus dans leur contexte méthodologique, en identifiant les principales limites de l'approche employée et en proposant des pistes d'amélioration pour la suite du projet. Nous revenons notamment sur l'usage des fréquences alléliques comme critère de classification, ainsi que sur l'impact du déséquilibre de liaison non contrôlé. Enfin, nous discutons l'intérêt d'approches alternatives, telles que les modèles bayésiens hiérarchiques, afin de palier aux limites des modèles linéaires dans l'estimation des effets complexes.

6.1 Limites de l'étude des fréquences alléliques

L'utilisation de la fréquence allélique pour discriminer les variants présente plusieurs avantages, notamment son interprétabilité, sa simplicité et sa facilité d'utilisation. Toutefois, en convertissant directement les génotypes en fréquences alléliques, nous perdons un niveau d'information essentiel lié à la distribution des génotypes homozygotes et hétérozygotes. Par exemple, une VAF de 0,50 indique qu'en moyenne, un individu possède un allèle alternatif sur deux à une position donnée. Cependant, **cette même valeur peut refléter des situations génétiquement très différentes** : (1) une répartition équilibrée entre homozygotes référence (0/0), hétérozygotes (0/1) et homozygotes alternatifs (1/1), (2) une majorité d'individus hétérozygotes avec très peu de variabilité, ou encore (3) un cas extrême où tous les individus d'une race (par exemple Landrace) sont homozygotes alternatifs et ceux de l'autre race (Large White) homozygotes référence.

Ces scénarios, bien que produisant la même VAF globale, traduisent des architectures génétiques et des implications biologiques très différentes pour l'expression génique, mais ne sont pas distingués dans notre étude actuelle. Vouloir différencier ces différents cas nécessiteraient une modélisation plus fine et probablement une redéfinition complète du protocole de simulation, permettant à la fois de modéliser la fréquence des variants, leur variabilité globale et leur variabilité intra-races.

Une piste explorée en fin de stage a consisté à utiliser, non plus uniquement la VAF ou la VAF balance, mais directement la variance génotypique des variants comme critère de catégorisation. Cette approche, combinée à l'analyse des fréquences alléliques, offre une vision plus complète de la diversité génétique observée. Elle permet notamment de mieux capturer les différentes configurations évoquées précédemment que la VAF seule ne permet pas de discriminer.

6.2 Impact du déséquilibre de liaison et exploration du regroupement des variants

Dans le cadre de cette étude de simulation, nous avons volontairement choisi de ne pas contrôler les faux positifs induits par le déséquilibre de liaison (en anglais, *linkage disequilibrium*, **LD**). Ce phénomène biologique, lié à la structure chromosomique et au taux de recombinaison, se traduit par une ségrégation fréquente de variants génétiquement proches au sein des individus. En conséquence, deux variants situés à proximité dans le génome peuvent présenter **un génotype presque identique**.

Ainsi, lorsqu'un modèle statistique identifie un variant causal, il est fréquent que plusieurs variants adjacents soient également déclarés significatifs, non pas en raison d'un effet propre, mais du fait de leur corrélation avec le vrai eQTL. Ce type de faux positif est attendu dans les GWAS et peut rapidement amplifier le nombre de faux positifs.

Dans notre cas, l'objectif principal était de mesurer la capacité des modèles à identifier les vrais positifs. Par conséquent, un taux plus élevé de faux positifs, dû au LD, n'était pas considéré comme problématique dans l'évaluation comparative des modèles.

Afin d'améliorer la robustesse de l'évaluation pour la suite du projet, une approche explorée durant le stage a consisté à regrouper les variants détectés en sous-ensembles plutôt qu'à les analyser individuellement. Ce regroupement s'appuie sur deux critères complémentaires : la proximité génomique (variants séparés par une distance inférieure à un seuil) et la corrélation génotypique, afin d'éviter d'associer des variants proches mais non co-hérités, ce qui pourrait augmenter le taux de faux négatifs.

