# Sélection génomique chez le pin maritime

Victor Papin

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

# DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SCIENCES ET ENVIRONNEMENTS

SPÉCIALITÉ Sciences agronomiques et forestières

Par Victor PAPIN

## Sélection génomique chez le pin maritime

Sous la direction de : Laurent Bouffier et Leopoldo Sanchez

Soutenue le 14 décembre 2023

Membres du jury :

| | | |
|---|---|---|
| M. DOMEC, Jean-Christophe | Professeur, Bordeaux Sciences Agro | Président |
| Mme. GILBERT, Hélène | Directrice de recherche, INRAE Toulouse | Rapportrice |
| Mme. LAPERCHE, Anne | Maîtresse de conférences, Institut Agro Rennes-Angers | Rapportrice |
| M. CROS, David | Chargé de recherche, CIRAD Montpellier | Examinateur |
| M. FLUTRE, Timothée | Chargé de recherche, INRAE Le Moulon | Examinateur |
| M. SANCHEZ, Leopoldo | Directeur de recherche, INRAE Orléans | Co-Directeur |
| M. BOUFFIER, Laurent | Chargé de recherche, INRAE Pierroton | Co-Directeur (invité) |

# Titre : Sélection génomique chez le pin maritime


## Résumé

Le programme d'amélioration génétique du pin maritime (*Pinus pinaster* Ait.) a été initié dans les années 1960. Il vise à développer des variétés améliorées en terme de croissance, de rectitude et d'adaptation au milieu. Les valeurs génétiques des arbres sont classiquement estimées à partir de l'observation des performances phénotypiques et des données de pedigree.
L'amélioration génétique des espèces animales et végétales fait actuellement face à un nouveau paradigme avec l'avènement des technologies de caractérisation de leur génome. Ainsi, la sélection génomique permet de prédire la valeur génétique des individus à partir d'un grand nombre de marqueurs moléculaires et d'un modèle calibré sur une population de taille limitée. Le potentiel de cette approche est considérable pour les arbres forestiers afin de raccourcir la durée des cycles de sélection et de diminuer les coûts associés au phénotypage de caractères complexes. Toutefois, les modèles de sélection génomique actuellement proposés dans le domaine forestier présentent une précision de prédiction insuffisante, qui est de plus rarement évaluée au niveau intrafamiliale. Il s'agit pourtant d'un préalable nécessaire afin de réaliser une sélection efficace et de permettre une gestion explicite de la diversité. Egalement, l'absence de considération pour les données environnementales dans les processus d'évaluation génétique fait cruellement défaut dans l'optique de proposer des variétés adaptées aux conditions futures.

Les travaux réalisés au cours de cette thèse visent à définir les conditions de mise en œuvre de la sélection génomique chez le pin maritime. Dans un premier temps, il s'agit d'évaluer la capacité prédictive des modèles de sélection génomique, au niveau global et intrafamiliale. Pour cela, 833 individus issus de 39 familles de pleins-frères ont été génotypés pour 8234 marqueurs SNP (Single Nucleotid Polymorphism) et phénotypés pour les caractères de croissance et d'écart à la verticalité. La structure originale de la population a permis de révéler que malgré un niveau de précision global plutôt satisfaisant, la précision en intra-famille était en moyenne nulle. Une approche de simulation complémentaire a permis d'identifier que la taille de notre population de calibration, pourtant classique dans le domaine forestier, se situe en dessous d'un seuil critique à partir duquel la sélection génomique peut pleinement révéler son potentiel en captant la ségrégation mendélienne au sein des familles.
La deuxième partie de cette thèse s'attache à étendre les modèles précédemment décrits en introduisant une dimension environnementale. Des données longitudinales de croissance radiale obtenues pour 628 nouveaux individus ont été modélisées au regard de variables environnementales à l'aide d'une régression aléatoire. L'intégration de données génomiques dans ce modèle (génotypage sur 3832 SNP) permet d'estimer les valeurs génétiques de façon continue le long d'un gradient environnemental. Prendre en compte la plasticité phénotypique des arbres via l'inférence des normes de réaction apparait clé pour la sélection génomique dans un contexte de changement climatique.

L'ensemble de ces résultats démontre que l'implémentation de la sélection génomique chez le pin maritime, et plus généralement chez les arbres forestiers, est pleinement envisageable, mais que la réflexion sur la construction de la population de calibration et la prise en compte de la plasticité phénotypique sont des prérequis essentiels afin d'en démontrer tout le potentiel.



**Mots clés :** normes de réaction, pin maritime, plasticité phénotypique, programme d'amélioration, sélection génomique, prédiction intra-famille

# Title: Genomic selection in maritime pine


## Abstract

A breeding program for maritime pine (*Pinus pinaster* Ait.) was implemented in the 1960s. The aim of this program is to develop improved varieties in terms of growth, stem straightness and adaptation to the environment. Genetic values of trees are classically estimated on the basis of phenotypic performance and pedigree data.
The genetic improvement in animal and plant species is currently facing a new paradigm with the advent of genome characterization technologies. Genomic selection makes it possible to predict the genetic value of individuals based on a large number of molecular markers and a model calibrated on a population of limited size. This approach has considerable potential for forest trees, as it can shorten selection cycles and reduce the costs associated with phenotyping complex traits. However, the genomic selection models currently proposed for forestry have insufficient predictive accuracy, which is also rarely assessed within families. Yet this is an essential prerequisite for effective selection and explicit diversity management. In addition, the failure to include environmental information in genetic evaluation processes is also sorely lacking when it comes to proposing varieties adapted to future conditions.

The aim of this PhD is to define the conditions for implementing genomic selection in maritime pine. The first step was to evaluate the predictive ability of genomic selection models, both on a global and within-family level. To this end, 833 individuals from 39 full-sib families were genotyped for 8234 SNP (Single Nucleotid Polymorphism) markers and phenotyped for growth traits and stem deviation to verticality. The original population structure revealed that, despite a rather satisfactory overall level of accuracy, within-family accuracy was on average zero. A complementary simulation approach enabled us to identify that the size of our calibration population, although classic in the forestry field, is below a critical threshold at which genomic selection can fully reveal its potential by capturing Mendelian segregation.
The second part of this thesis extends the models described above by introducing an environmental dimension. Longitudinal wood growth data obtained for 628 new individuals were modeled against environmental variables using random regression. The integration of genomic data in this model (genotyping on 3832 SNP) enables genetic values to be estimated continuously along an environmental gradient. Taking into account the phenotypic plasticity of trees via the inference of reaction norms appears to be key for genomic selection in a context of climate change.

Taken together, these results show that the implementation of genomic selection in maritime pine, and more generally in forest trees, is fully conceivable, but that reflection on the construction of the calibration population and consideration of phenotypic plasticity are essential prerequisites for demonstrating its full potential.



**Keywords:** reaction norms, maritime pine, phenotypic plasticity, breeding program, genomic selection, within-family prediction

# Unité de recherche et financement

Cette thèse a été réalisée au sein de l'UMR Biodiversité Gènes et Communautés (BIOGECO) au centre INRAE de Pierroton.

*INRAE – Institut National de la Recherche Agronomique, UMR 1202 – Biodiversité, Gènes et Ecosystèmes, Site de Recherches Forêt Bois de Pierroton -Domaine de l'Hermitage, 69, route d'Arcachon 33612 CESTAS Cedex-France*

L'allocation doctorale (N°2020-CK-126) a été obtenue auprès de l'Université de Bordeaux par le concours de l'Ecole doctorale Sciences et environnements (ED 304). Le financement de cette allocation a été assuré par Bordeaux Sciences Agro.

Ce travail a également été soutenu par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de l'accord de subvention n°773383 (Projet B4EST https://b4est.eu/).

Un séjour de 2 mois (mai et juillet 2023) a été réalisé au cours de la thèse au Roslin Institute (University of Edinburgh) au sein du Highlander lab dirigé par Gregor Gorjanc. Ce séjour a été financé par l'Ecole doctorale Sciences et environnements, le département ECODIV de l'INRAE ainsi que par les UMR BIOGECO (Bordeaux) et BioForA (Orléans).

**Dates de la thèse**

Novembre 2020 – Novembre 2023

# Remerciements

*Je remercie en premier lieu mes encadrants, Laurent Bouffier et Leopoldo Sanchez, avec qui j'ai adoré partager ces 3 années de recherche. Merci pour votre disponibilité, votre soutien et vos conseils. Cela a été un réel plaisir de travailler dans ces conditions et d'apprendre à vos côtés. J'espère que ce n'est que le début de notre collaboration et que nous nous recroiserons prochainement.*

*Je remercie l'Université de Bordeaux et Bordeaux Sciences Agro pour le financement de mon allocation doctorale. Je remercie également le projet B4EST d'avoir financé le fonctionnement de cette thèse. Merci pour les évènements organisés, d'abord en visio, puis enfin en présentiel à Lisbonne, qui m'ont permis de présenter mon travail et d'échanger avec les scientifiques européens du projet.*

*Je tiens à remercier l'UMR BIOGECO et Christophe Plomion pour l'accueil chaleureux que j'ai reçu au cours de ces trois années.*
*Je remercie également l'équipe XYLOMES dirigée par Jean-Marc Gion et Grégoire Le Provost pour l'accueil et la convivialité.*
*Merci à Florence Le Pierres, Léa Peypelut, Ophélie Lacaule et Sandrine Gardet pour le soutien administratif. Mention spéciale à Florence et Ophélie pour votre aide lors de l'organisation de ma mobilité en Ecosse.*
*Merci à Loïc pour son soutien infaillible en informatique.*
*Merci à l'ensemble de la plateforme PGTB, et en particulier Christophe Boury, pour m'avoir formé et permis de réaliser mes manips labo. Merci également à Céline Lalanne pour m'avoir formé aux extractions d'ADN et pour son aide de manière générale.*
*Merci à Mathilde Flores pour les innombrables manips de labo que tu as réalisées et pour ta précieuse contribution lors des échantillonnages sur le terrain.*
*Merci à Raphaël Segura pour son aide régulière lors du travail de terrain et sa bonne humeur permanente.*
*Merci à Fréderic Lagane pour tout son travail sur les carottes de bois et les profils densitométriques.*
*Merci à Alexandre Bosc pour sa disponibilité et sa grande aide dans le volet ecophysio de cette thèse.*
*Merci à Ludovic Duvaux de m'avoir appris à dompter le CBiB*

*Merci au GIS « Groupe Pin Maritime du Futur » et à l'Unité Expérimentale (UEFP) pour l'installation et la gestion des dispositifs mais aussi pour les collectes de données. Je remercie plus particulièrement Rémi Dourthe et Christophe Gauvrit pour leur aide lors des échantillonnages.*

*I would also like to extend my warmest thanks to Gregor Grojanc, Ivan Pocrnic and the whole Highlander Lab for their hospitality during my stay. I won't forget the good atmosphere and the precious help I received during these two months.*

*Je remercie chaleureusement les membres du jury de cette thèse, David Cros, Jean-Christophe Domec, Timothée Flutre, Hélène Gilbert et Anne Laperche, pour avoir consacré du temps à l'évaluation de mes travaux de recherche et pour leur présence au moment de la soutenance.*

*Un grand merci également aux membres de mon comité de suivi de thèse, Sophie Bouchet, Oliver Brendel, Marie Denis, Santiago Gonzalez-Martinez, Gwendal Restoux et Renaud Rincent, pour leur écoute, leurs conseils avisés et leur bonne humeur.*

*Je remercie mes plus proches collaborateurs doctorants pour les bons moments passés. Tout d'abord Arnaud Chevalier double tiret Mairet, dit Nono, en lui souhaitant un bon courage pour la fin de sa thèse et en espérant qu'il se calme un peu sur les desserts de la cantine. Egalement Adélaïde Theraroz, dit Adé, ou la fada de la canebière, en espérant que sa troisième année de thèse se déroulera au mieux (mais j'ai aucun doute là-dessus) et qu'elle supportera ses nouveaux collègues du bureau (j'ai plus de doutes là-dessus). Et enfin ma colloc de bureau pendant 3 ans, Domitille Coq Maison Haute, dit Dom, en lui souhaitant le meilleur dans son futur post-doc, avec ou sans chêne mais au moins je l'espère avec beaucoup de blé.*
*Je remercie également pour leur bonne humeur tous les stagiaires, CDD et post-docs passés par BIOGECO avec qui j'ai pu échanger de près ou de loin. Notamment : Alex, Amandine, Anwar, Audrey, Benoît, Beurre salé, Clément, Coralie, GeoGeo, Greg, Laura, Mathilde, Nastasia, Su, Thomas et bien d'autres !*

*Merci à Charlotte, capitaine de l'équipe, pour ton soutien sans faille, ta gentillesse, et ta patience qui m'ont sans nul doute permis de réaliser cette thèse dans les meilleures conditions possibles.*

*Pour finir, je remercie profondément ma famille pour les encouragements et pour m'avoir toujours permis de réaliser ce que je souhaitais. Je suis très reconnaissant envers ma maman et ma sœur pour leur soutien inconditionnel, et ce depuis toujours. J'ai enfin une pensée particulière pour mon papa.*

# Liste des figures

## Partie 1



## Partie 2



## Partie 3

## Partie 4



## Partie 5

# Listes des tableaux

## Partie 2



## Partie 4

# Listes des encadrés

## Partie 1



## Partie 2



## Partie 3

# Table des matières

# Abréviations

**A :** Matrice d'apparentement calculée à partir des données de pedigree
**ABLUP :** Pedigree-based Best Linear Unbiased Predictor
**ADN :** Acide Désoxyribonucléique
**AGC :** Aptitude Générale à la Combinaison
**AIC :** Critère d'information Akaike
**ASC :** Aptitude Spécifique à la Combinaison
**BLUE :** Best Linear Unbiased Estimator
**BIOGECO :** Unité mixte de recherché Biodiversité Gènes et Communautés
**BioForA :** Unité mixte de recherche Biologie intégrée pour la valorisation de la diversité des arbres et de la forêt
**BLUP :** Best Linear Unbiased Predictor
**B4EST :** Projet européen Adaptive breeding for productive, sustainable and resilient forests under climate change
**CV :** Cross-validation
**DL :** Déséquilibre de liaison
**EBV :** Estimated breeding value
**FAO :** Food and Agriculture Organization of the United Nations
**FLD :** Fisher Linear Discriminant
**FS :** Full-Sib
**G :** Matrice d'apparentement calculée à partir des données de génotypage
**G0 :** individu de première génération (population de base)
**G1 :** individu de deuxième génération
**G2 :** individu de troisième génération
**GBLUP :** Genomic Best Linear Unbiased Predictor
**GEBV :** Genome based Estimated Breeding Value
**GIEC :** Groupe d'Experts Intergouvernemental sur l'Evolution du Climat
**GIS PMF :** Groupement d'Intérêt Scientifique Pin Maritime du Futur
**GxE :** Interaction Genotype x Environnement
**Het-so :** Heterozygous Strength Offset
**IBD :** Identity By Descent
**INRAE :** Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement
**LOD :** Likelihood Odds Ratio
**MAF :** Minor Allele Frequency
**OCS :** Optimum Contribution Selection
**PIB :** Produit Intérieur Brut
**QC :** Quality Control
**QTL :** Quantitative Trait Loci
**SAM :** Sélection Assistée Par Marqueurs
**SNP :** Single Nucleotide Polymorphism
**UMR :** Unité Mixte de Recherche

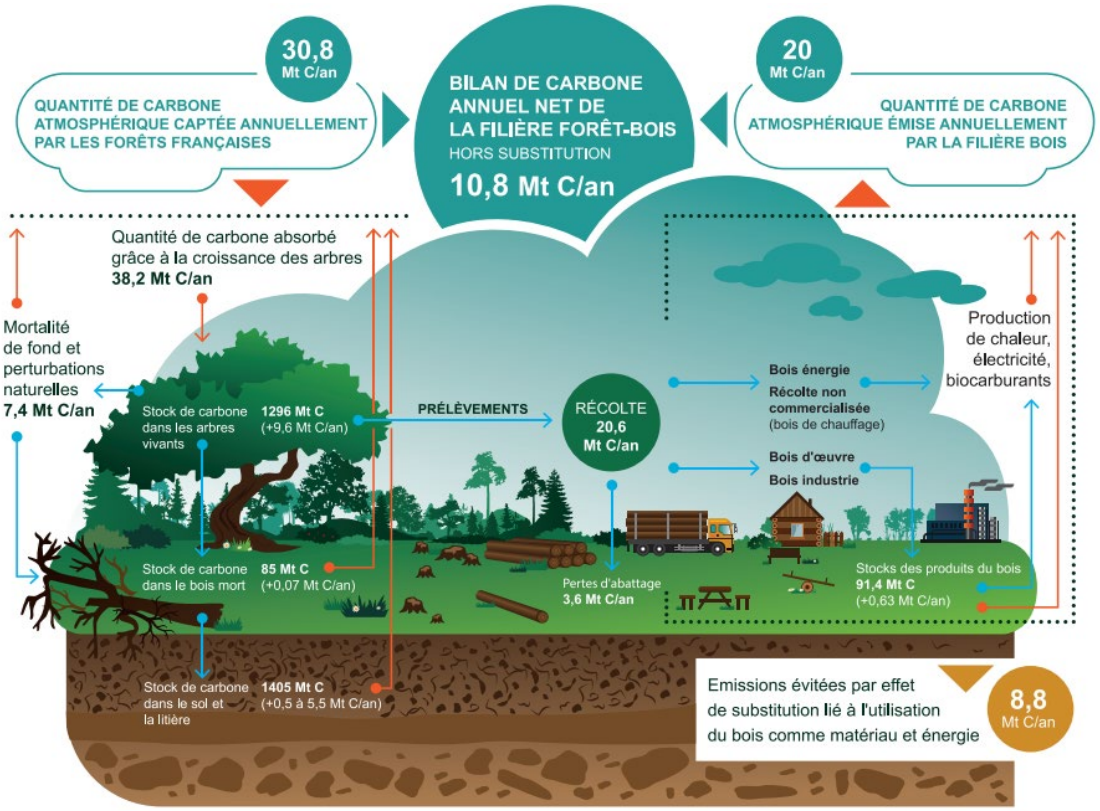



***Figure 1-1 : Bilan de carbone des écosystèmes forestiers (à gauche) et des produits du bois (à droite) pour le territoire métropolitain. (Source : ADS 2023)***

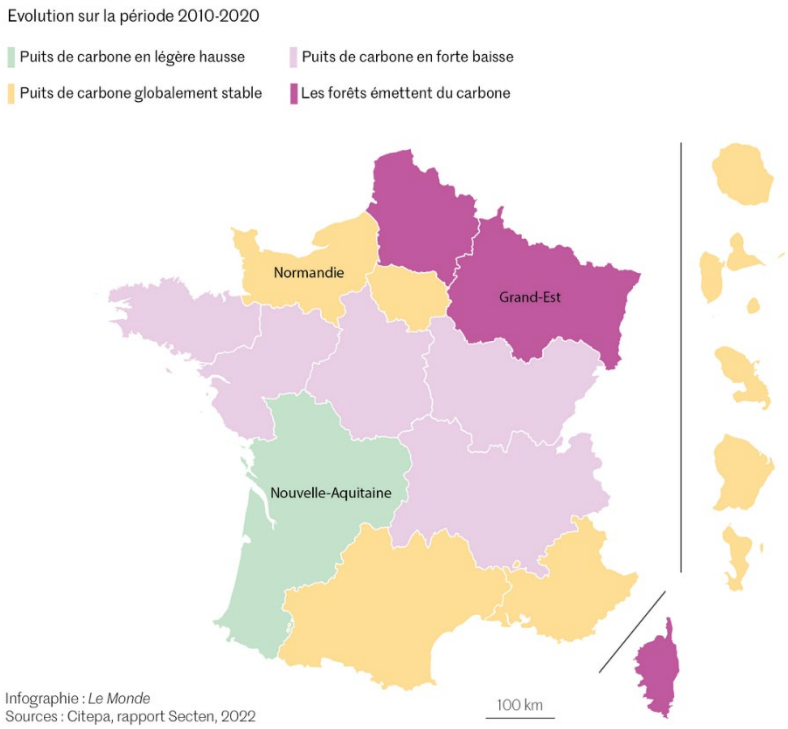



***Figure 1-2: Stockage ou émission de carbone par les forêts françaises de métropole en fonction des régions***

# 1. Introduction générale

## 1.1. Enjeux autour de la ressource bois et des forêts de plantation

### 1.1.1. Le rôle stratégique des forêts

Le sixième rapport du GIEC publié en mars 2023 réaffirme le potentiel des forêts pour faire face aux enjeux du changement climatique (IPCC 2023). Outre les services écosystémiques rendus tels que la régulation de la température et du microclimat local, ou la formation et la stabilisation des sols, ce sont les capacités des forêts à capter et à stocker le carbone atmosphérique qui sont aujourd'hui au centre de l'attention. Le bilan net du stockage de carbone par les forêts est estimé à 2.15 GtC/an dans le monde (Pugh et al., 2019) et 10.8 MtC/an en France (ADS 2023). Il se fait dans les parties végétatives aériennes et souterraines, mais aussi dans le sol, la litière et le bois mort (**Fig. 1-1**). Toutefois, les capacités de stockage des forêts sont très hétérogènes selon l'âge de la forêt, sa gestion, les essences présentes, la fertilité des sols ou encore les conditions climatiques (ADS 2023). En France, la capacité de stockage des forêts a été divisée par deux entre 2010 et 2020 (Citepa 2023). Les forêts des Hauts-de-France, du Grand Est et de la Corse deviennent même émettrices de CO2 certaines années (**Fig. 1-2**). En cause notamment, une croissance ralentie et une mortalité forte du fait des sécheresses à répétitions, des attaques de ravageurs récurrentes, ou encore des tempêtes et incendies.

En plus de leur capacité à capter le carbone atmosphérique, la production de bois est un autre atout majeur des forêts dans un contexte de changement climatique. La substitution du bois à des matériaux tels que l'acier ou le béton, dont la production génère une importante émission de gaz à effet de serre, contribue également à réduire l'impact du changement climatique. C'est une émission d'environ 8.8 MtC qui est évitée en France chaque année par l'utilisation préférentielle du bois, notamment pour des projets de construction et de fabrication (**Fig. 1-1**). Le caractère renouvelable de la ressource bois ainsi que ses avantages liés à son exploitation restent cependant conditionnés à une gestion durable des forêts.
En 2022, les prélèvements annuels de bois rond[1] s'élevaient à plus de 4 milliards de mètres cubes à l'échelle mondiale (FAO 2022). La moitié de ce bois est utilisée comme bois d'œuvre (construction, ameublement, menuiserie) et d'industrie (palettes, cartons, pâte à papier), et l'autre moitié comme bois de chauffage. Les progrès technologiques ont permis de développer

[1] : bois abattu et façonné, avant la première transformation industrielle : grume, bille, rondin ou bûche (Insee)


***Figure 1-3 : Taux annuel d'expansion de la forêt et de déforestation, 1990-2020 (source : FAO 2020)***

***Figure 1-4 : Plantation forestière dans les Landes de Gascogne (Dispositif A – Le Barp)***

des traitements et des techniques améliorant la résistance et la durabilité du bois, ou de produire de grandes pièces pour la construction à partir de petits diamètres (i.e. bois lamellé-collé), ce qui justifie aujourd'hui le foisonnement de ses usages. Le volume récolté chaque année dans le monde a augmenté de 20% depuis les années 2000 et de près de 60% depuis 1960 (FAO 2022). Ces prélèvements se font encore pour plus de 70% dans des forêts naturelles, principalement au Brésil et en Indonésie, mais aussi en Russie, au Canada et aux Etats-Unis (FAO 2022). Si la déforestation est principalement due à une conversion en terres agricoles, ces prélèvements contribuent nécessairement à la réduction de la superficie forestière, qui est encore estimée à 10 millions d'ha par an dans le monde (**Fig. 1-3**). L'exploitation effrénée des forêts naturelles épuise la ressource en bois, favorise l'émission de gaz à effet de serre et réduit drastiquement la biodiversité dans ces zones. Tous ces éléments contrastent avec la notion de durabilité du bois et l'attente sociétale sur la gestion responsable des forêts. Face à l'épuisement des ressources fossiles et la diversification des usages du bois, l'expansion et la productivité des forêts de plantation apparait essentielle pour limiter les prélèvements dans les forêts naturelles (McEwan et al., 2020; Payn et al., 2015).

### 1.1.2. Les forêts de plantation au service de la production de bois

#### *Définition et superficie des forêts de plantation dans le monde*

La FAO définit une forêt comme un couvert arboré occupant plus de 10% d'une surface de 0.5 hectares (ha), et avec des arbres atteignant une hauteur supérieure à 5m à l'âge mature. Les forêts recouvrent près de 4 milliards d'ha dans le monde, soit près d'un tiers des terres émergées (FAO 2022). Environ 294 millions d'ha, soit 7% de la surface forestière totale, correspondent à des forêts plantées, c'est-à-dire des forêts principalement composées d'arbres établis par plantation et/ou semis délibéré. Les objectifs de ces forêts plantées peuvent être multiples, allant de la restauration des milieux à la séquestration du carbone en passant par la génération de produits ligneux et non-ligneux. Au sein des forêts plantées, on distingue la sous-catégorie des forêts de plantation dont l'objectif premier est la production de bois. Ce type de forêt plantée se compose généralement d'une ou deux essences coexistantes, les plantations se faisant avec des arbres de même âge selon un espacement régulier (**Fig. 1-4**).

Ces forêts de plantation représentent 131 million d'ha, soit 45% des forêts plantées et environ 3% de la superficie forestière mondiale (FAO 2022). L'Asie possède la plus grande superficie de plantations forestières avec 79 millions d'ha, suivie par l'Amérique du sud et l'Amérique

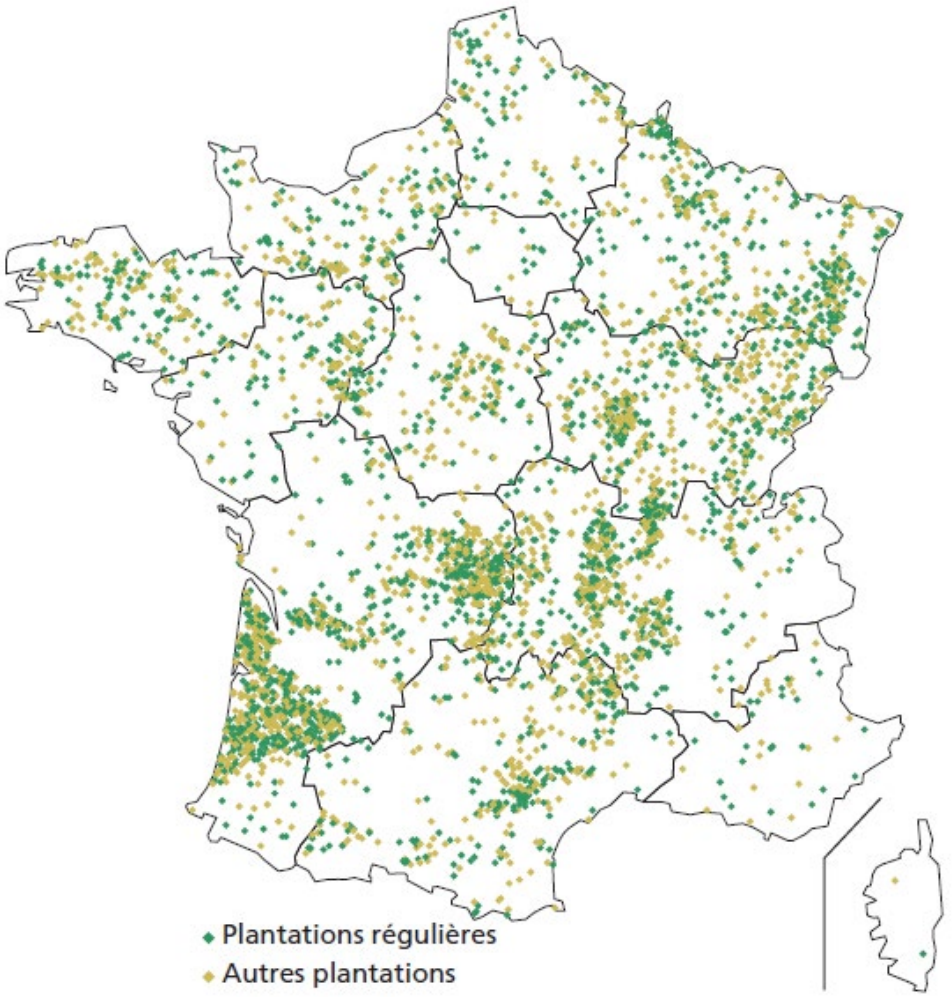



*Figure 1-5 : Distribution des forêts de plantation en France (source : IGN 2017)*

centrale et du Nord avec respectivement 20 et 15 millions d'ha. La superficie mondiale de plantation a augmenté de 55.8 millions d'ha entre 1990 et 2020, principalement grâce aux plantations massives effectuées en Chine qui réalise à elle seule plus de la moitié des boisements du monde. La hausse des surfaces forestières de plantation tend cependant à se réduire au fil des années.

Si l'Europe est le continent qui affiche le plus de forêts plantées (71 millions d'ha), le nombre d'ha en forêt de plantation n'est que de 4.5 millions. Après une forte augmentation entre 1990 et 2010, ces surfaces de plantation ont connu un léger recul de 17 700 ha entre 2010 et 2020.

### *Caractéristiques des forêts de plantation en France*

La France est le 4$^{e}$ pays européen en terme de surface forestière. En expansion depuis les années 1850, la forêt française recouvre aujourd'hui 31% du territoire métropolitain, soit 16.5 millions d'ha. Les forêts de plantation couvrent 2.1 millions d'ha représentent 13% des surfaces forestières en France (IGN 2017). A l'exception des Landes, les principales zones de plantation se situent dans la moitié nord du territoire français, allant de la Bretagne jusqu'aux Alpes en passant par la Sologne, le Morvan, le Jura et les Vosges (**Fig. 1-5**).

Environ 70% de la forêt française est privée et détenue par plus de 3 millions de propriétaires. Les 30% restants sont gérés par les collectivités territoriales ou l'Etat. Ce morcèlement rend la gestion forestière complexe et pose des difficultés d'accès et de logistique. Ainsi, seules les parcelles de plus de 10ha, réparties entre 200 000 propriétaires, sont considérées comme des forêts gérables et utiles pour la production de bois, soit une surface totale de 11.4 millions d'ha.

Les essences feuillues représentent plus de 66% des surfaces forestières, avec en premier lieu des chênes pédonculés et sessiles. Les forêts de résineux, dominées par les pins maritime et sylvestre, représentent 12% des forêts françaises tandis que les 22% restants correspondent à des forêts mixtes. La production biologique brute des arbres en France est estimée à 88 $Mm^3$/an sur la période 2012-2020 tandis que le prélèvement annuel représente environ 50 $Mm^3$ (IGN 2022).

Si l'ensemble du secteur industriel français compte pour 12.7% du PIB, le secteur forestier compte pour 1.1%, soit environ 25 milliards d'euros. Cette filière génère 400 000 emplois directs, soit plus que dans les secteurs du nucléaire ou de l'aéronautique (AAF 2014a). Malgré

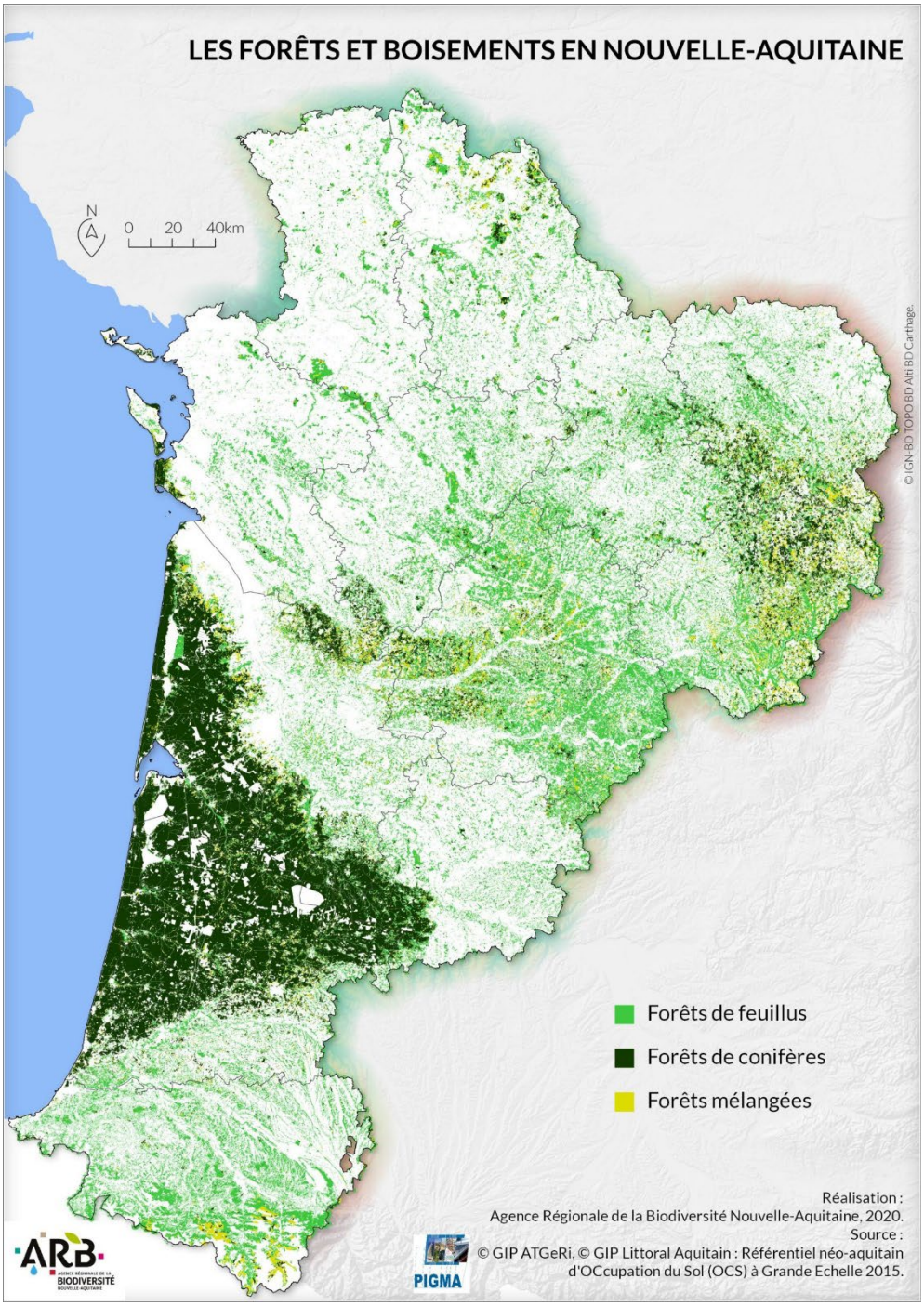



*Figure 1-6 : Forêts et essences en Nouvelle-Aquitaine*

une diminution au cours des 5 dernières années, la filière forêt-bois représente encore près de 10% du déficit de la balance commerciale (AAF 2014b). L'exportation de bois ronds et panneaux ne comble pas les importations massives en produits forestiers transformés tels les pâtes à papier, les cartons ou l'ameublement.

### 1.1.3. Les défis à relever pour les forêts de plantation : exemple de la forêt des Landes de Gascogne

#### *Une forêt de plantation hors norme pour relever le défi de la productivité*

La forêt des Landes de Gascogne est la plus grande forêt de plantation en Europe. Elle recouvre 1.3 millions d'ha sur les départements de la Gironde, des Landes et du Lot-et-Garonne (**Fig. 1- 6**). On y retrouve les caractéristiques classiques des massifs forestiers de plantation, avec la présence d'une essence ultra-majoritaire, le pin maritime, qui représente environ 90% des plantations. Cette monoculture permet de récolter chaque année entre 5 et 6 Mm3 de bois (FIBOIS 2022), utilisé pour les charpentes, la menuiserie de bâtiment, l'emballage et la papeterie. Cette forêt assure un quart de la production de bois en France alors qu'elle ne représente que 7% (0.8 millions d'ha) de la surface forestière nationale. Le morcellement modéré de cette forêt entre propriétaires privés et l'accès généralement facile aux parcelles ont simplifié la mise en place de ce système de production intensif. Ce massif est intégré dans une filière générant chaque année plus de 3 milliards d'euros de chiffre d'affaires et mobilise 34 000 emplois directs (FBF 2019), soit près de 2% de l'emploi salarié dans la région (Agreste 2021).

#### *Le défi du changement climatique*

Le réchauffement climatique est la première menace qui pèse sur les forêts de plantation. En tant qu'espèces pérennes, les arbres vont expérimenter différents environnements au cours de leur existence. Face à la rapidité des bouleversements climatiques, les mécanismes naturels d'évolution et d'adaptation des arbres pourraient s'avérer insuffisants (Kremer et al., 2014). La hausse des températures combinée à une baisse de la pluviométrie engendre des situations de stress hydrique qui affectent le fonctionnement des arbres et ralentit leur croissance (Rennenberg et al., 2006). L'augmentation en fréquence des sécheresses peut conduire à des taux de mortalité conséquents (Thomas et al., 2004). Egalement, des hivers trop doux sont à même de provoquer des perturbations dans le processus de dormance des bourgeons et des

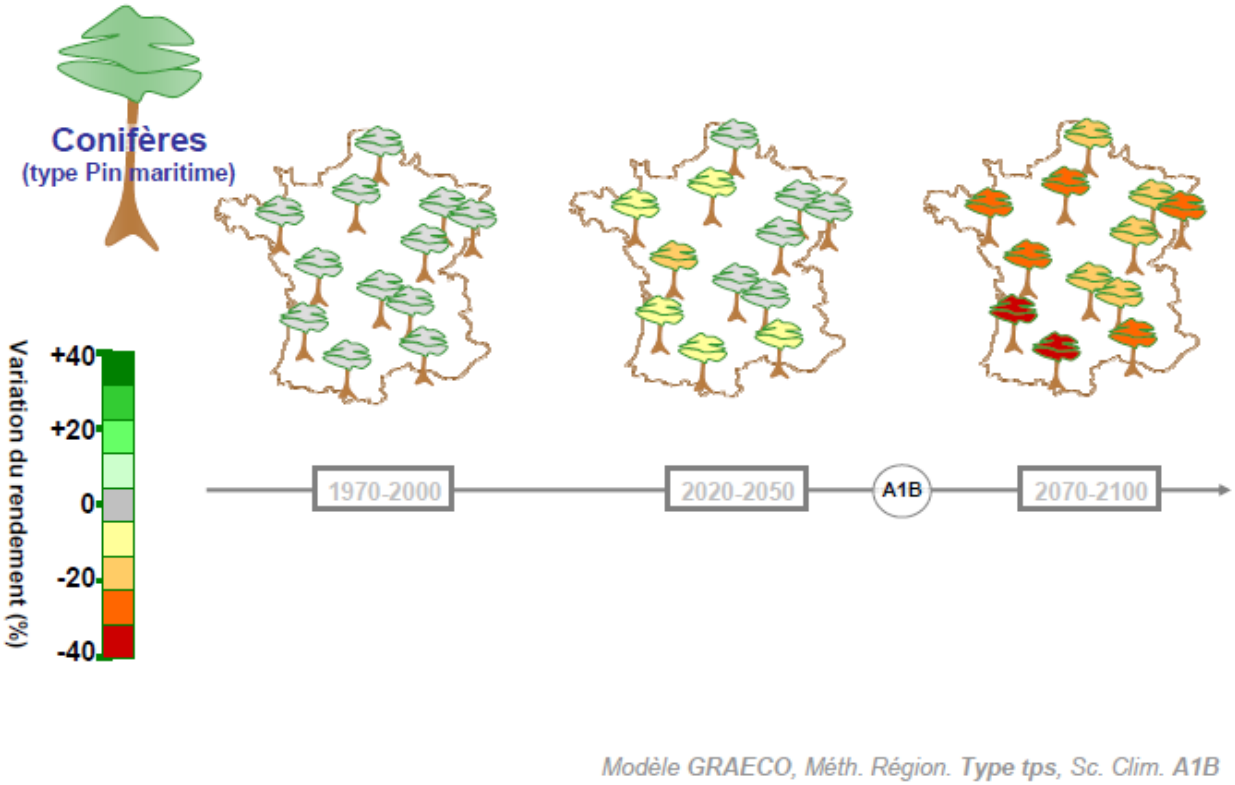



***Figure 1 7 : Evolution du rendement du pin maritime en France : modèle GRAECO ; méthode de régionalisation de type tps, scénario climatique A1B (source : A. Bosc 2011)***

***Figure 1-8 : La forêt des Landes touchée par des attaques de scolytes en 2010 (source : AFP/Archives - Nicolas Tucat)***

graines (Campoy et al., 2011). L'ensemble de ces pressions environnementales pose question quant à la résilience et la productivité des forêts de plantation.

