Comparaison d'approches linéaires et non-linéaires pour le scoring de SNPs et la prédiction de caractères complexes chez Populus nigra.

Placier Moïse & Fabrice Traore

2026-01-19

Table des Matières

1 Introduction

Notre travail a été réalisé en partenariat avec l'unité BioForA (Biologie intégrée pour la valorisation de la diversité des arbres et de la forêt) du centre INRAE Val de Loire. Il prend place dans la continuité directe des travaux de thèse menés par Alexandre Duplan, sous l'encadrement de Harold Duruflé et Leopoldo Sanchez-Rodriguez. Le projet vise à comprendre les mécanismes d'adaptation des arbres forestiers, spécifiquement le peuplier noir (Populus nigra), face aux changements environnementaux.

Pour ce faire, nous exploitons des données multi-omiques (génomique, épigénomique, transcriptomique) issues des projets ANR SYBIOPOP et EPITREE. L'objectif de notre travail est de construire des modèles capables de prédire des traits phénotypiques complexes (croissance, phénologie, résistance aux maladies) à partir des données génomiques.

1.1 L'héritage méthodologique

Les travaux conduits précédemment au sein de l'unité ont permis la constitution et la curation d'un jeu de données de référence regroupant 199 individus de peupliers noirs (Populus nigra). Sur cette base, les recherches d'Alexandre Duplan ont exploré diverses stratégies de modélisation pour la prédiction de caractères complexes à partir de données multi-omiques. Sa méthodologie repose sur une intégration précoce (early integration) consistant en la concaténation des différentes couches omiques, traitées ensuite par des algorithmes de régression Ridge ou de Random Forest. Ce choix technique permet d'appréhender le système biologique de manière unifiée, en traitant l'ensemble des l'information sur le même plan.

Néanmoins, cette approche augmente considérablement la dimensionnalité du jeu de données, soulevant plusieurs défis méthodologiques. Premièrement, l'hétérogénéité du nombre de variables entre les couches omiques induit un déséquilibre structurel : les couches les plus denses (nottament les 3 contextes épigénétiques) surreprésentent l'information, biaisant ainsi la contribution relative de chaque niveau omique au sein du modèle. Deuxièmement, l'augmentation du nombre de variables complexifie l'interprétation des résultats. La présence de variables non informatives, sans lien biologique avec le caractère étudié, dilue le signal causal. Cela rend difficile l'identification précise des locus d'intérêt ainsi que la quantification de la part de variance expliquée par ces derniers.

Pour pallier ces limites, Alexandre Duplan a développé un filtre de sélection de variables visant à ne conserver que les marqueurs les plus informatifs de chaque couche. Dans cette optique de réduction de dimension, il a mis en œuvre un modèle mixte multi-locus (MLMM), qui est une approche de GWAS (Genome-Wide Association Study), afin d'évaluer l'association statistique de chaque SNP avec les phénotypes. En sélectionnant les marqueurs présentant les p-values les plus significatives (inférieures à un seuil défini), il a pu isoler un ensemble de 31 754 SNPs associés de manière significative aux différents caractères, réduisant ainsi la dimensionalité du jeu de donné génomique avant l'étape d'intégration.

Cette stratégie de filtrage à eu un effet variable sur les performances de prédictions des différents phénotypes : tantôt positif, tantôt négatif. Alexandre Duplan suggère que ce filtre à pu éliminer du bruit, mais également des variables explicatives réelles, ce qui s'est traduit parfois par une dégradation des performances du modèle. Les analyses comparatives ont également révélé un résultat contre-intuitif;

la performance prédictive obtenue à partir d'un sous-ensemble de marqueurs génomique de même taille mais sélectionnés aléatoirement s'avère équivalente à celle obtenue via une sélection ciblée par MLMM/GWAS.

1.2 Objectifs de l'étude et hypothèses de travail

/ ici. je souhaite modifier l'approche point de départ = 1 constat : performances prédictives : Selection MLMM (initialement employé pour réduire les faux positifs et améliorer l'explicabilité des résultats) = selection aléatoire + Notre Objectif : améliorer les performances prédictives et améliorer l'explicabilité des modèles.