6.3 Ouverture sur les modèles hiérarchiques bayésiens

Les résultats obtenus au cours de cette étude mettent en évidence les limites des approches classiques basées sur des modèles linéaires mixtes, en particulier pour la détection de variants à effets hétérogènes ou opposés entre populations. Dans cette perspective, des méthodes récentes telles que **mashr** [7] représentent une alternative prometteuse.

Le cadre de *mashr* repose sur une modélisation bayésienne hiérarchique, spécifiquement conçue pour tirer parti de la redondance ou de la corrélation des effets entre conditions expérimentales (ici, les races). Ce type de modèle s'articule en deux étapes :

1. **Apprentissage des structures de covariance**
   À partir d'un ensemble de signaux (effets estimés, avec leurs erreurs standards), *mashr* infère une base de covariance représentative des structures d'effet possibles (effets partagés, spécifiques, opposés...). Cela permet de générer différentes configurations plausibles d'effets inter-populations.
2. **Régression bayésienne adaptative**
   Chaque nouveau signal (par exemple un triplet gène-variant-race) est ensuite projeté dans cette base, et une inférence bayésienne est effectuée. Le modèle génère une estimation *à posteriori* plus robuste des effets, en pondérant les structures identifiées selon leur vraisemblance.

Contrairement aux modèles linéaires classiques, **cette approche ne suppose pas que les effets soient identiques entre races, ni qu'ils soient totalement indépendants**. Elle permet donc de mieux capturer les effets partagés ou partiellement corrélés. Bien que cette approche n'ait pas pu être testée lors du stage, elle constitue une piste méthodologique pertinente à explorer dans la suite du projet.

Chapitre 7

Conclusion et perspectives

Ce travail de stage a permis de poser les bases méthodologiques et expérimentales pour évaluer la capacité de modèles statistiques linéaires à détecter des associations génétiques complexes dans un contexte multi-populations structuré. En s'appuyant sur des données génomiques réelles de porcs issues du projet GENE-SWitCH et sur une stratégie de simulation rigoureuse, nous avons mis en évidence les forces et les limites de deux types d'approches en modèles eGWAS, les modèles *globaux* et *intra-races*.

Les résultats obtenus montrent 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 de 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. Enfin, les modèles hiérarchiques bayésiens apparaissent comme des pistes prometteuses pour dépasser ces limitations, en combinant robustesse et flexibilité dans la prise en compte de l'hétérogénéité inter-populations.

Dans le cadre de ce stage, nous avons développé un système de simulation réplicable, modulaire et réaliste, qui a non seulement permis d'évaluer l'impact des variants complexes sur les résultats des modèles GWAS, mais aussi de mieux comprendre l'effet de paramètres tels que la fréquence allélique, le nombre de eQTLs par gène ou les covariables. Cette approche méthodologique constitue une stratégie puissante pour approfondir l'étude de la diversité génétique des espèces agricoles, en particulier dans une optique d'agroécologie et de sélection durable.

Ce travail s'inscrit pleinement dans les objectifs du projet AgroDiv et, plus largement, du PEPR Agroécologie et Numérique, en contribuant à une meilleure caractérisation de la diversité génétique pour accompagner la transition vers une agriculture plus résiliente et durable. Par ailleurs, les travaux réalisés au cours de ce stage seront poursuivis par un doctorant dans le cadre d'une thèse intitulée « Intégration de données multi-omiques appariées à grande échelle ».

Annexe A

Bilan personnel d'apprentissage

A.1 Bilan des compétences techniques

Ce stage m'a permis de consolider et d'approfondir mes compétences techniques en programmation dans l'environnement `R`, largement utilisé dans le domaine de la bioinformatique. Dès les premières étapes du projet, j'ai été amené à modifier plusieurs fois l'ensemble de la chaîne de traitement des données de simulation, pour la rendre plus robuste, généralisable et automatisée.