Dans la forêt des Landes, le pin maritime a initialement été choisi pour ses capacités d'adaptation aux sols landais acides et pauvres et pour sa tolérance à la fois aux sécheresses estivales et à l'engorgement hydrique hivernal. Toutefois, le GIEC prévoit une hausse des températures de +1.5°C dans le sud-ouest de la France d'ici à 2050 (scénario « moyen » A1B), avec dans le même temps une diminution de 100mm du niveau moyen annuel des précipitations (Terray et al., 2010). Ainsi, c'est une baisse de 10% à 20% de la productivité du massif aquitain qui est prévue en 2050 à cause du changement des conditions environnementales (**Fig. 1-7**, Mora et al., 2012).

A cela s'ajoute une augmentation de l'occurrence d'aléas climatiques extrêmes, tels que les tempêtes ou les périodes de sécheresse prolongées, et d'autres phénomènes tout aussi dramatiques mais plus ou moins liées au climat, tels que les incendies, auxquels les forêts sont particulièrement vulnérables. Si des actions de prévention et de gestion active sont mises en place, le niveau de contrôle du risque reste faible. La forêt des Landes a notamment subi deux tempêtes de grande ampleur en 1999 (tempête Martin) et 2009 (tempête Klaus). En 1999, ce sont 130 000 ha qui ont été détruits à plus de 50% (IFN 2003), tandis qu'en 2009 ce sont environ 600 000 ha touchés dont un tiers à plus de 40% (IFN 2009). De fortes tensions sur la ressource en bois sont apparues à la suite de ces deux tempêtes. La forêt de Landes a également été marquée par de terribles incendies dans les années 1940 ravageant 40% de la surface du massif (Traimond, 1980). Plus récemment, 30000 ha ont brûlé lors des incendies de 2022 partis des communes de Landiras et de La Teste-de-Buch (Région Nouvelle-Aquitaine 2022). Même s'ils sont d'origine humaine, la situation de sécheresse associée à des très fortes températures a rendu la végétation particulièrement inflammable.

### *L'augmentation des pressions biotiques*

Les dégâts occasionnés par les tempêtes et incendies favorisent les attaques de bioagresseurs. Après la tempête de 2009, près de 4 millions de $m^3$ de bois ont été perdus suite à des attaques de scolytes (DSF Aquitaine 2010). Le bois couché au sol a favorisé le développement de cet insecte qui s'est ensuite attaqué aux arbres encore sur pied (**Fig. 1-8**). De manière plus générale, les nouvelles conditions environnementales liées au changement climatique, et notamment les températures élevées, tendent à faciliter la propagation de champignons pathogènes et

d'insectes ravageurs dans les forêts, notamment vers des zones où les arbres n'ont pas encore développé de défenses (Candau, 2008; Nageleisen, 2018).

Dans les Landes, des attaques de chenilles processionnaires (*Thaumetopoea pityocampa*), de pyrales du tronc (*Diorycria sylvestrella*) ou encore des maladies fongiques (rouille courbeuse, armillaire, fomes) sont par exemple régulièrement identifiées sur le pin maritime. En tant qu'essence locale, ce dernier démontre une certaine résistance face à ces attaques, avec une propagation limitée au sein de la forêt. Une menace majeure plane toutefois sur le massif aquitain avec l'arrivée du nématode (*Bursaphelenchus xylophilus*) en Europe. Originaire d'Amérique du nord, ce ver microscopique est actuellement présent au Portugal et localement en Espagne (Sousa et al., 2011) . Attaqués par le nématode, les pins maritimes perdent leurs aiguilles et meurent en quelques semaines. Le nématode est véhiculé par un insecte vecteur, le coléoptère *Monochamus galloprovincialis,* qui est présent dans le sud-ouest de la France. Les mesures préventives de quarantaine sont à l'heure actuelle le seul moyen pour restreindre les propagations.

### 1.1.4. Vers l'amélioration génétique pour assurer l'avenir des forêts de plantation

Les progrès dans les pratiques sylvicoles et l'utilisation de variétés issues des programmes d'amélioration génétique ont permis d'améliorer les rendements des massifs forestiers tout en limitant certaines pressions biotiques. Historiquement, ces programmes d'amélioration chez les arbres forestiers se sont focalisés sur la quantité et la qualité du bois fourni. Les critères encore utilisés aujourd'hui pour la sélection du pin maritime sont le volume, estimé par des mesures de circonférence et de hauteur, et la rectitude basale du tronc, qui conduit à éliminer le bois de compression et faciliter l'exploitation lors du sciage. Ces critères visent à répondre en partie aux exigences de la filière forêt-bois. Ainsi, la plantation de variétés améliorées dans les Landes amène aujourd'hui à un gain en volume et en rectitude de +30-40% par rapport à des lots non-améliorés (GIS PMF 2014). D'autres paramètres relatifs à la qualité du bois peuvent être intégrés en sélection, comme sa densité moyenne, corrélée à sa résistance mécanique et à son rendement en pâte à papier, ou encore sa qualité de branchaison, déterminée afin de limiter le nombre et la taille des nœuds dans le bois.

L'amélioration génétique apparait comme un levier majeur pour permettre aux forêts de plantation de relever les défis liés à la diversification des usages du bois mais aussi face au changement climatique. Les critères de sélection évoluent sans cesse et vont nécessiter une adaptation rapide des schémas d'amélioration. Bien que complexe à caractériser, la tolérance à

la sécheresse et la résistance aux bioagresseurs sont amenées à devenir des critères majeurs en amélioration forestière.

## 1.2. Amélioration génétique chez les arbres forestiers

### 1.2.1. Approche théorique de la sélection

#### *Principes de base de la génétique quantitative*

L'amélioration génétique repose en grande partie sur la discipline de la génétique quantitative. La compréhension de ce processus d'amélioration nécessite donc la présentation de certains éléments de cette discipline. Focalisons-nous d'abord sur un caractère évalué dans une population d'arbres forestiers, tous de la même espèce et de la même génération. La distribution des valeurs phénotypiques observées au sein de cette population peut être discrète ou continue. Dans le premier cas, on parle de caractère qualitatif. Ce type de caractère est souvent contrôlé par un nombre limité de gènes. Dans le second cas, le caractère est dit quantitatif et généralement contrôlé par un très grand nombre de gènes à effet faible. La majorité des caractères étudiés en amélioration forestière, comme la hauteur, la circonférence ou la rectitude, sont des caractères quantitatifs. D'après le modèle polygénique infinitésimal (Fisher, 1918) qui s'applique pour ce type de caractère, la valeur phénotypique observée résulte d'une composante génétique et d'une composante environnementale (Falconer & Mackay, 1996) :

$$P = G + E \tag{1.1}$$

Où P désigne la valeur phénotypique, G la valeur génotypique[2] et E la déviation environnementale. Au niveau populationnel, en supposant les termes G et E indépendants et sans interaction, on peut écrire : $\sigma_P^2 = \sigma_G^2 + \sigma_E^2$, avec $\sigma_P^2$, $\sigma_G^2$ et $\sigma_E^2$ les variances phénotypique, génotypique et environnementale, respectivement.

La valeur génotypique G peut elle-même être décomposée en une valeur additive (A), de dominance (D) et d'épistasie (I), qui peuvent être considérées comme indépendantes les unes des autres (Falconer & Mackay, 1996) :

$$G = A + D + I \tag{1.2}$$

La valeur additive A résulte des effets additifs des allèles, D de l'interaction entre allèles au même locus, et I de tous les autres interactions alléliques, notamment celles entre allèles à des locus différents. Ainsi, la variance génotypique peut être décomposée par $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$, avec $\sigma_A^2$, $\sigma_D^2$ et $\sigma_I^2$ les variances additive, de dominance et d'épistasie. Le sélectionneur

[2] Un génotype est considéré ici comme une combinaison génétique unique trouvée dans un seul individu ou dans plusieurs copies végétatives génétiquement identiques.

s'intéresse tout particulièrement à la valeur A, aussi appelée « breeding value », qui représente la part du mérite génétique transmise de manière additive à la descendance (chaque parent transmet la moitié de sa valeur additive à ses descendants) ainsi que la part d'épistasie additive-additive.

Pour quantifier la proportion de la variation d'un caractère au sein d'une population qui est due à des différences génétiques de type additif, le paramètre d'héritabilité ($h^2$) est classiquement utilisé en génétique quantitative. Il est calculé au sens strict par :

$$h^2 = {\sigma_A^2}\Big/{\sigma_P^2} \tag{1.3}$$

Cette mesure standardisée de la variabilité génétique intervient dans l'équation du sélectionneur (voir ci-dessous l'équation **1.5**) afin d'estimer le gain génétique (R) à partir du différentiel de sélection (S) (Falconer & Mackay, 1996) :

$$R = S \times h^2 \tag{1.4}$$

Le gain génétique ou réponse à la sélection R est défini comme la différence entre la moyenne phénotypique des descendants et celle de la population initiale à laquelle appartiennent les parents. Quant au différentiel de sélection S, il s'agit de la différence entre la moyenne des individus sélectionnés comme parents et celle de la population initiale à laquelle ils appartiennent. Ce dernier différentiel peut être décomposé sous la forme $S = i \times \sigma_p$, avec i l'intensité de sélection. L'intensité de la sélection est donc le différentiel de sélection S exprimé en écart-type phénotypique. Elle peut aussi être calculée à partir du taux de sélection (rapport entre le nombre d'individus sélectionnés sur le nombre total d'individus évalués). Le gain génétique attendu par unité de temps $R_t$ s'exprime finalement sous la forme :

$$R_t = \frac{i \times h \times \sigma_A}{t} \tag{1.5}$$

Dans cette équation, communément appelée équation du sélectionneur, $t$ désigne la durée d'un cycle de sélection et $h$ la racine carrée de l'héritabilité correspond à la précision d'approximation des valeurs génotypiques à partir des observations phénotypiques (elle est aussi notée $r$). L'intensité de sélection $i$ et la variance génétique additive apparaissent clairement comme des potentiels de sélection.

## *Réponse à la sélection chez les arbres forestiers*

La durée des cycles de sélection (t) est très longue chez les arbres forestiers (**Fig. 1-9**). La maturité sexuelle, c'est-à-dire l'âge à partir duquel les arbres sont en mesure de se reproduire par voie sexuée, est d'environ 6 ans chez le genre *Eucalyptus*, 8 ans chez le pin maritime. Même si des croisements sont théoriquement possibles à partir de cet âge, la sélection des reproducteurs se fait généralement sur des caractères mesurés plus tardivement afin de maximiser les corrélations avec les caractères ciblés à l'âge de coupe. Chez le pin maritime, les critères de sélection correspondent à la rectitude basale du tronc et au volume, mesurés respectivement à 8 et 12 ans, et qui apparaissent très bien corrélés avec les caractères finaux d'intérêt économique visés à l'âge de coupe à 40-45 ans (Kremer, 1992; Zas et al., 2004). Ces caractères sont généralement très polygéniques et présentent des héritabilités faibles, comprises entre 0.1 et 0.4 pour le volume (fonction de la croissance en hauteur et du diamètre) et la rectitude (Pâques, 2013). Ces paramètres t et h interviennent directement dans l'équation du sélectionneur et vont affecter le gain génétique réalisable par unité de temps.

La diversification des critères de sélection peut également impacter la réponse à la sélection. La sélection pour un caractère 1 induit une réponse à la sélection dite corrélée ($CR_2$) pour un caractère 2 :

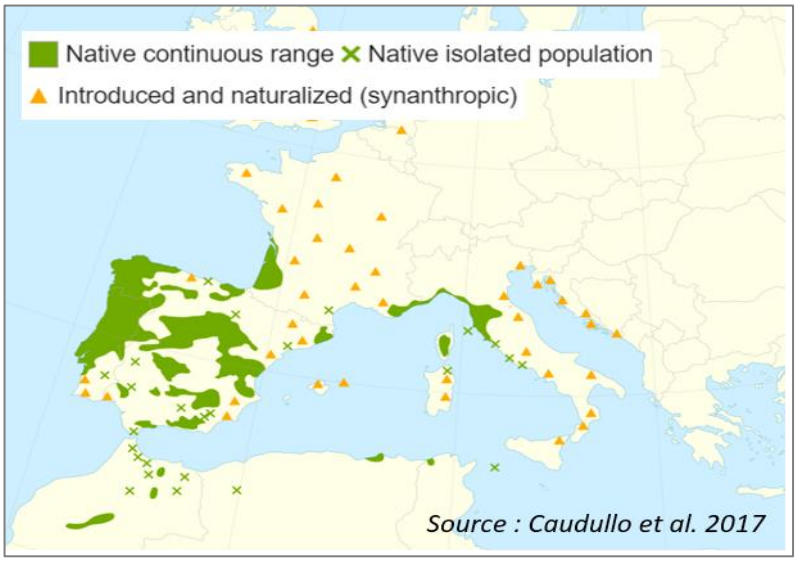
$$CR_2 = i \times (h_1 \times h_2 \times r_A) \times \sigma_{P2}^2 \qquad (1.6)$$

avec $h_1$ et $h_2$ les racines carrées des héritabilités du premier et du second caractère respectivement. Le terme $r_A$ correspond à la corrélation génétique additive entre les deux caractères. Cette corrélation peut être non nulle et défavorable, c'est-à-dire antagoniste entre les deux caractères, une augmentation de l'un entraînant une diminution de l'autre. Dans ce cas défavorable, une sélection simultanée sur les deux caractères corrélés pour, par exemple, augmenter leurs valeurs sera moins efficace qu'une sélection individuelle sur chaque caractère. C'est le cas notamment pour le pin maritime, chez qui les caractères de volume de bois et de rectitude basale du tronc présentent généralement une corrélation défavorable. Malgré ces différentes contraintes, les stratégies de sélection en amélioration forestière ont démontré une grande efficacité au cours des dernières décennies (Pâques, 2013) car les variétés améliorées permettent des gains génétiques très significatifs pour les caractères de volume et de qualité du bois.

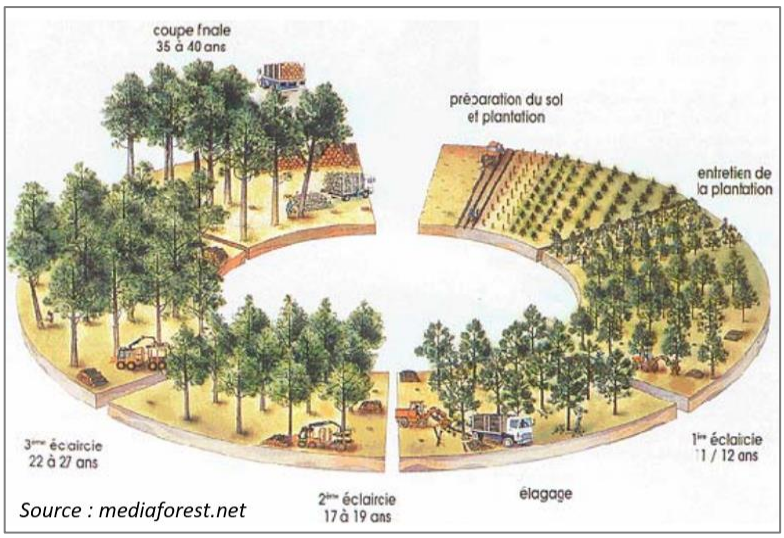
***Encadré 1-1: Le pin maritime dans les Landes de Gascogne (source : CNPF Nouvelle-Aquitaine & GIS PMF)***

Le pin maritime (*Pinus pinsater* Ait) appartient à l'embranchement des confières et au genre *Pinus*. Il s'agit d'une espèce monoïque et anémogame (pollen principalement dipersé par le vent). Son aire de distribution naturelle s'étend en Europe sur la façade atlantique (France, Espagne, Portugal) et le contour méditerranéen (France, Espagne, Italie), mais aussi dans le nord du Maghreb (Maroc, Tunisie). Introduit en Australie, Nouvelle-Zélande, Chili et Afrique du sud, il recouvre aujourd'hui 4.4 millions d'hectares à l'échelle mondiale. La diversité des conditions environnementales entre ces différents pays témoigne de la remarquable capacité d'adaptation du pin maritime.

A partir du 19e siècle, le pin maritime est massivement planté dans le sud-ouest de la France, afin de contribuer à la fixation des dunes, à l'assainissement des marécages et à l'amélioration des conditions d'hygiènes. Il devient largement dominant dans les Landes de Gascogne qui abandonnent alors leur système agro-pastoral. Cette essence autochtone apparait comme la seule adaptée aux sols landais sableux, acides et pauvres, et pouvant être humides ou secs. Le pin maritime est également capable de supporter à la fois la sécheresse estivale et l'engorgement hivernal. Dans des conditions de bonne fertilité, sa croissance est rapide et permet une production efficace de bois.




Les Landes de Gascogne correspondent aujourd'hui à une monoculture de pin maritime organisée en futaies régulières. La plantation se fait généralement après un travail du sol et une fertilisation en phosphore. Des plants âgés d'environ 6 mois sont installés avec un espacement régulier amenant à une densité d'environ 1250 arbres/ha et garantissant un peuplement homogène. Entre trois et quatre éclaircies sont réalisées au cours de la vie du peuplement afin de limiter la compétition entre les arbres. A l'âge de coupe entre 40 et 50 ans, la densité finale est généralement de 300 arbres/ha.




Le programme d'amélioration du pin maritime est géré par le groupement d'intérêt scientifique (GIS) « Pin Maritime du Futur ». En plus de l'INRA et du FCBA (Institut technologique) qui ont débute la sélection du pin maritime dès 1960, ce groupement inclue l'ONF (Office National des Forêts), le CPFA (Centre de Productivité Forestière d'Aquitaine) et le CRPF d'Aquitaine (Centre Régional de la Propriété Forestière Aquitaine).

### 1.2.2. Organisation des programmes d'amélioration chez les arbres forestiers

Bien que les programmes d'amélioration forestiers puissent varier dans leur organisation, l'exemple de la sélection récurrente chez le pin maritime est bien représentatif des pratiques couramment employées, en particulier chez les conifères (**Encadré 1-1**).

#### *Stratégie de sélection récurrente dans la population d'amélioration*

La stratégie de sélection récurrente est largement adoptée pour l'amélioration forestière (Bouvet et al., 1992; Durel, 1992; Pichot & Teissier Du Cros, 1988). Cette stratégie repose sur des cycles successifs composés chacun d'une phase d'évaluation, de sélection et de croisements, afin d'enrichir progressivement la population d'amélioration en allèles favorables pour les caractères ciblés (Namkoong et al., 1988). Ces schémas sont très utilisés notamment chez les conifères qui dominent les forêts de plantation en Europe.

Les programmes d'amélioration débutent généralement par une sélection massale d'individus au sein de peuplements non-améliorés. Ces arbres remarquables ou arbres « plus » constituent la population de base ou les individus fondateurs du programme.
Pour le pin maritime, cette population a été obtenue dans les années 1960 avec la sélection de 635 arbres « plus ». Ces derniers présentaient une rectitude et un volume au moins 10% supérieurs à ceux de leurs voisins (Illy, 1966). Echantillonnés en balayant l'ensemble de la forêt des Landes, ils sont considérés comme non-apparentés et représentatifs de la diversité génétique du massif. Ils forment la première génération du programme d'amélioration, nommée G0 (**Fig. 1-9**).

Pour chaque génération, l'aptitude générale à la combinaison des arbres (AGC) est évaluée par des tests de descendance *polycross*. Il s'agit de croisements entre une mère identifiée et un mélange de pollens issus plusieurs pères. L'AGC est la performance moyenne d'un parent dans sa descendance en combinaison avec d'autres parents et équivaut à la « breeding value » mentionnée ci-dessus. L'évaluation d'un grand nombre de descendants permet une estimation fiable de l'AGC. Les meilleurs arbres selon ce critère peuvent alors être utilisés comme reproducteurs pour générer une nouvelle génération. Dans ce cas, nous appelons ce type de sélection « *backward* ». Ces parents sélectionnés peuvent ensuite être croisés selon des schémas biparentaux, tout en conservant une grande diversité dans le nombre de parents retenus. Classiquement, c'est un plan de croisement double paire qui est mis en place dans lequel chaque

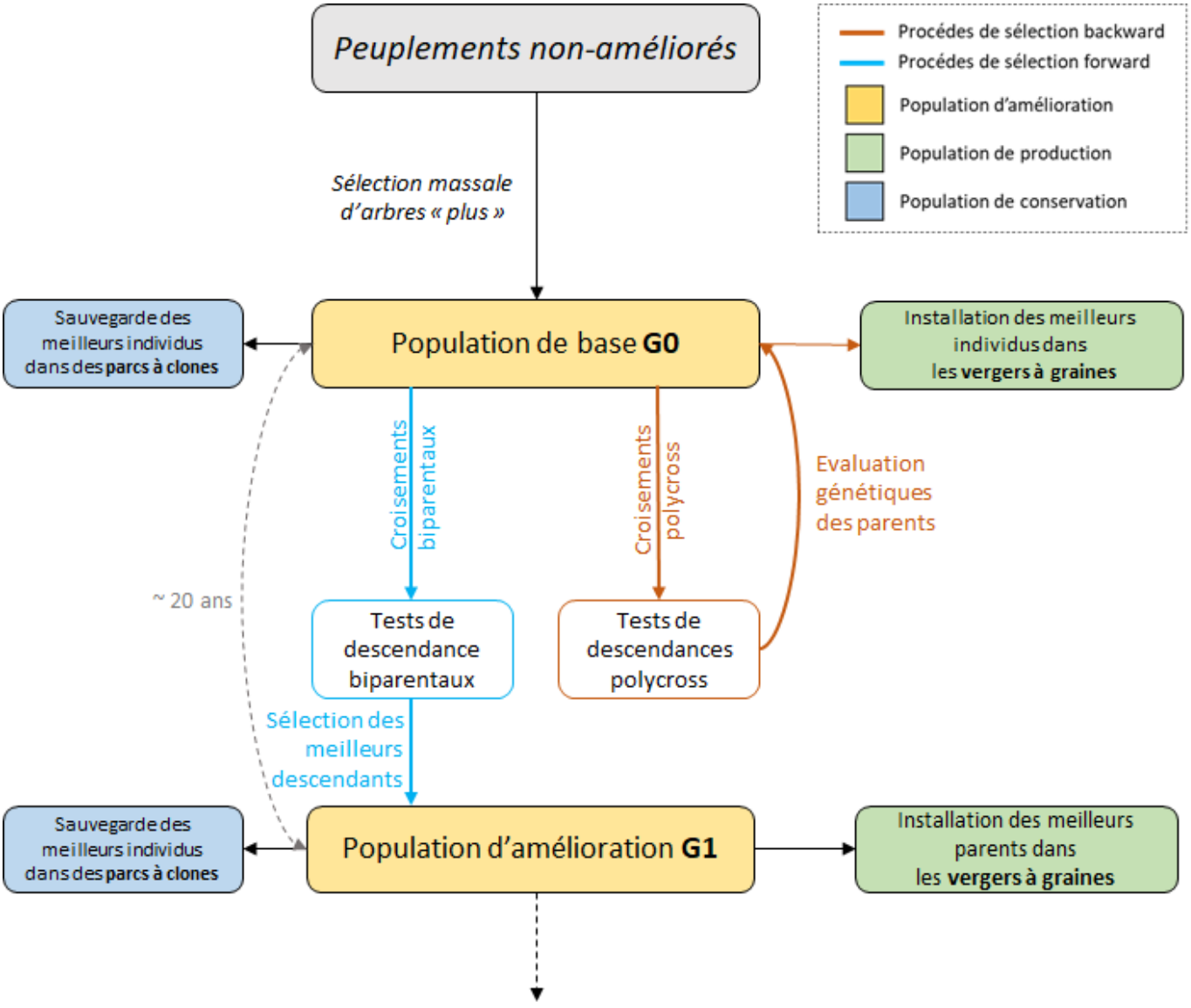



***Figure 1-9 : Schéma simplifié de la stratégie de sélection récurrente chez le pin maritime***

parent contribue à deux croisements différents. L'évaluation génétique des descendants permet d'identifier parmi eux les meilleurs individus qui constitueront la nouvelle génération de la population d'amélioration. Il s'agit ici d'un procédé de sélection « *forward* ».
La deuxième génération d'amélioration pour le pin maritime, nommée G1, est ainsi constitué de 2600 arbres issus des croisements entre 250 individus G0 sélectionnés. Evalués dans un nouveau cycle, c'est environ 400 individus G1 qui ont à leur tour été sélectionnés pour les croisements à l'origine de la troisième génération d'amélioration G2.

### *Déploiement du gain génétique avec les populations de production*

Le mode de déploiement du gain génétique est assez variable en fonction des caractéristiques de chaque espèce. La bonne aptitude à la propagation végétative pour les espèces du genre *Eucalytpus* ou *Populus* amène par exemple à la diffusion de variétés clonales obtenus par bouturage des individus élites. Pour ces deux mêmes genres, des variétés hybrides sont également proposées à la suite de croisements biparentaux contrôlés. Ces variétés sont à même de valoriser des effets génétiques de dominance et d'épistasie en plus des effets additifs de chaque parent.

La difficulté à la propagation végétative chez les espèces conifères et le coût du processus amène à considérer d'autres stratégies pour les sorties variétales. Pour ces espèces, les meilleurs individus issus de la sélection *backward* sont généralement installés dans des vergers à graines. L'installation peut se faire par greffage, ou par semis après une étape de croisement entre ces individus sélectionnés. Cette dernière approche est la plus facile à mettre en œuvre et la moins couteuse (car elle évite le passage par des greffages complexes), toutefois le greffage permet d'obtenir une production de graines plus précoce. Les croisements libres au sein des vergers permettent de produire les graines améliorées. La variété qui en découle est dite synthétique et commercialisée, après élevage en pépinière, sous forme de jeunes plants. Ce type de déploiement ne permet de valoriser que les effets additifs des parents et explique pourquoi seuls les effets d'AGC sont pris en compte lors de l'évaluation des arbres. De plus, il a été démontré notamment pour le genre *Pinus* que les effets d'aptitude spécifique à la combinaison (ASC) étaient limités par rapport aux effets d'AGC (Cotterill et al., 1987).
Trois générations de vergers à graines ont été installées pour le pin maritime, respectivement sur les périodes 1962-1978 (VF1), 1986-1995 (VF2) et 2002-2012 (VF3) (GIS PMF 2016). La génération VF4 est en cours d'installation. Chaque génération est représentée par des vergers implantés sur plusieurs sites pour répondre à la demande très importante en graines. Il n'y pas

de correspondance parfaite entre les générations du programme d'amélioration et les générations des vergers à graines. Le gain, estimé à partir des valeurs génétiques des individus des vergers, augmente en moyenne de +10% dans chaque génération de verger pour le volume et la rectitude par rapport à des lots non-améliorés. Aujourd'hui plus de 90% des plantations dans le massif des Landes se font avec des variétés améliorées de pin maritime issues des verges à graines (GIS PMF 2016).

### *Archivage grâce à la population de conservation*

Les individus sélectionnés dans chaque génération d'amélioration sont conservés par greffage dans des parcs à clones. Chaque individu est en général représenté par une dizaine de plants greffés (ramets). Ces parcs à clones jouent le rôle de réservoir de la diversité génétique ; ils sont utilisés pour constituer les verges à graines et réaliser les croissements contrôlés pour les futures générations.

### *Maintien de la variabilité génétique dans les programmes d'amélioration*

Maintenir la variabilité génétique au fil des générations est clé afin d'assurer le gain génétique sur le long terme (cf équation du sélectionneur). La conservation de la diversité est particulièrement importante dans un contexte de diversification des critères de sélection et de défis environnementaux. Au regard des différentes espèces animales et végétales améliorées, la variabilité génétique des populations d'amélioration forestières est généralement forte. Cela s'explique par la domestication récente des espèces forestières, mais aussi par la grande taille de ces populations d'amélioration. Cette variabilité diminue cependant grandement sous l'effet de la sélection.

Dans les populations d'amélioration, le nombre de reproducteurs sélectionnés à chaque génération impacte la diversité dans les générations suivantes. Les sélectionneurs cherchent donc un équilibre entre la maximisation du gain génétique et la conservation d'un certain niveau de variabilité génétique afin d'assurer un gain génétique sur le long terme. Les 250 individus G0 sélectionnés comme géniteurs dans la population de base du programme d'amélioration du pin maritime constitue un niveau de variabilité initiale très important. Des études moléculaires ont montré que ces individus, y compris les fondateurs et jusqu'aux dernières générations, constitue une population avec une taille effective d'environ 100, ce qui est cohérent avec les recommandations faites pour l'amélioration forestière et même supérieur aux tailles classiques

des populations d'amélioration des autres programmes forestiers (GIS PMF 2014, White et al., 2007). Par l'effort des sélectionneurs, ce niveau de variabilité génétique a été maintenu au cours des générations d'amélioration (Bouffier et al., 2008). Le maintien et même l'augmentation de la variabilité génétique peut être également assurée par l'introduction d'individus issus de nouvelles provenances. Des provenances corses et marocaines du pin maritime ont notamment été identifiées pour leur capacité à supporter la sécheresse.

Alors que la population d'amélioration constitue une forme de réservoir de la diversité génétique, les populations de production, qui ne regroupent que les meilleurs individus de chaque génération, présentent une variabilité génétique beaucoup plus réduite. Les plantations issues des variétés améliorées affichent une homogénéité forte, avantageuse pour les opérations sylvicoles et la valorisation du bois, mais qui peut poser question quant aux capacités d'adaptation des peuplements. Cette question est particulièrement importante dans la mesure où les arbres vont expérimenter un grand nombre de contraintes environnementales au cours de leur longue existence. Si la variabilité génétique est nulle pour des peuplements issus d'une variété clonale comme chez l'*Eucalyptus*, la diffusion de variétés synthétiques chez les conifères assure une certaine diversité génétique. Cette dernière est dépendante du nombre de géniteurs non-apparentés retenus pour la formation des vergers à graines. Compris entre 21 et 65 pour les vergers à graines du pin maritime, ce nombre apparait particulièrement élevé au regard des autres contextes forestiers (GIS PMF 2014). Il doit permettre de limiter la dépression de consanguinité et d'assurer un certain niveau de diversité génétique dans les variétés synthétiques améliorées.

### 1.2.3. Evaluation génétique des individus en sélection

#### *Exploitation de la covariance génétique entre individus*

Selon l'équation **1.1**, la ressemblance entre individus au niveau phénotypique est due à des ressemblances au niveau génétique et/ou au partage d'un environnement commun. Autrement dit, des individus partageant des liens de parenté vont se ressembler pour des caractères héréditaires car ils partagent statistiquement plus d'allèles en commun que deux individus tirés au hasard. La covariance phénotypique entre deux individus i et j s'écrit :

$$cov(P_i, P_j) = A_{ij} \times VarG \qquad (1.7)$$

$P_i$ et $P_j$ désignent les valeurs phénotypiques des individus $i$ et $j$, et $VarG$ la variance génétique. $A_{ij}$ est le coefficient d'apparentement entre les deux individus. Ce coefficient est classiquement calculé à partir des données de pedigree, traçant les parentés entre tous les individus d'un programme d'amélioration. Il traduit la fraction du génome théoriquement partagée entre deux individus et due à une ascendance commune.

A partir d'observations phénotypiques, l'utilisation du modèle mixte (Henderson, 1975) permet d'estimer les 'breeding values' en exploitant ces covariances génétiques entre individus. Ce type de modélisation initialement développé pour l'amélioration génétique animale est aujourd'hui largement utilisé chez les espèces végétales. Utilisée en routine depuis 2011 dans le cadre de l'amélioration du pin maritime, cette approche permet d'évaluer efficacement les candidats à la sélection en valorisant les données phénotypiques et génétiques de plus de 90 dispositifs, soit près de 445 000 arbres répartis sur trois générations.
Les méthodes de calcul de l'apparentement et le modèle mixte sont détaillées dans la **partie 2**.

### *Introduction des marqueurs moléculaires*

L'utilisation du pedigree ne permet toutefois d'estimer qu'un apparentement attendu au regard des probabilités de partage des allèles. Les données de pedigree peuvent de plus être incomplètes ou contenir des erreurs. Les individus issus de croisements *polycross* sont par exemple de père inconnu, ce dernier ne pouvant pas être identifié parmi les différents pères ayant contribué au mélange pollinique (Doerksen & Herbinger, 2010). Des erreurs de notation ou de manipulation interviennent aussi au cours des différentes étapes du programme d'amélioration, par exemple lors de l'étiquetage ou du greffage des plants, et peuvent impacter la précision des valeurs génétiques estimées (Beaulieu et al., 2022; Munoz et al., 2014).

L'introduction des marqueurs moléculaires parait particulièrement valorisable dans ce contexte. Les marqueurs moléculaires sont des variations génétiques détectées au niveau de la séquence d'ADN et servant notamment de balise pour caractériser les génotypes. Ils ont été introduits en amélioration forestière pour reconstruire le pedigree et corriger les erreurs. Chez le pin maritime, Vidal et al. (2015) ont montré que 63 marqueurs moléculaires finement sélectionnés étaient suffisants pour retrouver les paternités de l'immense majorité des individus d'un dispositif *polycross*. Si, dans ce cas, la précision des valeurs génétiques de la population parentale reste similaire, qu'elles soient estimées avec un pedigree incomplet ou reconstruit,

***Encadré 1-2 : L'échec de la sélection assistée par marqueurs chez les arbres forestiers***

La sélection assistée par marqueurs (SAM) consiste à sélectionner les individus à partir de leur génotypage au marqueur. Ce dernier renseigne sur la présence ou non d'allèles favorables pour les traits ciblés. Cette approche se base donc sur une étude préalable des associations entre marqueurs et QTLs. L'information de génotypage peut être connue très tôt, ce qui permet d'envisager une réduction des cycles de sélection tout en limitant les coûts de phénotypage. Cette stratégie de sélection a démontré tout son potentiel au cours des années 2000 notamment chez les bovins (Fritz et al., 2007), le maïs (Moreau et al., 2001) ou encore le riz (Jena & Mackill, 2008).

Chez les arbres forestiers, malgré de nombreuses études de cartographie des QTL publiées à cette même période (Bradshaw & Stettler, 1995; Devey et al., 2004; Markussen et al., 2002), la SAM n'est pas utilisée en routine. Plusieurs raisons expliquent cet échec (Muranty et al., 2014). En premier lieu, la SAM se base principalement sur la détection de QTLs à effet fort. La variabilité génétique capturée est assez limitée, notamment pour les traits très polygéniques contrôlés par un grand nombre de gènes à effet faible. La détection des QTLs n'est pas aisée notamment chez les conifères qui présentent des tailles de génome immenses (20 à 30Gb) et un déséquilibre de liaison faible souvent dans les populations étudiées (Neale & Savolainen, 2004). De plus, les associations entre QTLs et marqueurs ont été détectées à l'aide d'un nombre limité de familles et demeurent spécifiques du fond génétique dans lequel elles ont été mises en évidence. L'exploitation de ces associations est ainsi limitée dans le cadre de programmes d'amélioration exploitant une base génétique large.

l'étape de reconstruction ouvre toutefois la porte à une stratégie de sélection *forward* à partir de croisements *polycross*. La réalisation simultanée des sélections *backward* et *forward* peut s'avérer particulièrement intéressante pour réduire la durée des cycles d'amélioration (Bouffier et al., 2019).

Le développement des approches de génotypage au cours des dernières décennies permet désormais de caractériser à moindre coût les individus en sélection pour un grand nombre de marqueurs moléculaires. La sélection sur la base du génotypage au marqueur (SAM) a été proposée chez les arbres forestiers mais n'a jamais été utilisée en routine (**Encadré 1-2**). Plus généralement, le génotypage dense donne accès à une variation génétique continue et sans a priori sur la position et l'effet des QTL. Ces observations moléculaires permettent de calculer des apparentements « réalisés » entre individus, traduisant la fraction du génome qu'ils partagent réellement. Plusieurs études ont montré que ces apparentements réels suivent en fait une distribution normale centrée sur la valeur d'apparentement théorique obtenue avec les données de pedigree (Isik et al., 2016; Vidal et al., 2015). En captant les processus aléatoires de recombinaison génétique et de partage des allèles, le génotypage sur un grand nombre de marqueurs moléculaires peut ainsi améliorer la précision des valeurs génétiques estimées, et ouvre la voie à la sélection génomique. Face à la demande croissante en bois et aux contraintes liées au changement climatique, cette approche constitue un des piliers majeurs d'innovation qui doit permettre de réduire la durée de sélection et d'intégrer de nouveaux critères plus complexes en amélioration forestière.

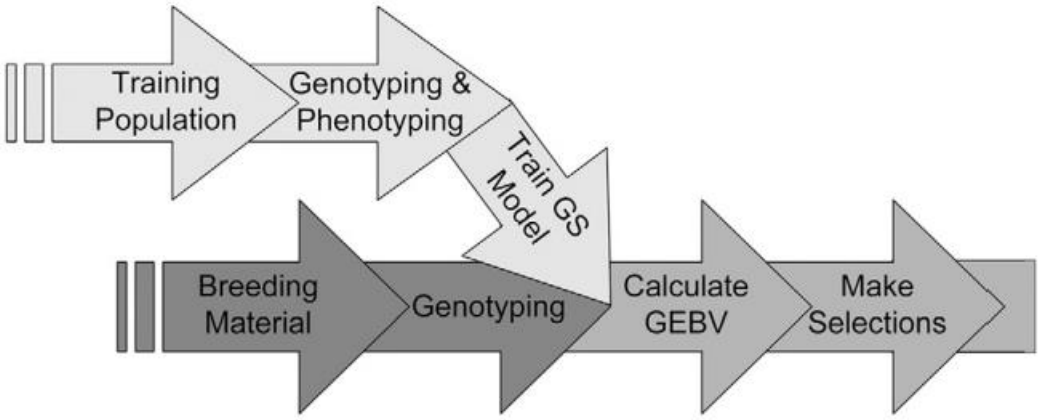



***Figure 1-10 : Principes de la sélection génomique (source : Heffner et al., 2009)***

## 1.3. Développement de la sélection génomique chez les arbres forestiers

### 1.3.1. La sélection génomique

*Principe de la sélection génomique*

La sélection génomique vise à prédire la valeur génétique des individus pour un caractère agronomique d'intérêt à partir de leur information moléculaire (Meuwissen et al., 2001). Chaque individu est caractérisé pour un très grand nombre de marqueurs moléculaires couvrant l'ensemble du génome. Les QTL contrôlant le caractère d'intérêt se trouvent ainsi en déséquilibre de liaison avec un ou plusieurs marqueurs. Contrairement à la SAM, l'objectif est ici d'estimer simultanément l'effet de l'ensemble des marqueurs, afin d'expliquer la variance génétique totale. Si dans le modèle infinitésimal les caractères quantitatifs sont contrôlés par un très grand nombre de QTL à effet faible, cette approche vise à capter l'information de chacun d'entre eux par une saturation du génome en marqueurs (B. Hayes & Goddard, 2010). La valeur génétique estimée à partir des données génomiques d'un individu est noté GEBV (Genome Based Estimated Breeding Values) et est égale à la somme des effets individuels de chaque marqueur pour cet individu. C'est sur la base de ces GEBV prédits que la sélection génomique propose de choisir les meilleurs individus.

La mise en place de la sélection génomique nécessite donc au préalable d'estimer l'effet des marqueurs, autrement dit de construire un modèle de prédiction. Cette étape s'appuie sur une population dite de calibration ou d'entrainement regroupant des individus génotypés et phénotypés pour le caractère d'intérêt. Différentes méthodes statistiques permettent ensuite de faire le lien entre les observations phénotypiques et les profils moléculaires des individus. Le modèle de sélection génomique ainsi construit va être capable prédire les GEBV pour des individus non-phénotypés, uniquement à partir de leurs données de génotypage. L'objectif pour les programmes d'amélioration est donc de construire un modèle de prédiction en utilisant une population de calibration de taille limitée afin ensuite de pouvoir prédire les GEBV pour un grand nombre d'individus en sélection (**Fig. 1-10**).

Deux métriques sont couramment utilisées dans la littérature pour évaluer la performance de la sélection génomique : la capacité de prédiction (« predictive ability » ou « predictive performance ») ou la précision de prédiction (« prediction accuracy »). Elles sont déterminées à l'aide d'une population dite de validation. Les individus de cette population sont génotypés

et phénotypés pour le caractère d'intérêt mais seule l'information de génotypage est utilisée dans le modèle précédemment construit pour prédire les GEBV. La capacité de prédiction est alors égale à la valeur de corrélation entre les GEBV et les phénotypes des individus de la population de validation. La précision de prédiction correspond quant à elle à la corrélation entre les GEBV et les valeurs génétiques vraies, et s'approxime en divisant la capacité de prédiction par la racine carrée de l'héritabilité du caractère (Legarra et al., 2008).