Ce que l'on sait : population très structuré, la biblio dit : la précision d'un modèle génomique est une résultante composite de : L'apparentement global (Structure/Kinship). Le déséquilibre de liaison (LD) avec les marqueurs. Les effets spécifiques des QTLs (Loci de Traits Quantitatifs).

=> Il y a une tension direct entre les l'objectif améliorer performance de prédictions et améliorer l'explicabilité des modèles(= identifier les variants causaux spécifiques des caractères).

Les analyses antérieures présentaient une incohérence structurelle dans le traitement de l'information : la sélection des SNPs était effectuée via un modèle MLMM (corrigé pour la structure), tandis que l'évaluation des performances de prédiction s'appuyait sur le phénotype brut. il est impossible de déterminer si la performance prédictive observée (R^2) provient de la pertinence biologique des marqueurs sélectionnés ou d'un simple effet de rappel de la structure de population.

Hypothèses sur les observations précédentes

Les bonnes performances obtenues avec de larges sous-ensembles aléatoires de SNPs s'expliqueraient principalement par la capacité de ces ensembles à capturer la structure de population et la parenté, qui constituent un proxy puissant du phénotype mesuré. Cette information, bien que prédictive, n'est pas directement informative sur les mécanismes biologiques sous-jacents.

Le plafonnement du R^2 observé dans les analyses précédentes suggère que la sélection initiale de SNPs n'incluait pas l'ensemble des variants contribuant effectivement au phénotype, soit en raison d'un nombre de SNPs sélectionnés insuffisant, soit en raison de limites méthodologiques dans l'identification de leur importance.

Nous faison donc le choix méthodologique de travailler sur les résidus déflatés des effets de structures car : 1) moins de faux négatifs sur les variants causaux identifiés Price et al. (2006) et Yu et al. (2006). 2) L'utilisation des variants causaux pour la prédictions rends plus robuste car Selon Spiliopoulou et al. (2015), dans un contexte de prédiction « across-cohort » (entre cohortes différentes ou populations non corrélées), les modèles « sparses » (parcimonieux) ciblant spécifiquement les variants causaux ont une meilleure capacité de généralisation.

Hypothèses :

Hypothèses de travail

Chaque phénotype est supposé être déterminé par un nombre variable mais limité de régions fonctionnelles, reflétant la complexité biologique du trait.

Une sélection de variables basée sur un protocole de Random Forest avec subsampling aléatoire et agrégation des scores d'importance permettrait d'identifier ces SNPs en capturant des effets non additifs et des interactions locales échappant aux modèles linéaires. Cette sélection devrait enrichir l'ensemble de SNPs en variants biologiquement pertinents, indépendamment de la structure de population. (Genome-wide prediction with non-additive models (González-Recio & Forni, 2011))

L'utilisation de cet ensemble restreint de SNPs dans des modèles de prédiction finale, qu'il s'agisse d'un modèle ridge ou d'un Random Forest, devrait permettre d'améliorer ou de maintenir les performances prédictives tout en augmentant la robustesse et l'interprétabilité biologique des résultats. Les différences de performances entre le ridge et le RF final fourniront en outre une indication sur le degré de non-linéarité et d'interactions résiduelles dans l'architecture génétique du trait.

1. Contexte et Héritage Méthodologique

Les travaux antérieurs de l'unité ont permis la constitution d'un jeu de données de référence sur 199 individus de peuplier noir (Populus nigra). Sur cette base, les premières stratégies de prédiction génomique (A. Duplan) ont reposé sur une intégration précoce (early integration) des données multi-omiques, traitées par Ridge Regression ou Random Forest. Bien que cette approche unifiée permette de saisir le système biologique dans son ensemble, elle se heurte à une dimensionnalité excessive qui dilue le signal causal dans un bruit de fond massif.

Pour tenter de réduire cette dimension, une présélection de variables par modèle mixte (MLMM/GWAS) a été testée. Cependant, les résultats ont révélé un "paradoxe de la sélection" : la performance prédictive obtenue via une sélection ciblée par GWAS s'avère équivalente à celle d'un sous-ensemble aléatoire de SNPs de même taille. Ce résultat suggère que les modèles actuels sont saturés par un signal global non spécifique au phénotype étudié.