J'ai notamment découvert et pris en main l'écosystème `tidyverse`, un ensemble cohérent de paquets conçus pour la manipulation, le nettoyage et la transformation de données. Cet environnement propose une grammaire unifiée (`dplyr`, `tidyr` etc.) qui s'est révélée particulièrement adaptée pour manipuler les données dans le pipeline informatique.

Par ailleurs, j'ai pu développer mes compétences en visualisation de données grâce à l'utilisation du paquet `ggplot2`. Ce dernier m'a permis de produire des graphiques complexes, paramétrables, esthétiques et facilement adaptables en fonction des publics.

Enfin, la mise en place d'un pipeline m'a permis de mieux appréhender les bonnes pratiques en programmation scientifique : séparation des fonctions et du code d'exécution, utilisation de structures modulaires et réutilisables, documentation des scripts, et mise en place d'une logique de reproductibilité. J'ai également pris conscience de l'importance de la gestion des versions et de l'organisation rigoureuse des fichiers, indispensables pour assurer la lisibilité, la maintenance et la pérennité d'un projet de simulation complexe.

De manière générale, ce stage a renforcé ma capacité à concevoir, tester et adapter des scripts pour répondre à des problématiques scientifiques réelles, tout en intégrant des dimensions de robustesse et d'efficacité dans le code.

A.2 Développement des compétences scientifiques

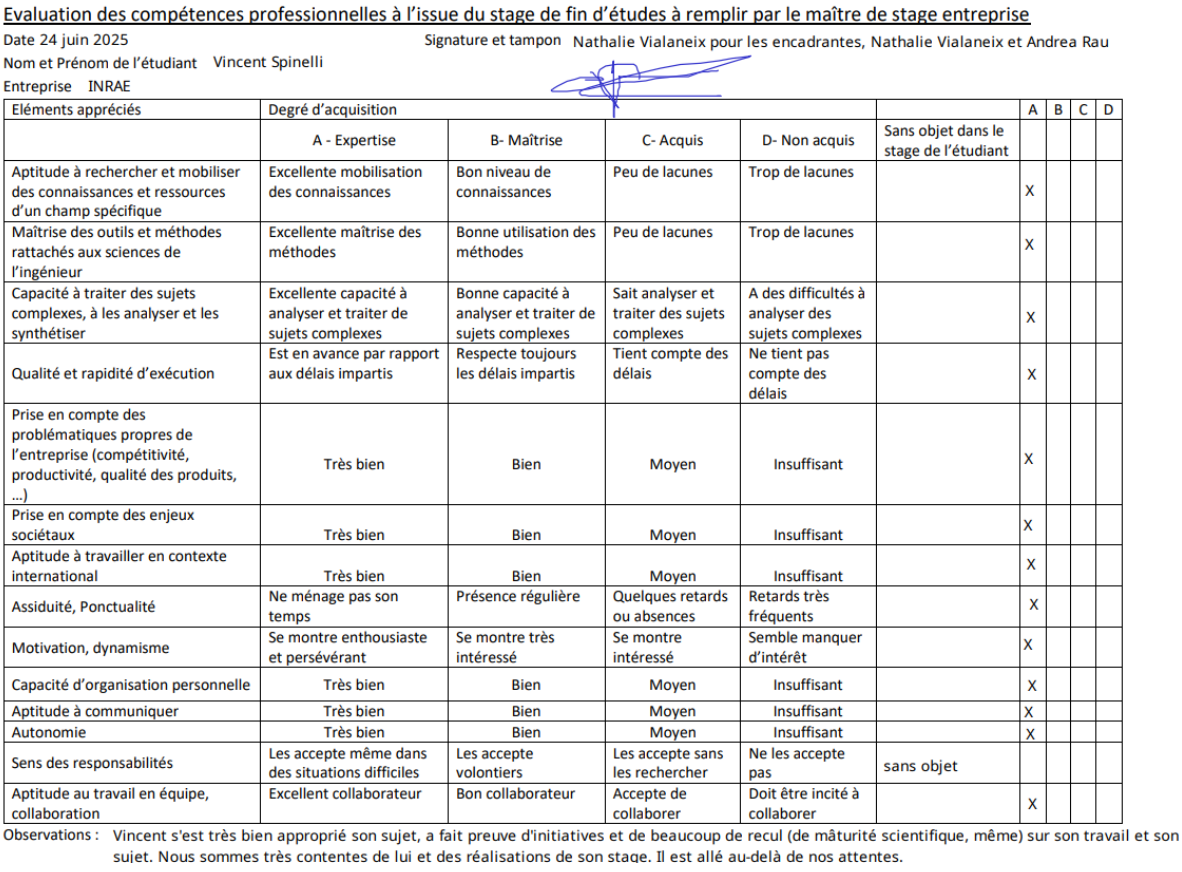
La réalisation d'une étude de simulation au cours de ce stage a constitué une expérience particulièrement formatrice d'un point de vue scientifique. Concevoir un système théorique entièrement contrôlé, dans lequel chaque paramètre est maîtrisé, permet de garantir que tout résultat contre-intuitif peut être expliqué de manière rigoureuse. Ce principe fondamental m'a conduit à interroger à plusieurs reprises la logique interne du système, à formuler des hypothèses, et à tester de nouvelles pistes de réflexion. Cette démarche a grandement contribué à améliorer ma compréhension des modèles statistiques utilisés, ainsi que ma capacité à raisonner sur leurs limites.