### *Facteurs influençant la précision de prédiction du modèle*

La précision de prédiction (r) intervient directement dans l'équation du sélectionneur (équation **1.5**). Elle doit être maximisée pour assurer un gain génétique fort. Cette précision est une grandeur complexe qui résulte d'une multitude de facteurs dépendants les uns des autres. Toutefois, plusieurs facteurs critiques ont pu être identifiés (Daetwyler et al., 2008; B. J. Hayes et al., 2009):

- L'**héritabilité** et l'**architecture génétique** (il s'agit ici du nombre et des effets des QTL sous-jacents) du caractère ciblé. De nombreuses études ont démontré que les caractères fortement héritables, autrement dit les caractères dont la variance phénotypique s'explique en grande partie pour des effets génétiques additives, sont mieux prédits par les modèles de sélection génomique. S'il est plus difficile de tirer des enseignements généraux concernant l'architecture génétique des caractères, étant donné la complexité de leur caractérisation, souvent simplifiée par le nombre et les effets des QTL, Grattapaglia & Resende (2011) ont montré, par exemple, par simulation que les précisions sont meilleures pour des caractères contrôlés par un nombre faible de QTL, de quelques dizaines de loci.
- La **densité de marqueurs moléculaires**. La densité de marquage doit être suffisante pour que l'ensemble des QTL contrôlant le caractère d'intérêt soit en déséquilibre de liaison avec au moins un marqueur. C'est la taille efficace de la population qui détermine en grande partie l'étendu du DL à l'échelle du génome, et donc le nombre de marqueurs nécessaires (B. Hayes & Goddard, 2010). L'augmentation de la densité de marquage augmente donc la précision de la sélection génomique, mais ce jusqu'à atteindre un plateau maximal. La densité de marqueurs nécessaire pour atteindre ce plateau dépend de la taille du génome et du déséquilibre de liaison dans la population étudiée (Grattapaglia & Resende, 2011).
- La **taille de la population de calibration**. L'augmentation du nombre d'individus inclus dans la population de calibration réduit les variances des effets estimés pour chaque marqueur et permet une meilleure précision de prédiction. Comme pour le nombre de

marqueur, la taille de la population de calibration n'augmente plus la précision de la sélection génomique passé un certain seuil (Grattapaglia & Resende, 2011). La diversité contenue dans la population de calibration par rapport à la diversité totale de la population sélectionnée serait également importante, et intrinsèquement liée à la taille. Une calibration riche en diversité fournira des prédictions plus robustes dans une variété de populations à prédire.

- Le **niveau d'apparentement** entre la population de calibration et la population de validation (Habier et al., 2007). Le déséquilibre de liaison entre marqueurs et QTL doit être similaire entre la population de calibration et la population de validation pour assurer une prédiction cohérente des valeurs génétiques. Plus l'apparentement entre ces deux populations est fort, plus les précisions seront élevées. Le choix des individus constituant la calibration peut être optimisé par différents algorithmes, comme le PEV ou le CD mean (Goddard et al., 2011; Rincent et al., 2012).

### *Méthodes statistiques pour la prédiction génomique*

L'estimation des effets aux marqueurs en sélection génomique s'avère complexe d'un point de vue technique dans la mesure où le nombre de variables explicatives, correspondant au nombre de marqueurs moléculaires utilisés, est généralement très supérieur au nombre de variables à expliquer, c'est-à-dire au nombre d'individus avec une observation phénotypique. Les modèles de régression classique ne sont ainsi pas utilisables dans ces conditions en raison d'un nombre de degrés de liberté bien trop faible pour estimer l'effet de chaque marqueur et de la multi-colinéarité entre ces derniers (Lorenz et al., 2011). Des méthodes de régression pénalisée et des méthodes semi-paramétriques ont alors été proposés. Outre les différences dans les algorithmes implémentés et le type d'inférence (fréquentiste ou bayésienne), ces méthodes diffèrent principalement en termes d'hypothèses sur la distribution des effets aux marqueurs. Le choix de la méthode dépendra donc de la connaissance a priori de l'architecture génétique du caractère d'intérêt, mais aussi du temps disponible et de la complexité de calcul envisageable :

- L'approche *Genomic Best Linear Unbiased Prediction* (GBLUP) est la méthode la plus couramment employée. Elle intègre les apparentements réalisés entre individus (calculés à partir des données de génotypage) dans un modèle mixte pour prédire directement les GEBV des individus sans passer par une estimation des effets aux marqueurs. L'hypothèse sous-jacente de cette approche est que tous les marqueurs utilisés pour le calcul des apparentements sont de même effet faible. La dimension des variables explicatives est ainsi limitée au nombre

d'individus inclus dans le modèle. Cette approche est similaire à l'évaluation génétique décrite en **1.2.3**, à l'exception près que les apparentements utilisés pour le GBLUP sont réalisés et non théoriques. Il apparait donc également possible de réaliser une prédiction des valeurs génétiques à partir d'apparentements théoriques calculés avec les informations de pedigree. Ce type d'approche est alors noté ABLUP. Toutefois, les apparentements théoriques ne prennent pas en compte les déviations dues aux effets de la ségrégation mendélienne, ce qui conduit théoriquement à des précisions de prédiction plus faibles pour le ABLUP que pour le GBLUP. Les modèles ABLUP et GBLUP sont détaillés en **partie 2**.

- L'approche *Ridge Regression Best Linear Unbiased Prediction* (RR-BLUP) s'appuie sur un modèle mixte dans lequel les effets aux marqueurs sont estimés simultanément en tant que variables aléatoires. Cette approche se base sur l'hypothèse que les effets de chaque marqueur sont non-nuls et suivent tous une loi normale centrée sur 0, autrement dit tous les marqueurs sont équitablement pénalisés vers 0.
- L'approche *Least Absolute Shrinkage Selection Operator* (LASSO) est une approche de sélection de variables avec une pénalisation forte qui peut réduire à zéro l'effet de certains marqueurs afin de mettre en avant l'effet d'autres marqueurs.
- Les approches bayésiennes (BayesA, BayesB, BayesC, Bayesian LASSO) permettent d'intégrer à priori des distributions d'effets aux marqueurs plus complexes. Les différences entre les différentes variantes proviennent de la distribution a priori supposée pour ces effets. Ces méthodes bayésiennes sont beaucoup plus lourdes au niveau informatique et nécessitent généralement des temps de calcul significativement plus longs que les approches GBLUP et RR-BLUP, par exemple.

Les différences de précision de prédiction en sélection génomique sont faibles entre ces divers approches (Heslot et al., 2012). L'approche GBLUP, particulièrement adaptée pour les caractères très quantitatifs, est donc généralement privilégiée en raison de sa rapidité.

Notons que d'autres méthodes ont été développées pour prendre en compte les effets génétiques non-additifs dans les prédictions génomiques. C'est le cas de la méthode de régression pénalisée EG-BLUP (extension du GBLUP pour prendre en compte les effets épistatiques), de la méthode semi-paramétrique RKHS (Reproducing Kernel Hilbert Space), ou encore les méthodes basées sur le *Machine Learning* comme le SVM (Support Vector Machine) ou le RF (Random Forest).

### *Application en sélection*

Par rapport à une approche de sélection conventionnelle, la sélection génomique devrait théoriquement permettre :

- D'améliorer la précision (r) des valeurs génétiques en traçant l'ensemble des QTL contrôlant le caractère d'intérêt,
- De réduire l'intervalle de temps (t) entre générations grâce à une prédiction précoce des valeurs génétiques à partir de données génomiques disponibles très tôt dans la vie des individus,
- D'augmenter l'intensité de sélection (i) car le modèle de prédiction permet d'évaluer un plus grand nombre d'individus sans augmenter les coûts de phénotypage

Si le génotypage des individus induit un coût supplémentaire pour la sélection, ce dernier a été drastiquement réduit au cours des 20 dernières années avec l'avènement des nouvelles techniques de séquençage (NGS). La sélection génomique a été implémentée en premier lieu chez les espèces animales et notamment chez les bovins laitiers. Dans les années 2000, l'évaluation d'un taureau à travers les performances laitières de ses descendantes était associée à un coût d'environ 30 000€ (Moreau, 2017). Avec le développement d'une puce de 50 000 SNP[3] (Single Nucleotid Polymorphism) en 2008, le génotypage d'un taureau prédiction de sa valeur génétique, a pu être réalisé pour environ 120€. Ce coût a encore diminué depuis et se situe aujourd'hui entre 20€ et 30€ par individu. L'évaluation d'un plus grand nombre de taureaux a permis d'augmenter l'intensité de sélection, ce qui, combiné à une amélioration de la précision des valeurs génétiques et à la réduction de la durée des cycles de sélection d'environ 3 ans, a amené un doublement du gain génétique chez les bovins laitiers au cours des dernières décennies (Wiggans et al., 2017). La sélection génomique est aujourd'hui implémentée pour une grand nombre d'espèces animales et végétales, notamment chez le maïs et le blé où elle a permis d'augmenter significativement les gains génétiques liés au rendement (Heffner et al., 2010).

[3] Un SNP est une variation génétique qui se produit au niveau d'un seul nucléotide de l'ADN et qui contribue à la diversité génétique au sein des populations.

### 1.3.2. Etat actuel de la sélection génomique chez les arbres forestiers

#### *Des avantages théoriques considérables*

La sélection de la plupart des arbres forestiers n'a débuté que depuis quelques décennies. Les grandes tailles des populations utilisées et les cycles de sélection particulièrement lents résultent aujourd'hui en une domestication relativement faible de ces espèces forestières. De plus, la sélection sur les arbres induit de nombreuses contraintes logistiques, techniques et financières : de très grandes superficies sont nécessaires pour mettre en place les tests de descendance, l'installation et l'entretien des dispositifs expérimentaux est couteuse, les mesures et opérations de croisements sont complexes en raison de la grande taille des arbres.

Si l'ensemble de ces éléments expliquent en partie le « retard » de la sélection forestière par rapport à la sélection animale, ils représentent également le contexte d'amélioration dans lequel la sélection génomique pourrait avoir le plus d'impact (Isik, 2014). Comme pour toutes les espèces pérennes, la réduction de la durée des cycles de sélection doit permettre d'augmenter significativement le gain génétique par unité de temps. Avec un génotypage possible dès le stade plantule, la sélection génomique permet de prédire, dès la première année de vie de l'arbre, sa valeur génétique pour des caractères finaux. Si la durée du cycle de sélection reste conditionnée à l'âge de maturité sexuelle des arbres, atteinte entre 5 et 10 ans selon les espèces, des recherches sur l'accélération de l'induction forale ont également été menées afin de pouvoir réaliser encore plus rapidement les nouveaux croisements (Greenwood et al., 1991; Hasan & Reid, 1995). Réduire la durée des cycles de sélection permet d'assurer une meilleure réactivité face à la demande croissante en bois et face aux changements d'environnements.
La sélection sur une valeur prédite, et donc sans phénotypage des candidats à la sélection, supprime les coûts liés à la mise en place et à l'entretien de tests de descendance, écarte les risques de perte des dispositifs à cause de tempêtes ou d'incendies. Les coûts de phénotypage étant limités à l'évaluation de la population de calibration, un avantage majeur de la sélection génomique réside aussi dans la possibilité d'introduire de nouveaux critères de sélection, plus complexes, et qui auraient été trop coûteux à mesurer en routine sur la totalité des candidats à la sélection. C'est le cas notamment des critères liés aux propriétés physico-chimiques du bois ou à la résistance aux insectes et aux maladies. Un autre avantage de la centralisation de l'évaluation et de la calibration sur une seule entité de population est qu'elle peut être plus facilement mutualisée entre différents acteurs, publiques ou privés, comme c'est le cas pour les

animaux, ce qui permet de réduire les coûts et de partager les avantages d'une population plus importante, sans nécessiter de partage de données entre les partenaires.

Outre ces avantages de prédiction, le passage à la génomique offre plusieurs avantages liés à la richesse de l'information générée. La saturation du génome en marqueurs moléculaires permet de capter par déséquilibre de liaison l'effet de l'ensemble des QTL contrôlant les caractères d'intérêts, même ceux avec un effet très faible (Manolio et al., 2009; Yang et al., 2010). Les approches génomiques apparaissent ainsi particulièrement pertinentes dans le contexte forestier pour trouver la « missing heritability » des caractères de volume très polygéniques, ou pour des caractères complexes de résistance aux maladies dont l'architecture génétique n'a pas pu être identifiée par les études classiques de cartographique des QTL (M. D. V. Resende et al., 2012).

Egalement, les données génomiques doivent permettre une gestion plus explicite de la diversité au sein des programmes d'amélioration forestiers. La seule prise en compte d'une information de pedigree, parfois incomplète et entachée d'erreurs, peut amener à la co-sélection d'individus apparentés et à l'augmentation de la consanguinité dans les populations d'amélioration. Les approches génomiques doivent permettre de représenter la ségrégation mendélienne au sein des familles de pleins-frères et de valoriser la sélection intrafamiliale (Isik, 2014). Cet aspect est particulièrement important chez les espèces végétales et forestières qui présentent généralement de très grands nombres d'individus par famille comparés aux espèces animales (Allier et al., 2019).

### *Un développement récent*

Alors que la période 2000-2010 marque l'avènement de la sélection génomique pour beaucoup d'espèces animales, les ressources génomiques demeurent à ce moment-là encore limitées pour les programmes forestiers. La faisabilité de la sélection génomique pour les arbres forestiers a donc d'abord été étudiée avec des approches de simulations (Grattapaglia & Resende, 2011; Iwata et al., 2011). Par des formules déterministes, Grattapaglia & Resende (2011) suggèrent qu'une densité de marquage comprise entre 2 et 15 marqueurs/cM est suffisante pour obtenir une précision des valeurs génétiques équivalente à celle de l'évaluation génétique conventionnelle basée sur les tests de descendance et les données de pedigree.

Ce n'est qu'au cours des années 2010 que des projets de grande envergure chez les arbres forestiers permettent le développement des premières puces de génotypage, avec principalement des marqueurs de type SNP (Chancerel et al., 2013; Eckert et al., 2010; Geraldes et al., 2013). L'assemblage d'un grand nombre d'échantillons par des groupes communautaires d'utilisateurs a été le facteur clé de la réduction des coûts de génotypage (Grattapaglia 2022).

On dénombre aujourd'hui plus de 80 études de sélection génomique pour une trentaine d'espèces d'arbres forestiers. Deux synthèses récentes sont proposées par Lebedev et al (2020) et Grattapaglia (2022). Ces études concernent en premier lieu les genres *Eucalyptus, Picea* et *Pinus,* qui représentent à eux seuls près de 80% des études. Les caractères étudiés sont principalement liés au volume de bois produit. Toutefois, espèces et caractères d'études tendent à se diversifier avec la publication ces dernières années d'articles sur des arbres forestiers moins répandus comme le frêne (Stocks et al., 2019) ou le cèdre japonais (Nagano et al., 2020), et la considération de nouveaux caractères liés aux propriétés physico-chimiques du bois (Lenz, Nadeau, Azaiez, et al., 2020; Mphahlele et al., 2020) à l'architecture des arbres (Ballesta et al., 2019) ou encore à la résistance aux ravageurs (Beaulieu et al., 2020).

### *Des précisions de prédiction limitées*

Les premières approches empiriques menées chez le pin taeda (Resende Jr et al., 2012; Zapata-Valenzuela et al., 2012) et l'*Eucalyptus* (Grattapaglia et al., 2011; M. D. V. Resende et al., 2012), ont encouragé le développement de la sélection génomique chez les arbres forestiers. Dans ces études, la précision des GEBV est similaire ou légèrement inférieure à celle des valeurs génétiques (EBV) obtenues par la sélection conventionnelle. Cependant, en l'absence de tests de descendance, la réduction de l'intervalle de temps entre générations permise par la sélection génomique assure des gains génétiques par unité de temps nettement supérieurs. Ainsi, pour le pin taeda, Resende Jr et al (2012) suggèrent que la sélection génomique apportera un avantage de 53% à 112% en gain génétique par unité de temps grâce à une réduction par deux de la durée des cycles de sélection.

Rappelons que malgré son attractivité, la sélection génomique nécessite un coût supplémentaire de génotypage. Comme évoqué plus haut, l'approche ABLUP, basée sur l'information de pedigree disponible dans les programmes d'amélioration forestiers, permet également de prédire des valeurs génétiques pour des individus non-phénotypés. Cette approche ABLUP

constitue donc une référence plus appropriée pour évaluer les gains réels apportés par la génomique.

Au regard des études qui ont clairement intégré un approche ABLUP comme point de comparaison, la précision de prédiction des modèles génomiques apparait en fait similaire ou même inférieure à celle des modèles ABLUP. Les exemples sont multiples notamment pour les genres *Pinus* (Bartholomé et al., 2016; Calleja-Rodriguez et al., 2019; Ukrainetz & Mansfield, 2019; Zapata-Valenzuela et al., 2013), *Picea* (Beaulieu, Doerksen, Clément, et al., 2014; Beaulieu, Doerksen, MacKay, et al., 2014, p. 201; Beaulieu et al., 2020; Chen et al., 2018, 2019; El-Dien et al., 2018; Lenz et al., 2017; Lenz, Nadeau, Mottet, et al., 2020), ou *Pseudotsuga* (Thistlethwaite et al., 2017, 2019, 2020). La précision des modèles génomiques apparait supérieure à celle du ABLUP dans quelques études sur les espèces du genre *Eucalyptus* (Estopa et al., 2023; Kainer et al., 2018; Suontama et al., 2019; Thavamanikumar et al., 2020), mais que très rarement chez les conifères (El-Dien et al., 2018; Zhou et al., 2020). De plus, ces gains de précisions sont souvent limités.

Un nombre non négligeable de ces études proposent toutefois des conclusions très optimistes pour la sélection génomique, en l'associant à des cycles de sélection plus courts (Beaulieu, Doerksen, Clément, et al., 2014; Beaulieu, Doerksen, MacKay, et al., 2014; Chen et al., 2018; Lenz et al., 2017; Zapata-Valenzuela et al., 2013). Par exemple, dans l'étude de Chen et al (2018), la précision de prédiction du modèle GBLUP chez *Picea abies* est inférieure à celle obtenue par le modèle ABLUP pour des caractères de hauteur et de qualité du bois. Le calcul de l'efficacité relative de la sélection génomique dans cette étude souligne la moins bonne performance de l'approche génomique. Cependant, ABLUP et GBLUP sont ensuite associés à des durées de cycles de sélection de 25 et 12.5 années, respectivement. Le calcul de l'efficacité relative par année met alors en évidence un net avantage de l'approche génomique et l'article conclut quant à l'attractivité cette dernière.

Ce type de raisonnement ne semble pas cohérent dans la mesure où la prédiction ABLUP peut permettre une réduction de la durée des cycles de même envergure que la prédiction GBLUP. La sélection génomique ne permet une réduction de la durée des cycles de sélection que par rapport à des approches conventionnelles réalisant des évaluations génétiques à partir d'observations phénotypiques sur les candidats à la sélection.

De plus, l'utilisation d'un pedigree incomplet ou non corrigé pénalise les approches ABLUP et fausse quelque peu la comparaison avec les modèles génomiques (El-Dien et al., 2018;

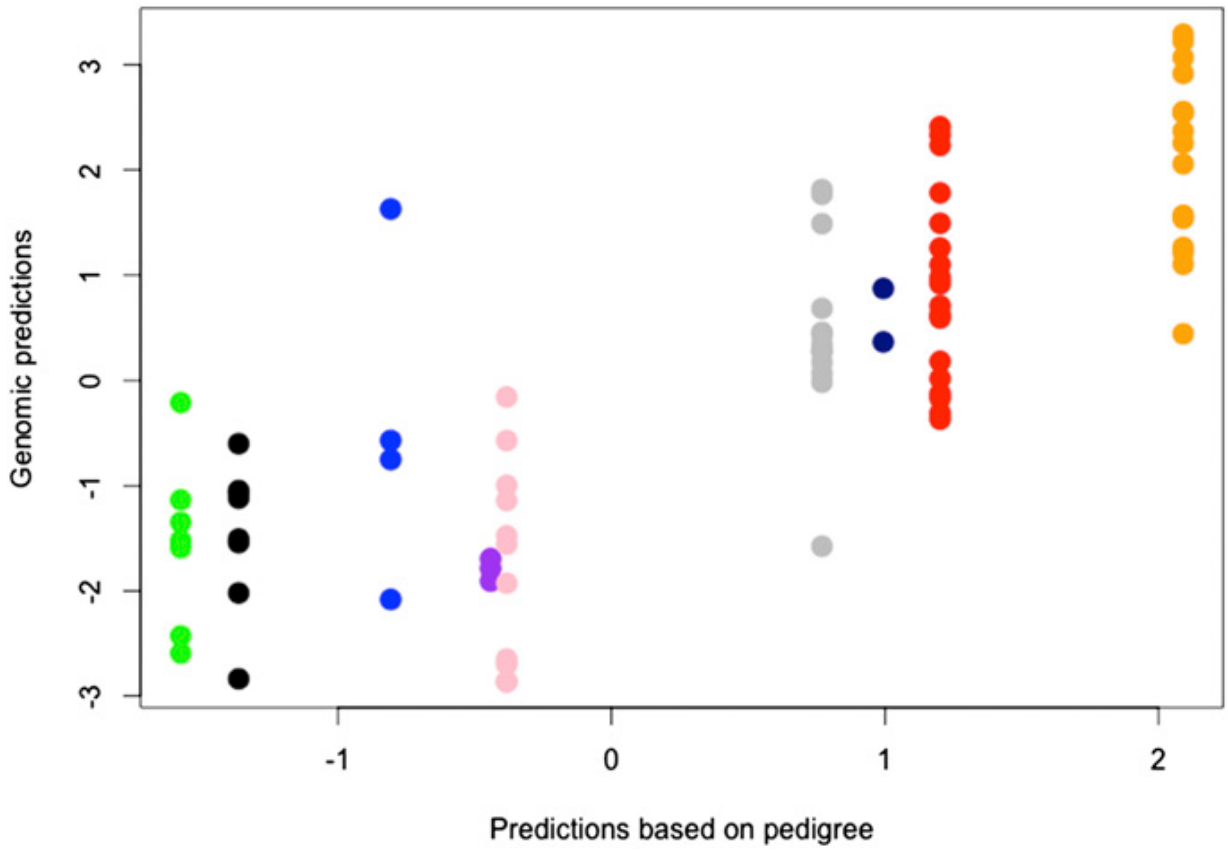



***Figure 1-11: GEBV prédits pour les individus de 9 familles de pleins-frères, en utilisant les données pedigree (axe X) ou les données génomiques (axe Y) (Source : Zapata-Valenzuela et al. 2013).***

Lenz, Nadeau, Azaiez, et al., 2020; Li et al., 2019; Tan et al., 2017). Le gain de précision obtenu par les modèles génomiques peut être assimilé à une reconstruction du pedigree plutôt qu'à un réel avantage dans les capacités de prédiction du modèle (Lenz, Nadeau, Azaiez, et al., 2020). Si cette reconstruction correspond tout de même à un des avantages de la génomique, notons que ce même travail peut être fait avec un nombre faible de marqueurs (Vidal 2017).

Dans ce contexte, il est assez difficile d'affirmer l'avantage de la sélection génomique par rapport aux approches basées sur le pedigree. Plusieurs études affichent des conclusions plus réservées et remettent en cause les tailles des populations de calibration et les densités de marquage, jugées insuffisantes pour démontrer la faisabilité de la sélection génomique chez les arbres forestiers (Bartholomé et al., 2016; Beaulieu, Doerksen, MacKay, et al., 2014; Thistlethwaite et al., 2017).

## *Prédiction intrafamiliale en sélection génomique*

Un avantage théorique majeur des modèles génomiques par rapport au ABLUP réside dans leur capacité à capter la ségrégation mendélienne. Au sein d'une même famille de pleins-frères, le modèle ABLUP attribuera la même valeur génétique à l'ensemble des individus non-phénotypés, à savoir la moyenne des valeurs génétiques parentales. A l'inverse, par la prise en compte d'un apparentement réalisé ou par l'estimation des effets au marqueur, les modèles génomiques prédisent des valeurs génétiques différentes pour les pleins-frères non-phénotypés d'une même famille (**Fig. 1-11**).

Fuentes et al. (2017) ont développé des modèles de sélection génomique chez *Picea stichensis* en se focalisant sur une unique famille de pleins-frères regroupant 500 clones, tous établis sur le même site. Pour des caractères de hauteur et de débourrement, la précision intrafamiliale atteint respectivement 0.38 et 0.54 (corrélation avec le phénotype) lorsque 200 pleins-frères sont inclus dans la calibration. Ces précisions tombent à 0.25 et 0.33 avec 50 pleins-frères dans la calibration. Dans la même idée, Cros et al. (2019) évaluent la précision intrafamiliale chez *Hevea brasiliensis* à l'aide d'une population de 330 clones issus d'un même croisement et répartis sur deux sites. Les meilleures précisions intrafamiliales, comprises entre 0.5 et 0.6, sont atteintes lorsque 175 individus sont inclus dans la calibration. Elles demeurent assez variables en fonction du caractère, du nombre de marqueurs et ou encore du scénario de validation croisée considéré (au sein d'un site, croisé entre les deux sites).

En se focalisant sur une unique famille, ces premières études se placent dans une situation très favorable pour la sélection génomique. Les familles biparentales présentent en effet un

déséquilibre de liaison élevé et une absence de structuration. L'apparentement entre calibration et validation est maximisé. De plus, les précisions les plus intéressantes sont atteintes pour tailles de calibration conséquentes. En pratique, l'évaluation génétique chez les conifères se fait par exemple au moyen d'un grand nombre de familles dont les effectifs ne dépassent généralement pas les 30 pleins-frères (Lebedev et al., 2020). Par opposition à ces modèles famille-spécifique, la sélection génomique chez les conifères requiert plutôt le développement d'une population de calibration intégrant plusieurs familles (Grattapaglia, 2017). A taille de calibration constante, l'augmentation du nombre de familles représentées dans la calibration peut toutefois diminuer les précisions dans la mesure où l'apparentement entre calibration et validation se réduit (Crossa et al., 2017; Lenz et al., 2017).

A notre connaissance, en plus de ces deux premières approches (Fuentes et al. (2017), Cros et al., (2019)), seules trois autres études de sélection génomique intégrant plusieurs familles ont brièvement évalué la précision intrafamiliale dans le domaine forestier. Resende et al. (2017) utilisent une population composée de 856 individus hybrides *Eucalyptus grandis* × *Eucalyptus urophylla* issus de 37 familles de pleins-frères interconnectées. Chaque famille constitue successivement le set de validation. Selon les scénarios et les caractères, la précision intrafamiliale moyenne est comprise entre 0.34 et 0.61. Dans chaque cas, la dispersion des précisions est très importante, ces dernières variant par exemple entre -0.03 et +0.78 dans pour le rendement en pâte à papier (caractère SPY, scenario *CV+Relatedness*, moyenne à 0.47). Chez *Pinus contorta* Douglas, Ukrainetz & Mansfield (2019) utilisent des familles de demi-frères d'un premier cycle de tests pour prédire simultanément les GEBV dans 42 familles de pleins-frères issues d'un second cycle de tests. La précision intrafamiliale moyenne, évaluée dans chaque famille sur environ 19 individus, est comprise entre 0.32 et 0.59 sur l'ensemble des caractères étudiés. Le détail de la précision par famille n'apparait pas. Mentionnons enfin sans détailler l'étude de Pégard et al. (2019) sur peuplier qui affiche des précisions intrafamiliales positives mais encore une fois très variables ('predictive ability' moyennne comprise entre -0.1 et 0.4).

Si ces quelques approches démontrent la possibilité de prédire la variabilité intra-famille en sélection génomique, le faible nombre d'études traitant de cet aspect soulève des questions. L'augmentation des gains génétiques et la meilleure gestion de la diversité en sélection génomique dépendent de cette précision intrafamiliale, car si cette dernière est trop faible, cela revient à faire une sélection aléatoire au sein des familles.

## 1.4. La sélection génomique au défi du changement climatique

Le développement de la sélection génomique est motivé par la possibilité d'accélérer et de simplifier les procédés de sélection. Les gains génétiques attendus dans les années à venir restent toutefois grandement conditionnés à la capacité des arbres à maintenir leur fonctionnement dans un contexte environnemental changeant. Pour assurer la résilience des forêts de plantation, une évaluation génétique intégrant l'information environnementale parait désormais indispensable (A. B. Nicotra et al., 2010).

### 1.4.1. Le phénotype comme fonction de l'environnement

#### *Plasticité phénotypique*

La plasticité phénotypique est définie comme la capacité d'un génotype à produire différents phénotypes en réponse à des variations environnementales (Stearns, 1989). Elle est associée à un phénotype donné, comme la croissance ou la phénologie, et non à un organisme dans son ensemble. Certaines réponses plastiques sont des exemples de plasticité adaptative parce qu'elles procurent un avantage sélectif transmissible à la descendance, tandis que d'autres sont des réponses inévitables à des facteurs physiques ou à des limitations de ressources (Windig et al., 2004). La plasticité phénotypique est sous contrôle génétique (Bradshaw, 1965; Schlichting & Smith, 2002). Plusieurs théories expliquent son fonctionnement. Via et al. (1993) considèrent que les allèles ont des effets variables en fonction des environnements. Un même caractère dans deux environnements différents correspond ainsi à deux caractères distincts selon cette théorie. Les loci impliqués dans le contrôle de ce caractère sont alors vus comme pléiotropiques. Bradshaw et al. (1965) suggèrent plutôt l'idée de l'existence de gènes régulant la forme de la fonction qui relie le phénotype à l'environnement. Ces deux théories sont aujourd'hui considérées comme valides. Aux côtés de procédés d'adaptation par sélection naturelle ou de migration, la plasticité phénotypique apparait comme un moyen d'ajustement des espèces végétales à de nouvelles conditions environnementales (A. B. Nicotra et al., 2010).

La plasticité phénotypique implique qu'il existe plusieurs états phénotypiques par génotype, correspondant soit à différentes mesures prises au fil du temps, soit à différents sites d'essai à un stade de maturation donné. Dans ce dernier cas, nous parlons de plasticité spatiale ou liée au site. C'est la situation la plus courante dans les essais sur le terrain, où les génotypes sont souvent partagés entre les parcelles et les sites. Lorsque les mesures sont répétées dans le temps

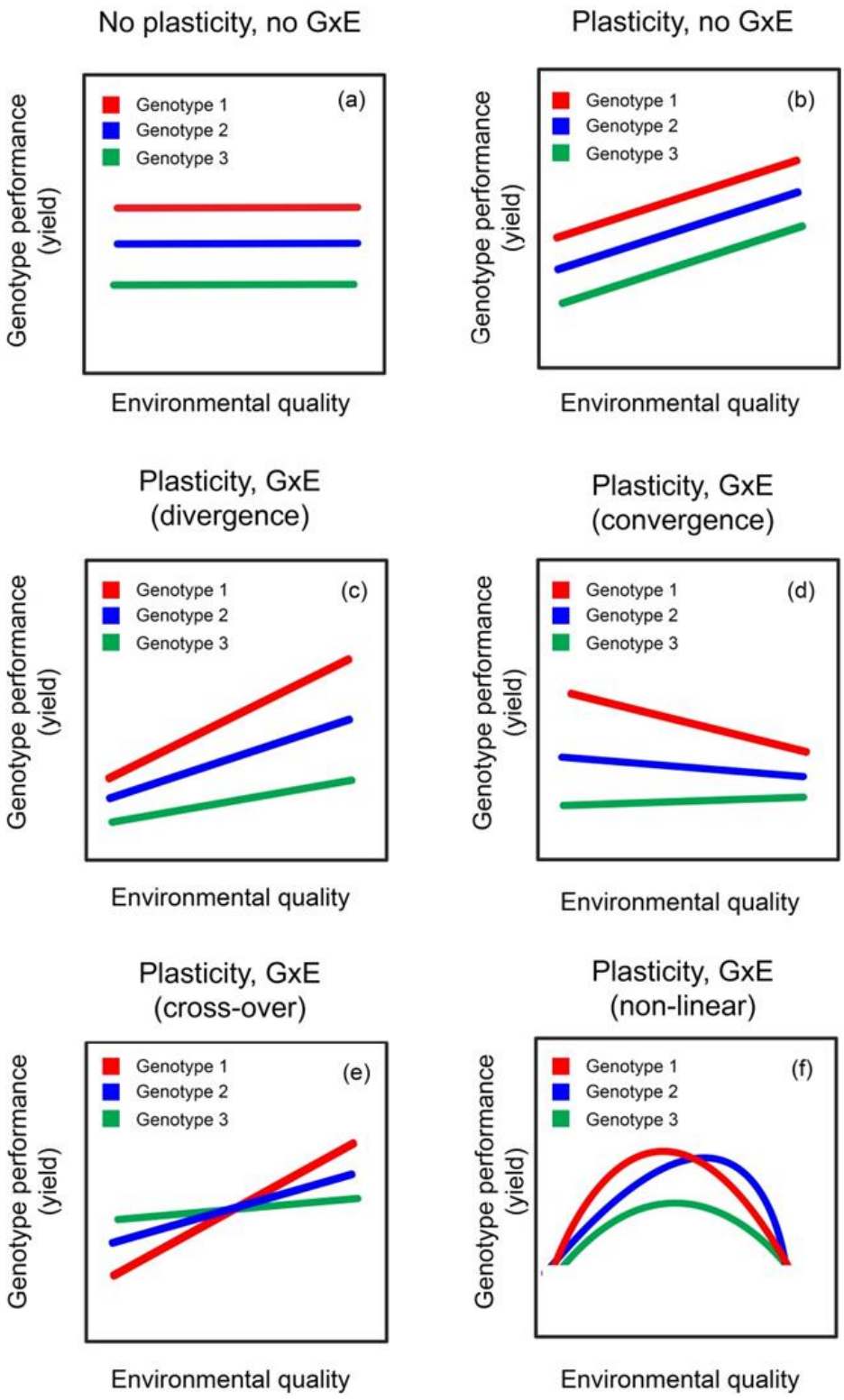



***Figure 1-12 : Normes de réaction pour trois génotypes illustrant plusieurs formes de plasticité et d'interactions GxE (source : van Eeuwijk et al. 2016)***

pour le même individu, la réaction qui en résulte est appelée plasticité temporelle ou liée au temps.

Autrefois considérée comme une variation aléatoire, la plasticité intéresse de plus en plus les sélectionneurs (A. B. Nicotra et al., 2010). La sélection pour l'augmentation de la plasticité phénotypique en soi pourrait permettre de construire une résilience stratégique des espèces cultivées dans des environnements de plus en plus variables (Sambatti & Caylor, 2007). La sélection de plasticité pour l'efficience de l'utilisation de l'eau peut par exemple amener à un meilleur taux de survie et de meilleurs rendements (A. Nicotra & Davidson, 2010). Au-delà de ces aspects fonctionnels, c'est le maintien d'une croissance ou d'une production forte dans une large gamme environnementale qui est aujourd'hui ciblée (Aspinwall et al., 2015).

### *Interactions génotype x environnement*

Revenons sur les hypothèses de base de notre équation P = G + E (**1.1**). Tout d'abord, G et E sont supposés être indépendants, c'est-à-dire que l'expression de G n'est pas conditionnée par le niveau de E. Egalement, cette équation suppose l'absence d'interaction entre les génotypes et l'environnement, autrement dit que les variations d'environnement affectent uniformément l'ensemble des génotypes. Alors que ces hypothèses apparaissent très réductrices au vu du contexte décrit précédemment, l'équation **1.1** peut être étendue pour prendre en compte les réponses différentielles des génotypes aux changements d'environnement :

$$P = G + E + G \times E \tag{1.8}$$

Où le terme $G \times E$ désigne les interactions entre génotypes et environnements (Comstock RE, Moll RH (1963). Les représentations des performances de chaque génotype en fonction des environnements sont appelées 'normes de réaction' et permettent de mettre en évidence différents types d'interactions GxE impliquant ou non un reclassement des génotypes (**Fig. 1-12**). Les interactions GxE sont dites qualitatives lorsque le classement des génotypes change en fonction des environnements, autrement dit lorsque les courbes se croisent (**Fig. 1- 12e, f**). Elles sont dites qualitatives lorsque le classement reste inchangé mais que l'écart entre les performances des génotypes varie (**Fig. 1-12c, d**) (van Eeuwijk et al., 2016).

Les observations phénotypiques répétées dans plusieurs environnements peuvent être analysées en étendant le modèle individuel en un modèle « multi-trait ». Ce type de modèle considère les phénotypes de chaque environnement comme des caractères distincts mais pouvant admettre un certain niveau de corrélation entre eux, notamment au niveau de leur déterminisme génétique

(Falconer & MacKay 1996). Ces modèles permettent ainsi d'estimer des valeurs génétiques individuelles pour les différents environnements testés et même, dans un cadre de sélection génomique, de prédire la performance d'un génotype dans un environnement où il n'a pas pu être observé (Grattapaglia, 2017).

### *La sélection en présence d'interactions GxE*

Les interactions GxE sont essentiellement interprétées par les sélectionneurs comme des manques de cohérence dans les performances relatives des individus lorsqu'ils sont évalués dans des environnements différents (Grattapaglia, 2017). Plusieurs études dans le domaine forestier ont démontré que de fortes interactions GxE biaisent l'estimation de l'héritabilité et peuvent réduire le gain génétique attendu par le sélectionneur (Sierra-Lucero et al., 2003; Xie, 2003). L'efficacité de la sélection d'un caractère dans un environnement donné pour obtenir un gain dans un autre environnement est proportionnelle à la corrélation génétique entre les deux environnements et à l'héritabilité globale dans les deux environnements (Falconer & Mackay, 1996).

Dans ce contexte, deux grandes stratégies de sélection ont été proposées (Li et al. 2017). La première consiste à sélectionner les individus dont la performance est stable entre les différents environnements. Cette stratégie est applicable lorsque les interactions GxE sont relativement faibles. Elle doit permettre de sélectionner des génotypes dotés d'une grande capacité d'adaptation, assurant un certain niveau de croissance dans un large éventail d'environnements. La seconde stratégie consiste à tirer avantage de ces interactions GxE en sélectionnant, pour chaque environnement, les génotypes les plus adaptés. Si cette stratégie peut permettre de maximiser les gains par environnement, elle implique de développer des populations d'amélioration spécifiques pour chaque type d'environnement, ce qui induit des contraintes logistiques et financières conséquentes.

## 1.4.2. La prise en compte de l'information environnementale en amélioration forestière

### *L'évaluation conventionnelle et génomique sur des dispositifs multi-sites*

Les critères de sélection chez les arbres forestiers correspondent à des mesures ponctuelles, réalisées à un âge avancé (Pâques, 2013). Les interactions GxE liées à ces phénotypes sont évaluées au moyen de dispositifs multi-sites. Ces derniers permettent de tester la stabilité des

génotypes dans plusieurs contextes environnementaux. Pour les espèces présentant une bonne aptitude à la propagation végétative, comme celles du genre *Eucalyptus* ou *Populus*, ce sont les mêmes génotypes qui sont évalués dans des sites contrastés (Oliveira et al., 2020; Osorio et al., 2001; Silva et al., 2022). Hardner et al. (2011) ont par exemple mis en évidence de fortes interactions GxE pour la circonférence mesurée à 3 ans chez *Eucalyptus camaldulensis x E. globulus and x E. grandis*, par l'évaluation de 896 clones issus de 10 familles et répartis sur 22 sites expérimentaux représentant une très large gamme environnementale en Australie. Pour les autres espèces, même si des dispositifs clonaux peuvent exister, c'est généralement l'évaluation d'individus apparentés qui permet d'estimer la composante génétique liée à la plasticité (Valladares et al., 2006). L'importance des interactions varie en fonction de l'espèce, des caractères et des sites étudiés (Grattapaglia, 2017). Chez le pin taeda, Sykes et al. (2006) ont détecté des interactions GxE faibles à modérées pour des caractères liés aux propriétés chimiques du bois à 11 ans, en se basant sur 14 familles de pleins-frères connectées et évaluées sur 4 sites proches aux Etats-Unis. Avec 5 sites fortement contrastés au Portugal évaluant 30 populations de pin maritime, Correia et al. (2010) ont quant à eux détecté de fortes interactions GxE pour les caractères de volume mesurés à 10 ans. Les études multi-sites pour les arbres forestiers ont été synthétisées par Li et al. (2017).

