Rapport

Table des Matières

1 Introduction et Contexte Scientifique

Notre travail a été réalisé 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 antérieurs au sein de l'unité ont permis de constituer et de traiter une base de données conséquente regroupant 199 individus de peupliers noirs génotypés et phénotypés. Alexandre Duplan à développé, entre autres, plusieurs approches de modélisations pour prédire les phénotypes à partir des données multi-omiques. Il base son approche sur une concaténation des différentes couches omiques à partir de laquel il applique différents modèles de prédictions comme la régression Ridge ou le random Forest. Le choix d'intégration précoces des différentes couches omiques par concaténation est motivé par le fait que cela permet d'avoir une vue unifiée du système système biologique. Cependant, la concaténation

Dans une optique de réduction de dimension et d'identification des régions génomiques clés, une stratégie de pré-filtrage a été initialement adoptée. Elle reposait sur l'utilisation de modèles mixtes multi-locus (MLMM), une approche classique de type GWAS (Genome-Wide Association Study). L'hypothèse sous-jacente était que la sélection des marqueurs présentant les associations statistiques les plus fortes avec le phénotype permettrait de maximiser la précision prédictive tout en réduisant le bruit.

1.2 Le paradoxe de la sélection : GWAS vs Sélection Aléatoire

Les analyses comparatives menées ont révélé un résultat contre-intuitif majeur : la performance prédictive (R^2) obtenue à partir d'un sous-ensemble de marqueurs sélectionnés aléatoirement s'avère équivalente à celle obtenue via une sélection ciblée par MLMM/GWAS. Ce phénomène suggère que l'approche GWAS classique, en se focalisant exclusivement sur les signaux additifs les plus robustes, échoue à capturer la complexité du paysage génétique nécessaire à une prédiction précise. Ce constat nous impose de repenser la sélection de variables non plus comme un simple filtrage statistique, mais comme une recherche de signaux complémentaires.

1.3 Objectifs Scientifiques

À la demande de Harold Duruflé, ce projet explore 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), plutôt que de l'utiliser uniquement comme outil de prédiction final. Cette stratégie répond à une double ambition :

1.3.1 Identification de l'architecture épistatique et gain en explicabilité

Le premier objectif est d'identifier des interactions complexes (gène x gène) qui constituent une part importante de l'héritabilité manquante. En utilisant des algorithmes capables de capter des dépendances non linéaires, nous cherchons à découvrir des effets épistatiques à l'échelle des SNPs afin de faciliter l'interprétation biologique.(**papier1?**)

1.3.2 Raffinement de la sélection pour une meilleur prédiction

Le second objectif vise à aboutir à une sélection de marqueurs plus fine et plus robuste. En confrontant les signaux issus de méthodes linéaires (GWAS) et d'approches non linéaires de scoring des SNP (interactions), nous espérons d'un part : filtrer le bruit de fond et ne conserver que les variants porteurs d'une information unique ou synergique. D'autre part : maximiser la variance expliquée en combinant des sources de signaux complémentaires pour construire un modèle de prédiction plus performant que ceux basées sur une sélection aléatoire ou strictement GWAS de SNP.

2 Enjeux et Architecture de l'Approche en "Deux Bras"

La prédiction de phénotypes complexes se heurte à deux obstacles majeurs : la structure de population hautement hiérarchisée chez le peuplier noir (*Populus nigra*) et l'existence à priori d'interactions non linéaires (épistasie). Pour répondre à ces défis, nous avons développé une approche de **"Double Tamisage"** combinant des modèles linéaires mixtes et la flexibilité du machine learning.

2.0.1 Le Bras Linéaire

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.

2.0.2 Le Bras Non-Linéaire (Ranger/RF sur résidus)

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.

2.0.3 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.

3 Déflation Phénotypique et Gestion de la Structure

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

3.0.1 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 (PCs - effets fixes) :** Elles agissent comme des "interrupteurs" puissants retirant les différences massives entre grandes populations géographiques.

3.0.2 Justification de la correction hybride

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.

4 Pourquoi l'approche sur résidus est-elle "le meilleur des deux mondes" ?

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).

4.0.1 Chercher les exceptions à la règle

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.

4.0.2 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.

5 Quantification de la Divergence du Signal : Une Preuve par le Transport Optimal

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.

5.1 Divergence des listes de candidats (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)**.

5.2 L'Earth Mover's Distance (EMD) dans l'espace du LD

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.

5.2.1 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.

5.3 Mise en évidence de la "Matière Noire" Génomique

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 ($r_{max}^2 < 0.1$).

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

5.3.1 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.

6 Modèle Final : Intégration par Régression Ridge

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

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

6.0.1 1. Feature Engineering Intelligent

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).

6.0.2 2. Arbitrage par la Ridge

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.