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

Les hypothèses que nous posons dans le cadre de ce travail cherchent a répondre simultanément à plusieurs objectifs :

1- Comprendre pourquoi la selection d'un subset aléatoire de taille identique au subset identifié par le MLMM permet d'obtenir des performances de prédictions similaires.

2- Améliorer les performances de prédictions.

3- Améliorer l'explicabilité des modèles de prédictions.

4- Réduire la dimensionnalité de l'espace des prédicteurs génétiques pour permettre une concaténation de données (data integration) efficace avec d'autres couches omiques (transcriptomique, métabolomique), sans saturer les modèles par le nombre de variables.

À la demande de Harold Duruflé, ce projet explore spécifiquement des alternatives non linéaires pour pallier les limites des approches linéaire dans la prédiction de phénotypes complexes. Nous avons pris le parti méthodologique d'intégrer cette non-linéarité dès l'étape de sélection des variables (scoring) via un Random Forest. Cette stratégie repose sur des hypothèses que nous chercherons a tester :

1.2.1 Complémentarité des signaux : Épistasie vs Additivité

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.

1.2.2 Parsimonie et Spécificité du signal

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.

1.2.3 Gain prédictif par hybridation

Enfin, nous chercherons à montrer que la combinaison de ces sources de signaux complémentaires permet de maximiser la variance expliquée. En confrontant les scores issus du GWAS (linéaire) et de la RF (interactions), nous espérons filtrer le bruit de fond pour ne conserver que les variants porteurs d'une information soit unique, soit synergique. L'objectif est de démontrer qu'un modèle hybride surpasse les modèles "mono-méthode" en termes de pouvoir prédictif (R^2).

Afin de confronter ces hypothèses à la réalité biologique, ce projet s'appuie sur l'étude d'une population de peupliers noirs (Populus nigra). Le peuplier est caractérisé par un génome très structuré par population et constitu donc un terrain d'étude idéal pour tester la robustesse du scoring par Random Forest face aux approches linéaires classiques.

Le chapitre suivant détaille les ressources génomiques utilisées ainsi que le pipeline bio-informatique — allant de la déflation des phénotypes par modèles mixtes à la construction de modèles de prédiction hybrides — mis en œuvre pour valider notre approche.

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.

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**. - **Tamis 1 (LMM) :** Enlève la structure globale. - **Tamis 2 (RF) :** Cherche les pépites (interactions) dans ce qui reste.

3.2 Extraction et validation des résidus phénotypiques

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.
2. **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.
3. **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 :

$$Co\hat{u}t(i,j) = 1 - r_{ij}^2$$

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