Ce système d'évaluation présente deux contraintes majeurs :

- La mise en place de dispositifs multi-sites demeurent particulièrement couteuse et complexe dans le domaine forestier. L'évaluation se fait donc classiquement sur un nombre limité de sites. Bien qu'ils soient choisis pour représenter les contrastes environnementaux d'une certaine zone géographique, les environnements les plus extrêmes ne se sont généralement pas retenus afin d'assurer tout de même une productivité suffisante de la plantation. Autrement dit, les dispositifs multi-sites ne sont pas nécessairement construits pour être représentatifs de l'ensemble des environnements pertinents dans un contexte de changement climatique (Ray et al., 2022).
- L'évaluation sur des caractères mesurés à un âge avancé suppose que l'environnement de chaque site a été constant au cours de la croissance des arbres. Cette hypothèse parait peu réaliste au vu des longues périodes considérées (caractères de sélection mesurés entre 3 et 10 ans selon les espèces) et encore plus dans un contexte de changement climatique. Les caractères ciblés apparaissent très intégrateurs dans le mesure où ils correspondent en fait à une réaction cumulative des arbres au fil des années, face aux variations environnementales. Il n'est donc pas possible d'identifier les facteurs environnementaux contribuant au

phénotype final, l'environnement ne pouvant être considéré que de manière globale sur l'ensemble de la période de croissance.

La prédiction génomique pourrait a priori présenter un grand intérêt pour limiter les coûts associés à l'évaluation systématique des génotypes sur chaque site. Dans ce cas, la prise en compte et l'analyse des interactions GxE dans un contexte de sélection génomique suivraient les mêmes procédures que pour une évaluation classique (Grattapaglia, 2017). Pourtant, la plupart des modèles génomiques chez les arbres forestiers ne se focalisent que sur un unique environnement (Lebedev et al., 2020). Dans certaines des rares études où plusieurs sites sont pris en compte, il est démontré que l'augmentation de la précision des GEBV est possible, mais qu'elle dépend fortement du contexte et des caractères (Gamal El-Dien et al., 2015; Souza et al., 2019). Si les précisions de prédiction dans un même site sont généralement bonnes, globalement la capacité des modèles à prédire les performances des génotypes dans un environnement donné à partir d'une calibration composée de génotypes évalués dans un autre environnement reste encore limitée (Chen et al., 2018; Resende Jr et al., 2012).

### *La dendroplasticité comme évaluation efficace au niveau multi-environnemental*

Dans un arbre, l'évaporation de l'eau au niveau des stomates crée une aspiration dans les vaisseaux du xylème et permet le transport d'eau des racines jusqu'aux feuilles. L'augmentation de la température de l'air intensifie la transpiration foliaire ce qui accélère le transport de l'eau. Cependant, en cas de sécheresse ou de stress hydrique, les stomates se ferment partiellement afin de maintenir la colonne d'eau tout en minimisant les pertes. Cette fermeture des stomates, tout comme la chute des feuilles ou le flétrissement, constitue une réponse plastique de l'arbre au stress hydrique. De même, la baisse de l'activité photosynthétique limite notamment les ressources de la croissance radiale (Rennenberg et al. 2006).

En zone tempérée, les températures douces du printemps combinée à la forte disponibilité en eau favorisent la croissance des arbres. Le cambium[4] produit alors du bois peu dense appelé bois initial et constitué de larges cellules aux parois fines permettant un transfert facile de l'eau dans l'arbre (Domec & Gartner, 2002). Au cours de la période estivale, la croissance des arbres ralentit en raison de la moindre disponibilité en eau. Le cambium produit alors des cellules plus étroites et aux parois plus épaisses ce qui a pour conséquence de réduire les pertes en eau et de

[4] Aussi appelé « écorce secondaire ». Le cambium est une fine couche de cellules méristématiques indifférenciés et capables de se diviser.

limiter les risques de cavitation du xylème. Le bois formé au cours de cette période, appelé bois final, est ainsi caractérisé par une forte densité. La croissance s'arrête pour la plupart des espèces au cours de l'hiver en raison des faibles températures.

La plasticité relative à la croissance radiale des arbres est appelée dendroplasticité et demeure « enregistrée » dans le bois via l'alternance de bois initial et de bois final. Elle constitue une formidable opportunité pour évaluer à postériori les performances d'un même individu au regard de l'ensemble des environnements qu'il a rencontré au cours de son existence. Les techniques d'analyse de cerne, très largement maitrisées dans le domaine forestier, peuvent ainsi fournir des longues séries répétées de phénotypes, beaucoup plus explicatives que les caractères intégratifs mesurés ponctuellement et à un âge souvent avancé, mais aussi plus facilement corrélables avec des variables climatiques (Martinez-Meier et al., 2008; Zas et al., 2020). Ces phénotypes qui réagissent plastiquement à l'environnement et de manière différentielle entre les individus sont sous contrôle génétique (Dalla-Salda et al., 2009; Sánchez-Vargas et al., 2007).

Dans un contexte de sélection génomique, l'inférence de normes de réaction à partir de la dendroplasticité apparait comme particulièrement prometteuse afin d'évaluer un grand nombre d'individus dans une large gamme environnementale et ce à moindre coût.

## 1.5. Objectifs de la thèse

Les critères actuels et les schémas conventionnels de l'amélioration forestière contrastent avec le besoin de rapidité et de flexibilité en sélection induit par la demande croissante en bois et l'accélération du changement climatique. L'objectif général de cette thèse est d'explorer de nouvelles méthodologies intégrant l'usage de la génomique ainsi qu'une meilleure prise en compte de l'environnement afin de proposer des applications concrètes aux sélectionneurs des arbres forestiers pour relever ces défis.

L'**objectif n°1** vise à évaluer les capacités des modèles de sélection génomique à prédire la variabilité intra-famille et cherche à identifier les facteurs clés de cette prédiction. La capture de la ségrégation mendélienne est essentielle afin d'assurer l'efficacité de la sélection mais aussi de permettre une gestion plus explicite de la diversité. Cet objectif s'inscrit plus généralement dans une volonté de déterminer les conditions d'application de la sélection génomique afin de révéler tout son potentiel par rapport aux approches basées sur le pedigree.

Pour cela, cette partie s'appuie sur :

- La construction d'un modèle de sélection génomique à partir de données recueillies sur un dispositif d'évaluation du pin maritime (**A**). Ce dispositif a été choisi afin de maximiser le nombre d'individus par famille, au-delà de ce qui a été fait jusqu'à présent dans les études précédentes sur l'espèce. Il s'agit d'une structure originale dans le domaine forestier qui va permettre d'évaluer explicitement la précision de la sélection génomique au niveau intrafamiliale.
- Le développement d'un modèle de simulation adapté au cas du pin maritime et cohérent avec les données réelles. Les simulations vont permettre d'explorer une multitude d'autres scénarios afin d'identifier les déterminants de la prédiction génomique au niveau global et intrafamiliale.

Un modèle de prédiction génomique efficace ouvre la porte à un gain génétique rapide pour des critères variés. Cependant, les gains espérés sont tributaires de la capacité des arbres à maintenir leur productivité face aux évolutions d'environnements.

L'**objectif n°2** vise à étendre les modèles de sélection génomique classiques en intégrant une dimension environnementale lors de l'évaluation des performances individuelles. Plus précisément, c'est une modélisation de normes de réaction génomiques par régression aléatoire qui est proposée ici. L'adoption de 'traits-fonction' représente un changement de paradigme

dans l'amélioration des arbres forestiers et doit fournir des phénotypes plus facilement utilisables pour les projections face aux changements globaux.
Pour cela, cette partie s'appuie sur des données recueillies pour une population échantillonnée dans un second dispositif d'évaluation du pin maritime (**B**). Ce dispositif a été choisi en raison de son âge avancé permettant de générer une longue série de phénotypes annuels pour l'ensemble des individus. La modélisation de ces séries phénotypiques en fonction de variables environnementales peut permettre d'analyser plus finement la dendroplasticité dans un contexte de changement environnemental.

Le succès de l'intégration des normes de réaction reste toutefois dépendant de l'efficacité des modèles génomiques préalablement établis. Le lien entre les deux objectifs de thèse résulte donc de la possibilité de travailler sur des phénotypes plus fins grâce à la sélection génomique car le phénotypage est concentré sur une population de calibration de taille réduite par rapport à la sélection conventionnelle. Ces deux objectifs de cette thèse permettent d'approcher deux piliers majeurs de l'innovation en amélioration forestière.

Le programme d'amélioration du pin maritime en France constitue un exemple idéal pour étudier l'implémentation de la sélection génomique et l'utilisation des 'traits-fonction' dans contexte forestier. Le massif aquitain cristallise en effet les principaux enjeux de l'amélioration forestière, à savoir assurer une productivité forte en bois dans un contexte de changement climatique. Le système d'évaluation et la stratégie de sélection récurrente du programme est très commune notamment chez les conifères, ce qui donnera un caractère transposable aux différentes conclusions établies. Enfin, ce programme dispose de ressources considérables. Ayant atteint aujourd'hui sa troisième génération d'amélioration, il s'agit d'un des programmes les plus avancés à l'échelle internationale. L'évaluation génétique s'appuie sur un pedigree d'une grande profondeur et sur un large réseau de dispositifs expérimentaux. Des puces de génotypage denses ont été développées au cours des dernières années et permettent d'envisager une obtention de données génomiques en routine.

Une partie des résultats de l'objectif n°1 est présentée sous forme d'un article « *Unlocking genomic selection potential: within-family prediction in conifers* » (**article n°1 – cf partie 3.2**), de même qu'une partie des résultats de l'objectif n°2 « *Integrating environmental gradients into breeding: application of genomic reactions norms in a perennial species* » (**article n°2 – cf partie 4.2**)

*Tableau 2-1 : Caractéristiques principales des dispositifs expérimentaux étudiés dans cette thèse*

| | Objectif de thèse n°1 | Objectif de thèse n°2 |
|---|---|---|
| **Nom du dispositif** | **A**<br>**Test de descendance biparental** | **B**<br>**Test de descendance polycross** |
| **Sites expérimentaux** | 1 site : Le Barp (2011) | 2 sites : Cestas et Escource (1997) |
| **Objectif d'étude** | Sélection génomique sur traits classiques | Normes de réaction sur croissance annuelle |
| **Structure génétique** | 90 familles de **pleins-frères** x<br>48 individus par famille | 196 familles de **demi-frères** x<br>35 individus par famille et par site |
| **Population échantillonnée** | Effectif de 1000 individus<br>30 familles x 20 individus<br>10 familles x 40 individus | Effectif de 650 individus<br>25 familles x 26 individus (13 par site) |

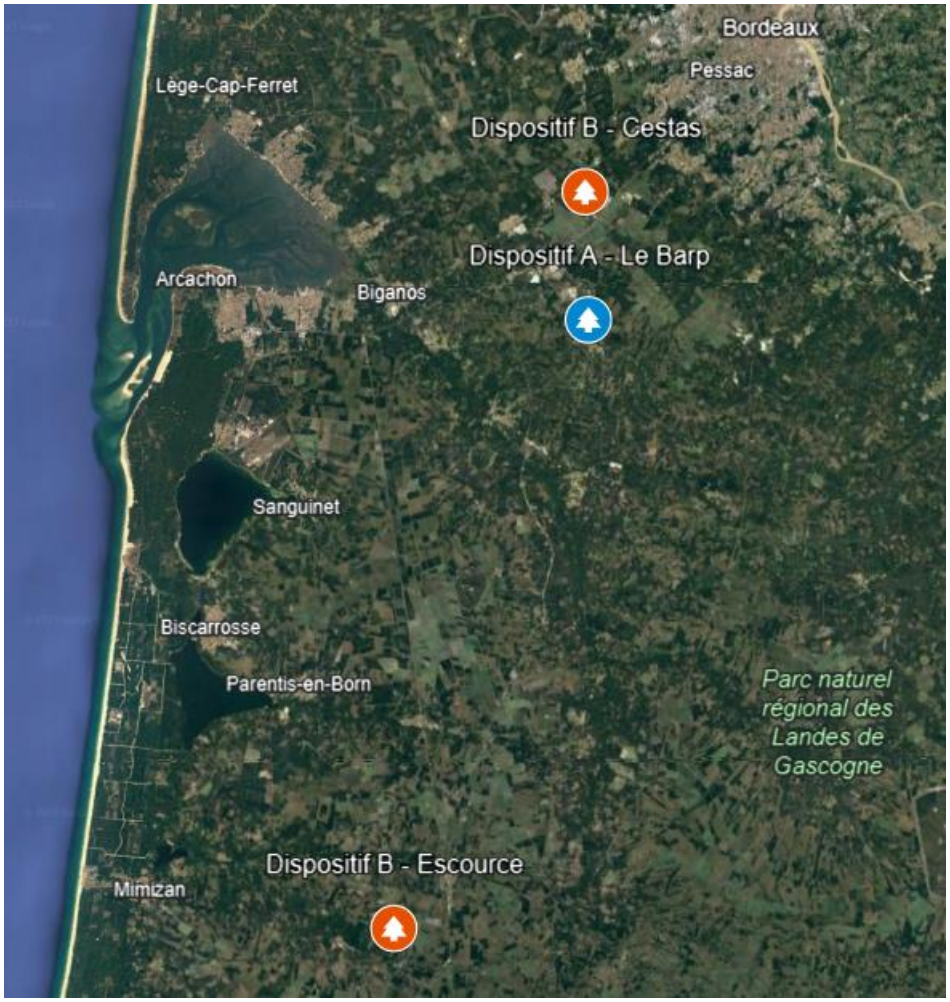



*Figure 2-1 : Localisation des dispositifs expérimentaux étudiés dans cette thèse*

# 2. Matériel et méthodes

Les deux axes de cette thèse se basent sur des dispositifs d'évaluation génétique mis en place dans le cadre du programme d'amélioration du pin maritime. La structure de ces dispositifs est décrite en **partie 2.1**. Les différentes données génétiques utilisées sont présentées dans la **partie 2.2.1** tandis que les mesures et le prétraitement des données phénotypiques sont détaillés en **partie 2.2.3**. Les données environnementales exploitées pour l'objectif n°2 sont décrites en **partie 2.2.4**. Enfin, les approches statistiques utilisées pour intégrer ces différents types de données sont présentées en **partie 2.3**.

## 2.1. Dispositifs expérimentaux étudiés et échantillonnages réalisés

Les dispositifs expérimentaux du pin maritime ont pour but d'évaluer les performances des candidats à la sélection ou celles de leurs descendants (**Fig. 1-9**). Dans chaque cas, une évaluation systématique sur 3 sites (lande humide, mésophile et sèche) a été envisagée. Toutefois, la difficulté à trouver des sites réellement contrastés au sein des Landes de Gascogne et les pertes régulièrement causées par des aléas climatiques extrêmes (tempêtes, incendies) amènent aujourd'hui à des évaluations sur un nombre souvent plus réduit de sites. Les trois générations d'amélioration pour le pin maritime ont conduit à la mise en place d'environ 90 dispositifs d'évaluation, dont deux d'entre eux sont étudiés dans le cadre de cette thèse (**Tab. 2- 1**).

### *Dispositif A*

Le dispositif « A » est étudié dans le cadre de l'**objectif n°1** (**Fig. 1-4**). Il ne comprend aujourd'hui qu'un seul site, localisé au Barp et classifié comme lande humide (**Fig. 2-1**). Ce dispositif correspond à un test de multiples descendances biparentales établi en 2011 et s'inscrit dans un procédé de sélection *forward*. 155 individus parmi les meilleurs de la génération G1, eux-mêmes descendant de 144 individus G0, ont été utilisés dans des croisements biparentaux afin de générer 90 familles de pleins-frères, peu connectées entre elles. Chaque famille de pleins-frères est représentée sur le site par 48 individus, aux côtés de lots témoins (améliorés et non-améliorés). Ces 4320 individus forment une population de taille efficace $N_s$ (Lindgren et al., 1996) égale à 53. L'évaluation de ces individus doit permettre de sélectionner ceux qui intégreront la génération G2 du programme d'amélioration.

L'évaluation d'un grand nombre d'individus par famille de pleins-frères est peu commune mais particulièrement intéressante pour notre étude visant à capter et prédire la ségrégation mendélienne. De plus, l'échantillonnage d'une population caractérisée sur un unique site pour tous les individus assure une homogénéité de l'information phénotypique et permet d'éviter l'utilisation des pseudo-phénotypes de type EBV souvent plus complexes à traiter en sélection génomique.

En raison de contraintes de temps et de coût, il n'est pas envisageable d'inclure tous les individus du dispositif dans notre étude. Par une approche de simulation, nous avons défini puis réalisé un échantillonnage de 1000 individus sur ce dispositif (**Suppl. 3-1**). Notre échantillon regroupe 40 familles, sélectionnées pour couvrir au mieux la diversité génétique présente dans le dispositif. Dans 30 de ces familles, nous avons échantillonné 20 individus, tandis que pour les 10 familles restantes nous avons échantillonné 40 individus. La population ainsi sélectionnée présente une taille efficace $N_s$ égale à 25.

### *Dispositif B*

Près de 1000 individus de la génération G1 ont été évalués dans un procédé de sélection *backward* à partir des performances de leurs descendants. Les meilleurs G1 sont ensuite installés dans les vergers à graines. Le dispositif « B », étudié dans le cadre de l'**objectif n°2**, correspond à un test de descendance *polycross* servant à évaluer les descendants de 188 de ces individus G1 candidats à la sélection. Ces candidats, utilisés comme mère, ont été croisés avec un mélange pollinique issu de plusieurs pères. Plus précisément, deux mélanges différents ont été utilisés regroupant respectivement le pollen issu de 42 et de 43 pères distincts. Chaque mère n'est croisée qu'avec un seul mélange pollinique, à l'exception de 8 d'entre elles qui sont croisées avec les deux mélanges polliniques. Les 196 unités génétiques ainsi obtenues correspondent à des familles de demi-frères. Pour 174 d'entre elles, 35 individus par famille ont été installés dans chacun des deux sites de ce dispositif. Ces sites sont localisés à Cestas et Escource et sont respectivement classés comme lande humide et sèche (**Fig. 2-1** et **Tab. 4-1**). Les 22 unités restantes n'ont été évaluées que sur le site d'Escource, toujours avec 35 individus par famille. La population générée (6440 arbres à Cestas et 7240 arbres à Escource) présente une taille efficace $N_s$ (Lindgren et al., 1996) de 23.

Ces tests de descendance *polycross* sont très communs en amélioration forestière. La représentation des mêmes familles sur chacun de ces deux sites doit permettre d'estimer

efficacement la composante génétique dans les modèles statistiques. Plantés en 1996, les arbres de ce dispositif étaient âgés de 24 ans au début de cette étude en 2019, ce qui permet une analyse de la dendroplasticité sur un grand nombre d'années.

En 2019 (en amont de la thèse), 25 familles de demi-frères ont été sélectionnées de manière à couvrir au mieux la variabilité phénotypique de la croissance observée sur le dispositif (pour les traits de hauteur et de circonférence). Pour ces 25 familles, 13 individus par site ont été échantillonnés, choisis principalement en raison de leur faible écart à la verticalité et de l'absence de maladie ou blessure. La population de base pour cette objectif n°2 est donc composée de 650 individus et présente une taille efficace $N_s$ (Lindgren et al., 1996) égale à 21.

***Figure 2-2 : Prélèvement d'aiguilles à l'aide d'un échenilloir***

## 2.2. Acquisition et prétraitement des données

### 2.2.1. Données génétiques

#### 2.2.1.1. Acquisition des données génomiques

*Préparation et groupement des échantillons d'ADN*

Des aiguilles ont été prélevées sur les différents arbres sélectionnés au sein des dispositifs A et B. Ces prélèvements ont été réalisés à l'aide d'un échenilloir dans le dispositif A (**Fig. 2-2**), et par des tirs au fusil dans le dispositif B en raison de la grande taille des arbres (20m en moyenne en 2019). Des extractions d'ADN ont ensuite été réalisées à partir de ces aiguilles (**cf Articles 1 et 2 - Material & Methods**). Les échantillons sont finalement envoyés à l'entreprise ThermoFisher (Santa Clara, CA, USA) qui réalise le génotypage avec la puce multi-espèces « 4TREE » développée dans le cadre du projet B4EST (Guilbaud et al., 2020). Parmi les plus de 45892 marqueurs SNP intégrés à cette puce, 13407 ont été spécialement désignés pour le pin maritime.

Dans le cadre de cette thèse, trois groupes d'échantillons ont été constitués :

- Un premier groupe composé des 1000 échantillons issus du dispositif A, et pour lesquels l'entièreté de la démarche de préparation été réalisée au cours de la thèse, du prélèvement des aiguilles jusqu'à l'envoi pour le génotypage.
- Un second groupe composé des 650 échantillons issus du dispositif B, et pour lesquels les données génomiques étaient déjà disponibles en début de thèse. Toutefois, un taux d'échec anormalement élevé a été constaté dans les résultats de génotypage de ces échantillons. Ainsi, une nouvelle préparation et un nouveau génotypage pour 200 échantillons ont été réalisés au cours de la thèse.
- Un dernier groupe composé de 1426 échantillons issus de précédentes études (Bartholomé et al., 2016; Isik et al., 2016). Ce groupe comprend la quasi-totalité des individus des générations G1 et G0 du programme d'amélioration. Ces échantillons disponibles en début de thèse sous forme d'ADN dilué ont déjà été génotypés sur des précédentes puces. Le travail de thèse a consisté à en contrôler la qualité de l'ensemble des échantillons et à relancer un génotypage avec la nouvelle puce 4TREE. Ils seront utilisés pour reconstruire et vérifier le pedigree des individus issus des dispositifs A et B (**cf partie 2.2.1.2**).

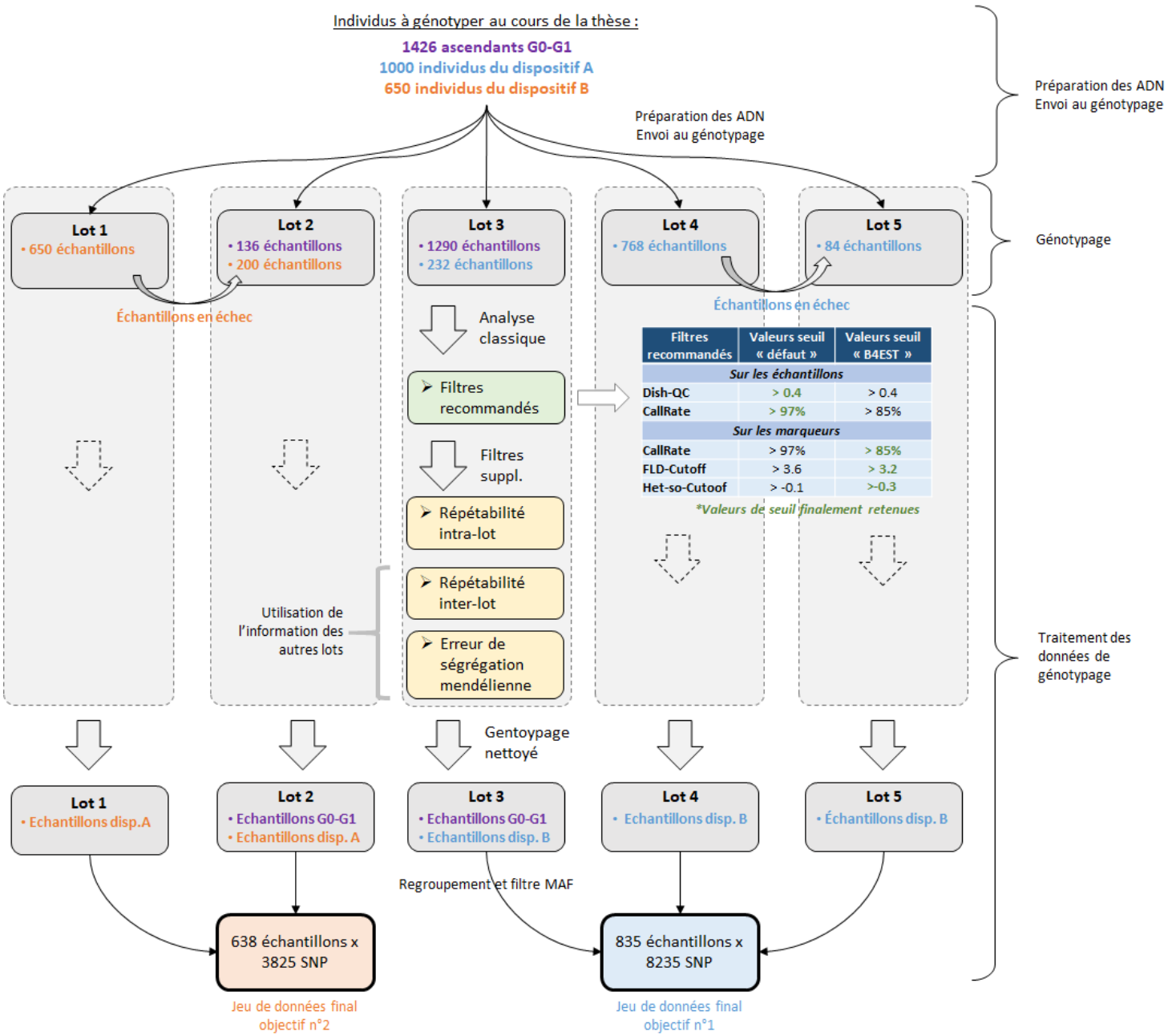



***Figure 2-3 : Résumé de la procédure de traitement des données de génotypage.***
*La puce de génotypage contient des sondes spécifiques qui sont conçues pour s'associer sélectivement avec l'ADN qui porte le SNP d'intérêt. Chaque association génère l'émission d'un signal de fluorescence spécifique qui renseigne sur l'allèle détecté. Le Dish-QC est une métrique évaluée avec l'ensemble des échantillons. Elle mesure le signal capté sur des sites non-polymorphiques afin d'estimer le « bruit de fond » dans les données. Sa valeur est comprise entre 0 et 1. Plus elle est proche de 1, plus le signal d'intérêt sera facilement séparable du signal « bruit de fond ». Le CallRate est calculé par échantillon et correspond au pourcentage de marqueurs ayant une donnée disponible pour cet échantillon. Un CallRate est également calculé pour chaque marqueur est correspond au pourcentage d'échantillons ayant une donnée disponible pour ce marqueur. Le FLD Cutoff (Fisher's Linear Discriminant) mesure la distance entre les signaux des deux groupements homozygotes. Le Het-so-Cutoff (Heterozygous Strength Offset) mesure la distance entre le signal du groupement hétérozygote et celui du groupement homozygote. Plus les valeurs de FLD et de Het-so sont grandes, meilleure est la résolution du SNP.*

La préparation des échantillons s'est étalée sur plusieurs mois. Pour des raisons pratiques, les échantillons ont ensuite été répartis en 5 lots, chacun d'entre eux ayant été génotypé par ThemoFisher dans des sessions distinctes (**Fig. 2-3**).

## *Analyse des données de génotypage*

Une démarche classique de prétraitement des données de génotypage est proposée par l'entreprise ThemoFisher. Elle consiste à appliquer une série de filtres sur les fichiers de génotypage bruts afin d'écarter les échantillons et les marqueurs considérés de qualité insuffisante. Deux types de seuils sont envisageables pour ces filtres (**Fig. 2-3**) :

- Les seuils notés « défaut », proposés par ThermoFisher et définis par défaut pour l'ensemble des espèces animales et végétales.
- Les seuils notés « B4EST », proposés par les membres du projet B4EST, moins stricts et considérés comme plus adaptés aux espèces forestières

L'utilisation de ces deux types de seuils amène des résultats très contrastés en terme de nombre d'échantillons et de marqueurs exclus. Notamment, l'application des seuils 'défauts' dans le lot n°1 amène à exclure 30% des échantillons et 25% des marqueurs, alors que ces pourcentages sont respectivement réduits à 5% et 10% avec les seuils 'B4EST'. Bien que moins conservateurs, ces seuils 'B4EST' posent question vis-à-vis de la qualité de l'information génomique retenue. En l'absence de procédure claire et consensuelle, et s'agissant des tout premiers résultats de génotypage obtenus avec la puce 4TREE, nous avons choisi de développer notre propre démarche d'analyse, basée sur les seuils 'B4EST' mais incluant des vérifications additionnelles afin de s'assurer d'une qualité du génotypage pour la suite nos analyses.

**<u>Filtres au niveau des échantillons</u>**

Initialement, deux filtres proposés par ThermoFisher ont été appliqués au niveau des échantillons en utilisant les seuils 'B4EST' (**Fig. 2-3** : Dish-QC>0.4 et QC-CallRate>85%). Peu de vérifications complémentaires sont possibles à ce niveau. Cependant, la qualité des données génomiques ainsi obtenues est mise en doute par des analyses ultérieures. Brièvement, le coefficient d'apparentement génomique attendu entre pleins-frères est de 0.5, mais il peut varier approximativement entre 0.4 et 0.6 (VanRaden, 2007; Visscher et al., 2006). Avec les 1000 individus génotypés pour le dispositif A, nous avons pu vérifier cet attendu (**Fig. 2-4**). Toutefois, l'existence d'individus faiblement apparentés à leurs pleins-frères (coefficient

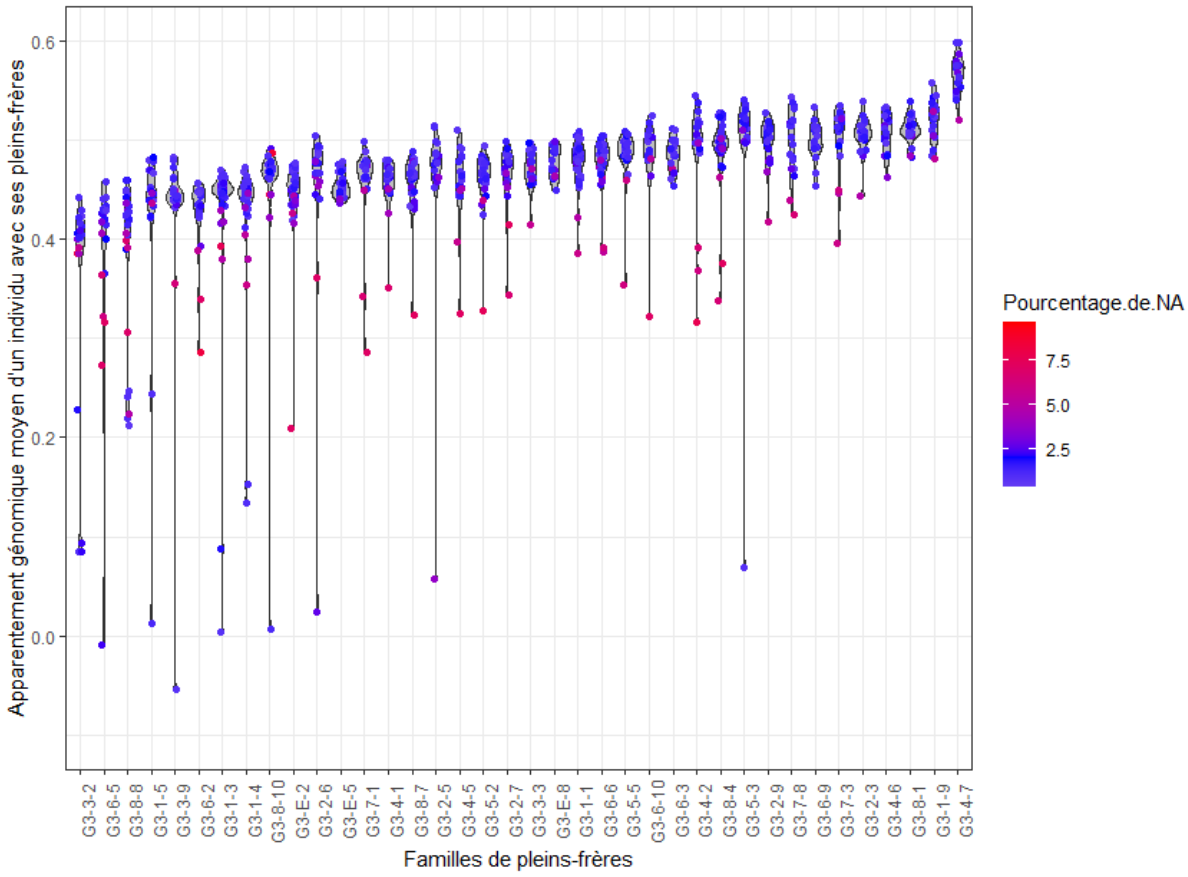



***Figure 2-4 : Apparentement génomique moyen des individus avec leurs pleins-frères***

*Les coefficients d'apparentement sont calculés à partir des données de génotypage de l'ensemble des individus échantillonnés dans le dispositif A et en utilisant la formule 1 de VanRaden, 2008. Chaque point représente la moyenne des coefficients d'apparentement entre un individu et le reste de sa famille. Pour chaque famille, la densité des points est représentée par un violin plot. Les couleurs de chaque point représentent le pourcentage de données manquantes dans les données de génotypage de l'individu concerné. Seuls les individus ayant moins de 15% de données manquantes ont été conservés ici suite à l'application des seuils 'B4EST'(CallRate=85%). Les seuils 'défaut' auraient conduit à l'élimination de tous les individus ayant plus de 3% de données manquantes (CallRate=97%)*

d'apparentement moyen compris entre ~0.30 et ~0.40) pose question. L'interprétation par une erreur de pedigree est ici exclue. En effet, en cas d'erreur, l'apparentement attendu d'un individu avec la famille est de 0.25 si un parent renseigné dans le pedigree est faux, ou 0 si les deux parents sont faux. Le fait que la totalité de ces individus atypiques coïncident avec des échantillons qui auraient été exclus avec les seuils 'défauts' (Taux de Na > 3%) remet plutôt en cause la qualité de l'information génomique conservée avec les seuils 'B4EST'. Pour la suite, seuls les seuils 'défauts' ont finalement été retenus pour les filtres s'appliquant au niveau des échantillons (Dish-QC>0.4 et QC-CallRate>97%).

**Filtres au niveau des marqueurs**

Les filtres recommandés par ThermoFisher au niveau des marqueurs ont été appliqués pour chaque lot avec les seuils 'B4EST' (**Fig. 2-3** : CallRate>85%, FLD-cutoff>3.2 et Het-so-cutoff>-0.3). Cependant, plusieurs filtres complémentaires ont été mis en place ici.

- La qualité des marqueurs est premièrement estimée par le calcul d'une répétabilité se basant sur des échantillons dupliqués au sein chaque lot. Selon les lots, le nombre de duplicats varie entre 15 et 72. Les marqueurs présentant plus d'une erreur de répétabilité sont exclus.

- L'information des autres lots est valorisée dans un second temps pour appliquer deux nouveaux filtres. La présence d'échantillons répétés entre lots permet de calculer une répétabilité 'inter-lot' des marqueurs. A nouveau, les marqueurs présentant plus d'une erreur de répétabilité sont exclus. Egalement, la présence de trios (2 parents – 1 descendant) au travers des différents lots permet de vérifier les conditions de la ségrégation mendélienne. Les marqueurs affichant un taux d'erreur de ségrégation supérieur à 5% sont exclus.

- Enfin, les individus d'intérêt pour notre étude sont extraits de chaque lot. Ils forment deux jeux de données associés aux objectifs n°1 et n°2. Dans chaque jeu de données, seuls les marqueurs présentant une fréquence de l'allèle mineur supérieure à 1% sont finalement conservés.

Pour l'objectif n°1, le jeu de données final regroupe 833 individus (sur 1000 individus initiaux) caractérisés pour 8234 SNP. 80% de ces SNP sont de type « PolyHighResolution[5] » et 20% de type « NoMinorHom [6] ». Les données manquantes restantes dans ce jeu de données (<1%),

[5] Marqueurs très résolutifs affichant pour la population considéré trois classes alléliques (AA, AB et BB)
[6] Marqueurs affichant pour la population considérée deux classes alléliques (AA et AB)

***Encadré 2-1 : Qualité de génotypage hétérogène et imputation des données manquantes entre lots***

La procédure de traitement des données de génotypage amène à conserver plus de 8000 marqueurs dans les lots n°2, 3, 4 et 5. Le lot n°1 constitue un ensemble atypique dans lequel 200 des échantillons sont amenés à être exclus (30%) et seuls 3832 marqueurs environ sont conservés. Si ces 200 échantillons problématiques ont pu être à nouveau génotypés via le lot n°2, la qualité du génotypage des 450 échantillons restants (correspondant à 450 individus du dispositif B) est faible. Le lot n°1 est le seul à regrouper des échantillons d'ADN extraits à partir d'aiguilles séchées (par opposition aux aiguilles fraîches utilisés pour tous les autres lots), ce qui a pu affecter la qualité des ADN. De plus, des problèmes techniques lors du génotypage par l'entreprise ne sont pas exclus.

Sans carte physique ou génétique pour les marqueurs de la puce 4TREE, nous avons envisagé une imputation de données manquantes dans ce lot n°1 grâce aux informations de génotypage des autres lots, via le logiciel LinkImpute (Money et al., 2015). Ce logiciel se base sur une méthode dite LD-kNNI visant dans un premier temps à sélectionner les plus proches voisins de l'individu à imputer. Ces voisins sont identifiés sur la base des SNPs les plus en LD avec le SNP à imputer. La valeur de génotypage manquante est ensuite obtenue par la moyenne modale pondérée des valeurs de génotypage des individus voisins.
Dans notre jeu de données, le logiciel réalise l'imputation et affiche une précision globale de 85%. Cette précision est calculée en masquant aléatoirement 10000 valeurs de génotypage (soit 0.005% des données) et en comparant valeur imputée et valeur vraie. Nous avons appliqué une procédure supplémentaire de contrôle consistant en un même type de cross-validation, mais cette fois-ci par SNP. Pour chaque SNP, 10% des données sont masquées puis imputées à l'aide de l'ensemble du jeu de données. Nous avons alors obtenu un précision d'imputation moyenne de 41% (variant entre 20% et 55%). Autrement dit, l'imputation apparait moins bonne qu'un tirage aléatoire.
Ces résultats sont assez surprenants étant donné la large utilisation de ce logiciel, notamment chez les arbres forestiers (Chen et al., 2018; Lenz et al., 2020). Contacté à ce moment, l'auteur du package rappelle que ce logiciel, initialement développé pour des variétés de pommes, n'a pas été conçu pour fonctionner sur des individus fortement apparentés bien que cela ait pu fonctionner dans certaines études. La présence de parents et descendants dans le même jeu de données peut donc poser problème tout comme la répartition assez atypique des données manquantes. Finalement, les SNPs exclus d'un lot n'ont pas été imputés par l'information disponibles dans les autres lots.

réparties de manière aléatoire, ont été imputées avec la fréquence allélique de la population au marqueur associé.
Pour l'objectif n°2**,** le jeu de données regroupe 628 individus (sur 650 individus initiaux) caractérisés par 3832 SNP, dont 68% sont de type PolyHighResolution. Chaque valeur génotypique manquante a été imputée avec la fréquence allélique au sein de la famille de pleins-frères correspondante pour le marqueur associé.

Des qualités de génotypage très hétérogènes entre les lots sont à l'origine de ces écarts dans les nombre de marqueurs finalement présents dans les jeux de données finaux (**Encadré 2-1**).

#### 2.2.1.2. Correction et reconstruction des informations de pedigree

Pour évaluer l'intérêt des données génomiques, nous avons choisi de les comparer à des informations de pedigree complètes et corrigées. Pour l'ensemble des individus échantillonnés sur les dispositifs A et B, une étape préliminaire consiste donc à retrouver ou à vérifier les liens de parenté entre ces individus et les ascendants de la génération G1. Les liens de parenté entre les individus G1 et les individus G0 ont déjà été validés dans une étude précédente.

Pour les 833 individus échantillonnés sur le dispositif A (familles de pleins-frères), l'objectif est de vérifier l'identité des deux parents renseignés dans le pedigree. Pour chaque marqueur caractérisant un individu et ses deux parents théoriques[7], la compatibilité de la ségrégation mendélienne est testée avec le package R pedtools (Dehli Vigeland, 2022). Une erreur de pedigree est décrétée lorsque plus de 1% des marqueurs indiquent une incompatibilité de ségrégation dans ce trio. Une analyse par duo (1 parent – 1 descendant) permet ensuite de déterminer plus précisément le ou les liens de parenté qui pose problème[8].
Pour les individus présentant une erreur de pedigree (42 individus ayant au moins un parent faux), une recherche de parenté « naïve » a été réalisée. La compatibilité de la ségrégation mendélienne a été testé avec l'ensemble des autres parents du dispositif A. Un nouveau parent a pu être assigné à un individu lorsque moins de 1% d'incompatibilité de ségrégation est détecté entre eux.