2. La Tension Méthodologique des travaux précédents

Les analyses antérieures présentaient une incohérence structurelle dans le traitement de l'information : la sélection des SNPs était effectuée via un modèle MLMM (corrigé pour la structure), tandis que l'évaluation des performances de prédiction s'appuyait sur le phénotype brut.

il est impossible de déterminer si la performance prédictive observée (R^2) provient de la pertinence biologique des marqueurs sélectionnés ou d'un simple effet de rappel de la structure de population. En effet, comme le soulignent Habier et al. (2007) et Spiliopoulou et al. (2017), la précision d'un modèle génomique est une résultante composite de :

L'apparentement global (Structure/Kinship). Le déséquilibre de liaison (LD) avec les marqueurs. Les effets spécifiques des QTLs (Loci de Traits Quantitatifs).

En prédisant sur le phénotype brut, la composante "Apparentement" sature le signal, rendant invisible la contribution réelle des QTLs et expliquant pourquoi une sélection aléatoire (qui capture aussi l'apparentement) performe aussi bien qu'une sélection ciblée.

Dans une population fortement structurée comme celle du peuplier noir, un grand nombre de SNPs aléatoires capture efficacement la composante "Apparentement", agissant comme un proxy du phénotype sans pour autant identifier les mécanismes biologiques (Daetwyler et al., 2013). Cette prédiction par "proximité génétique" présente une limite majeure : elle masque le signal des régions fonctionnelles et empêche l'interprétabilité biologique. Pour dépasser ce plafond et isoler les variants causaux, il est impératif de neutraliser les effets de confusion liés à la démographie (Astle & Balding, 2010). Conformément aux principes établis par Price et al. (2006) et Yu et al. (2006), la correction de la structure (via PCA ou modèles mixtes Q+K) est un prérequis pour distinguer les associations fortuites des véritables signaux biologiques.

3. Objectifs de l'étude

L'objectif central de cette étude est de passer d'une prédiction fondée sur la structure de population à une prédiction fondée sur l'architecture génétique du trait.Nous cherchons à identifier un ensemble restreint et pertinent (parsimonieux) de SNPs, localisés dans des régions fonctionnelles impliquées dans le déterminisme du trait, puis à exploiter cet ensemble pour la prédiction finale. Cette démarche vise deux buts conjoints :

Améliorer ou maintenir la précision prédictive (R^2) en maximisant la composante liée aux effets de QTL spécifiques (LD-based prediction).

Garantir l'interprétabilité biologique en s'assurant que les marqueurs sélectionnés pointent vers des régions régulatrices ou codantes causales, et non vers des artefacts de structure.

4. Hypothèses de travailPour répondre à ces objectifs, nous formulons les hypothèses suivantes :

Hypothèse 1 : Spécificité fonctionnelle et architecture du trait

Chaque phénotype est déterminé par un nombre variable mais limité de régions fonctionnelles (architecture oligogénique ou polygénique modérée). L'identification précise de ces zones, une fois le signal de structure déflaté, permet de capturer la variabilité phénotypique réelle, là où les approches globales sont saturées par le bruit de fond.

Hypothèse 1bis : Influence de la colinéarité "Adaptation-Structure" Nous supposons que l'impact de la déflation sur les performances de prédiction dépendra du degré de stratification adaptive du trait :

- Pour les traits liés à l'adaptation locale (ex: phénologie), la déflation devrait réduire le R^2 car le signal biologique est intrinsèquement colinéaire à la structure de population (Berg & Coop, 2014).
- Pour les traits moins structurés, la déflation devrait au contraire augmenter les performances en supprimant le bruit de fond structural, permettant ainsi aux modèles (notamment Random Forest) de se focaliser sur le signal fonctionnel pur.

Hypothèse 2 : Supériorité de la sélection non-linéaire (Random Forest)

Les modèles linéaires classiques (type MLMM) échouent à capturer une part de l'héritabilité due aux interactions complexes (épistasie). Une sélection de variables basée sur un protocole de Random Forest,

via l'agrégation des scores d'importance, permet d'identifier des SNPs impliqués dans des effets non-additifs échappant aux GWAS standards (González-Recio & Forni, 2011).