Au-delà de l'analyse statistique elle-même, ce stage de recherche m'a amené à développer un ensemble de compétences complémentaires, essentielles dans une démarche scientifique :

- une maîtrise accrue de la recherche bibliographique, nécessaire pour situer mon travail dans un état de l'art en constante évolution ;
- la capacité à restituer oralement les éléments clés de mon travail lors de réunions de suivi, en sélectionnant les informations pertinentes et adaptant mon discours selon les interlocuteurs ;
- la capacité à rédiger de documents scientifiques structurés et codifiés, comme ce rapport, avec l'apprentissage et l'utilisation du langage LaTeX.

Par ailleurs, j'ai eu l'opportunité d'assister à une conférence organisée à INRAE Versailles sur le projet *AgroDiv*. Cette journée m'a permis d'échanger avec des chercheurs venus de partout en France et de présenter mon sujet de stage. Les questions ayant découlées de la présentation m'ont permis d'explorer de nouvelles pistes d'analyse que je n'avais pas encore envisagées.

A.3 Grille d'évaluation ENSAT



Evaluation des compétences professionnelles à l'issue du stage de fin d'études à remplir par le maître de stage entreprise

Date 24 juin 2025 Signature et tampon Nathalie Vialaneix pour les encadrantes, Nathalie Vialaneix et Andrea Rau