[7] Le nombre de marqueurs utilisés varie entre 6532 et 8110 selon le trio considéré. Pour être comptabilisé, un marqueur ne doit présenter de données manquantes pour aucun des membres du trio.
[8] En l'absence de données de génotypage pour un parent, ses liens de parenté renseignés dans le pedigree ne peuvent pas être vérifiés et ont été considérés comme vrais par défaut.

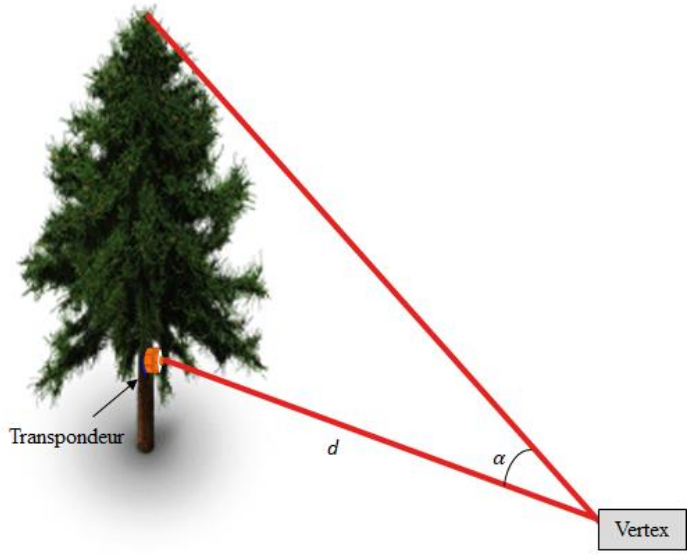



***Figure 2-5 : Principe de la mesure de la hauteur d'un arbre au vertex***
*Un transpondeur est placé contre le tronc à 1.30m. L'utilisateur se positionne ensuite à distance équivalente à la hauteur de l'arbre. Avec le vertex (un dendromètre à ultrasons), il vise le transpondeur, afin d'estimer la distance d, puis la cime de l'arbre, afin d'estimer l'angle ∝. La hauteur de l'arbre est ensuite déduite par les règles de trigonométrie. (Source : Vertex – Haglöf Sweden AB)*

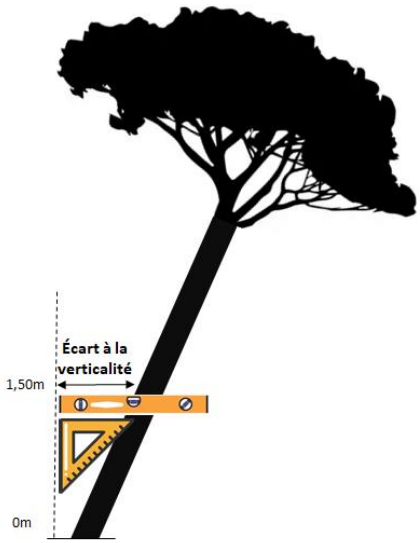



***Figure 2-6 : Principe de la mesure de l'écart à la verticalité***

Cette même démarche a été appliquée pour vérifier l'identité des mères renseignée dans le pedigree des 628 individus échantillonnés sur le dispositif B (familles de demi-frères). Mais pour ces individus, une étape supplémentaire est nécessaire pour retrouver leur père parmi ceux ayant contribués au mélange pollinique. Une recherche de parenté « naïve » serait trop fastidieuse ici au vu du nombre d'individus concernés. Une démarche plus efficace en 3 étapes a été définie :

- Un jeu de 161 marqueurs SNP a été sélectionné au sein de nos données génomiques. Ces marqueurs sont considérés de très haute qualité (CallRate >99%) et très informatifs dans notre population étudiée (MAF>40% et LD entre eux ($r^2$) < 0.1).
- Une recherche de parenté a ensuite été réalisée avec le logiciel Cervus (Kalinowski et al., 2007; Marshall et al., 1998) sur la base de ces 161 SNP. Ce logiciel implémente une approche du maximum de vraisemblance qui permet d'identifier facilement les parents les plus probables pour un individu.
- Parmi les pères candidats proposés par le logiciel, un père est retenu si son apparentement avec l'individu au regard de la matrice génomique (calculée avec l'ensemble des marqueurs disponibles) est compris entre 0.4 et 0.6

### 2.2.2. Données phénotypiques

#### *Mesures phénotypiques classiques*

Les critères de sélection correspondent à la circonférence et la hauteur mesurées à 12 ans ainsi qu'à l'écart à la verticalité mesurée à 8 ans. Toutefois, des mesures complémentaires peuvent être réalisés plus tôt, afin d'assurer une donnée phénotypique en cas de perte ultérieure du dispositif, ou plus tard, afin de maximiser la corrélation avec le trait final ciblé à l'âge de coupe à 45 ans. Ainsi, l'ensemble des individus du dispositif A, âgés de 9 ans au début de notre étude, ont déjà été phénotypés à 8 ans pour ces trois traits. Plus âgés, les arbres du dispositif B ont été phénotypés pour la circonférence à 8, 12, 16 et 22 ans, pour la hauteur à 8, 12 et 19 ans, ainsi que pour l'écart à la verticalité à 8 ans.

La circonférence du tronc est mesurée à 1.30m du sol à l'aide d'un mètre ruban. Jusqu'à 8 ans, la hauteur est déterminée à l'aide d'une perche télescopique. Au-delà, les arbres dépassent généralement les 10m et la perche est remplacée par le vertex (**Fig. 2-5**). L'écart à la verticalité


Carotte de bois prélevée à 1.30m avec une tarière de 5mm de diamètre puis découpée avec une scie à lames jumelles donnant des planchettes de 2mm d'épaisseur


Extraction des résines par un trempage de 24h dans du pentane puis sèchage
Radiographie aux rayons X

Les variations dans les tons de gris reflètent des disparités de densité (sombre = faible densité clair= densité élevée)

**Logiciel Windendro®**

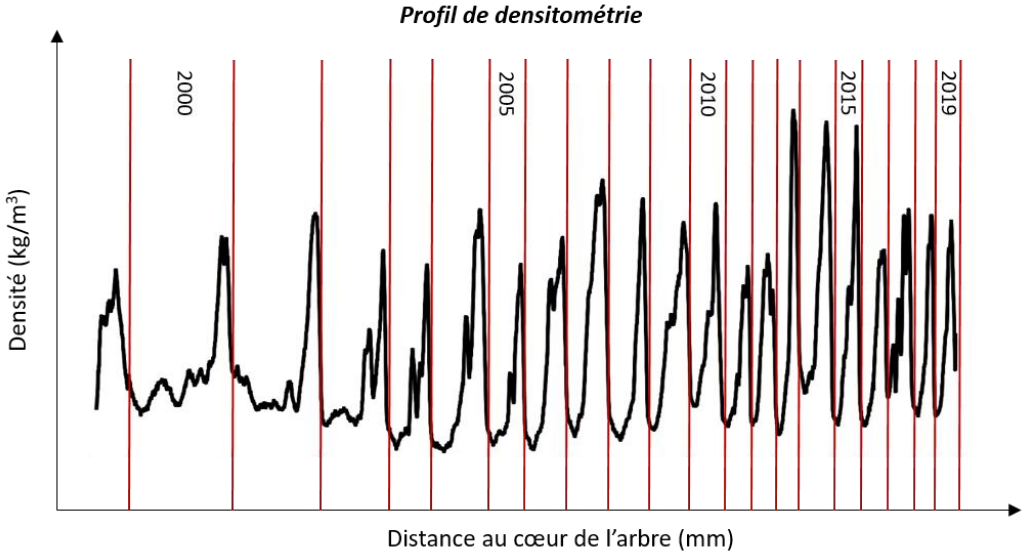



***Figure 2-7 : Caractérisation des cernes de croissance par densitométrie pour un arbre du dispositif B.***
*Les premiers cernes ne sont pas toujours visualisables sur les profils car la carotte ne passe pas systématiquement par le cœur de l'arbre.*

représente la distance du tronc à la verticale à 1.50m. Il est mesuré en cm à l'aide d'une équerre incluant un niveau à bulles (**Fig. 2-6**).

L'objectif n°1 se focalisera en sélection génomique sur la prédiction de ces traits classiques. Comme évoqué en introduction, ces mesures phénotypiques demeurent très intégratives et ne permettent pas de mettre en relation la croissance avec des variables environnementales. Si l'augmentation en fréquence des mesures de croissance semble peu envisageable en raison du temps et du coût nécessaire, l'objectif n°2 vise plutôt à générer de longues séries phénotypiques via l'étude de la dendroplasticité.

### *Dendroplasticité pour les individus échantillonnés dans le dispositif B*

Pour les 650 arbres sélectionnés dans le dispositif B, une carotte de bois a été prélevée dans le tronc, puis radiographiée (**Fig. 2-7**). Cette radiographie permet d'évaluer la densité du bois, du cœur de l'arbre jusqu'à l'écore, et de tracer un profil densitométrique pour chaque individu. L'alternance entre le bois initial et final permet facilement de placer les limites de cernes sur chaque profil. La largeur, la surface, la densité moyenne ou encore le pourcentage de bois initial-final sont autant de paramètres qui permettent de caractériser chaque cerne correspondant à une année de croissance. La surface de cerne est un indicateur intéressant dans le cadre de l'amélioration du volume du bois, renseignant sur la biomasse produite à hauteur de prélèvement de la carotte (ici 1.30m). Par opposition à la largeur de cerne qui diminue avec l'âge de l'arbre, la surface présente l'avantage de ne pas être dépendante avec la position du cerne. En effet, pour une même biomasse annuelle produite, un cerne proche du cœur de l'arbre présentera une largeur plus importante qu'un cerne éloigné, alors que les surfaces seront identiques. Notre étude se focalisera donc sur les surfaces de cernes annuels que nous allons chercher à modéliser en fonction de variables environnementales définies au même pas de temps. Notons que l'étude de la variabilité intra-cerne, non-considérée dans ce travail, est bien plus complexe car elle implique de faire le lien entre une distance (densité exprimée en fonction d'une distance au cœur de l'arbre) et une échelle temporelle intra-annuelle (échelle de mesure de la variable environnementale choisie).

Tous les arbres du dispositif B ont été plantés la même année. Ainsi, les effets « âge de l'arbre » et « environnement de l'année » sur les surfaces de cernes sont confondus. Ce n'est pourtant que ce deuxième effet que nous souhaitons modéliser.

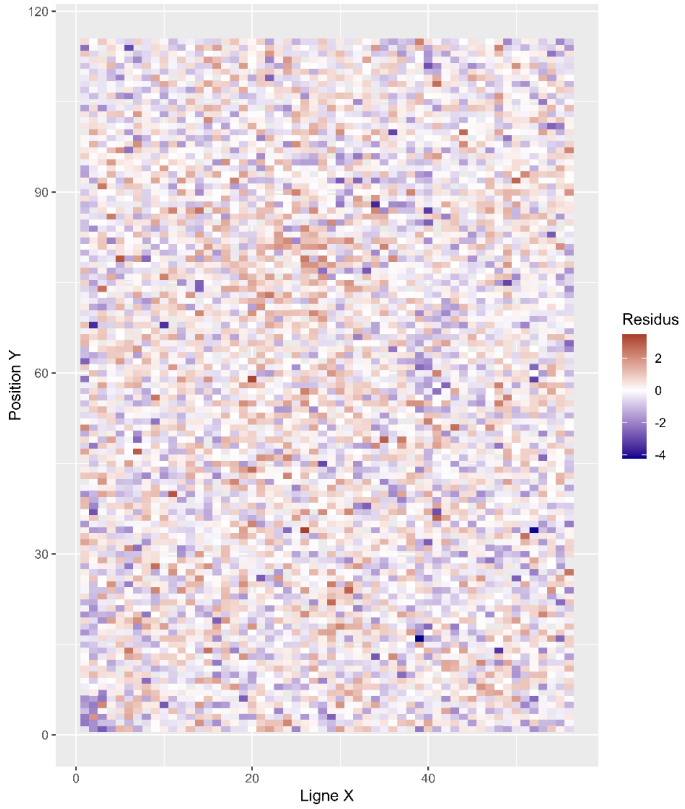



***Figure 2-8 : Représentation spatiale des résidus issus d'un modèle analysant les hauteurs (m) mesurées à 9 ans de l'ensemble des arbres du site du Barp.***

*Le modèle est de la forme* $\boldsymbol{Ht_9 = \mu + Zu + e}$ *avec* $\boldsymbol{Ht_9}$ *le vecteur des hauteurs individuelles,* $\boldsymbol{\mu}$ *la hauteur moyenne dans ce site, u le vecteur de solutions pour l'effet génétique additif aléatoire, et e le vecteur des résidus. On assume* $\boldsymbol{u \sim N(0, A\sigma_A^2)}$ *avec A la matrice des coefficients d'apparentement calculés à partir du pedigree et* $\boldsymbol{\sigma_A^2}$ *la variance génétique additive. Chaque individu est défini par des coordonnées x et y sur le terrain. La représentation des résidus met clairement en évidence un effet spatiale avec des résidus distribués de manière très hétérogènes.*

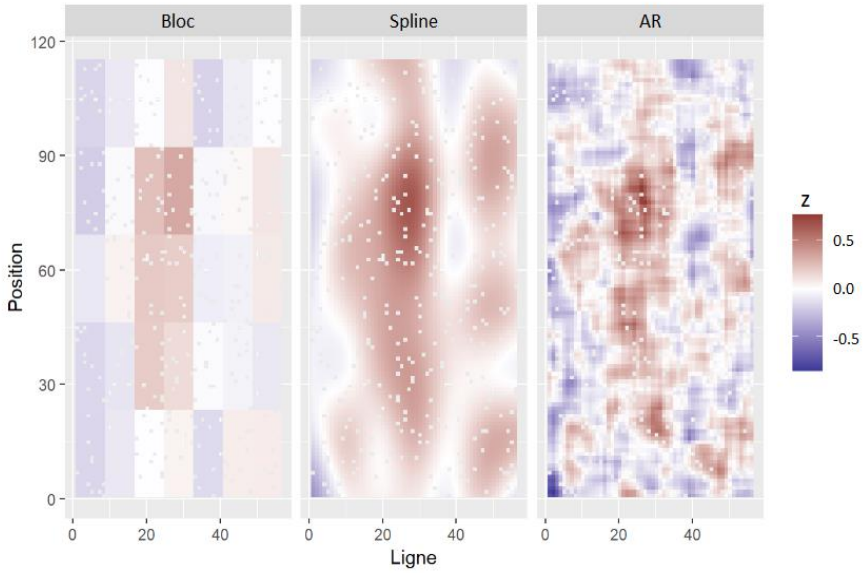



***Figure 2-9 : Représentation des différentes composantes aléatoires spatiales intégrées au modèle pour analyser les hauteurs (m) mesurées à 9 ans sur le site du Barp***

L'évolution de la surface de cerne moyenne dans notre population (650 individus) peut être décomposée en deux phases (**Article2 - Fig. 4-S1**). La première phase (1999-2005) correspond à une augmentation régulière et rapide de la surface de cerne moyenne. Cette courte période, assimilable à la phase juvénile des arbres, semble ainsi fortement marquée par l'effet âge et a été exclue de nos analyses. A l'inverse, au cours de la seconde période (2006-2019), assimilable à la longue phase mature des arbres, la surface de cerne moyenne présente une évolution stable malgré de fortes variations interannuelles. Pour cette dernière période qui définira notre plage d'étude, nous assumons un effet négligeable de l'âge ainsi qu'un fonctionnement physiologique des arbres homogène au cours des années.

Notons qu'un phénotype dit « nettoyé » de l'effet âge a été proposé. Pour cela, une trajectoire moyenne de l'effet âge a été estimée pour l'ensemble de la population grâce à une régression aléatoire modélisant la surface de cerne en fonction des années. La soustraction des valeurs annuelles moyennes aux valeurs phénotypiques initiales a permis d'obtenir ces phénotypes « nettoyés ». Par la suite, l'équivalence entre analyses utilisant les phénotypes initiaux et analyses utilisant les phénotypes « nettoyés » nous conforte dans l'idée que cet effet est négligeable sur notre plage d'étude. Par simplicité, seul le traitement des phénotypes initiaux est présenté dans la suite.

### *Ajustement des effets spatiaux dans chaque site*

Les dispositifs expérimentaux chez les arbres forestiers recouvrent généralement de très grandes surfaces (4.1 ha pour Le Barp, 5.2 ha pour Cestas et 5.8 ha pour Escource) pouvant présenter de fortes hétérogénéités de terrain. Des autocorrélations spatiales ont été mises en évidence dans les résidus de modèles simples analysant les phénotypes dans chaque site (**Fig. 2- 8**). Nous avons fait le choix d'estimer ces effets spatiaux et de corriger les valeurs phénotypiques. Cette correction a été réalisée en amont de nos analyses afin de ne pas complexifier les modèles de sélection génomique et de régression aléatoire utilisés par la suite.

Pour l'objectif n°1, nous considérons des traits phénotypiques mesurées en routine sur tous les arbres du dispositif A (circonférence, hauteur, écart à la verticalité). L'effet des hétérogénéités intra-site sur les phénotypes a été successivement modélisé par un facteur aléatoire « bloc », un processus autorégressif ou des fonctions B-splines (**Fig. 2-9**). Sur la base du critère AIC (Akaike, 1974), le modèle incluant un effet spatial représenté par des fonctions B-splines a été

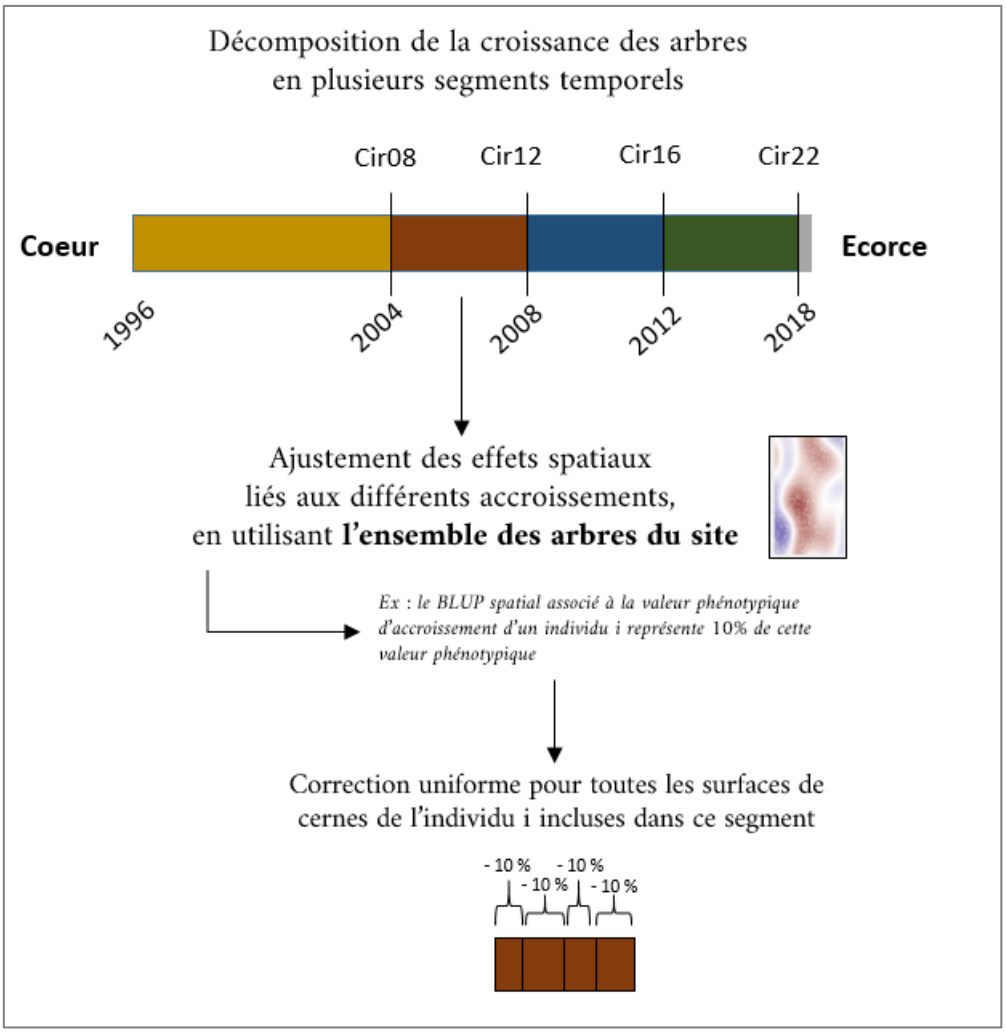



***Figure 2-10 : Ajustement des effets spatiaux associés aux surfaces de cernes mesurées sur 650 arbres du dispositif B***

retenu. Les phénotypes dit « corrigés » des effets spatiaux sont obtenus en soustrayant le BLUP spatial aux valeurs phénotypiques initiales.

Pour l'objectif n°2, nous considérons des surfaces de cerne mesurées par densitométrie uniquement sur les 650 individus de notre échantillon (325 individus dispersés par site). L'absence de telles mesures pour plus de 95% des arbres du dispositif rend difficile l'estimation des effets spatiaux liés à ces phénotypes. Toutefois, les mesures répétées de circonférence sur l'ensemble des arbres du dispositif ont été valorisées en estimant des accroissements de surface de bois sur plusieurs segments temporels (**Fig. 2-10**). Pour chaque segment, les effets spatiaux associés aux accroissements ont été estimés, à nouveau par l'intermédiaire de fonctions B-splines. Une correction uniforme est ensuite appliquée aux surfaces de cernes du segment concerné.

Par la suite, nous ne considérons que les phénotypes dits « corrigés » des effets spatiaux.

### 2.2.3. Données environnementales

La zone géographique couverte par le programme d'amélioration du pin maritime affiche des conditions climatiques homogènes (climat océanique avec des étés secs et des hivers pluvieux) et de faibles variations dans la nature des sols (sols acides avec podzol sableux). Ainsi, les différents dispositifs expérimentaux sont généralement considérés comme des répétitions lors de l'évaluation de l'écart à la verticalité et de la croissance à 8 et 12 ans respectivement.

Pour intégrer l'information environnementale dans le processus d'évaluation, nous proposons plutôt d'évaluer la croissance annuelle via les surfaces de cerne afin d'exploiter les contrastes environnementaux entre années (objectif n°2). Cette approche nécessite de définir des indices caractérisant l'environnement de chaque année dans chaque site et expliquant la plasticité de la croissance annuelle. Pour cela, plusieurs données environnementales sont disponibles pour le Dispositif B :

- Des données climatiques fournies au pas de temps horaire par des stations météo proches de chaque site : température, précipitations, vitesse et direction du vent, rayonnement photosynthétique actif (PAR), rayonnement global.
- Des analyses de composition des 60 premiers cm du sol (quantité de matière organique, pourcentage de sables fins/grossiers) réalisés en 2012 sur les deux sites.

- Des données d'humidité du sol obtenues grâce à 16 sondes TDR (Time Domain Reflectometry) par site, permettant un suivi continu de la hauteur de nappe entre 2015 et 2019.

Ces données permettraient une exploration complète des facteurs climatiques pertinents pour la croissance des arbres. Cependant, cette exploration en elle-même impliquerait une étude très complexe qui dépasserait les objectifs initiaux de la thèse. Par exemple, l'optimisation du choix des variables climatiques en fonction de la performance des modèles résultants, qui impliquerait une recherche sans doute très chronophage. Pour cette raison, nous avons préféré nous appuyer sur les connaissances préalables des sélectionneurs et sur la littérature existante. Comme évoqué en introduction, la température et la quantité d'eau disponible sont des facteurs majeurs conditionnant la croissance des arbres. Deux séries d'indices ont été proposés afin d'intégrer ces deux composantes (**cf Article 2 – Material & Methods**).

Brièvement, les premiers indices sont dérivés de la formule de Martonne (de Martonne, 1926) qui intègrent les relevés de températures et de précipitations afin caractériser l'aridité globale de chaque année dans chaque site. Dans un second temps, des données climatiques plus complètes (ajout de la vitesse du vent et de paramètres liés au rayonnement solaire) et des données de caractérisation du sol (fertilité et hauteur de nappe) ont été intégrées dans le modèle de croissance GO+ (Moreaux et al., 2020). Ce modèle permet de simuler les processus biophysique et biogéochimique associés au fonctionnement d'une plantation forestière de pin maritime. En particulier, il modélise les variations du potentiel hydrique dans les arbres. Un certain niveau de potentiel hydrique est nécessaire pour assurer la division et l'élongation cellulaire, et donc la croissance. Combiné avec la température du microclimat de la plantation, il forme une nouvelle série d'indices caractérisant le potentiel de croissance des arbres chaque année dans chaque site. Avec ces deux indices climatiques de construction très contrastée, avec la simplicité du premier et la complexité du second, nous voulions deux scénarios possibles pour caractériser la moyenne annuelle, sans entrer dans l'objectif trop ambitieux d'optimiser le choix de l'indice.

## 2.3. Estimation des paramètres et valeurs génétiques

L'estimation des valeurs et des paramètres génétiques est réalisée au moyen du modèle individuel (Henderson, 1975). En amont, cette approche nécessite d'estimer les similarités génétiques entre individus qui sont représentées dans des matrices d'apparentement.

### 2.3.1. Matrices d'apparentement

Une matrice d'apparentement est une matrice carré et symétrique, de dimension égale au nombre d'individus dans la population étudiée. Chaque élément hors diagonale, noté $a_{ij}$, correspond au coefficient d'apparentement entre deux individus $i$ et $j$. Ce coefficient d'apparentement est égal au double du coefficient de parenté $f_{ij}$, lui-même défini comme la probabilité que deux allèles tirés aléatoirement chez ces deux individus $i$ et $j$ soient identiques par descendance (IBD). La matrice d'apparentement a pour terme diagonale $1 + F_i$, avec $F_i$ le coefficient de consanguinité de l'individu $i$. Ce coefficient mesure la probabilité que deux allèles d'un locus donné soient identiques par descendance pour cet individu. Ces matrices d'apparentement peuvent être estimées à partir de données de pedigree ou de données génomiques.

#### 2.3.1.1. Matrice A calculée à partir des informations de pedigree

La matrice d'apparentement calculée à partir de données de pedigree est nommée « *numerator relationship matrix* » et notée A. Son calcul suppose de l'existence d'une population de référence dans laquelle les individus, appelés « fondateurs », sont non apparentés. Concrètement, c'est la population de base du programme d'amélioration (ici la génération G0) qui est généralement utilisée comme référence. Les apparentements pour les descendants sont calculés relativement à cette population de référence : $a_{ij}$ vaut 0.5 pour deux pleins-frères issus d'un croisement entre deux fondateurs, 0.25 pour deux demi-frères, 0 pour deux individus sans ancêtre commun... Toutefois, l'hypothèse de non-apparentement entre les fondateurs n'est pas systématiquement vérifiée en pratique ce qui peut biaiser le calcul des apparentements, tout comme l'utilisation d'un pedigree incomplet ou contenant des erreurs.

#### 2.3.1.2. Matrice G calculée à partir de données génomiques

La matrice d'apparentement obtenue à partir des données moléculaires est nommée « *realized relationship matrix* » et notée G. Elle peut être calculée par la formule de VanRaden (2008):

$$G = \frac{(M - P)(M - P)'}{2\Sigma\, p_i(1 - p_i)} \quad (2.1)$$

M est une matrice de dimension $n \times m$ (nombre d'individus × nombre de marqueurs) spécifiant les allèles au marqueur (codés en -1, 0, 1 respectivement pour l'homozygote, l'hétérozygote et l'autre homozygote), et P est une matrice de même dimension contenant les fréquences alléliques exprimées sous la forme $2(p_i\text{-}0.5)$, avec $p_i$ la fréquence de l'allèle minoritaire au locus $i$. Cette méthode de calcul de la matrice G donne plus de poids aux allèles rares dans l'estimation des relations génomiques.

Le calcul de la matrice G se fait à l'échelle de la population étudiée et ne nécessite pas d'hypothèse concernant l'existence d'une population antérieure de référence. Cette matrice présente des coefficients d'apparentements dits « réalisés » ou « réels » car ils se basent sur une observation directe du génome par l'intermédiaire des marqueurs moléculaires. La matrice G offre ainsi une meilleure résolution génétique que la matrice A car elle prend en compte les déviations autour de la parenté théorique dues à la ségrégation mendélienne lors de la méiose. De plus, les données moléculaires permettent de passer outre les problématiques de pedigree incomplet ou incorrect et permettent de calculer des apparentements entre individus a priori non connectés dans le pedigree. Les coefficients d'apparentements peuvent être négatifs, indiquant que les deux individus concernés sont plus différents génétiquement que ce qui est attendu en tirant aléatoirement deux individus dans la population.

### 2.3.2. Modèles statistiques pour l'évaluation génétique

#### 2.3.2.1. Modèle individuel

Le modèle individuel, ou modèle animal, est une application du modèle mixte défini par Henderson (1975). Développé en premier lieu pour la génétique animale, ce modèle est aujourd'hui largement utilisé en amélioration forestière. Les performances phénotypiques des individus sont exploitées afin d'estimer des effets fixes, correspondant généralement à des effets environnementaux contrôlés (site, bloc, traitement), et de prédire les effets génétiques

considérés comme des effets aléatoires. Le modèle de base peut s'écrire sous la forme développée (Mrode & Thompson, 2005) :

$$y_i = \mu_i + u_i + e_i \tag{2.2}$$

où $y_i$ désigne le phénotype de l'individu $i$, $\mu_i$ les effets environnementaux fixes de liés à l'individu $i$, $u_i$ la valeur génétique additive de l'individu $i$, et $e_i$ les effets résiduels affectant le phénotype de l'individu $i$. On retrouve plus généralement ce modèle écrit sous une forme matricielle :

$$y = Xb + Zu + e \tag{2.3}$$

Avec :

- $y$ le vecteur des observations phénotypiques
- $b$ le vecteur des effets fixes environnementaux et $X$ sa matrice d'incidence associée
- $u$ le vecteur des effets aléatoires génétiques et $Z$ sa matrice d'incidence associée
- $e$ le vecteur des effets d'erreurs résiduelles.

Ce modèle suppose que $a$ et $e$ suivent des lois normales centrées sur 0 et dont les variances s'expriment par :

$$Var(u) = A.\sigma_u^2 = \Gamma \text{ et } Var(e) = I_d.\sigma_e^2$$

avec $\sigma_u^2$ et $\sigma_e^2$ les variances associées aux effets génétiques et résiduels respectivement, $I_d$ la matrice identité et A est la matrice d'apparentement calculée précédemment.
On déduit $E(y) = Xb$ et $Var(y) = Z\Gamma Z' + R = V$.

La résolution de l'équation **2.3** passe d'abord par une estimation des composantes de la variance $\sigma_u^2$ et $\sigma_e^2$, classiquement via la méthode du maximum de vraisemblance restreint ou ReML (Patterson & Thompson, 1971). Puis, les solutions pour les effets fixes $\hat{b}$, appelées BLUE (Best Linear Unbiased Estimates), et les solutions pour les effets aléatoires $\hat{u}$, appelées BLUP (Best Linear Unbiased Predictors), sont calculées à partir des équations du modèle mixte établies par Henderson (1963) :

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\dfrac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \tag{2.4}$$

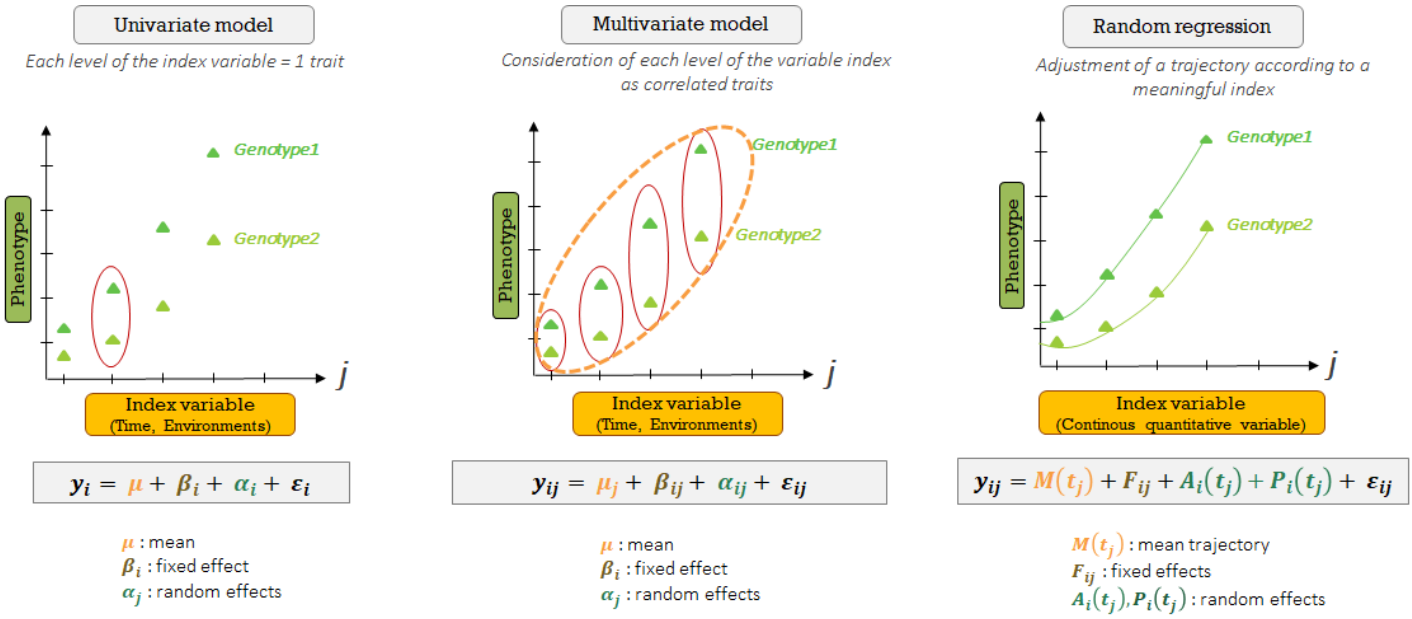



***Figure 2-11 : Différents modèles pour analyser des données longitudinales (Workshop B4EST 2021 – Papin, Bouffier, Sanchez)***

Les effets étant obtenus par :

- $\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y$
- $\hat{u} = \Gamma Z'V^{-1}(y - X\hat{b})$

Le vecteur $\hat{u}$ regroupe les valeurs génétiques prédites pour l'ensemble des individus inclus dans la matrice A. Par abus de langage, ces valeurs génétiques sont dites « estimées » et sont notées EBV (Estimated Breeding Values). Le terme « prédit » est généralement réservé aux valeurs génétiques obtenues pour les individus inclus dans le modèle via la matrice A mais sans observation directe de leur performance phénotypique.

Cette méthode d'évaluation BLUP est très utilisée car elle propose notamment :

- Une estimation simultanée des effets environnementaux fixes et des effets génétiques aléatoires, ce qui permet de combiner des informations provenant de plusieurs dispositifs tout en garantissant un classement adéquat des individus via leur valeur génétique,
- Et une exploitation efficace de la covariance additive entre l'ensemble des individus inclus dans la matrice A, autrement dit l'estimation de la valeur génétique d'un individu bénéficie de l'observation des performances phénotypiques de ses apparentés.

Le modèle décrit précédemment est de type ABLUP. La matrice d'apparentement A peut être remplacée par la matrice G, ce qui définit alors le modèle GBLUP.

#### 2.3.2.2. Modèle de régression aléatoire

Plusieurs extensions du modèle individuel ont été proposées afin de valoriser les données longitudinales, très communes en génétique animale et de plus en plus présentes pour les espèces végétales (**Fig. 2-11**). Ces données correspondent généralement à des mesures de croissance ou de productivité répétée au cours du temps comme par exemple le suivi de la production laitière chez les bovins (Jensen, 2001), ou l'étude du développement racinaire par phénotypage haut-débit chez les espèces de grandes cultures (Campbell et al., 2018; Rincent et al., 2019) .

Le modèle simple de répétabilité (Simple Repeatability Model) considère les différentes mesures d'un même individu comme des répétitions. Ce modèle suppose une variance constante pour les différentes dates de mesures successives ainsi qu'une corrélation fixe entre elles (Rutkoski et al., 2012; Sun et al., 2017). Il n'est donc par exemple pas adapté lorsque les mesures s'étalent sur plusieurs stades de croissance associés à fortes évolutions de variance. Le modèle multi-trait considère plutôt les mesures répétées de chaque individu comme des traits

phénotypiques distincts, en supposant un certain niveau de corrélation entre eux, transcrit par les matrices de covariance des effets additifs et résiduelles. Cette approche devient cependant particulièrement lourde lorsque que le nombre de mesures répétées est grand et que la corrélation entre elles est forte (Meyer & Hill, 1997). Enfin, le modèle de régression aléatoire apparait comme l'alternative la plus adaptée pour modéliser de manière continue la croissance ou la productivité au cours du temps, et ce, avec une paramétrisation réduite par rapport au modèle multi-trait et sans hypothèse de variance constante comme pour le modèle simple de répétabilité (Meyer, 2000). La formulation développée de ce type de modèle est de la forme (Mrode & Thompson, 2005) :

$$y_{it} = \sum_{k=0}^{k_m} \Phi_{\mathrm{itk}}\, \mathrm{m_k} + \sum_{k=0}^{k_u} \Phi_{\mathrm{itk}}\, u_{ik} + \sum_{k=0}^{k_p} \Phi_{\mathrm{itk}}\, p_{ik} + e_{it}$$

Avec :

- $y_{it}$ la performance de l'individu $i$ au temps $t$
- $m_k$ le $k$èmes coefficient de régression fixe pour la trajectoire moyenne
- $u_{ik}$ et $p_{ik}$ les $k$èmes coefficients de régression aléatoires pour les effets génétiques et d'environnement permanents, respectivement. L'environnement permanent regroupe tous les effets propres à un individu mais qui ne sont pas d'origine génétique additive.
- $e_{it}$ les résidus aléatoires
- $\Phi_{\mathrm{ijk}}$ est un coefficient de covariation temporelle défini par une fonction de base évalué au temps $t$ pour l'individu $i$.

L'estimation des paramètres et valeurs génétiques est ainsi possible de manière continue sur toute la période d'étude. L'originalité de l'approche proposée dans l'objectif n°2 consiste à remplacer cette échelle temporelle par un gradient environnemental. Les surfaces de cerne ne sont pas vues comme des mesures de croissance répétées chaque année, mais plutôt comme des mesures de croissance répétées dans différents environnements (**cf Article 2 – Material & Methods**).

# 3. Prédiction de la variabilité intra-famille en sélection génomique

## 3.1. Introduction

Si la sélection génomique est décrite comme particulièrement valorisable chez les arbres forestiers, les précisions de prédiction sont encore limitées et apparaissent généralement équivalentes à celles des modèles ABLUP basés sur le pedigree. Un avantage théorique des modèles génomiques réside dans leur capacité de capter la ségrégation mendélienne. Cependant, la précision de prédiction en intra-famille n'a pour le moment été évalué que dans un nombre très limité de cas.

Deux études ont évalué la faisabilité de la sélection génomique chez le pin maritime (Bartholomé et al., 2016; Isik et al., 2016). Si Isik et al. (2016) présentent des capacités de prédiction comprises en moyenne entre 0.43 et 0.49 pour les traits de rectitude et de croissance, l'absence de comparaison avec une approche ABLUP empêche d'évaluer clairement le gain apporté par les données génomique. Cette comparaison apparait explicitement dans l'étude de Bartholomé et al. (2016) qui démontre un avantage significatif des modèles ABLUP par rapport aux modèles génomiques en terme de précision de prédiction. Avec des effectifs moyens d'environ 2 individus par famille pleins-frères, ces deux études n'ont pas pu évaluer la précision de prédiction au niveau intrafamilial.

L'objectif de thèse n°1 est d'évaluer les capacités de prédiction des modèles génomiques au niveau intrafamilial et de définir plus généralement les conditions de mise en œuvre de la sélection génomique pour assurer son avantage par rapport aux modèles ABLUP. La démarche proposée dans cette partie s'articule en trois étapes :

1. Construction d'un modèle de sélection génomique à partir d'un échantillonnage comprenant de larges effectifs intrafamiliaux (20 à 40 arbres / famille). L'échantillonnage réalisé sur le dispositif A comprend initalement 1000 individus répartis dans 40 familles de pleins-frères. Ces individus ont été génotypés à l'aide de la puce 4TREE qui intègre un nombre de SNP significativement supérieurs à celui des puces utilisées dans les deux précédentes études chez le pin maritime. La précision des modèles génomiques a été explicitement évaluée par rapport aux modèles ABLUP intégrant des données de pedigree complètes et corrigées. Ce design original a permis d'évaluer clairement la précision intrafamiliale.