Nous postulons que l'utilisation de la Random Forest (RF) permet d'identifier des interactions complexes (gène x gène) qui échappent aux modèles linéaires de type MLMM. Là où le MLMM capture la contribution additive infinitésimale, le RF détecterait des dépendances non linéaires et des effets épistatiques. Nous supposons donc que ces deux méthodes ne voient pas le même signal génétique et que les deux méthodes ne sont pas redondantes mais complémentaires.

Nous émettons l'hypothèse que le scoring par RF aboutit à une sélection de SNPs plus restreinte mais plus dense en information. En isolant les combinaisons de variants les plus prédictives, la RF permettrait de s'affranchir des signaux redondants liés au Déséquilibre de Liaison (LD) et à la structure de population. Le scoring par RF aboutirait à une sélection de SNPs qui seraient plus spécifique des processus biologiques du caractère étudié et situés sur des locus fonctionnels.

Si cette hypothèse se confirme, l'intérêt majeur réside dans un gain direct en explicabilité. En réduisant le nombre de SNPs candidats à une liste de locus hautement spécifiques au caractère étudié, nous facilitons l'interprétation biologique grâce à l'identification de gènes candidats ou des voies métaboliques spécifiques.

Hypothèse 3 : Le gain par la parsimonie

L'utilisation d'un subset de SNPs "nettoyé" de la structure et enrichi en signal fonctionnel permettra, dans les modèles de prédiction finale (Ridge ou RF), d'atteindre une robustesse supérieure. Si les performances du modèle Random Forest final dépassent celles du modèle Ridge sur ce subset sélectionné, cela confirmera l'existence d'une architecture génétique complexe (non-linéaire) que nous aurons réussi à capturer.

1.3 Résultats :

Résultats :

1) Pour les les traits très corréllés à la strucuture de la population => R^2 diminue sur le phénotype déflaté : Pour les selections aléatoires de subset de SNP => le R^2 tombe à presque 0 !

On vérifie l'hypothèses : A population genetic signal of polygenic adaptation (Berg & Coop, 2014) / Convergent adaptation of human lactase persistence (Tishkoff et al., 2007) / Pour les traits liés à la structure, la déflation devrait réduire le R^2 car le signal biologique est intrinsèquement colinéaire à la structure de population (Berg & Coop, 2014)

Par contre cool car les traits très peux corrélés a la pop : R^2 augmente considérablement !!! vérifie : - Pour les traits moins structurés, la déflation devrait au contraire augmenter les performances en supprimant le bruit de fond structural, permettant ainsi aux modèles de se focaliser sur le signal fonctionnel pur.

2) Egalement, Lorsque l'on compare les performances de prédictions sur le phénotype Brut ou déflaté pour un même caractère : Lorsque le caractère est peu lié a la population, le R^2 reste sensiblment

le même Lorsque le caractère est très lié a la pop, Le R^2 sur le phéno Brut s'éffondre (pas à 0 mais -40% quoi)

confirme : A population genetic signal of polygenic adaptation (Berg & Coop, 2014) / Convergent adaptation of human lactase persistence (Tishkoff et al., 2007). Les bonnes performances obtenues avec de larges sous-ensembles aléatoires de SNPs s'expliqueraient principalement par la capacité de ces ensembles à capturer la structure de population et la parenté, qui constituent un proxy puissant du phénotype mesuré.

3) Pour certains phénotypes très peu corréllés à la pop (anle insertion des branches), on réussit a identifier un top 20-30 SNP qui a des performances de prédictions énormes !

Par contre, lorsque le trait est complex il faut un nombre bcp plus importants de SNP : entre 1000 et 5000 SNP !

Beaux résultats !!!!