Nom et Prénom de l'étudiant Vincent Spinelli

Entreprise INRAE

Eléments appréciés	Degré d'acquisition					A	B	C	D
	A - Expertise	B- Maîtrise	C- Acquis	D- Non acquis	Sans objet dans le stage de l'étudiant				
Aptitude à rechercher et mobiliser des connaissances et ressources d'un champ spécifique	Excellente mobilisation des connaissances	Bon niveau de connaissances	Peu de lacunes	Trop de lacunes		X			
Maîtrise des outils et méthodes rattachés aux sciences de l'ingénieur	Excellente maîtrise des méthodes	Bonne utilisation des méthodes	Peu de lacunes	Trop de lacunes		X			
Capacité à traiter des sujets complexes, à les analyser et les synthétiser	Excellente capacité à analyser et traiter de sujets complexes	Bonne capacité à analyser et traiter de sujets complexes	Sait analyser et traiter des sujets complexes	A des difficultés à analyser des sujets complexes		X			
Qualité et rapidité d'exécution	Est en avance par rapport aux délais impartis	Respecte toujours les délais impartis	Tient compte des délais	Ne tient pas compte des délais		X			
Prise en compte des problématiques propres de l'entreprise (compétitivité, productivité, qualité des produits, ...)	Très bien	Bien	Moyen	Insuffisant		X			
Prise en compte des enjeux sociétaux	Très bien	Bien	Moyen	Insuffisant		X			
Aptitude à travailler en contexte international	Très bien	Bien	Moyen	Insuffisant		X			
Assiduité, Ponctualité	Ne ménage pas son temps	Présence régulière	Quelques retards ou absences	Retards très fréquents		X			
Motivation, dynamisme	Se montre enthousiaste et persévérant	Se montre très intéressé	Se montre intéressé	Semble manquer d'intérêt		X			
Capacité d'organisation personnelle	Très bien	Bien	Moyen	Insuffisant		X			
Aptitude à communiquer	Très bien	Bien	Moyen	Insuffisant		X			
Autonomie	Très bien	Bien	Moyen	Insuffisant		X			
Sens des responsabilités	Les accepte même dans des situations difficiles	Les accepte volontiers	Les accepte sans les rechercher	Ne les accepte pas	sans objet				
Aptitude au travail en équipe, collaboration	Excellent collaborateur	Bon collaborateur	Accepte de collaborer	Doit être incité à collaborer		X			

Observations : Vincent s'est très bien approprié son sujet, a fait preuve d'initiatives et de beaucoup de recul (de maturité scientifique, même) sur son travail et son sujet. Nous sommes très contentes de lui et des réalisations de son stage. Il est allé au-delà de nos attentes.

FIGURE A.1 – **Grille d'évaluation ENSAT.**

Bibliographie

[1] Jeong Hwan Ko, Andrea Rau et Nathalie Vialaneix. "Analyse de la spécificité des associations génétiques dans les études multi-population". In : *Journées de Statistique de la SFdS*. Bordeaux, France : Société Française de Statistique, mai 2024. URL : https://hal.science/hal-04593233 (visité le 30/06/2025).

[2] The 1000 Genomes Project Consortium. "A global reference for human genetic variation". In : *Nature* 526.7571 (2015), p. 68-74. ISSN : 0028-0836. DOI : 10.1038/nature15393. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478/ (visité le 30/06/2025).

[3] Daniel Crespo-Piazuelo et al. "Identification of transcriptional regulatory variants in pig duodenum, liver, and muscle tissues". In : *GigaScience* 12 (juin 2023), giad042. ISSN : 2047-217X. DOI : 10.1093/gigascience/giad042. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10290502/ (visité le 30/06/2025).

[4] Hua He et al. "Effect of population stratification analysis on false-positive rates for common and rare variants". In : *BMC Proceedings* 5.Suppl 9 (nov. 2011), S116. ISSN : 1753-6561. DOI : 10.1186/1753-6561-5-S9-S116. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3287840/ (visité le 30/06/2025).

[5] Andrey A. Shabalin. "Matrix eQTL : ultra fast eQTL analysis via large matrix operations". eng. In : *Bioinformatics (Oxford, England)* 28.10 (mai 2012), p. 1353-1358. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/bts163.

[6] Cristen J. Willer, Yun Li et Gonçalo R. Abecasis. "METAL : fast and efficient meta-analysis of genomewide association scans". In : *Bioinformatics* 26.17 (sept. 2010), p. 2190-2191. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btq340. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2922887/ (visité le 30/06/2025).

[7] Sarah M. Urbut et al. "Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions". In : *Nature genetics* 51.1 (jan. 2019), p. 187-195. ISSN : 1061-4036. DOI : 10.1038/s41588-018-0268-8. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6309609/ (visité le 30/06/2025).