2. Développement d'un modèle de simulation afin d'évaluer la précision de la sélection génomique au niveau global et intrafamiliale dans différentes conditions d'héritabilité, de nombre de marqueurs et de taille de calibration, qui n'étaient pas disponibles dans les études empiriques.
3. Identification par simulation des caractéristiques familiales qui expliquent l'hétérogénéité des niveaux de précision observée entre les familles.

Les points 1 et 2 sont présentés au sein de l'article n°1 « Unlocking genomic selection potential: within-family prediction in conifers » qui sera prochainement soumis à *Annals of Forest Science.* Cet article correspond à la **partie 3.2** de ce manuscrit. Les premières recherches réalisées pour le point 3 sont présentées dans la **partie 3.3**.

## 3.2. Article n°1

# Unlocking genomic selection potential: within-family prediction in conifers

**Victor Papin[1], Gregor Gorjanc[2], Ivan Pocrnic[2], Laurent Bouffier[1] and Leopoldo Sanchez[3]**

[1] **INRAE, BIOGECO**, UMR 1202, 33610 Cestas, France.
[2] **The Roslin Institute** and Royal School of Veterinary, The University of Edinburgh, Edinburgh, UK
[3] **INRAE-ONF, BioForA**, UMR 0588, 45075 Orléans, France.

**Key message:**

The often-highlighted equivalence between genome-based and pedigree-based prediction accuracies of breeding values in forest trees is associated with a zero accuracy of genome-based prediction within-families, which can be explained by the still insufficient size of the genomic training sets.

**Abstract:**

Genome-based prediction and selection hold substantial promise for forest tree breeding. However, its advantage in terms of prediction accuracy over the conventional pedigree-based model is unclear. To shed light on this phenomenon, we compared accuracy of predictions from pedigree-based model (ABLUP) with complete and correct pedigree data and from genome-based model (GBUP) across and within families. The data comprised 833 individuals across 39 full-sib families, each with 10 to 40 individuals. Prediction accuracies with ABLUP and GBLUP were comparable and accuracy with GBLUP within families was on average zero with large variation between families. Simulations showed that the number of individuals in the training set is the main limiting factor of GBLUP accuracy in our study and likely in many forest tree breeding programmes. Accurate within-family prediction is possible with 40-65 individuals per full-sib family included in the genomic training set, out of a total of 1600-2000 individuals in the training set. Such conditions lead to a significant advantage of GBLUP over ABLUP in terms of prediction accuracy and more clearly justify the switch to genome-based prediction and selection in forest trees.

**Key words:**

Breeding programme, pedigree prediction, genomic prediction, genomic selection, maritime pine, progeny validation, stochastic simulation, within-family prediction.

**Abbreviations:**

**ABLUP**: pedigree-based best linear unbiased prediction

**BLUP**: best linear unbiased prediction

**CV**: cross-validation

**DEV**: stem deviation to verticality

**DNA**: deoxyribonucleic acid

**GBLUP**: genome-based best linear unbiased prediction

**GEBV**: genomic estimated breeding values

**GS**: genomic selection

**HT**: height

**LD**: linkage disequilibrium

**$nT_{set}$**: training set size

**nSNP**: number of SNP

**OCS**: optimum contribution selection

**$POP_R$**: trees sampled for this study

**$POP_s$**: simulated version of $POP_R$

**QTL**: quantitative trait loci

**SNP**: single nucleotide polymorphism

**$T_{set}$**: training set

**$V_{set}$**: validation set

# Introduction

The use of pedigree information in populations with genealogy records has revolutionized the improvement of selection programmes for many species. Pedigree information can be used to infer the expected relatedness between each pair of individuals, and this is the key to gauge the extent to which phenotypic values of individuals in the studied population have a genetic basis. This pedigree-based model has been the basis for the development of predictions of the individual additive genetic value, or the breeding value, via BLUP methodology implemented using the mixed models equations (Henderson, 1975, Mrode & Pocrnic, 2023). Breeding value of each individual can be decomposed into parent average and Mendelian sampling terms. Parent average term captures variation between families, it represents expected breeding value of progeny given its parents. Mendelian sampling term captures variation within families, it represents deviation of each individual's breeding value from the parent average due to recombination and segregation of parental genomes. With pedigree-based model, we need phenotypic values on an individual or its progeny to estimate the Mendelian sampling term of the individual's breeding value. Hence, pedigree-based prediction of breeding values for non-phenotyped individuals (forward prediction) captures only parent average term. By using genome data, we observe outcome of recombination and segregation of parental genomes as well as recent or past mutations, meaning that we can in principle estimate parent average and Mendelian sampling terms of breeding value even for non-phenotyped individuals (VanRaden, 2008, Hill & Weir, 2011).

The use of genetic markers potentially allows a more accurate estimation of the reproductive values of candidates and has greatly facilitated the implementation of genomic selection (GS). Genome-based prediction took off with the work of Meuwissen et al. (2001) which showed how regressing individuals' phenotypic values onto their genome-wide marker genotypes captured variation between individuals' breeding values by leveraging linkage-disequilibrium between quantitative trait loci (QTL) affecting traits of interest and the genome-wide markers. Using a training set of individuals that have been phenotyped and genotyped, the model estimates associations between variation in genome-wide markers and variation in phenotypic values. This means that the associations can be used to predict breeding values for non-phenotyped individuals which have genomic information. Such genome-based predictions have revolutionized many breeding programmes, enabling an efficient and early selection of candidates individuals and leading to significant genetic and economic gain per unit of time (Crossa et al., 2017; Hayes, Bowman, et al., 2009; Pryce et al., 2011).

Genome-based prediction is of particular interest in forest trees, as it could reduce the long duration of breeding cycles and also cut the cost of phenotyping complex traits such as drought tolerance or disease resistance (Grattapaglia & Resende, 2011, Isik 2014). Encouraged by promising simulations and first empirical approaches (Grattapaglia et al., 2011; Grattapaglia & Resende, 2011; Iwata et al., 2011), experimental studies with genome-based predictions have appeared in recent years for a large number of forest tree species (see Lebedev et al., 2020 for a recent review). Many of the studies emphasize attractiveness of genome-based predictions by reporting moderate to high prediction accuracies (Durán et al., 2017; Isik et al., 2016; J. Resende M. F. R. et al., 2012), and by reporting improved genetic gain per unit of time due to 20-50% shorter generation interval for GS (Chen et al., 2018; Lenz et al., 2017; Ratcliffe et al., 2015; Resende Jr et al., 2012). However, this reduction in generation interval is also possible with pedigree-based prediction (if considering forward selection) and it is not clear if higher genetic gain per unit of time with GS is due to genome-based predictions enabling shorter generation interval, higher accuracy, or both. In fact, several studies in forest tree breeding report that pedigree-based predictions and genome-based predictions have similar accuracy to genome-based predictions (Beaulieu et al., 2014; Lenz, Nadeau, Azaiez, et al., 2020; Thistlethwaite et al., 2017, 2019; Zapata-Valenzuela et al., 2012, 2013; Zhou et al., 2020). Furthermore, using a pedigree that is incomplete or that contain errors tends to distort the comparison with genomic data (El-Dien et al., 2018; Li et al., 2019). Such errors may be common in forestry breeding and penalize pedigree-based evaluation (Doerksen & Herbinger, 2010; Munoz et al., 2014). In this sense, part of the advantage of genomic selection may come from error-laden pedigree-based evaluations (Lenz, Nadeau, Azaiez, et al., 2020). Clarifying the conditions in which genome-based models can deliver real benefits remains a prerequisite for full exploitation of the advantages of GS in forest trees.

The access provided by molecular markers to within-family variability should facilitate the advantage of genomic selection over pedigree-based selection and potentially allow better management of diversity. Indeed, accurate within-family prediction would allow better exploitation of within-family genetic variability rather than inter-familial variability, preventing over-representation of certain lineages during selection and subsequent drift (Allier et al., 2019; Jannink, 2010; Rauf et al., 2010). However, little attention has been paid to the within-family prediction accuracy of GS models in forest trees, mainly due to the rather limited number of individuals per half and full-sib families commonly used in the progeny trials. To our knowledge, only few studies have addressed this issue. The studies of Fuentes et al. (2017) and

Cros et al. (2019) involved a single large full sibling family, making extrapolation to more general cases difficult. In three other papers, (Pégard et al., 2020; R. T. Resende et al., 2017; Ukrainetz & Mansfield, 2019), the within-family accuracies were substantial, although variable.

Maritime pine (*Pinus pinaster* Ait.) covers 4.2 million hectares in south-western Europe (Abad Viñas et al., 2016). A breeding programme was initiated for this species in France in the 1960's and follows a recurrent selection scheme (C.-E. Durel, 1992, GIS 2002). From a base population (600 G0 individuals) selected for growth, environmental adaptation and stem straightness, two breeding cycles were performed using estimated breeding values from pedigree-based model (Bouffier et al., 2016).The potential of genome-based prediction in maritime pine has already been highlighted in two previous studies (Bartholomé et al., 2016; Isik et al., 2016), bus, as for most forest tree species, it is essential to investigate in greater depth the conditions in which genome-based prediction is clearly superior to the pedigree-based one.

The aim of this study was to evaluate the ability of genome-based prediction to capture Mendelian sampling term in a maritime pine breeding population with empirical and simulation approaches. In both cases, the accuracy of genome-based prediction was estimated at the population as well as at the within-family levels and was compared with that of pedigree-based prediction. The real data were obtained for a population of 39 full-sib families with family sizes ranging from 10 to 40 individuals per family. The simulation, designed to mimic the conditions of the maritime pine programme, added other scenarios not observed in the real population, including variations of heritability, training set size and marker density.

# Materials and Methods

## Exploring accuracy of predictions with real data

### *Maritime pine trial*

A maritime pine trial was established in 2011 in the Landes de Gascogne forest at Le Barp (Lat 44.62, Long -0.77). A complete block design was used, with 89 full-sib families and 10 checklots, each containing 48 individuals planted in six-tree plots (1,250 trees/ha). Full-sib families considered came from the third generation of the French maritime pine breeding programme (i.e. the pedigree of the trees is known back to the grand-parent level).

Preliminary simulations were performed to determine a relevant proportion of families and offspring per family among the total population in the trial, in order to maximize GS accuracy at both global and within-family levels (**Suppl. 3-1**). As a result of the simulation, an optimal sample of 40 families was obtained, with 30 of them containing 20 individuals and 10 of them containing 40 individuals. The selected families were representative of the genetic diversity present in the trial and the within-family samples were representative of the phenotypic variability of each family. The larger families corresponded to 5 of the best and 5 of the worst connected families. They are referred to in the following as large well-related and poorly-related families respectively, their average relatedness to the rest of the population being 0.03 and 0.01 (calculated with pedigree data). Considering families with large numbers of offspring is key to investigating within-family predictive ability. After genotyping, our study set, called $POP_R$, contained 833 individuals (see Results) with an effective population size equal to 25 (Lindgren et al., 1996). Thirty-nine families can be used to assess within-family predictive ability, including 9 families with more than 30 individuals.

### *Genomic and pedigree information for $POP_R$*

Genomic DNA was extracted from young needles collected on each individual of $POP_R$. Quantification and quality control of DNA were respectively carried out by fluorimetry (Qubit 2.0, Life Technologies, ThermoFisher Scientific, USA) and spectrophotometry (NanoDrop Technologies, Wilmington, DE, USA). Genotyping was performed by Thermo Fisher Scientific (Thermo Fisher Scientific, Santa Clara, CA, USA) with the 4TREE Axiom 50K SNP multi-species array (Guilbaud et al., 2020). Samples with a call rate of less than 97% were excluded from further analysis. In addition to the quality controls suggested by Thermo Fisher Scentific at the SNP level (CallRate ≥ 85%, fld-cutoff ≥ 3.2, het-so-cutoff: ≥ -0.3), we also excluded SNP with more than 5% Mendelian segregation errors, SNP with repeatability below 98% (estimated

with 42 duplicated samples), and SNP with a minor allele frequency (MAF) below 1%. Missing genotypes were imputed by assigning the average genotype within each full-sib family. Other more sophisticated imputation methods could not be applied due to the lack of genetic map. We computed a realized genomic relationship matrix (**G**) following VanRaden (2008) using the R-package AGHmatrix (Amadeu et al., 2016):

$$G = \frac{(M - P)(M - P)'}{2\Sigma p_i(1 - p_i)} \tag{3.1}$$

where $\boldsymbol{M}$ and $\boldsymbol{P}$ are matrices of dimension $n$ (number of individuals) x $p$(number of markers). $\boldsymbol{M}$ gives genotypes at each locus coded as -1 for one of the homozygotes, 0 for heterozygotes and 1 for the other homozygotes; and the $\boldsymbol{P}$ is a matrix of allele frequencies expressed as $2(p_i - 0.5)$, where $p_i$ is the observed allele frequency at marker $i$ for all genotyped individuals.

Additionally, pedigree information was available for $POP_R$ at the parental (41 seed parents and 40 pollen parents) and grand-parental (103 initial progenitors from the base population of the breeding programme) levels. Pedigree errors were detected using the R-package pedtools (Dehli Vigeland, 2022) by comparing the genotyping data of $POP_R$ with the genotyping data available for 78/81 of parents. A pedigree error was declared when more than 1% of mismatches on all SNP were detected between an individual and its recorded parent. Non-genotyped-parents were assumed to be correct. Where possible, pedigree errors were corrected by identifying a new parent with less than 1% mismatches with the descendant; otherwise, pedigree was noted as unknown. Complete and corrected version of the pedigree was used to calculate an additive relationship matrix **A**.

### *Phenotypic data*

All individuals in the original trial, including $POP_R$, were phenotyped at 8 years old for height (HT) and stem deviation to verticality (DEV). Phenotypic values were adjusted for within-site spatial effects using spline functions implemented in the R-package breedR (Muñoz & Sanchez, 2020) and pedigree information available for the whole trial. In the following, we will only consider corrected phenotypes for $POP_R$.

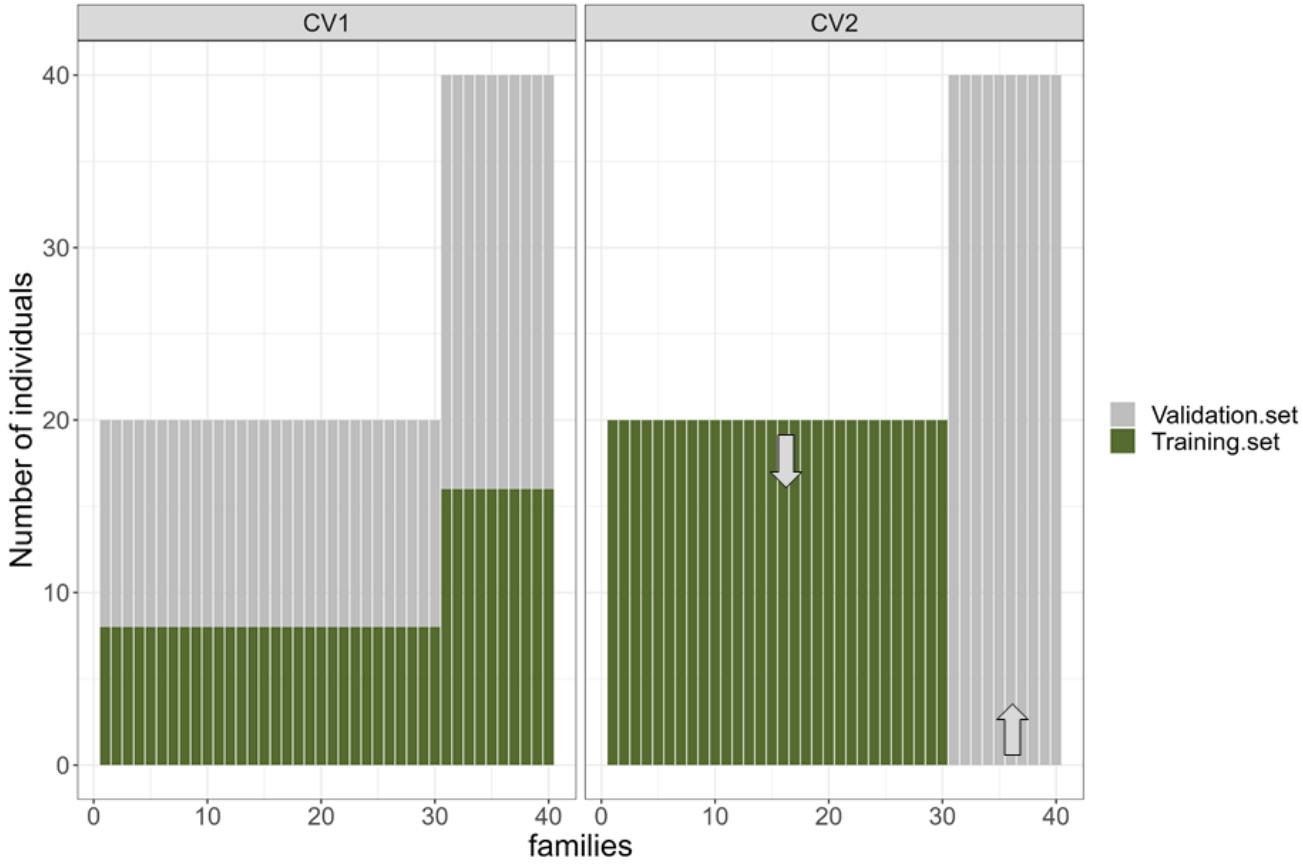



***Figure 3-1 : Cross-validation scenarios CV1 and CV2 performed with ABLUP and GBLUP models.*** *In CV1, the training set include 40% of each family, the remaining 60% being the validation set. In the first sub-scenario of CV2, all non-large families constitute the training set and all large families constitute the validation set. For other sub-scenarios of CV2, the contribution of non-large families to the training set decreases in favor of large families. Training sand validation set sizes remain constant between sub-scenarios of CV2.*

### *Genome-based and pedigree-based models*

Breeding values were estimated for each trait and for the n individuals using the model:

$$y = 1\mu + Zu + e \tag{3.2}$$

where $\boldsymbol{y}$ is the vector of adjusted phenotypes (dimension *n* x 1), **1** is the vector of 1s, $\boldsymbol{\mu}$ is the population mean associated with a vector **1** of dimension (*n* x 1), $\boldsymbol{Z}$ is the incidence matrix (dimension *n* x *n*) connecting the phenotypes to the vector of breeding values $\boldsymbol{u}$ (dimension *n* x 1) and *e* is the vector of residuals (dimension *n* x 1). The $\boldsymbol{u}$ and $\boldsymbol{e}$ are assumed to be independent from each other and to follow a normal distributions of the form $\boldsymbol{u} \sim N(0, \boldsymbol{X}\sigma_u^2)$ and $\boldsymbol{e} \sim N(0, \boldsymbol{I_n}\sigma_e^2)$, where $\boldsymbol{X}$ is either the genome-based (realized) relationship matrix $\boldsymbol{G}$ or pedigree-based (expected) relationship matrix $\boldsymbol{A}$, $\sigma_u^2$ is the associated variance of breeding values, $\boldsymbol{I}_n$ is the n-dimensional identity matrix and $\sigma_e^2$ is the variance of residuals effects. Mixed model equations were solved to predict the random genetic effects $\boldsymbol{u}$:

$$\begin{bmatrix} 1'1 & 1'Z \\ Z'1 & Z'Z + X^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} 1'y \\ Z'y \end{bmatrix} \tag{3.3}$$

where $\boldsymbol{X}^{-1}$ is the inverse of $\boldsymbol{X}$ and $\alpha = {\sigma_e}^2/{\sigma_u}^2$ (Henderson, 1975; Mrode & Pocrnic, 2023. These two model versions (GBLUP and ABLUP) respectively gave us genome-based (GEBV) and pedigree-based (EBV) estimates / predictions of breeding values. All model fitting was performed with the R.4.2.2 environment (R Core Team, 2022) using the R package breedR (Muñoz & Sanchez, 2020).

### *Cross-validation scenarios and assessment of prediction accuracy*

Two cross-validation scenarios were used to assess the prediction accuracies of the GBLUP and ABLUP models (**Fig. 3-1**): CV1 was designed to study within-family accuracy for the 39 families of $POP_R$ whereas CV2 focused on the 9 large families (i.e. families with 40 individuals sampled).

- Scenario **CV1:** the training set ($T_{set}$) included 40% of each family and the validation set ($V_{set}$) included the remaining 60% of each family.
- Scenario **CV2** was divided into 6 sub-scenario. For the first sub-scenario, the $T_{set}$ included the entire population $POP_R$ except the 9 large families and the $V_{set}$ included all the individuals from large families. The number of individuals from large families included in the $T_{set}$ was progressively increased in the other 5 sub-scenarios (from 5, 10, 15, 20 to 25 individuals per large family), while reducing accordingly the contribution of other families to $T_{set}$ to maintain

a constant $T_{set}$ size. Individuals from non-large families who are not included in the $T_{set}$ became part of the $V_{set}$, which therefore also maintained a constant size.

Each cross-validation sub-scenario was replicated 100 times. For each replicate, the predictive ability was calculated as the Pearson correlation between corrected phenotypes ($y$) and (G)EBV ($\hat{y}$) for individuals in the $V_{set}$. The predictive ability was calculated at the global level, i.e. using all individuals in the $V_{set}$ or at the within-family level, i.e. considering each family separately. Note that within-family prediction is only meaningful for the GBLUP model. Namely, ABLUP model will only predict parent average component of the breeding value for non-phenotyped full-sibs, meaning that the resulting EBV will be the same for all full-sibs giving a within-family predictive ability of 0.

As more common practice in forestry literature, the prediction accuracy, defined as the correlation between true ($g$) and predicted genetic values ($\hat{g}$) was obtained at the global level only to facilitate comparison with other studies (Legarra et al., 2008):

$$accuracy = r(g, \hat{g}) = \frac{r(y, \hat{y})}{\sqrt{h^2_{\hat{y}}}} \quad (3.4)$$

with $h_{\hat{y}}{}^2$ the heritability of the trait. We retained predictive ability as the metric for within-family analyses because dividing by the heritability defined at population level gives rise to values much higher than 1 or lower than -1. Instead, we focused on the deviation from 0.

## Identifying key parameters for GS accuracy with simulations

### *Simulation model description*

Stochastic simulations were carried out based on an allelic model using the R-package AlphaSimR (Gaynor et al., 2021). Details of simulations are provided in **Suppl. 3-2** and codes are available on Github. Briefly, from a base population built with parameters (genome size, demographic history, mutation rate) given in the literature, we simulated successive breeding populations based on a single trait (equivalent to HT) and considering the characteristics (population size, relatedness, selection intensity) of the real French maritime pine breeding programme. Simulated phenotypes and genotypes data for the final population $POP_S$ (the simulated version of $POP_R$) were used to fit GBLUP and ABLUP models as described for the real data. The analyses were performed with the scenario CV1 (40% of individuals in each family included in the $T_{set}$). Heritability of simulated phenotypes was set to 0.13 and the number

of markers used was 8234 to mimic the real-life data as closely as possible. The entire process described above was replicated independently 10 times to ensure the robustness of the results.

### *ABLUP and GBLUP prediction accuracy under different scenarios*

Stochastic simulations were used to extend the comparison between GBLUP and ABLUP prediction accuracy under different scenarios varying the trait heritability ($h^2$), training set size ($nT_{set}$) and number of markers (nSNP). For this purpose, $POP_S$ was extended by generating 100 individuals for each of the 40 initially sampled families. The prediction accuracy was assessed with a unique cross-validation scenario similar to CV1 (all families contributed in the same proportion to the $T_{set}$) but with a fixed-size $V_{set}$ of 1,200 individuals (evenly distributed between families). The values assumed by the three parameters mimic possibilities for the maritime pine breeding:

- The size of the $T_{set}$ was set to $nT_{set}$ = 400, 600, 1600 or 2600 individuals, corresponding respectively to 10, 15, 40 and 65 individuals per family. Such numbers are usually available in most forest trees breeding programmes. These numbers are also economically viable, because they require only 10, 15, 40 and 65% of the population to be phenotyped.
- The number of markers was set to nSNP = 8234, 17220 or 35000 SNP corresponding respectively to marker densities of 5.7, 12 and 24 markers/cM. Currently, 8234 SNP were already available in our real maritime pine data set, but this number could be increased by the development of new chips.
- The heritability of the trait was set to $h^2$ = 0.13, 0.33 or 0.50. Although the average heritability for HT in the maritime pine breeding programme is 0.33, it can vary between 0.13 and 0.50, depending on the trial considered.

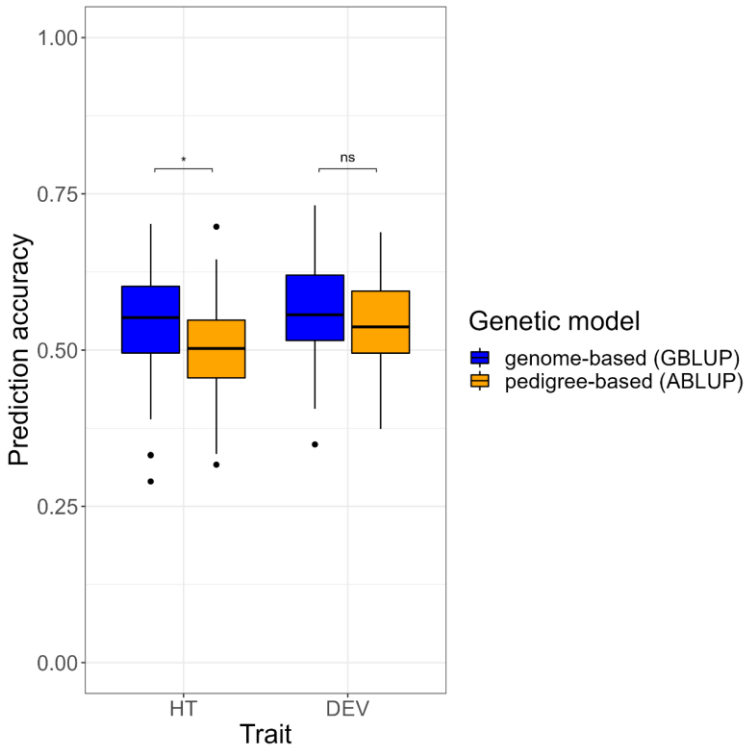



***Figure 3-2 : Global prediction accuracies obtained in the scenario CV1 with ABLUP and GBLUP models, for height (HT) and stem deviation to verticality (DEV)***

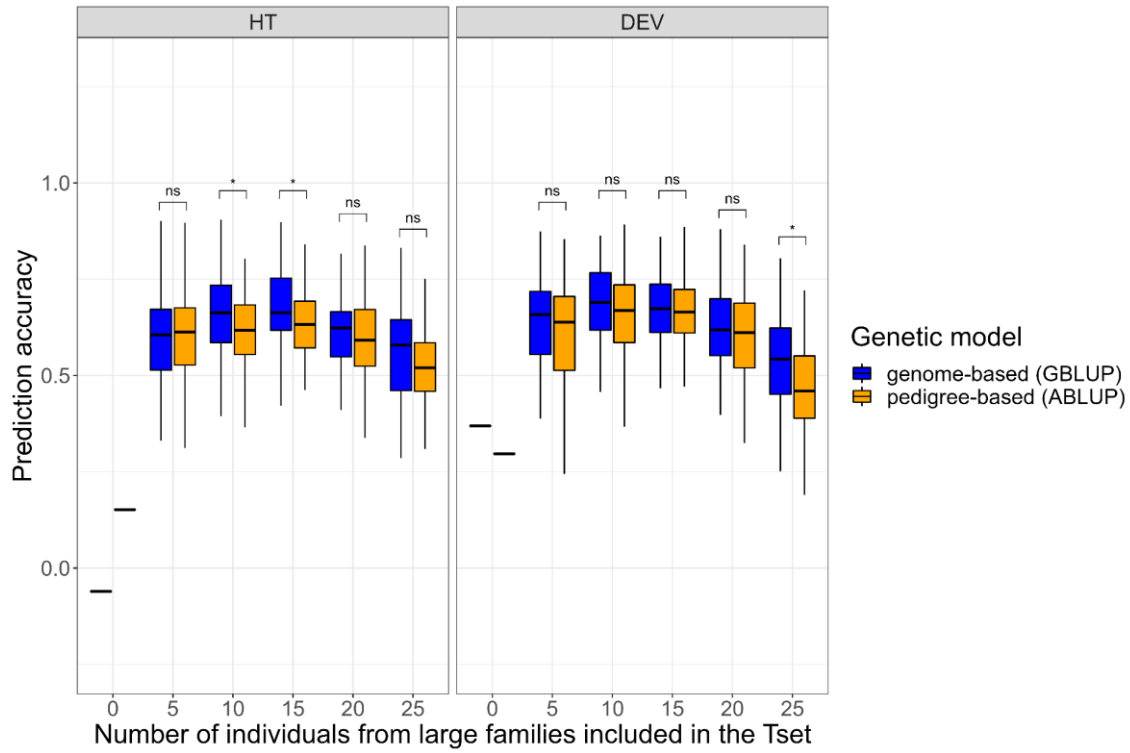



***Figure 3-3 : Global prediction accuracies obtained in the different sub-scenarios CV2 with ABLUP and GBLUP models, for height (HT) and stem deviation to verticality (DEV)***

# Results

## Genetic and phenotypic characterization of $POP_R$

After quality control, a total of 833 individuals characterized with over 8,235 SNP were available for this study. Out of the 42 individuals affected by the pedigree errors (5% of individuals), 27 were reassigned to another sampled family, subsequently reducing the number of families from 40 to 39, and the remaining 15 individuals were considered of unknown parentage. Within-family predictive ability was assessed for 39 full-sib families: 9 large-families with an average of 34 individuals per family (30 to 40) and 31 other families with an average of 17 individuals (10 to 20).

Heritability for HT and DEV within $POP_R$ was respectively 0.13 and 0.21 when estimated with thegenomic data, and 0.17 and 0.25 when estimated with the pedigree information.

## Global and within-family prediction accuracy with maritime pine data

### *GBLUP and ABLUP global prediction accuracies*

Global predictions accuracies estimated with the scenario CV1 are presented in **Figure 3-2**. Mean accuracies are similar for both traits, varying between 0.50 and 0.56. The prediction accuracies of the GBLUP model were slightly higher than that of the ABLUP model, on average +0.04 for HT and +0.02 for DEV. For this CV1 scenario, we also varied the percentage of individuals from each family included in the $T_{set}$, from 20% to 80% (40% being the modality presented in **Fig. 3-2**). Accuracy increases with this percentage, ranging from 0.45 to 0.62, but showing equivalence between ABLUP and GBLUP.

In the scenario CV2, global accuracy varies greatly depending on the structure of the $T_{set}$ (**Fig. 3-3**). By adding individuals from the large and from the other families, gradually increased the mean accuracy, reaching a maximum of 0.68 for both traits when 10-15 individuals from large families were included in the $T_{set}$. Beyond that point, global accuracy declines. In other words, the best accuracies are achieved when all families are equally represented in the training set, and the over-representation of the large families does not facilitate predictions. In most sub-scenarios, the differences between the GBLUP and ABLUP models were considered insignificant due to the similarity of the mean values and the large overlap of their standard deviations.

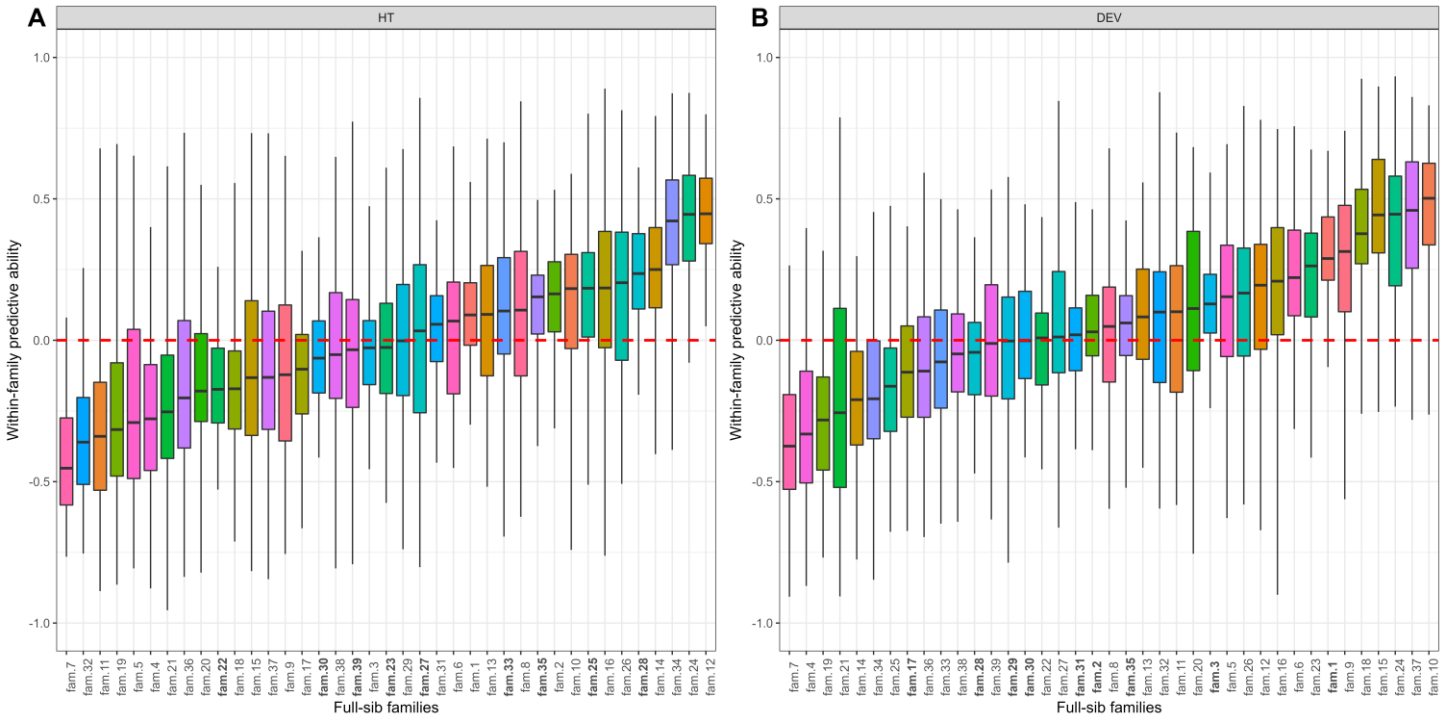



*Figure 3-4 : Genomic within-family predictive ability obtained in the scenario CV1 for each of the 39 full-sib families, for height (HT) and stem deviation to verticality (DEV)*

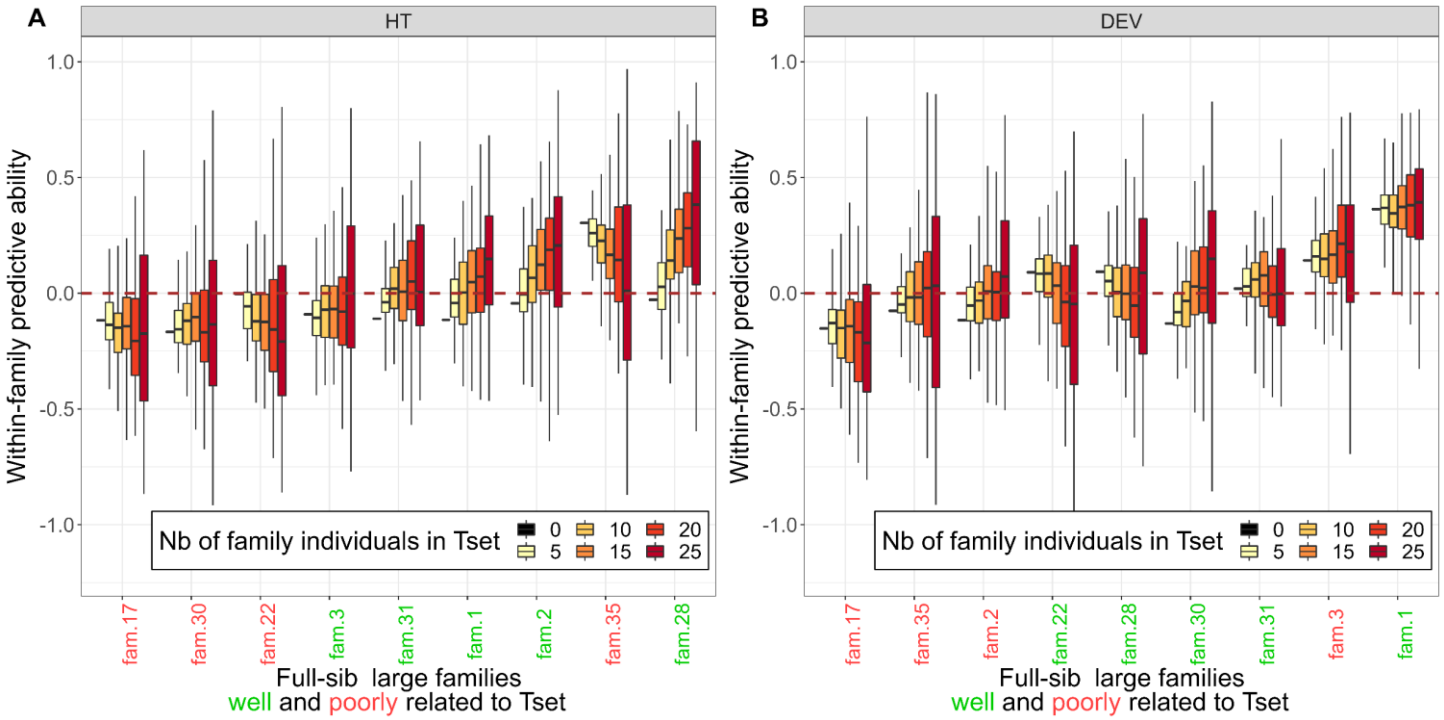



*Figure 3-5 : Genomic within-family predictive ability obtained in the sub-scenarios CV2 for each of the 9 large full-sib families, for height (HT) and stem deviation to verticality (DEV)*

### *Investigating the within-family genomic predictive ability*

The large number of individuals per family in our design enabled calculation of within-family prediction ability. **Figure 3-4** shows the predictive abilities obtained with the scenario CV1 when 40% of the individuals in the families were included in the $T_{set}$. Within-family predictive ability was therefore estimated with the 60% of individuals per family included in the $V_{set}$, i.e. an average of 20 individuals for the 9 large families and 10 individuals for the 30 other families. For each family, the variance of predictive ability was very high, indicating that the choice of individuals in the $T_{set}$ (and therefore in the $V_{set}$) has a major impact on the predictive ability. Mean predictive ability followed a contrasting gradient across families (**Fig. 3-4**), ranging from -0.43 to +0.46 for HT and from -0.35 to +0.46 for DEV. These mean values were centered around 0, with very low across families' averages (-0.02 for HT and +0.07 for DEV). Note that despite a similar distribution of within-family predictive abilities for the two traits, the ranking of families differs significantly between HT and DEV (Kendall correlation of +0.01).

Within-family predictive abilities was further investigated under the scenario CV2 (**Fig. 3-5**). As before, we can note that the variance of predictive ability significantly increased when the $V_{set}$ size decreased. Furthermore, the ranking of families with regards to the predictive ability was quite different between the two traits. Although predictive abilities remained low for both traits, families well-related to the training set had higher predictive ability compared to poorly-related families (on average +0.12 for HT and +0.16 for DEV). For families well related to training set, adding individuals in the $T_{set}$ either had no impact on the within-family mean predictive ability, or increased the accuracy, as in the case for instance for families 1, 2 and 28 for HT. For families poorly related to training set, adding individuals in the $T_{set}$ had no impact, increased, or, more surprisingly, decreased the within-family predictive ability, as in the case of families 22 for both traits. For both scenarios (CV1 and CV2), despite strong variation between the families, within-family predictive abilities remained close to 0 on average, and may explain the equivalence in terms of global predictive ability between GBLUP and ABLUP models.