4) Analyse divergence du signal capturé par les différentes méthodes pour un phénotype :

Résultats 1) : Le RF qui fait la selections de SNP sur le phéno pas déflaté => performance prédictives nulles R^2 diminue sur le phénotype déflaté. Il capture également des signaux transphénotypes (% overlap top 100 SNP pour différents phénotype trés elevé) non spécifiques des phénotypes : surement très liés a la population !

résultats 2) : De manière général, une selection par RF ou par LMM aboutit relativement aux même R^2 ! Etonnant car : résultats 3) La selection de SNP par RF sur phéno déflaté ne capture pas dutoût le même signal que les méthodes linéaires !
overlap du top 100 SNP pour un même phéno environ de 30% + Extraction du Top 5000 SNPs pour chaque méthode (GWAS et RF) Calcul du Déséquilibre de Liaison (LD) entre tous les SNPs de cet ensemble

=> Le coût de transport d_ij n'est plus une distance physique, mais une distance de corrélation Une partie de l'importance identifiée par la Random Forest se situe sur des blocs d'haplotypes distincts de ceux identifiés par le GWAS.

=> Un score de 0.42 signifie qu'en moyenne, il faut déplacer 40% de la "masse d'importance" vers des haplotypes totalement non corrélés pour faire correspondre les deux méthodes.

Ouverture possible : il serait génial d'investiguer par Gène Onthologie ou autre outils biostats : les QTL disjoints identifiés par chaque méthodes pour caractériser ce signal.

5) identification d'un Traits adaptatifs et effets masqués par la correction (snp du chromosome 13)

effet transphénotypique (top SNP pour plusieurs caractère complexe) capturé par la méthode de RF sur phénotype brut. Fréquence allèlique montr très bien la corrélation avec une pop en particulier et absent chez les autres. Souligne les limites de l'approche Tishkoff et al. (2007). Convergent adaptation of human lactase persistence. Nature Genetics. → Exemple canonique d'un variant fortement structuré, biologiquement causal mais confondu avec la population.

Mathieson et McVean (2012). Differential confounding of GWAS by population stratification. PLoS Genetics. → Explique pourquoi la correction peut éliminer de vrais signaux biologiques liés à l'adaptation locale.

2 Caractérisation des ressources génomiques et phénotypiques

La prédiction de phénotypes complexes se heurte à un obstacle majeur : la structure de population est hautement hiérarchisée chez le peuplier noir (*Populus nigra*)

2.1 Etude de la structure des populations de Populus nigra

SNPs et structure de la population : qlq ACP etc etc

2.2 Génération de phénotypes de contrôle (traits nuls)

3 Correction de la structure de population et de l'apparentement

Justification du modèle $Q + K$ (Yu & Price) pour obtenir des résidus propres :

3.1 Implémentation du modèle mixte (Q+K)

Pour isoler le signal épistatique du signal polygénique de fond, nous utilisons un modèle linéaire mixte (LMM) pour déflater le phénotype.

Distinction entre Kinship et PCs Dans notre modèle, nous utilisons deux niveaux de correction : - **La Kinship (effet aléatoire) :** Elle modélise la variance liée à l'apparentement "fin" (cousins, frères). Elle traite les individus comme issus d'une distribution continue. - **Les Composantes Principales de la Kinship et la Population (PCs - effets fixes) :** Elles agissent comme des "interrupteurs" puissants retirant les différences massives entre grandes populations géographiques.

formule de la kinship d'après Yang et al. (2011) :

$$K = \frac{ZZ^T}{N}$$

avec Z la matrice centrée et réduite de la matrice de génotypes brute $(0, 1, 2)$:

$$z_{ij} = \frac{x_{ij} - 2p_i}{\sqrt{2p_i(1 - p_i)}}$$

avec

$$p_i = \frac{\text{Nombre d'allèles de référence}}{2 \times \text{Nombre d'individus}}$$

et le dénominateur : l'écart-type théorique sous l'équilibre de Hardy-Weinberg.

$$\sqrt{2p_i(1-p_i)}$$

On utilise le facteur 2 car nos individus sont diploïdes. $2p$ représente l'espérance du nombre d'allèles au locus i

$$(ZZ^T)_{jk} = \sum_{i=1}^{N} z_{ij} \times z_{ik}$$

$$(ZZ^T)_{jk} = \sum_{i=1}^{N} \frac{(x_{ij} - 2p_i)}{\sqrt{2p_i(1-p_i)}} \times \frac{(x_{ik} - 2p_i)}{\sqrt{2p_i(1-p_i)}}$$

$$K_{jk} = \frac{1}{N} \sum_{i} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

L'utilisation conjointe des PCs (ici les 3 premières) et de la Kinship garantit que : 1. Le "gros" de la structure (histoire évolutive et géographie) est évacué mathématiquement. 2. La Kinship affine la correction pour la proximité familiale.