## Exploring prediction accuracy with simulations

### *A relevant simulation model*

The comparison of the simulated data with the real data is an interesting preliminary step in assessing the relevance of the simulation model that has been constructed. For both models, GBLUP and ABLUP, the simulated and real data gave very similar prediction accuracies

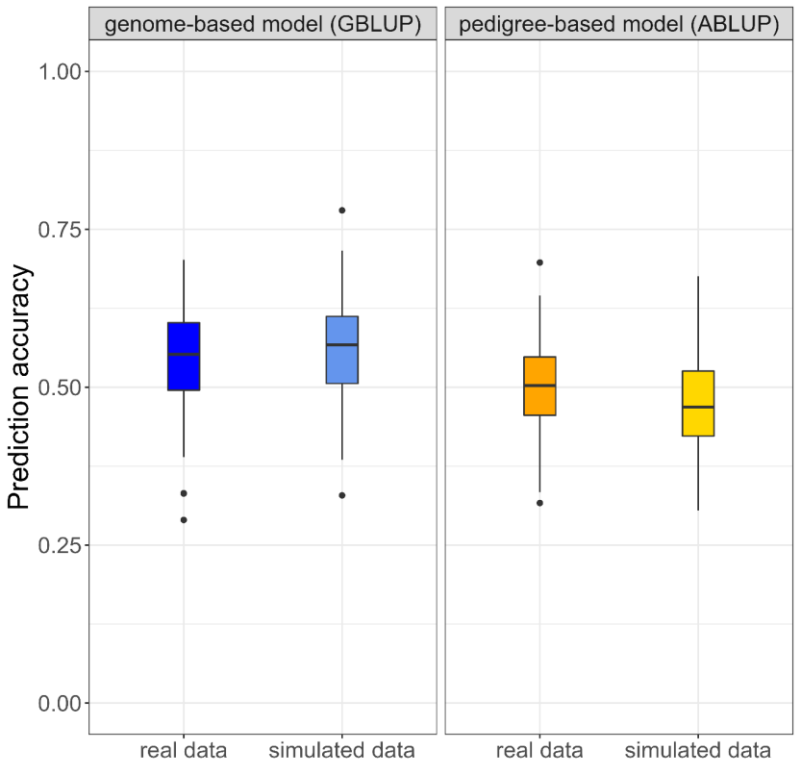



***Figure 3-6 : Global prediction accuracies obtained in the scenario CV1 (40% of individuals in each family included in $T_{set}$) for height (HT) with ABLUP and GBLUP models based on real or simulated data.***

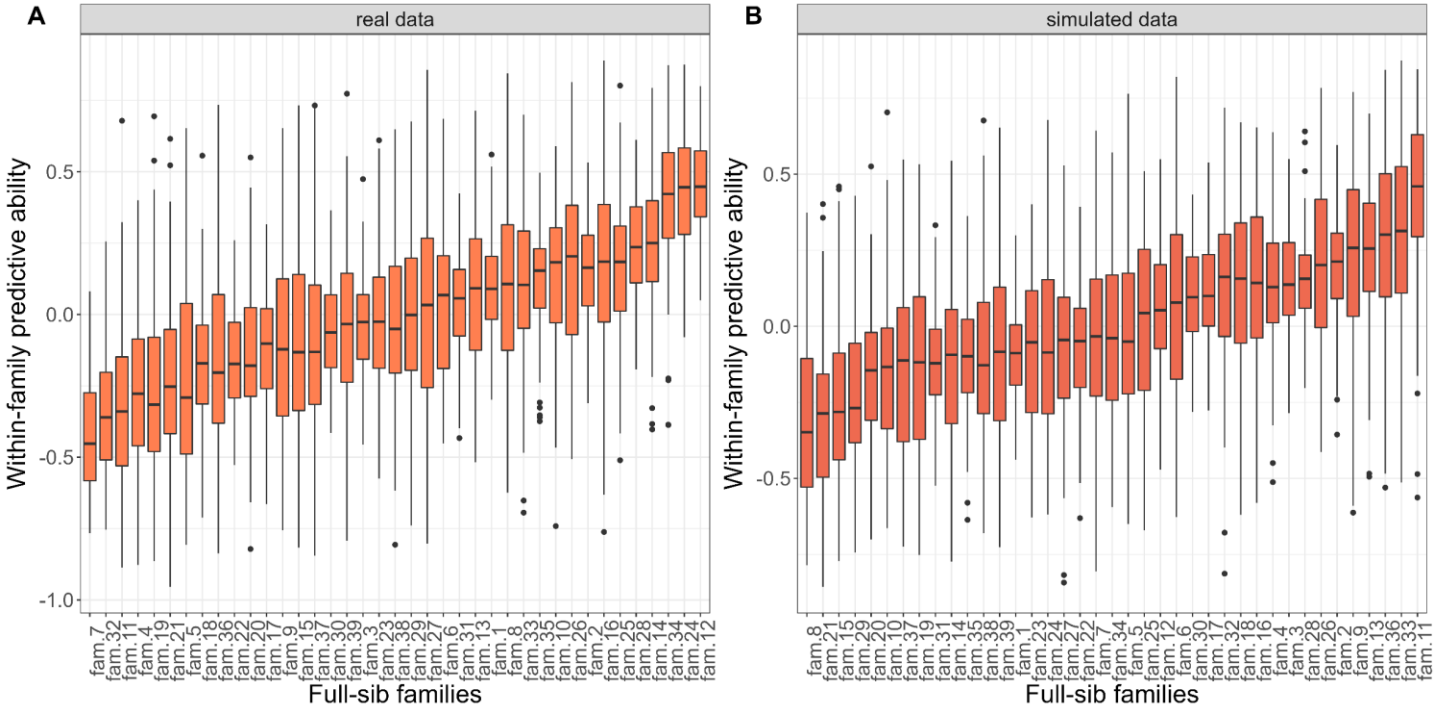



***Figure 3-7 : Genomic within-family predictive ability obtained in the scenario CV1 for height (HT) for each of the 42 full-sib families with real or simulated data***

(**Fig. 3-6**) in terms of mean (0.54 and 0.55 respectively for real and simulated data with GBLUP model, 0.49 and 0.47 with ABLUP model), and distribution of values (coefficient of variation of prediction accuracies being 15% and 14% respectively for real and simulated data with GBLUP model, 15% and 16% with ABLUP model). Again, with simulated data, we found a slight advantage in prediction accuracy for the GBLUP over the ABLUP models. The simulations also showed a high level of agreement with the real data when comparing accuracy performance within families (**Fig. 3-7**). The high variance at the level of each family in terms of within-family predictive abilities described previously with the real data was also observed with the simulated data, as it is the strong variation between family in terms of average prediction accuracy. Within-family predictive abilities ranged from -0.43 to +0.47 with the real data, and from -0.31 to +0.42 with the simulated data. It should be noted that the ranking of families in terms of within-family predictive ability is not the same between the real and the simulated data, because the genotypes were simulated independently of the real data.

### *Determining relevant conditions for GS implementation*

Starting from the initial conditions defined by the real data ($h^2$=0.13, nSNP=8234 and $nT_{set} \in [167:667]$ depending on the CV1 scenario modality), new scenarios with different conditions were explored using the simulations. **Figure 3-8** shows the overall prediction accuracy for different combinations of $h^2$, $nT_{set}$ and nSNP. The objective was not only to study the behavior of the accuracy as a function of the variation of these key parameters, but also to compare it with the prediction accuracy obtained using only the pedigree, i.e. the reference baseline for many forest tree improvement programmes.

First, marker density did not appear to be of a critical importance for the GS accuracy, since the blue, red and green curves overlap in most situations. The greatest benefits of higher marker density (12 and 24 markers/cM) in the tested scenarios were seen when large $T_{set}$ and high heritability were combined. Similarly, the same combination of parameters produced lower levels of variation in accuracy. Overall, the results for marker density show similarity at the two higher densities, suggesting a saturation of the accuracy with 17,000 markers.

Irrespective of the heritability, prediction accuracy appeared to be highly dependent on the $T_{set}$ size. The GS accuracy increased steadily for the first steps of $T_{set}$ size ($nT_{set}$=400, 600 and 1600) and then tended to stabilize around $nT_{set}$=2600, indicating the plateau, with inflexion points between $nT_{set}$=1600 and $nT_{set}$=2000. While the prediction accuracy of ABLUP models followed the same trend, the advantage of GBLUP models over ABLUP was greater with the larger $T_{set}$

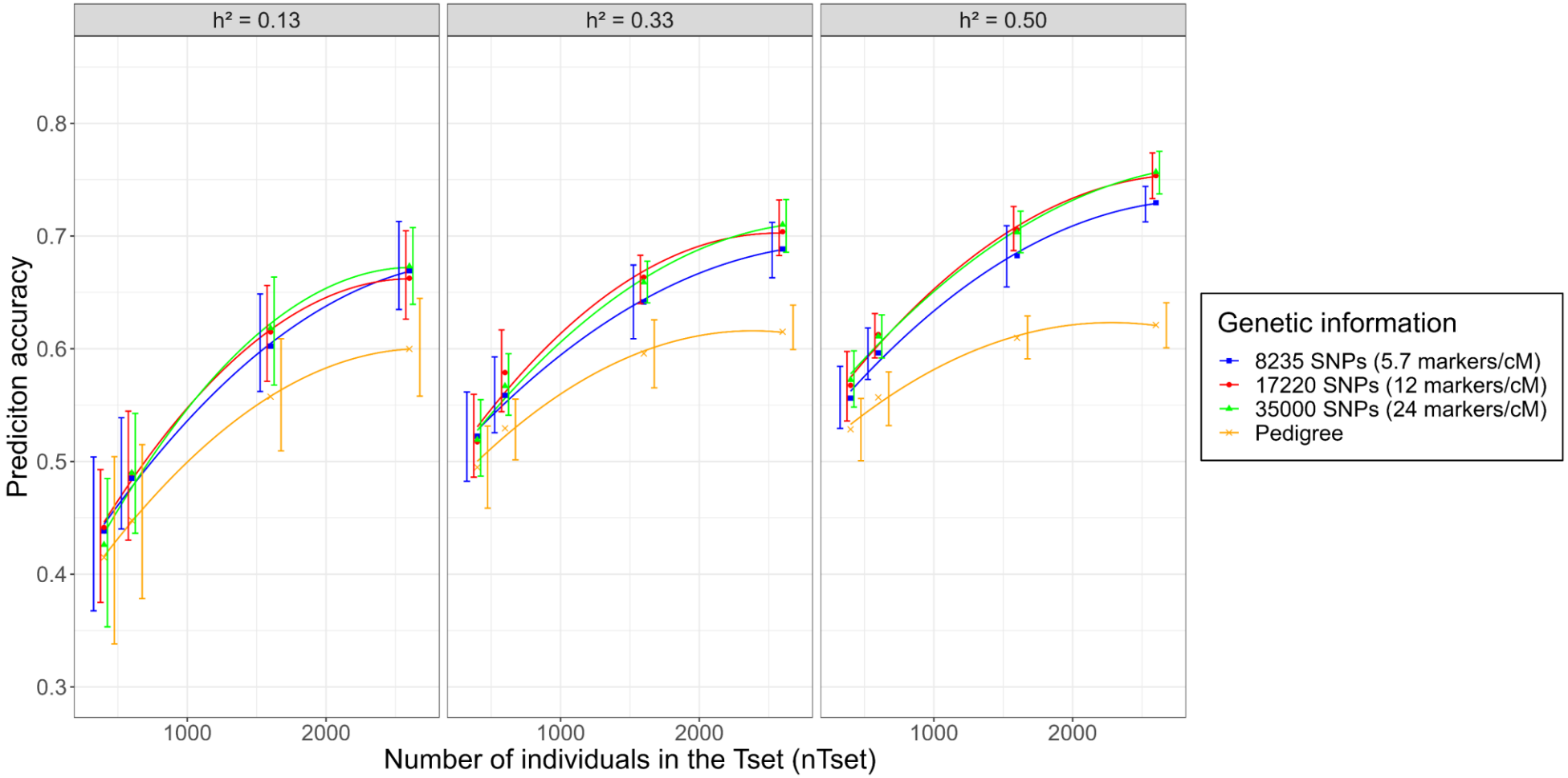



***Figure 3-8 : Global prediction accuracies for height (HT) with simulated data for different combinations of heritabilities ($h^2$), training set sizes ($nT_{set}$) and marker densities (nSNP).***

and the higher heritability. For an average heritability ($h^2$=0.33), the advantage of GBLUP models in terms of mean prediction accuracy was only +0.02 when $nT_{set}$=400 (or 600), whereas it reached +0.08 when $nT_{set}$=1600 (+0.13 with 2600). Clearly, the results of the simulations indicate that the conditions under which the real data were analyzed were not optimal to reveal an advantage of the GBLUP model over the ABLUP model. The GBLUP model would be much more advantageous, according to these simulations, with an increase in $nT_{set}$ ($nT_{set} \geq 1600$), with intermediate or higher heritabilities and, in addition, with a higher number of markers.

For two of the situations where GBLUP shows an advantage over ABLUP, where the difference was minimal (+0.05 in mean accuracy) and where it was maximal (+0.13), we present the within-family predictive abilities in **Figure 3-9**. As with all the within-family predictive abilities previously presented, the variances were high and the mean values varied between the families. However, in this case the average prediction levels were positive for most families, ranging from 0 to 0.45 for the first situation ($h^2$=0.33, $nT_{set}$=1600, nSNP=17220) and from 0.10 to 0.50 for the second situation ($h^2$=0.5, $nT_{set}$=2600, nSNP=35000). Considering all the families, average within-family predictive abbilities is respectively +0.18 and +0.29 for the two situations shown, indicating a clear advantage of GBLUP over the reference value of 0 associated with the ABLUP. The advantage of the GBLUP model over the ABLUP model in terms of overall prediction accuracy coincides with non-zero within-family predictive ability. This advantage of the GBLUP models was increasingly evident for higher values of within-family predictive ability.

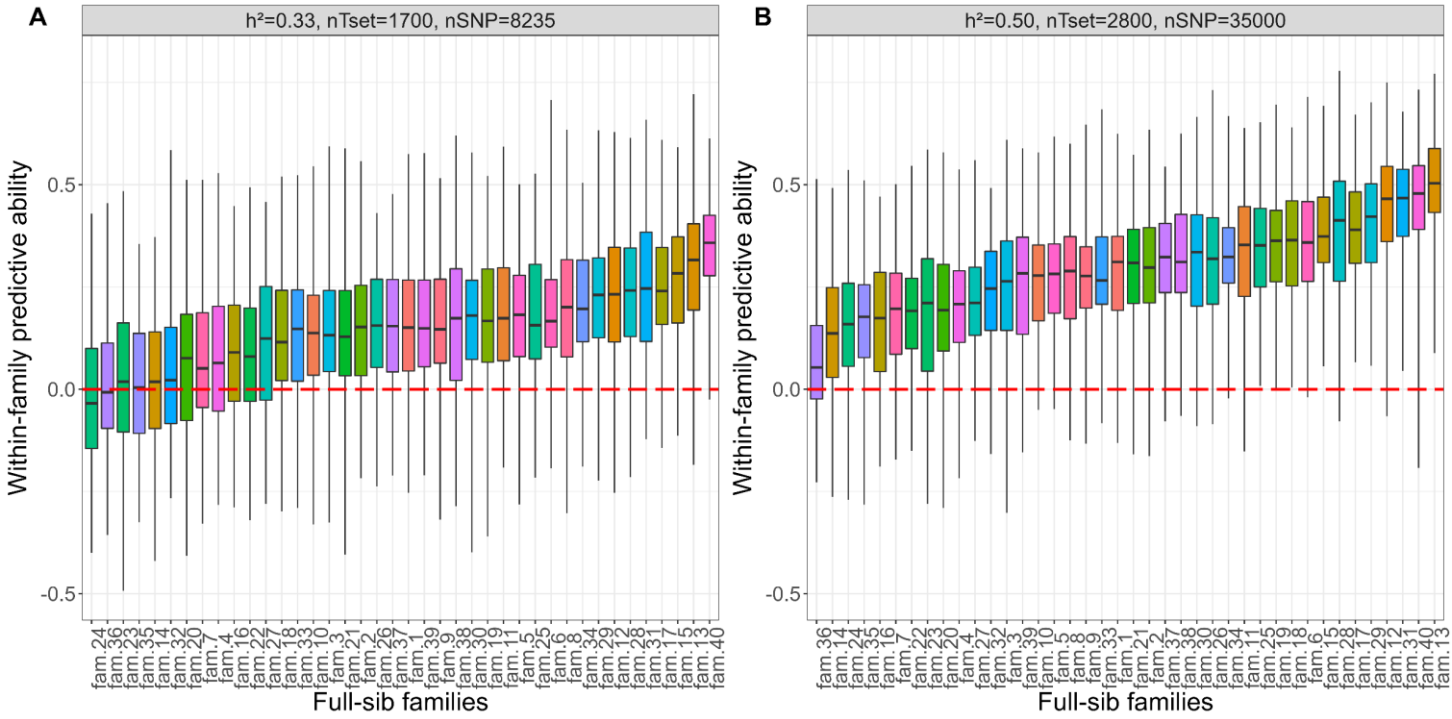



***Figure 3-9 : Genomic within-family predictive ability for each full-sib family for height (HT), with simulated data and for two different combinations of heritability ($h^2$), training set size ($nT_{set}$) and marker density (nSNP).***

# Discussion

Before implementing genome-based prediction and selection in maritime pine and more generally in forest trees, it is necessary to define more clearly the conditions that give this approach a real advantage over the conventional methods, particularly in the terms of prediction accuracy (Bartholomé et al., 2016; Beaulieu et al., 2014). In contrast to most previous genomic studies on the forest trees, we evaluated genome-based prediction accuracies against the corresponding pedigree-based prediction accuracies, and we investigated within-family accuracy. The reference we considered here is key as it allows a fair comparison between predictions based on genome and pedigree information, available at the same time (seedling stage), without the need to calculate a selection response considering the length of breeding cycles. We found that even with a test design that a priori favoured within-family variability, the advantage of GBLUP over ABLUP models was not clearly significant, and within-family predictive abilities was null on average across families. Complementing these results with stochastic simulations mimicking our selection scheme allowed us to identify that our real case was close to the tipping point in terms of recommended training set size, where genome-based predictions would start to show its full potential compared to conventional pedigree-based predictions.

## Equivalence of ABLUP and GBLUP prediction accuracy with a maritime pine dataset

### *Baseline for comparison between ABLUP and GBLUP models*

Bartholomé et al. (2016) presented the superiority of prediction accuracy with ABLUP over GBLUP models in maritime pine, and similar conclusions were founded in several studies in the forest trees (Thistlethwaite et al., 2017; Zapata-Valenzuela et al., 2012; Zhou et al., 2020). Most commonly, insufficient marker density has been pointed to justify the lack of advantage of genome-based predictions. Our study differs from the previous studies in maritime pine by a higher marker density (5.7 SNP/cM versus 2.4 and 1.7 SNP/cM respectively in Bartholomé et al., 2016 and Isik et al., 2016). Furthermore, the original design of $POP_R$ characterized by a limited number of full-sib families but relatively large numbers of trees per family was expected to favor the genome-based approach, the one able to capture the Mendelian sampling, terms that drive the within-family variation

However, the GBLUP models showed only a slight advantage over the ABLUP models in terms of prediction accuracy for some of the cross-validation sub-scenarios evaluated (**Fig. 3- 2 and 3-3**), and even this advantage may be partly due to the way the prediction accuracy is calculated. Prediction accuracy, obtained from the ratio of predictive ability to the square root of heritability of the predicted trait requires the estimation of variance components which are subject to additional errors (Legarra et al., 2008). As in other studies, GBLUP estimates of heritability were lower than the estimates obtained with the ABLUP model (El-Dien et al., 2018; Lenz, Nadeau, Mottet, et al., 2020; R. T. Resende et al., 2017). This resulted in higher prediction accuracies for the GBLUP model despite a predictive ability similar to that for the ABLUP model. However, prediction ability is less common in the literature and poses problems of interpretation in terms of genetic gain since it includes environmental effects, which justifies our choice to present prediction accuracies.

We also decided to base our comparison solely on the accuracy, without converting it into a response to selection. The predictions obtained with pedigree-based model takes into account only the pedigree information for the individuals in Vset (not their phenotypic records). The predictions of this model are therefore available for use in early selection, as for genome-based predictions. We believe that the assumption of a shorter selection cycle for genome-based predictions does not hold and that this assumption introduces a bias, increasing the perceived advantage of GS over traditional approaches (Lenz et al., 2017; Ratcliffe et al., 2015; Resende Jr et al., 2012). Genome-based prediction is clearly advantageous when it enables pedigree recovery and correction (Zapata-Valenzuela et al., 2012), but similar results can routinely be obtained with a very small number of carefully chosen markers across the genome (Vidal et al., 2017). For the use of genome-based prediction to be considered truly advantageous, it must, therefore provide a higher accuracy than pedigree-based prediction associated with a complete corrected pedigree.

### *Capture of linkage disequilibrium and relatedness with GBLUP models*

Theoretically, genome-wide markers used for genome-based predictions can provide additional information beyond that obtained from a simple pedigree, such as the ability to capture the effects of nearby QTL in linkage disequilibrium (Habier et al., 2007), and also to reveal more accurately the relationship between any two given individuals (Nejati-Javaremi et al., 1997; VanRaden, 2008). In this study, however, these advantages to the use of markers were not apparent. In our case, the equivalence between ABLUP and GBLUP (**Fig. 3-2 and 3-3**) could

be explained by the two main reasons, also pinpointed by other authors with similar outcomes (Beaulieu et al., 2014). First, the estimation of genomic effects may not be very accurate, which could come from an insufficient $T_{set}$ size to guarantee a good estimate (Habier et al., 2013). For example, with simulations, Hayes et al. (2009) showed that the advantage of GBLUP over ABLUP models was apparent from 100 individuals per full-sib family. Second, it is possible that genome-wide markers only capture existing genetic relationships (Legarra et al., 2008) and are not effective in capturing actual LD (Habier et al., 2007).

The lack of long-range LD has already been described for conifers (Eckert et al., 2010; Kujala & Savolainen, 2012), including maritime pine (Isik et al., 2016; Plomion et al., 2014), suggesting that a large number of markers may be required for genome-based predictions (R. T. Resende et al., 2017; Thistlethwaite et al., 2019). Using simulations, Grattapaglia et al. (2011) suggested about 10 markers/cM to increase the prediction accuracy of GBLUP above that of ABLUP when Ne=100. The small genetic map size compared to the large physical size of most conifer genomes [respectively 1435 cM (Chancerel et al., 2013) and 24Gb (Chagné et al., 2002) for maritime pine] suggests that large parts of the genome have very few recombinations. Therefore, rather than the number alone, it may also be the distribution of markers across the genome that counts. However, this can be difficult to improve when there is no physical map of the species, as is the case for maritime pine. Therefore, the densities used in our real case would be sufficient to capture pedigree-like relatedness, but still not the densities needed and probably not the good distribution to capture QTL information across the LD. This is what our simulations suggested, together with the fact of increasing the training set. One alternative not considered in our study would be the use of alternative statistical approaches that do Bayesian variable selection, such as Bayes-B. Although they have been shown to better capture the population LD and have more weight put on the causative SNP (Habier et al., 2007; Thistlethwaite et al., 2017), theirs benefits are often case and trait specific, and may even disappear especially when the training population is large enough (Karaman et al., 2016).

### *Genome-based within-family predictive ability*

Unlike ABLUP model, GBLUP models capture realized relatedness between and within family through the **G** matrix. Nevertheless, whatever the CV scenario tested in this study, within-family predictive ability in the real data was on average null when considering all full-sib families (**Fig. 3-4**) indicating that underlying genome-wide marker associations in GBLUP (Strandén & Garrick, 2009) were not estimated accurately. For both traits, within-family predictive abilities were not significantly different from zero in most families, but there was

considerable variation across the families, including some (15%) for which accuracy values were, surprisingly, significantly negative. For some relatively small families, the number of individuals included in the Vset was probably too small for a robust estimation of correlations, resulting in a very large standard deviation when all CV iterations were taken into account. By contrast, within-family predictive ability for large families was calculated with a mean of 20 individuals (in the Vset), resulting in zero or slightly positive values, but with a lower standard deviation. Genome-based prediction accuracy was, therefore, mostly driven by capturing parent average term rather than capturing the Mendelian sampling term of breeding values. This partly explains the observed equivalence with pedigree-based prediction models, as suggested in other studies with similar results (Thistlethwaite et al., 2019).

Some specific scenarios of cross-validation that imposed restrictions on relatedness between $V_{set}$ and $T_{set}$, particularly where $V_{set}$ did not contain relatives of individuals present in $T_{set}$, resulted in particularly low prediction accuracies at the global (**Fig. 3-3**) and within-family levels (**Fig. 3-5**). This suggests that relatedness was the main source of information in both GBLUP and ABLUP, with little or no additional information captured from LD to maintain prediction quality in the absence of relatedness. Globally, our population had low relatedness between the families, with parents and grandparents at the founding level of the population producing on average 1.1 and 2.4 full sib families, respectively. Such a structure may have created a challenging condition for GBLUP to outperform ABLUP. In general, it is always desirable to have a diversified training population to produce robust predictions and a validation set that is well related to the training set. The first condition could have been met by having different grandparents and parents equally represented. The second, however, is less clear as the sibs in $V_{set}$ had probably little number of collaterals other than the remaining sibs already present in $T_{set}$.

## Identifying conditions for the superiority of GBLUP over ABLUP prediction accuracy using simulations

The stochastic simulation model produced results comparable to those obtained empirically, giving a degree of confidence in the relevance of the results, both in terms of the observed trends and the absolute values shown. The effect of increasing the $T_{set}$ size and the number of genome-wide markers on the accuracy of GS was tested under three contrasting heritability scenarios, each with an explicit comparison with the ABLUP prediction accuracy. These simulations showed that the size of the $T_{set}$ seems to be the most important determinant of the

prediction accuracy in our study. The inflection point of the curves was between 1,500 and 2,000 individuals, which is also consistent with a deterministic approaches in other forest tree contexts (Grattapaglia et al., 2011; Grattapaglia & Resende, 2011). It is from this $T_{set}$ range that GBLUP begin to show significantly higher prediction accuracies compared to the ABLUP, with narrower confidence intervals that avoid overlapping distributions between the two types of models. Increasing the number of individuals per family, which occured in the simulation scenarios as $T_{set}$ increases, allowed breeding values to be estimated with greater accuracy (Habier et al., 2013). Therefore, the often-highlighted equivalence between the GBLUP and ABLUP accuracies in the forest trees can be explained by the insufficient size of the training sets, which rarely exceeds 1000 individuals, regardless of the species considered (Lebedev et al., 2020). On the contrary, increasing the number of markers did not seem to be a critical factor in improving the genome-based predictive accuracy, at least for the moderate heritability values, which could indicate that our initial marker density would have been sufficient when coupled with a larger training population. This conclusion could be readily extrapolated to other conifers, which share similar characteristics in terms of genome size and effective breeding population size, but is more difficult to extrapolate to deciduous species whose genome structure and LD profiles can be very different.
Complementary analysis of genome-based prediction accuracy precision made using deterministic approaches (Daetwyler et al., 2008) with the empirical parameters for maritime pine in our population (see **Suppl. 3-3**) corroborate our stochastic simulation results.

Within-family prediction accuracy is rarely considered in genomic studies but appears to be key for the superiority of genome-based predictions over pedigree-based predictions to be expressed. Simulations showed that more accurate within-family prediction was associated with a greater accuracy advantage of GBLUP over ABLUP. For our study design, the suggested Tset size for efficient within-family prediction corresponds to between 40 and 65 individuals per full-sib family. This is a very important requirement for the implementation of genome-based prediction and selection, as the use of GBLUP models with zero within-family predictive abilities can have several negative consequences.

In the short term, the effectiveness of selection and the response to selection would be reduced if one of the sources of genetic variation in the population, the within-family variation due to Mendelian sampling, was not exploited. While this source of variation can be measured with genome-wide markers at the genotype level, attaching quantitative value to these genotypes requires sufficiently powered training set of phenotyped and genotyped individuals. Ensuring

such sufficiently large training set is important to avoid longer-term consequence related to the fact that selection on underpowered genome-based predictions would be only leveraging variation between families. Such an approach would increase the risk of losing diversity due to the elimination of certain lineages and the co-selection of candidates from the same families. This loss of diversity due to a shift in the weighting between within-family and between-family selection would lead to long-term losses of genetic gain (Jannink, 2010) and an accumulation of inbreeding.

This tendency can be counteracted by selection methods based on the optimization of genetic contributions (Meuwissen, 1997; Woolliams et al., 2015)— so-called "optimal contribution selection" (OCS) — which allows a trade-off between short-term and longer-term gains through the application of constraints to the balance between parental genetic contributions (Gorjanc et al., 2018). Future studies should assess these optimal strategies which would, presumably, work better if genome-based predictions could discriminate between candidates within families more accurately (Hallander & Waldmann, 2009), with sufficiently large families in the training populations, thereby increasing the efficiency of selection and of the constraints imposed by the OCS.

# Conclusion

Despite the undeniable potential benefits for forest trees, examples are lacking for which genome-based approaches have clearly demonstrated superiority over pedigree-based approaches. Using an ABLUP model with full and corrected pedigree information as a reference, we evaluated the accuracy of GS in a maritime pine trial with the largest number of individuals per full-sib family to date. Prediction accuracy was found to be similar for the pedigree-based and genome-based models, and within-family genomic accuracy for forward predictions was close to zero. By constructing a relevant simulation model, we were able to demonstrate that the number of individuals per family, and thus the overall size of the training set, is a key parameter for accurately estimating marker associations and for detecting a clear advantage of genome-based approach. This conclusion can be extended to many forestry contexts in which the equivalence between ABLUP and GBLUP prediction accuracies can be explained by suboptimal training set sizes and structures. Increases in training-set size may be readily achievable in forestry due to the large numbers of individuals commonly used in breeding programmes and decreasing genotyping costs. Effective within-family prediction, based on well-scaled genome-based approaches, will be key to maintaining diversity in the long term and ensuring genetic gain in the challenging years ahead.

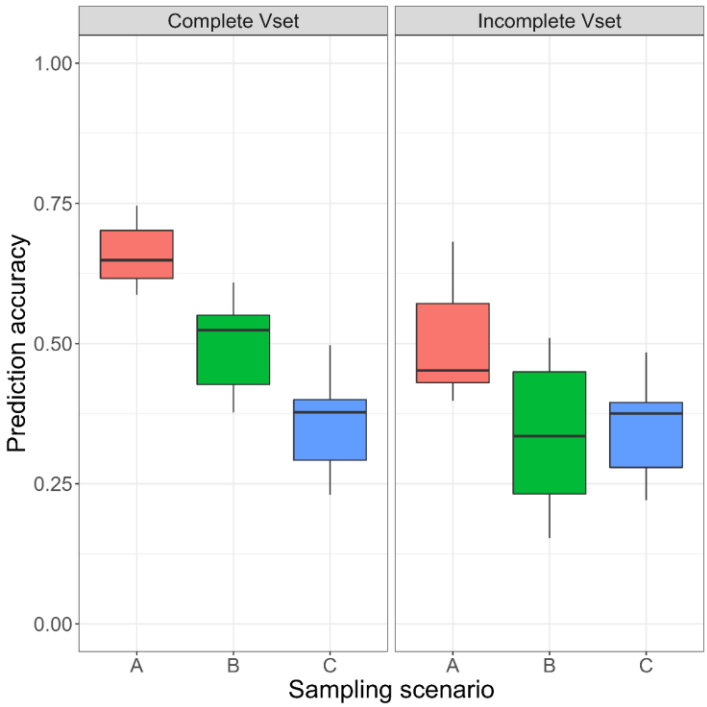



***Figure 3-S1: GS prediction accuracy at the global level for the different sampling scenarios (A: 40 families x 20 individuals, B: 32 families x 25 individuals, C: 20 families x 40 individuals).*** *"Complete Vset" indicates that all individuals in the Vset were considered to assess prediction accuracy, while "incomplete Vset" indicates that only individuals from families not included in the Tset were considered to assess prediction accuracy.*

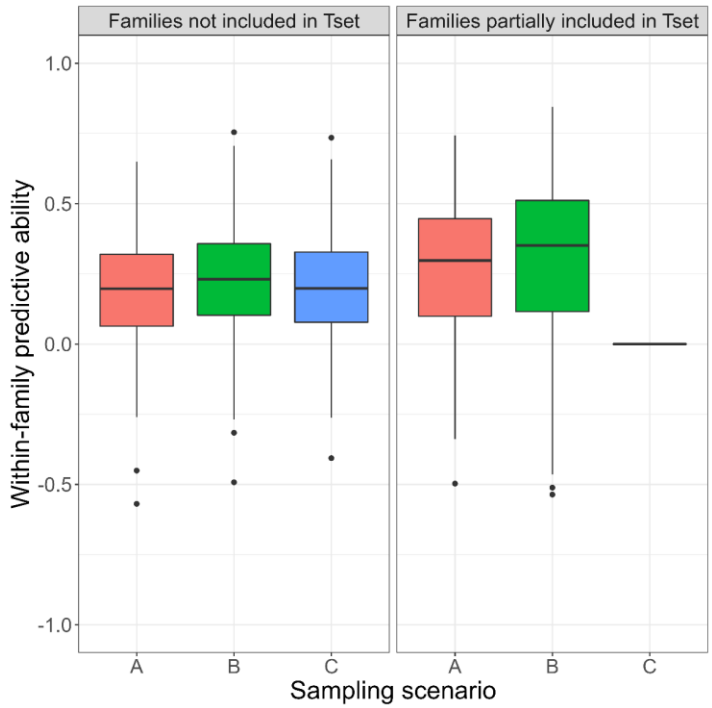



***Figure 3-S2 : GS within-family prediction accuracy for the different sampling scenarios (A: 40 families x 20 individuals, B: 32 families x 25 individuals, C: 20 families x 40 individuals).*** *On the left-hand side, assessing within-family accuracy for all the families not included in the Tset, and on the right-hand side only for families with some individuals in the Tset. In the latter modality, prediction accuracy is not available considering sampling scenario C since all the individuals in these families are included in the Tset.*

## Supplementary 3-1: definition of sampling within the trial for GS analysis

The initial maximum capacity for genotyping in this study was 800 individuals. Thus, preliminary simulations were performed to determine a relevant choice of these 800 individuals among the total population in the trial.

### Choice of sampling scenario and families

We compared 3 sampling scenarios to select 800 individuals among the 90 families present in the trial: 40 families with 20 individuals each (A), 32 families with 25 individuals each (B) and 20 families with 40 individuals each (C). The best scenario was selected on the basis of the prediction accuracy obtained, following a three-step procedure repliacted 10 times independently:

- Genotypes and phenotypes were simulated for the families present in the trial using the MoBPS software (Pook, 2022).
- For each scenario, the subset of 20, 32 or 40 families was selected with the goal to minimize the average genomic relationship among the subset (with genomic relationships calculated using the simulated genotypes) and with a simple trail-and-error optimization process. The aim was to select the subset that maximizes the genetic diversity.
- Genome-based models were run using the R-package breedR for each scenario using simulated data and the accuracy was assessed through a cross-validation routine (the 800 individuals selected were as the $T_{set}$ and the remaining ~2000 individuals from the trial were in the $V_{set}$). For each sampling scenario we assessed both global (**Fig. 3-S1**) and within-family accuracy (**Fig. 3-S2**).

Finally, the scenario using 40 families with 20 individuals each, was chosen since it achieved significantly higher genome-based prediction accuracy at the global level, and similar within-family accuracy compared to the other 2 scenarios.

### Choice of individuals within each family

Within each of the 40 selected families, we applied the Kennard-Stone algorithm (Kennard and Stone, 1969) implemented in the R-package prospectr (Stevens, 2022) to get the subset of 20 individuals that maximize phenotypic diversity (now based on the real phenotypic values) for the three traits of interest: circumference and height at 9 years, and deviation to verticality at 8 years.

**Addition of 200 individuals**

Subsequently, additional 200 individuals could be genotyped and added to the study. As the sampling of the 800 individuals was already performed, we chosen to add 20 additional individuals in the 10 out of the 40 selected families. These families (in the main manuscript labeled as the “large families”) were finally made up of 40 individuals each.

These ten families were chosen to provide a contrast in connectivity within the trial. They correspond to 5 of the best and 5 of the worst connected families, their average relatedness to the rest of the population being 0.03 and 0.01 (calculated with pedigree data). Note that this choice of large families was made a posteriori so that the gap in relatedness with the rest of the population between the two groups of large families may be not very pronounced.

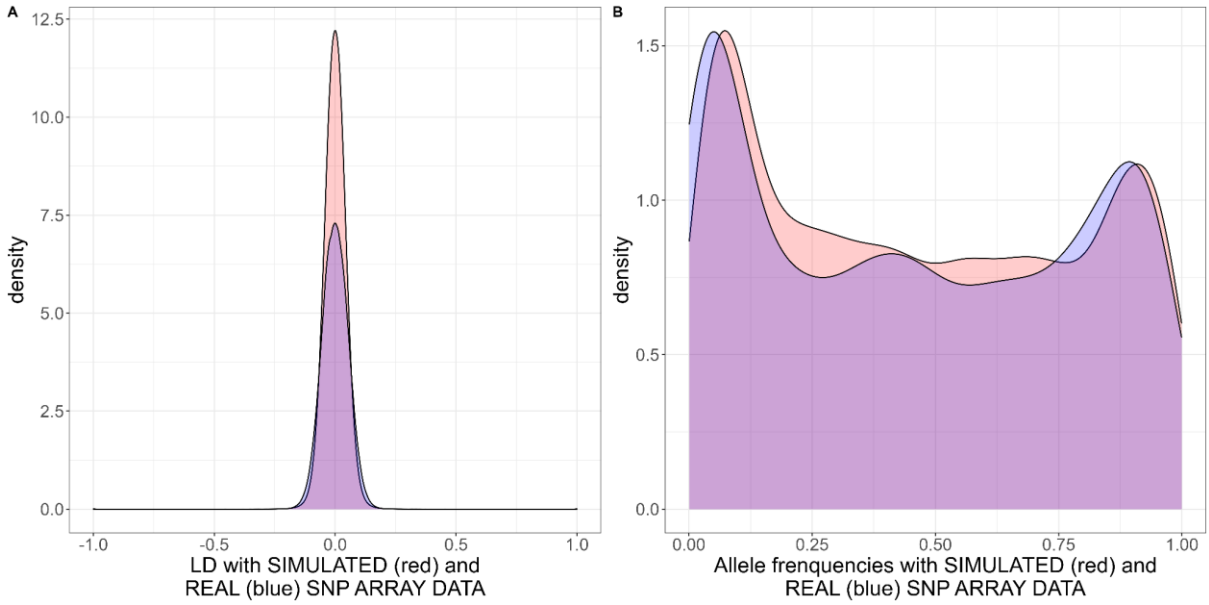



***Figure 3-S3 : Comparison of LD (A) and allele frequency (B) distributions between real and simulated SNP array data***

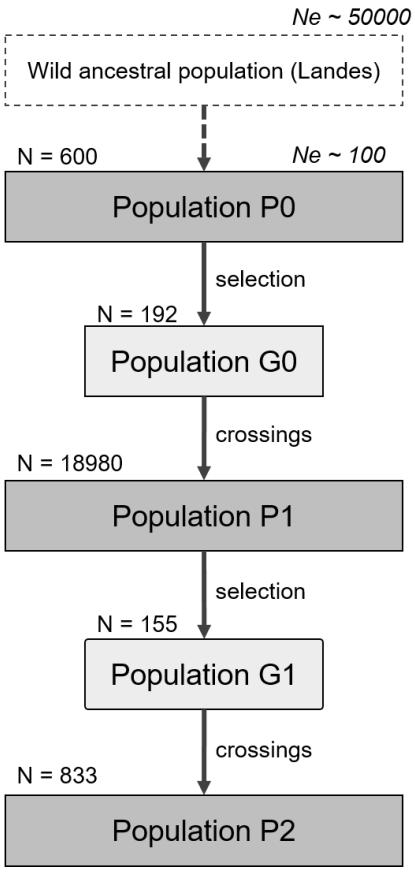



***Figure 3-S4 : Breeding cycles simulated to mimic the French maritime pine breeding program***

## Supplementary 3-2: simulation of the French maritime pine breeding programme

We simulated a diploid maritime pine genome composed of 12 chromosomes each with a physical length of 2.14e+09bp (Chagné 2002), a genetic length of 1.2M (Chancerel 2013) and a total number of segregating sites of 6910 (10 times the average number of SNP per chromosome available in the real dataset plus 50 theoretical QTL). A simple species-specific demography was used to mimic the selection in the natural environment of the breeding programme base population (popG0): from a wild ancestral population in the Landes forest, we simulated one significant bottleneck reducing effective population size from ~50000 (Milesi et al., 2023) to 100 (actual effective size for popG0) and with a mutation rate of 4e-18 per generation (Jaramillo-Correa 2020). We investigated the coherence in LD and allele frequency distributions for the 600 individuals of popG0, by comparing real genotyping data and simulated SNP array data (**Fig. 3-S3**). One phenotypic trait was simulated with reference values equal to those of the "height" trait targeted in the breeding programme (phenotypic average=6.81(m) and genetic variance=0.12m²).