Le MLM classique "lisse" la génétique pour éviter les faux positifs en ajustant les $p - values$ selon la ressemblance familiale. Cependant, il suppose que les effets s'additionnent simplement ($1+1=2$) et reste "aveugle" à l'épistasie (où l'effet du SNP A dépend du SNP B).

En fournissant les résidus au Random Forest, on lui transmet la part du phénotype que la parenté n'a pas pu expliquer. Si un individu est beaucoup plus performant que ce que sa "famille" laisse prévoir, la RF cherchera les combinaisons uniques de SNPs (chemins décisionnels) expliquant ce gain.

Le "Double Tamisage" Cette méthode résout le problème de la colinéarité. Une RF sur phénotype brut "redécouvrirait" simplement la parenté. Sur résidus, chaque point d'importance apporte une **information à priori nouvelle**. - **(LMM) :** Enlève la structure globale. - **(RF) :** Cherche les pépites (interactions) dans ce qui reste.

3.2 Extraction et validation des résidus phénotypiques

pour circ 2009 : Variance Génétique (σ_g^2) : u:ID = 0.079

Variance Résiduelle (σ_e^2) : units = 0.234

Héritabilité calculée : $h^2 = \frac{0.079}{0.079+0.234} \approx \mathbf{0.25}$

pour angle insertion

calcul de l'héritabilité (h^2) :

σ_g^2 (u:ID) : 0.182 σ_e^2 (units) : 0.198 h^2 : $\frac{0.182}{0.182+0.198} = \mathbf{0.48}$ Zratio (1.07) montre que cette estimation reste incertaine statistiquement

4 Analyse comparative des méthodes de scoring des SNPs

4.1 Estimation des effets additifs via le modèle MLMM

L'objectif est d'identifier les effets additifs principaux, considérés comme les "piliers" de l'architecture génétique du caractère.

Outils: MLM (Mixed Linear Model) et FarmCPU (Fixed and Adaptive Model for Mixed Probability).

Principe: Ces modèles reposent sur l'hypothèse de l'additivité, où chaque variant contribue de manière indépendante et linéaire à la valeur phénotypique. Si le MLM s'inscrit dans une vision infinitésimale (multiplicité de petits effets), FarmCPU utilise une stratégie itérative pour mieux isoler les QTLs majeurs tout en contrôlant la structure de population.

Sélection: Extraction du Top K SNPs basée sur la significativité statistique (p-value). Ce "tamisage" privilégie les marqueurs présentant un signal robuste et stable sur l'ensemble de la population étudiée.

4.2 Capture des effets non linéaires par Random Forest (Ranger)

Ce bras vise à capturer les effets fins, les interactions gène x gène (épistasie) et les effets à seuil que le modèle linéaire échoue à détecter.

- **Outils :** Forêt aléatoire (`ranger`) avec mesure d'importance corrigée.
- **Principe :** En travaillant sur les **résidus** du modèle nul, on force l'algorithme à faire abstraction de la structure de population et de l'apparentement moyen pour se concentrer sur les "exceptions à la règle" (déviations phénotypiques inexpliquées par la génétique additive).
- **Stratégie de "Feature Subspacing" :** Pour pallier le problème de la haute dimension ($p = 210\,000$ pour $n = 199$), nous avons mis en place un protocole d'échantillonnage itératif : 870 tirages avec remise de sous-ensembles de 5 000 SNPs.

4.2.0.1 Avantages méthodologiques du sous-échantillonnage de SNPs :

1. **Stabilisation du score d'importance :** Chaque SNP est sélectionné en moyenne 20 fois dans des contextes génomiques (voisinages de tirage) différents. Le score final d'importance est une moyenne pondérée de ces itérations, ce qui permet de lisser le "biais de contexte" et de stabiliser l'importance statistique de chaque marqueur :

$$E[Imp] = \frac{1}{N}\sum_{i=1}^{N} Imp_i$$

où $N \approx 20$ est le nombre de répétitions par SNP.

Équité statistique et couverture exhaustive : Avec 210 000 SNPs, une forêt aléatoire globale nécessiterait un nombre d'arbres démesuré pour garantir que chaque variable soit testée de manière significative. Notre approche garantit mathématiquement une couverture totale du génome. Chaque SNP passe un "examen" répété, assurant qu'aucun variant d'intérêt ne reste dans l'angle mort du modèle par simple malchance au tirage.

Réduction des biais de sélection et robustesse du scoring

Inspirée par les travaux de **Strobl et al. (2007)**, cette approche par sous-échantillonnage de variables (Feature Subspacing) limite les biais de sélection inhérents aux algorithmes d'arbres de décision en haute dimension. Elle transforme la forêt aléatoire d'un outil de prédiction "boîte noire" en un outil de scoring génomique grâce à plusieurs mécanismes :

- **Neutralisation du biais de catégorie par l'homogénéité d'échelle :** L'un des biais majeurs identifiés par Strobl est la tendance des forêts aléatoires à favoriser les variables offrant le plus grand nombre de points de coupure (split points). Dans notre étude, tous les SNPs sont codés de manière identique $(0, 1, 2)$. Cette homogénéité garantit qu'aucun marqueur n'est favorisé par sa structure mathématique ; seule sa capacité intrinsèque à expliquer la variance du résidu phénotypique détermine son score.
- **Décomposition du Déséquilibre de Liaison (LD) par le sub-sampling :** En génétique, la corrélation entre SNPs voisins (LD) crée une compétition où un "leader" statistique peut masquer l'importance de ses voisins. En ne présentant que ~2% du génome (5 000 SNPs) à chaque tirage, nous réduisons drastiquement la probabilité que plusieurs SNPs d'un même bloc de LD soient en compétition directe. Cela permet de "casser" temporairement ces corrélations et d'attribuer un score à chaque SNP de la région causale, plutôt que d'écraser les scores des voisins.
- **Stabilité statistique par l'espérance d'importance :** La répétition de l'expérience sur 870 itérations permet de passer d'une mesure ponctuelle et potentiellement instable à une **espérance mathématique d'importance** ($E[Imp]$). Avec chaque SNP testé en moyenne 20 fois dans des voisinages génomiques aléatoires différents, le score final est stabilisé et reflète la contribution robuste du marqueur :

$$E[Imp] \approx \frac{1}{N} \sum_{i=1}^{N} Score_i$$

où N est le nombre de tirages incluant le SNP. Cette convergence statistique assure que le "Top K" final est constitué de variables ayant prouvé leur importance de manière répétée et équitable, garantissant la fiabilité biologique de la sélection.

- **Sélection :** Extraction du Top K SNPs par score d'importance moyen.

4.3 Comportement des modèles de scoring face au signal aléatoire

4.4 Évaluation de la divergence des sélections

La mise en place de deux approches méthodologiques distinctes (**GWAS** vs **Random Forest sur résidus**) soulève une question fondamentale : ces méthodes capturent-elles la même information biologique via des marqueurs différents (redondance due au déséquilibre de liaison) ou explorent-elles des architectures génétiques réellement distinctes ?

Pour répondre à cette interrogation, nous avons déployé une stratégie de comparaison métrique.

4.4.1 Mesure de l'indice de Rand ajusté (ARI)

Dans un premier temps, nous avons évalué le chevauchement des "Top SNPs" sélectionnés par chaque méthode via l'**Adjusted Rand Index (ARI)**.

Bien que classiquement utilisé pour le clustering, l'ARI permet ici de mesurer la similarité entre les listes de variants priorisés. Les résultats préliminaires indiquent un ARI proche de zéro entre les sélections *FarmCPU* et *Random Forest*, suggérant une divergence quasi totale des cibles moléculaires brutes.

Cependant, en génétique des populations structurées, la distance physique ou l'identité stricte des marqueurs ne suffit pas. Deux SNPs différents peuvent porter la même information s'ils sont en fort **Déséquilibre de Liaison (LD)**.

4.4.2 Analyse de la distribution des scores par Earth Mover's Distance (EMD)

Pour s'affranchir des biais de position et mesurer la véritable distance biologique entre les signaux, nous avons appliqué l'algorithme de l'**Earth Mover's Distance (EMD)**, ou distance de Wasserstein.

Dans cette analyse, au lieu d'une distance physique en paires de bases, nous avons défini le coût de transport entre deux SNPs i et j comme fonction de leur corrélation génétique :

$$Coût(i, j) = 1 - r_{ij}^2$$

Nous avons normalisé les scores d'importance (RF) et les $-\log_{10}(p\text{-values})$ (GWAS) pour qu'ils forment des distributions de probabilité (somme égale à 1). L'EMD mesure alors l'effort minimal pour transformer la "carte d'importance" du GWAS en celle de la Random Forest.

4.4.2.1 Résultats et Interprétation

Les calculs d'EMD-LD révèlent une hiérarchie claire dans la divergence des signaux :

- **GWAS (FarmCPU) vs MLM classique :** $EMD \approx 0.02$.
  - *Interprétation :* La distance est négligeable. Les deux modèles linéaires identifient essentiellement les mêmes blocs d'haplotypes. C'est notre contrôle négatif.
- **GWAS (FarmCPU) vs Random Forest (sur Résidus) :** $EMD \approx 0.64$.
  - *Interprétation :* Ce score élevé indique une **rupture haplotypique**. Pour retrouver le signal de la RF à partir du GWAS, il faut "transporter" la masse d'importance vers des SNPs qui ne partagent en moyenne que **36%** de corrélation $(1 - 0.64)$ avec les signaux linéaires.

Cela démontre que la Random Forest ne se contente pas de sélectionner des "tags" alternatifs pour les mêmes QTLs, mais identifie des régions génomiques indépendantes des découvertes du GWAS.

Cette divergence métrique s'incarne biologiquement dans l'identification de variants **"exclusifs"**. En croisant les scores d'importance avec les matrices de LD, nous avons isolé des SNPs présentant une importance prédictive majeure (Score > 2000) tout en étant totalement décorrélés des pics GWAS ($r^2_{max} < 0.1$).

Ces résultats mettent en lumière une **"matière noire" génétique** : des locus essentiels pour la prédiction phénotypique mais invisibles pour l'inférence statistique classique.

4.4.3 Limite et Validation Conceptuelle

Nous reconnaissons que la p-value (inférence de l'effet moyen additif) et le score d'importance (contribution à l'architecture prédictive globale) représentent des propriétés statistiques distinctes.

Cependant, l'utilisation de l'EMD sur des distributions normalisées permet de s'affranchir des échelles de mesure pour se concentrer sur la **topographie de l'information**. La divergence observée (0.64) valide l'hypothèse selon laquelle la supériorité prédictive de la Random Forest provient de l'exploitation de cette "matière noire" — des interactions complexes et des effets non-linéaires que le modèle linéaire, par construction, ne peut percevoir.

5 Évaluation des performances de prédiction génomique

5.1 Validation de la baseline de prédiction

5.2 Comparaison des capacités prédictives (R^2)

L'étape finale consiste à fusionner les découvertes des deux bras dans un modèle unique :

$$Y \sim Ridge(SNP_{GWAS} + SNP_{RF})$$

Nous réalisons une intégration dirigée : - Nous forçons le modèle à considérer les SNPs linéairement importants (GWAS). - Nous injectons les SNPs porteurs d'information complexe (RF).

La Ridge Regression réalise l'arbitrage final grâce à sa pénalité L_2. Contrairement à un XGBoost qui pourrait accorder une importance démesurée à une interaction par pur hasard statistique (bruit) sur un faible effectif ($n = 199$), la Ridge stabilise les coefficients. Si une interaction trouvée par la RF n'est pas robuste, son coefficient sera réduit vers zéro, assurant ainsi la généralisation du modèle.

6 Discussions et perspectives