The 833 individuals of $POP_R$ considered in this study were simulated after the two breeding cycles, as closely as possible to the real-life conditions (**Fig. 3-S4**). Each individual in popG0 took the identity of a real individual in the pedigree, the one with which it shared the same rank in terms of EBV. The EBV were generated with a correlation of 0.97 with true breeding values (BV) since the accuracy of EBV in real programme for these individuals is very high due to progeny testing. Individuals of popG0 were crossed according to the actual crossing plan to generate popP1 FS families of 130 individuals. Individuals for popG1 were selected based on EBV in each family with an intensity of 1.07. The use of this selection intensity, estimated with real data, made it possible to mimic the multi-character aspect of the actual selection carried out, as well as the diversity constraint actually used. Finally, the families of $POP_S$ (simulated version of $POP_R$) were obtained by crossing the individuals of popG1 according to the actual pedigree.

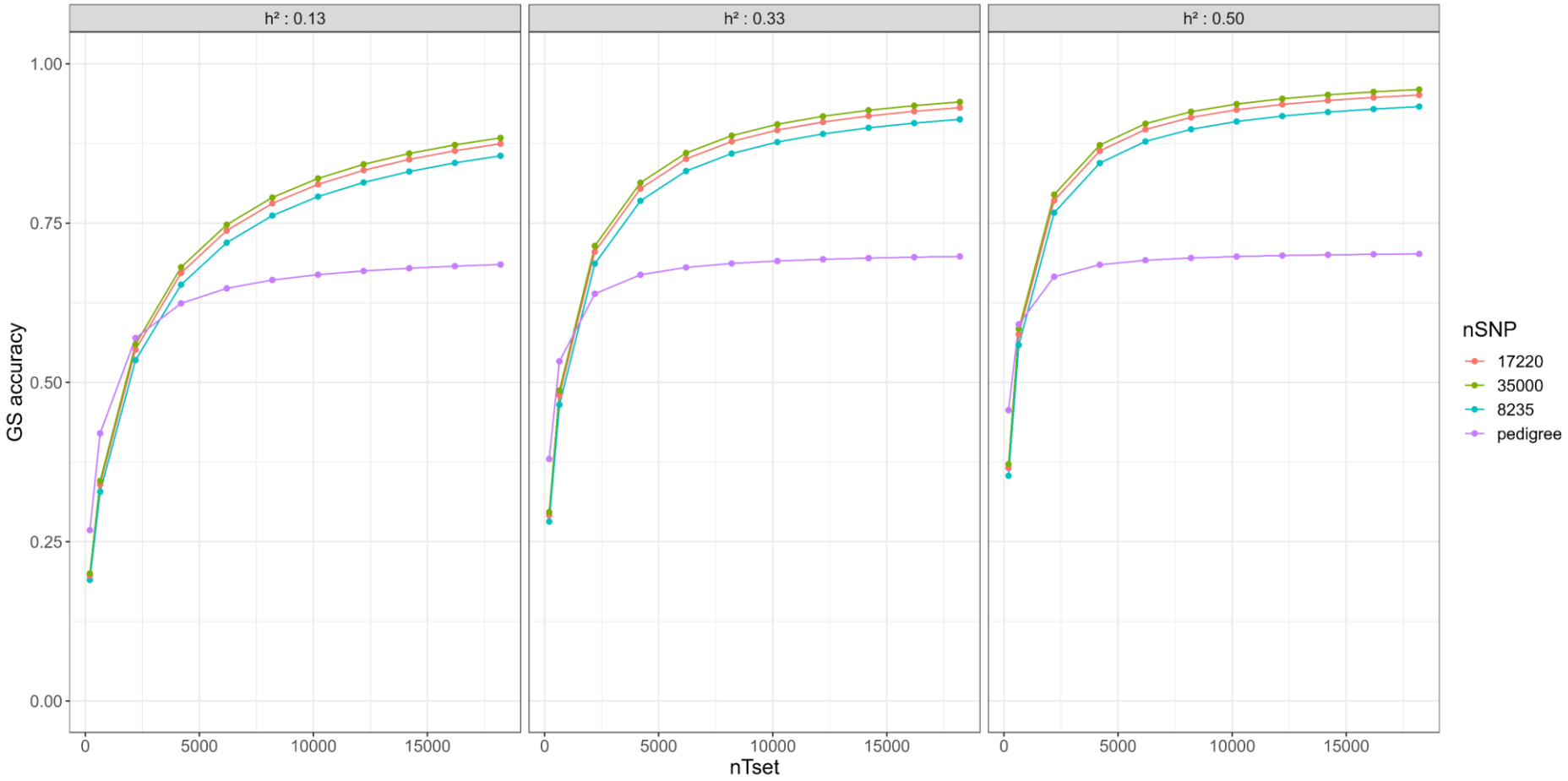



***Figure 3-S5: GS accuracy determined with deterministic formulas***

## Supplementary 3-3: Complementary precision inferences made using deterministic approaches

Mathematical modeling approaches are useful tools to predict the accuracy of genome-based prediction as a function of various population parameters (Daetwyler et al., 2008; Goddard, 2009; Hayes et al., 2009). Deterministic formulas have already been applied to forest trees (Grattapaglia & Resende, 2011) and are here adapted with input parameters from our maritime pine case study, following Gorjanc et al., 2015. The accuracy (square root of reliability) of GEBV was obtained by:

$$r_{EBV} = \sqrt{\frac{\theta}{1+\theta} b} \quad (3.5)$$

Where $\theta = nT_{set} b h^2 / M_e$ with $nT_{set}$ the size of the training set and $h^2$ the heritability of the trait. $M_e$ and $b$, respectively the effective number of chromosome segments and the proportion of genetic variance captured by markers, being defined by $M_e = 2N_e LC/\log(N_e L)$ and $b = nSNP/(nSNP + M_e)$, with $N_e = 100$ the effective size of our breeding population, $L = 1.2$ the average size of maritime pine chromosomes (in Morgans) and $C = 12$ the number of chromosomes. Based on progeny records, the prediction accuracy of non-phenotyped progeny with ABLUP depends solely on the accuracy of EBV in its parents. In case of no selection among parents, we have $r^2_{EBV_{pedigree}} = \frac{1}{4}\left(r^2_{EBV_{mother}} + r^2_{EBV_{father}}\right)$ with $r^2_{EBV_{mother}} = r^2_{EBV_{father}} = n/(n + \frac{4-h^2}{h^2})$, $n$ being the number of progenies with phenotypic values per parent.

The trends observed (**Fig. 3-S5**) are fully consistent with those obtained with stochastic simulations presented in the previous section (**Fig. 3-8**). Overall prediction accuracy is mainly impacted by $T_{set}$ size and to a much lesser extent by trait heritability. The advantage of GBLUP over ABLUP models in terms of forward prediction accuracy only becomes apparent above 2000 individuals in the $T_{set}$. Although deterministic approaches are very interesting for quickly revealing key parameters for genome-based prediction accuracy, they have been called into question and refined several times (Elsen, 2016, 2017). Especially when employing genome-based prediction within full-sibling families, as in this study, accurately forecasting the performance of corresponding accuracy for a specific trait within a particular family remains a challenging endeavor (Schopp et al., 2017).

## Author contributions

**Conceptualization**: Victor Papin, Laurent Bouffier, Leopoldo Sanchez; **Methodology**: Victor Papin, Gregor Gorjanc, Ivan Pocrnic, Laurent Bouffier, Leopoldo Sanchez; **Formal analysis and investigation**: Victor Papin; **Software**: Victor Papin, Gregor Gorjanc, Ivan Pocrnic; **Writing - original draft preparation**: Victor Papin; **Writing - review and editing**: Gregor Gorjanc, Ivan Pocrnic, Laurent Bouffier, Leopoldo Sanchez; **Funding acquisition**: Laurent Bouffier, Leopoldo Sanchez; **Resources**: Laurent Bouffier, Leopoldo Sanchez; **Supervision**: Gregor Gorjanc, Ivan Pocrnic, Laurent Bouffier, Leopoldo Sanchez.

## Acknowledgments

The authors would like to thank GIS "Groupe Pin Maritime du Futur" and INRAE - UEFP (https://doi.org/10.15454/1.5483264699193726E12) for the installation of the studied sites, the management of the sites, the help to collect data (HT and DEV measurements) and biological material (needles). Part of the experiments (DNA extraction, quantification and manipulation) were also performed at the PGTB (doi:10.15454/1.5572396583599417E12), with the help of Mathilde Flores, Christophe Boury and Céline Lalanne. Authors are also thankful to Christophe Plomion for his help during the conceptualization of this study. GG and IP acknowledge support from BBSRC (grants BBS/E/D/30002275, BBS/E/RL/230001A, and BBS/E/RL/230001C, and BB/P020488/1).

## Funding

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme Project under grant agreement n°773383 (B4EST). VP was awarded a doctoral fellowship (N°2020-CK-126) from Ecole Nationale Supérieure Des Sciences Agronomiques de Bordeaux-Aquitaine, 1 cours du Général de Gaulle, CS 40201 33175 Gradignan Cedex.

## Data availability

The data underlying this article will be made public and accessible to all on Data INRAE and scripts for simulations will be available on GitHub.

# References

Abad Viñas, R., Caudullo, G., Oliveira, S., & de Rigo, D. (2016). *Pinus pinaster in Europe : Distribution, habitat, usage and threats*.

Allier, A., Teyssèdre, S., Lehermeier, C., Claustres, B., Maltese, S., Melkior, S., Moreau, L., & Charcosset, A. (2019). Assessment of breeding programs sustainability : Application of phenotypic and genomic indicators to a North European grain maize program. *Theoretical and Applied Genetics*, *132*(5), 1321-1334. https://doi.org/10.1007/s00122-019-03280-w

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A. F., Resende Jr., M. F. R., & Muñoz, P. R. (2016). AGHmatrix : R package to construct relationship matrices for autotetraploid and diploid species : A blueberry example. *The Plant Genome*, *9*(3), 1-10. https://doi.org/10.3835/plantgenome2016.01.0009

Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., & Bouffier, L. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics*, *17*(1), 604. https://doi.org/10.1186/s12864-016-2879-8

Beaulieu, J., Doerksen, T. K., MacKay, J., Rainville, A., & Bousquet, J. (2014). Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genomics*, *15*(1), 1048. https://doi.org/10.1186/1471-2164-15-1048

Bouffier, L., Raffin, A. A., & Dutkowski, G. (2016, mars 14). *Using pedigree and trait relationships to increase gain in the French maritime pine breeding program*. IUFRO Conference « Forest Genetics for Productivity ». https://hal.inrae.fr/hal-02801580

Chagné, D., Lalanne, C., Madur, D., Kumar, S., Frigério, J.-M., Krier, C., Decroocq, S., Savouré, A., Bou-Dagher-Kharrat, M., Bertocchi, E., Brach, J., & Plomion, C. (2002). A high density genetic map of maritime pine based on AFLPs. *Annals of Forest Science*, *59*(5-6), 627-636. https://doi.org/10.1051/forest:2002048

Chancerel, E., Lamy, J.-B., Lesur, I., Noirot, C., Klopp, C., Ehrenmann, F., Boury, C., Provost, G. L., Label, P., Lalanne, C., Léger, V., Salin, F., Gion, J.-M., & Plomion, C. (2013). High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. *BMC Biology*, *11*(1), 50. https://doi.org/10.1186/1741-7007-11-50

Chen, Z.-Q., Baison, J., Pan, J., Karlsson, B., Andersson, B., Westin, J., García-Gil, M. R., & Wu, H. X. (2018). Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. *BMC Genomics*, *19*(1), 946. https://doi.org/10.1186/s12864-018-5256-y

Cros, D., Mbo-Nkoulou, L., Bell, J. M., Oum, J., Masson, A., Soumahoro, M., Tran, D. M., Achour, Z., Le Guen, V., & Clement-Demange, A. (2019). Within-family genomic selection in rubber tree (Hevea brasiliensis) increases genetic gain for rubber production. *Industrial Crops and Products*, *138*, 111464. https://doi.org/10.1016/j.indcrop.2019.111464

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., Campos, G. de los, Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic Selection in Plant Breeding : Methods, Models, and Perspectives. *Trends in Plant Science*, *22*(11), 961-975. https://doi.org/10.1016/j.tplants.2017.08.011

Dehli Vigeland, M. (2022). *pedtools : Creating and Working with Pedigrees and Marker Data* (R package version 1.3.0) [Logiciel]. https://github.com/magnusdv/pedtools

Doerksen, T. K., & Herbinger, C. M. (2010). Impact of reconstructed pedigrees on progeny-test breeding values in red spruce. *Tree Genetics & Genomes*, *6*(4), 591-600. https://doi.org/10.1007/s11295-010-0274-1

Durán, R., Isik, F., Zapata-Valenzuela, J., Balocchi, C., & Valenzuela, S. (2017). Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile. *Tree Genetics & Genomes*, *13*(4), 74. https://doi.org/10.1007/s11295-017-1158-4

Durel, C.-E. (1992). Gains génétiques attendus après sélection sur index en seconde génération d'amélioration du Pin maritime. *Revue forestière française*, *44*(4), 341-355. https://doi.org/10.4267/2042/26331

Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C., & Neale, D. B. (2010). Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (Pinus taeda L., Pinaceae). *Genetics*, *185*(3), 969-982. https://doi.org/10.1534/genetics.110.115543

El-Dien, O. G., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., & El-Kassaby, Y. A. (2018). Multienvironment genomic variance decomposition analysis of open-pollinated Interior spruce (Picea glauca x engelmannii). *Molecular Breeding*, *38*(3), 26. https://doi.org/10.1007/s11032-018-0784-3

Fuentes-Utrilla, P., Goswami, C., Cottrell, J. E., Pong-Wong, R., Law, A., A'Hara, S. W., Lee, S. J., & Woolliams, J. A. (2017). QTL analysis and genomic selection using RADseq derived markers in Sitka spruce : The potential utility of within family data. *Tree Genetics & Genomes*, *13*(2), 33. https://doi.org/10.1007/s11295-017-1118-z

Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR : An R package for breeding program simulations. *G3 Genes|Genomes|Genetics*, *11*(2), jkaa017. https://doi.org/10.1093/g3journal/jkaa017

Gorjanc G, Bijma P, Hickey JM (2015) Reliability of pedigree-based and genomic evaluations in selected populations. Genet Sel Evol 47:65. https://doi.org/10.1186/s12711-015-0145-1

Gorjanc G, Gaynor RC, Hickey JM (2018) Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor Appl Genet 131:1953–1966. https://doi.org/10.1007/s00122-018-3125-3

Grattapaglia, D., & Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genetics & Genomes*, *7*(2), 241-255. https://doi.org/10.1007/s11295-010-0328-4

Grattapaglia, D., Vilela Resende, M. D., Resende, M. R., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Takahashi, E. K., Zamprogno, K. C., & Kilian, A. (2011). Genomic Selection for growth traits in Eucalyptus : Accuracy within and across breeding populations. *BMC Proceedings*, *5*(7), O16. https://doi.org/10.1186/1753-6561-5-S7-O16

Guilbaud, R., Biselli, C., Buiteveld, J., Cattivelli, L., Copini, P., Dowkiw, A., Esselink, D., Fricano, A., Guerin, V., Jorge, V., & others. (2020). Development of a new tool (4TREE) for adapted genome selection in European tree species. *Proceedings of the Gentree Symposium*. Proceedings of the Gentree Symposium, Avignon, France.

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics*, *177*(4), 2389-2397. https://doi.org/10.1534/genetics.107.081190

Habier, D., Fernando, R. L., & Garrick, D. J. (2013). Genomic BLUP Decoded : A Look into the Black Box of Genomic Prediction. *Genetics*, *194*(3), 597-607. https://doi.org/10.1534/genetics.113.152207

Hallander, J., & Waldmann, P. (2009). Optimum contribution selection in large general tree breeding populations with an application to Scots pine. *Theoretical and Applied Genetics*, *118*(6), 1133-1142. https://doi.org/10.1007/s00122-009-0968-7

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review : Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, *92*(2), 433-443. https://doi.org/10.3168/jds.2008-1646

Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, *91*(1), 47-60. https://doi.org/10.1017/S0016672308009981

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., Plomion, C., & Bouffier, L. (2016). Genomic selection in maritime pine. *Plant Science*, *242*, 108-119. https://doi.org/10.1016/j.plantsci.2015.08.006

Iwata, H., Hayashi, T., & Tsumura, Y. (2011). Prospects for genomic selection in conifer breeding : A simulation study of Cryptomeria japonica. *Tree Genetics & Genomes*, *7*(4), 747-758. https://doi.org/10.1007/s11295-011-0371-9

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genetics Selection Evolution*, 42(1), 35. https://doi.org/10.1186/1297-9686-42-35

Karaman E, Cheng H, Firat MZ, et al (2016) An Upper Bound for Accuracy of Prediction Using GBLUP. PLOS ONE 11:e0161054. https://doi.org/10.1371/journal.pone.0161054

Kujala, S. T., & Savolainen, O. (2012). Sequence variation patterns along a latitudinal cline in Scots pine (Pinus sylvestris) : Signs of clinal adaptation? *Tree Genetics & Genomes*, *8*(6), 1451-1467. https://doi.org/10.1007/s11295-012-0532-5

Lebedev, V. G., Lebedeva, T. N., Chernodubov, A. I., & Shestibratov, K. A. (2020). Genomic selection for forest tree improvement : Methods, achievements and perspectives. *Forests*, *11*(11), 1190. https://doi.org/10.3390/f11111190

Legarra, A., Robert-Granié, C., Manfredi, E., & Elsen, J.-M. (2008). Performance of Genomic Selection in Mice. *Genetics*, *180*(1), 611-618. https://doi.org/10.1534/genetics.108.088575

Lenz, P. R. N., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M., & Bousquet, J. (2017). Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). *BMC Genomics*, *18*(1), 335. https://doi.org/10.1186/s12864-017-3715-5

Lenz, P. R. N., Nadeau, S., Azaiez, A., Gérardi, S., Deslauriers, M., Perron, M., Isabel, N., Beaulieu, J., & Bousquet, J. (2020). Genomic prediction for hastening and improving efficiency of forward selection in conifer polycross mating designs : An example from white spruce. *Heredity*, *124*(4), Article 4. https://doi.org/10.1038/s41437-019-0290-3

Lenz, P. R. N., Nadeau, S., Mottet, M.-J., Perron, M., Isabel, N., Beaulieu, J., & Bousquet, J. (2020). Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. *Evolutionary Applications*, *13*(1), 76-94. https://doi.org/10.1111/eva.12823

Li, Y., Klápště, J., Telfer, E., Wilcox, P., Graham, N., Macdonald, L., & Dungey, H. S. (2019). Genomic selection for non-key traits in radiata pine when the documented pedigree is corrected using DNA marker information. *BMC Genomics*, *20*(1), 1026. https://doi.org/10.1186/s12864-019-6420-8

Lindgren, D., Gea, L., & Jefferson, P. A. (1996). Loss of genetic diversity monitored by status number. *Silvae Genetica*, *45*, 52-59.

Meuwissen, T. H. E. (1997). Maximizing the response of selection with a predefined rate of inbreeding. *Journal of Animal Science*, *75*(4), 934-940. https://doi.org/10.2527/1997.754934x

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, *157*(4), 1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Mrode R, Pocrnic I (2023) Linear Models for the Prediction of the Genetic Merit of Animals, 4th Edition. CABI

Muñoz, F., & Sanchez, L. (2020). *breedR: statistical methods for forest genetic resources analysts* [Logiciel]. https://github.com/famuvie/breedR

Munoz, P. R., Resende Jr., M. F. R., Huber, D. A., Quesada, T., Resende, M. D. V., Neale, D. B., Wegrzyn, J. L., Kirst, M., & Peter, G. F. (2014). Genomic Relationship Matrix for Correcting Pedigree Errors in Breeding Populations : Impact on Genetic Parameters and Genomic Selection Accuracy. *Crop Science*, *54*(3), 1115-1123. https://doi.org/10.2135/cropsci2012.12.0673

Pégard, M., Segura, V., Muñoz, F., Bastien, C., Jorge, V., & Sanchez, L. (2020). Favorable Conditions for Genomic Evaluation to Outperform Classical Pedigree Evaluation Highlighted by a Proof-of-Concept Study in Poplar. *Frontiers in Plant Science*, *11*. https://www.frontiersin.org/articles/10.3389/fpls.2020.581954

Plomion, C., Chancerel, E., Endelman, J., Lamy, J.-B., Mandrou, E., Lesur, I., Ehrenmann, F., Isik, F., Bink, M. C., van heerwaarden, J., & Bouffier, L. (2014). Genome-wide distribution of genetic diversity and linkage disequilibrium in a mass-selected population of maritime pine. *BMC Genomics*, *15*(1), 171. https://doi.org/10.1186/1471-2164-15-171

Pryce, J. E., Daetwyler, H. D., Pryce, J. E., & Daetwyler, H. D. (2011). Designing dairy cattle breeding schemes under genomic selection : A review of international research. *Animal Production Science*, *52*(3), 107-114. https://doi.org/10.1071/AN11098

R Core Team. (2022). *R: A Language and Environment for Statistical Computing* [Logiciel]. https://www.R-project.org/

Ratcliffe, B., El-Dien, O. G., Klápště, J., Porth, I., Chen, C., Jaquish, B., & El-Kassaby, Y. A. (2015). A comparison of genomic selection models across time in interior spruce (Picea engelmannii × glauca) using unordered SNP imputation methods. *Heredity*, *115*(6), 547-555. https://doi.org/10.1038/hdy.2015.57

Rauf, S., da Silva, J. T., Khan, A. A., & Naveed, A. (2010). Consequences of plant breeding on genetic diversity. *International Journal of plant breeding*, *4*(1), 1-21.

Resende, J., M. F. R., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., & Kirst, M. (2012). Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). *Genetics*, *190*(4), 1503-1510. https://doi.org/10.1534/genetics.111.137026

Resende Jr, M. F. R., Muñoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., Resende, M. D. V., & Kirst, M. (2012). Accelerating the domestication of trees using genomic selection : Accuracy of prediction models across ages and environments. *New Phytologist*, *193*(3), 617-624. https://doi.org/10.1111/j.1469-8137.2011.03895.x

Resende, R. T., Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., & Grattapaglia, D. (2017). Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. *Heredity*, *119*(4), Article 4. https://doi.org/10.1038/hdy.2017.37

Strandén I, Garrick DJ (2009) Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci 92:2971–2975. https://doi.org/10.3168/jds.2008-1929

Thistlethwaite, F. R., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., Stoehr, M. U., & El-Kassaby, Y. A. (2017). Genomic prediction accuracies in space and time for height and wood density of Douglas-fir using exome capture as the genotyping platform. *BMC Genomics*, *18*(1), 930. https://doi.org/10.1186/s12864-017-4258-5

Thistlethwaite, F. R., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., Stoehr, M. U., & El-Kassaby, Y. A. (2019). Genomic selection of juvenile height across a single-generational gap in Douglas-fir. *Heredity*, *122*(6), Article 6. https://doi.org/10.1038/s41437-018-0172-0

Ukrainetz, N. K., & Mansfield, S. D. (2019). Assessing the sensitivities of genomic selection for growth and wood quality traits in lodgepole pine using Bayesian models. *Tree Genetics & Genomes*, *16*(1), 14. https://doi.org/10.1007/s11295-019-1404-z

Vidal, M., Plomion, C., Raffin, A., Harvengt, L., & Bouffier, L. (2017). Forward selection in a maritime pine polycross progeny trial using pedigree reconstruction. *Annals of Forest Science*, *74*(1), 21. https://doi.org/10.1007/s13595-016-0596-8

Woolliams, J. a., Berg, P., Dagnachew, B. s., & Meuwissen, T. h. e. (2015). Genetic contributions and their optimization. *Journal of Animal Breeding and Genetics*, *132*(2), 89-99. https://doi.org/10.1111/jbg.12148

Zapata-Valenzuela, J., Isik, F., Maltecca, C., Wegrzyn, J., Neale, D., McKeand, S., & Whetten, R. (2012). SNP markers trace familial linkages in a cloned population of Pinus taeda—Prospects for genomic selection. *Tree Genetics & Genomes*, *8*(6), 1307-1318. https://doi.org/10.1007/s11295-012-0516-5

Zapata-Valenzuela, J., Whetten, R. W., Neale, D., McKeand, S., & Isik, F. (2013). Genomic Estimated Breeding Values Using Genomic Relationship Matrices in a Cloned Population of Loblolly Pine. *G3 Genes|Genomes|Genetics*, *3*(5), 909-916. https://doi.org/10.1534/g3.113.005975

Zhou, L., Chen, Z., Olsson, L., Grahn, T., Karlsson, B., Wu, H. X., Lundqvist, S.-O., & García-Gil, M. R. (2020). Effect of number of annual rings and tree ages on genomic predictive ability for solid wood properties of Norway spruce. *BMC Genomics*, *21*(1), 323. https://doi.org/10.1186/s12864-020-6737-3

## 3.3. Comprendre et optimiser la précision de prédiction intrafamiliale

Dans chaque scénario de simulation présenté dans la partie précédente, on observe des niveaux de précision très contrastés selon les familles. Notamment, dans la situation où la précision globale de la sélection génomique est la plus forte (i.e. quand $h^2$=0.5, $nT_{set}$=2600 et nSNP=17220), la précision intrafamiliale moyenne varie entre 0.05 et 0.50 selon les familles (**Fig. 3-9B**). L'objectif de cette partie est d'identifier les facteurs expliquant cette hétérogénéité de précision intrafamiliale.

Afin d'assurer la robustesse des explications proposées dans la suite, notons que les recherches par simulation se sont faites sur 10 populations $POP_s$ générées indépendamment.

### 3.3.1. Facteurs explicatifs du niveau de précision intrafamiliale

Notre hypothèse de base dans cette nouvelle partie est la suivante : plus les frères d'une famille sont contrastés d'un point de vue génétique, meilleure sera la capacité du modèle à prédire les différences de performance en intra-famille. Autrement dit, la précision intrafamiliale repose sur le niveau de variabilité génétique au sein des familles.

#### 3.3.1.1. Précision intrafamiliale et valeurs génétiques vraies

La variabilité génétique au sein d'une famille peut d'abord être renseignée par la variance des valeurs génétiques vraies. Ces valeurs vraies sont uniquement disponibles dans un cadre de simulation et sont utilisés ici dans un but explicatif. Les variances des valeurs génétiques vraies ont été calculées pour chacune des 40 familles puis corrélées avec les précisions intrafamiliales moyennes de chaque famille. Cette procédure, réalisée pour chacune des 10 populations $POP_s$, amène à une corrélation moyenne de +0.62. La prédiction est donc significativement plus précise pour les familles avec une grande variance des valeurs génétiques vraies.

Rappelons que la précision intrafamiliale est obtenue par :

$$Précision_{intra-famille_i} = cor\left(GV_{pred_i}, GV_{real_i}\right) = \frac{covar(GV_{pred_i}, GV_{real_i})}{\sqrt{var(GV_{pred_i})}\sqrt{var(GV_{real_i})}}$$

Avec :

- $GV_{pred,i}$ les valeurs génétiques prédites par le modèle pour les individus de la famille *i* inclus dans le $V_{set}$,
- $GV_{real,i}$ les valeurs génétiques vraies pour ces mêmes individus.

Au premier abord, il n'est donc peut-être pas surprenant de voir la variance des valeurs génétiques vraies ($var(GV_{real_i})$) fortement corrélée avec la précision intrafamiliale étant donné qu'elle intervient dans le dénominateur de cette formule de précision. Toutefois, nous avons également vérifié le lien entre la précision intrafamiliale et $covar(GV_{pred_i}, GV_{real_i})$, la covariance entre valeurs génétiques prédites et valeurs génétiques vraies. Ces deux grandeurs apparaissent corrélées à +0.95, indiquant que la précision intrafamiliale est donc largement dépendante du numérateur de cette formule plutôt que des paramètres de standardisation du dénominateur. Le lien entre précision intrafamiliale et variances des valeurs génétiques vraies apparaît donc plutôt comme un lien de causalité réel et non dû à un biais méthodologique. Cette première information a permis de guider le choix des paramètres à tester par la suite.

#### 3.3.1.2. Précision intrafamiliale et paramètres classiques

Nous nous sommes intéressés dans un second temps à des paramètres pouvant être calculés dans un cas pratique réel pour caractériser chaque famille. Comme précédemment, l'idée est ensuite d'évaluer par corrélation le lien entre chaque paramètre et la précision intrafamiliale moyenne de chaque famille.

##### *Paramètres caractérisant les familles*

Premièrement, les familles ont chacune été caractérisées au niveau génétique :

- À partir des données de la matrice G, en calculant la **moyenne** et le **coefficient de variation** (CV) des coefficients **d'apparentement génomique** au sein de la famille,

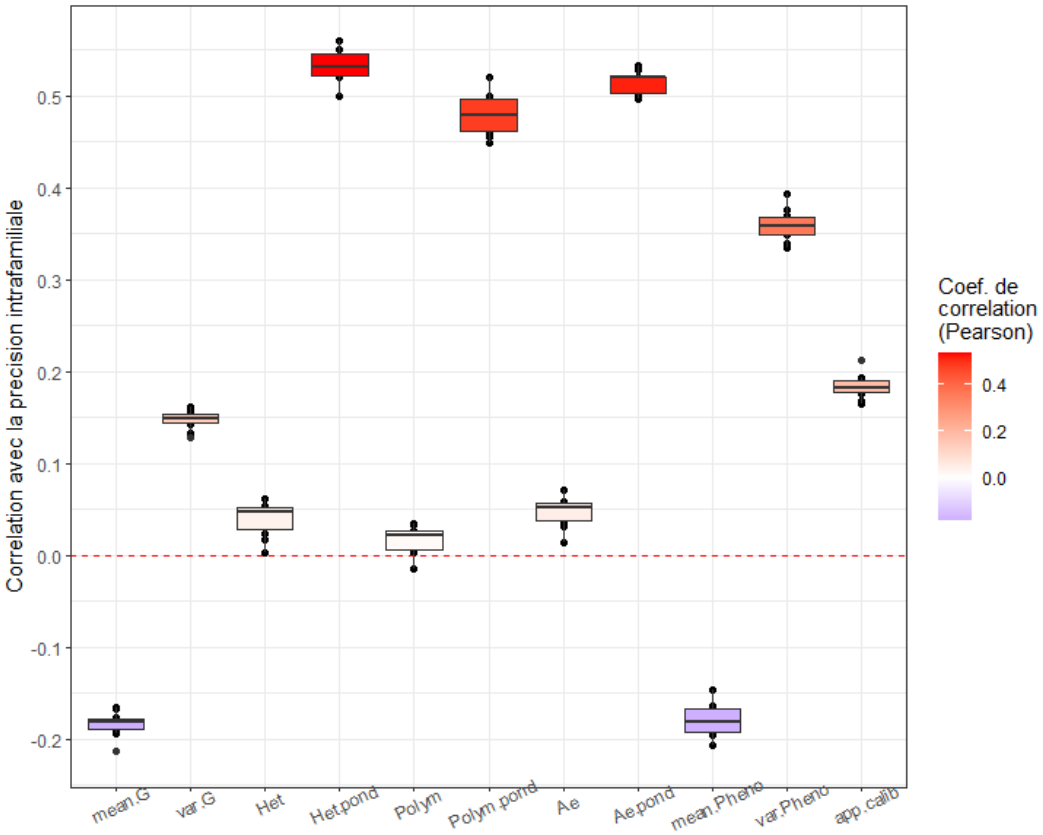



***Figure 3-10 : Niveau de corrélation entre les différents paramètres caractérisant les familles et la précision de prédiction intrafamiliale.*** *Chaque point correspond à un coefficient de corrélation obtenu entre les 42 valeurs du paramètre et les 42 valeurs de précision moyenne dans chacun des familles. Les boxplots regroupent chacun 10 points correspondant aux 10 coefficients de corrélation obtenus dans chacune des populations POPs.* ***Mean.G*** *: moyenne des coefficients d'apparentement génomique;* ***var.G*** *: coefficient de variation des coefficients d'apparentement génomique ;* ***Het*** *: taux d'hétérozygotie moyen ;* ***Polym*** *: taux de polymorphisme moyen ;* ***Ae*** *: nombre d'allèles efficaces moyen.* ***Het.pond, Polym.pond, Ae.pond*** *: similaire à Het, Polym, et Ae, respectivement, mais calculé en pondérant chaque locus par son effet ;* ***mean.Pheno*** *: moyenne phénotypique;* ***var.Pheno*** *: coefficient de variation des valeurs phénotypiques;* ***app.calib*** *: apparentement moyen à la calibration.*

- Et à partir des données de génotypage, en calculant le taux d'**hétérozygotie** et de **polymorphisme** moyens, ainsi que le nombre moyen d'**allèles efficaces**[9] (Kimura & Crow, 1964). Ces trois paramètres sont calculés successivement pour chaque locus simulé en considérant l'ensemble des individus de la famille, puis moyennés sur l'ensemble des loci. Dans cette approche, tous les loci sont donc considérés comme équivalent, comme c'est le cas dans les modèles de type GBLUP. Notons qu'une méthode de calcul alternative a été proposée. Dans celle-ci, le paramètre calculé à chaque locus est pondéré par l'effet de ce dernier ($\alpha^2$) avant d'être intégré dans la moyenne. Cette pondération vient de l'idée que plus un marqueur à un effet fort, plus il sera responsable d'une grande variance des valeurs génétiques et donc d'une bonne précision intrafamiliale. De tels effets pourrait être estimés dans le cadre d'un modèle RR-BLUP.

Deuxièmement, les familles ont été caractérisées au niveau phénotypique :

- Par la **moyenne** et le **CV** du trait étudié au sein des familles.

### *Lien avec la précision intrafamiliale*

Pour chaque paramètre, nous obtenons donc 40 valeurs familiales que nous avons ensuite corrélées avec la précision intrafamiliale moyenne de chaque famille (**Fig. 3-10**). Dans chacune des 10 populations $POP_s$, une valeur de corrélation a été obtenue, amenant à des boxplots de 10 points.

- La corrélation négative à -0.18 entre précision intrafamiliale et **moyenne familiale des apparentements génomiques** conforte à nouveau notre hypothèse initiale. Plus les individus d'une même famille se ressemblent d'un point de vue génétique, moins bonne est la précision dans cette famille. La corrélation positive à +0.15 entre précision intrafamiliale et **variance familiale des apparentements génomiques** indique quant à elle que l'existence de niveaux d'apparentements relativement variés au sein d'une famille facilite la prédiction dans cette famille. Des niveaux d'apparentement variables indiqueraient une certaine hétérogénéité dans la similarité génétique au sein d'une famille, ce qui à son tour indiquerait une variabilité mendélienne. Pour résumé, les familles les mieux prédites sont

[9] Défini comme $1/\sum {p_i}^2$ avec $p_i$ la fréquence du $i$ème allèle au locus considéré. Repris notamment par Nielsen et al. (2003)

celles qui présentent des moyennes d'apparentement faibles mais des variances d'apparentement fortes.

- Quand ils sont calculés sur tous les loci sans pondération, les paramètres d'**hétérozygotie**, de **polymorphisme** et d'**allèles efficaces** présentent une corrélation positive très faible avec le niveau de précision intrafamiliale. Ces corrélations augmentent significativement lorsque les paramètres sont calculés avec une pondération des loci par leur effet. Il s'agit donc bien de la variabilité génétique sur les loci à effet non-nul, induisant en partie plus grande variance des valeurs génétiques, qui permet une meilleure précision intrafamiliale.
- Notons enfin que la **moyenne phénotypique** des familles n'est pas corrélée avec le niveau de précision intra-famille, autrement dit que la précision du modèle est indépendante de la performance globale des familles pour le trait considéré. A l'inverse, la **variance des valeurs phénotypiques** d'une famille est très corrélée à la précision dans cette famille. Ce n'est pas surprenant au vu des résultats précédents, rappelons que la variance phénotypique est égale à la variance des valeurs génétiques plus une variance d'erreur.

Le signe de l'ensemble de ces corrélations fait sens au vu de notre hypothèse initiale. Toutefois, les niveaux de corrélation demeurent assez faibles, suggérant que les paramètres calculés n'expliquent qu'une partie réduite de l'hétérogénéité des précisions intrafamiliales. Le fait d'isoler chaque paramètre facilite la compréhension, mais la précision de prédiction est une grandeur complexe souvent décrite comme le résultat d'un processus multifactoriel (Grattapaglia & Resende, 2011). Ainsi, nous avons également identifié que l'apparentement moyen des familles à la calibration, calculé comme la moyenne des apparentements génomiques entre une famille et l'ensemble des autres familles, était corrélé à +0.20 avec la précision intrafamiliale. Contrairement aux paramètres précédents caractérisant distinctement chaque famille, la connectivité entre familles incluses dans le modèle de sélection génomique apparait aussi un facteur explicatif important (Beaulieu et al., 2014; Lenz et al., 2017).

### *Pistes d'analyses complémentaires*

Plusieurs analyses peuvent compléter ce premier travail d'exploration.

- La pondération par l'effet des loci pourrait être intégrée lors du calcul des paramètres de moyenne et de variance des apparentements génomiques. Cela impliquerait d'inclure le terme de pondération au moment du calcul de la matrice G. C'est notamment envisageable

**Encadré 3-1 : calcul de "l'indice génotypique" (IG)**

Nous avons défini dans cet étude un critère simple permettant de décrire la variabilité qu'un couple parental peut engendrer au niveau allélique dans sa descendance. Contrairement au taux d'hétérozygotie calculé pour chaque parent, ce critère prend en compte la complémentarité entre parents.
Notons P1 et P2 les deux parents d'une famille de pleins-frères.
La valeur de l'indice génotypique pour chaque locus $p$ (notée $IG_p$) dépend du génotypage de P1 et P2 :

- $IG_p = 0$ si :
  - P1 est homozygote A/A et P2 est homozygote A/A
  - P1 est homozygote A/A et P2 est homozygote B/B
  - P1 est homozygote B/B et P2 est homozygote A/A
  - P1 est homozygote B/B et P2 est homozygote B/B
- $IG_p = 1$ si :
  - P1 est homozygote A/A et P2 est hétérozygote A/B
  - P1 est homozygote B/B et P2 est hétérozygote A/B
  - P1 est hétérozygote A/B et P2 est homozygote A/A
  - P1 est hétérozygote A/B et P2 est homozygote B/B
- $IG_p = 2$ si :
  - P1 est hétérozygote A/B et P2 est hétérozygote B/B

L'indice génotypique (IG) pour un couple parental correspond à la moyenne des indices $IG_p$ pour l'ensemble des loci considérés.

avec la formule de VanRaden n°2 (VanRaden, 2008) par l'intermédiaire de la matrice diagonale D qui pondère déjà chaque locus par sa variance allélique attendue.

Ce dernier point amène l'idée que notre pondération pourrait de manière générale être étendue sous la forme $2pq\alpha^2$, autrement dit en intégrant les fréquences alléliques à chaque locus. Cette pondération approxime alors la variance additive du locus dans une situation d'équilibre de liaison et d'équilibre Hardy-Weinberg (Falconer & Mackay, 1996). Ce terme $2pq$ peut être calculé à l'échelle de la famille, afin de donner plus de poids aux loci présentant des fréquences alléliques équilibrées, sous-entendu qu'ils sont à même de générer plus de variabilité génétique dans la famille et donc une meilleure précision en sélection génomique.

- De manière générale, chaque paramètre testé ici est limité par le gradient des valeurs définis par les 40 familles étudiées. Les parents utilisés pour générer ces familles sont assez équidistants génétiquement (issus d'un grand nombre de G0 considérés non-apparentés) et ont pour la plupart tous contribué une seule fois lors de la création de ces 40 familles. Une possibilité serait de générer par simulation de nouvelles familles plus contrastées à partir d'un ensemble de parents dans lequel les distances génétiques sont plus variables.

### 3.3.2. Optimiser la précision de prédiction intra-famille

S'il parait désormais clair que la variabilité intra-famille joue un rôle dans le niveau de précision intrafamiliale, il faut désormais se demander comment raisonner les croisements pour assurer une précision la plus élevée possible. Nous allons nous intéresser dans cette partie à des paramètres caractérisant les parents des différents croisements, dans un but plus applicatif.

#### *Paramètres parentaux et lien avec la précision intra-famille*

Chaque parent a été caractérisé par :

- Son taux moyen d'**hétérozygotie**, de **polymorphisme** et son nombre moyen d'**allèles efficaces**. A chaque fois, ces paramètres sont calculés avec ou sans pondération des loci.

En plus de ces premiers paramètres, chaque couple parental ayant généré une famille de pleins-frères est caractérisé par :

- Son « **indice génotypique** » (**Encadré 3-1**). Ce dernier a été calculé avec l'information de génotypage de tous les loci, avec ou sans pondération.

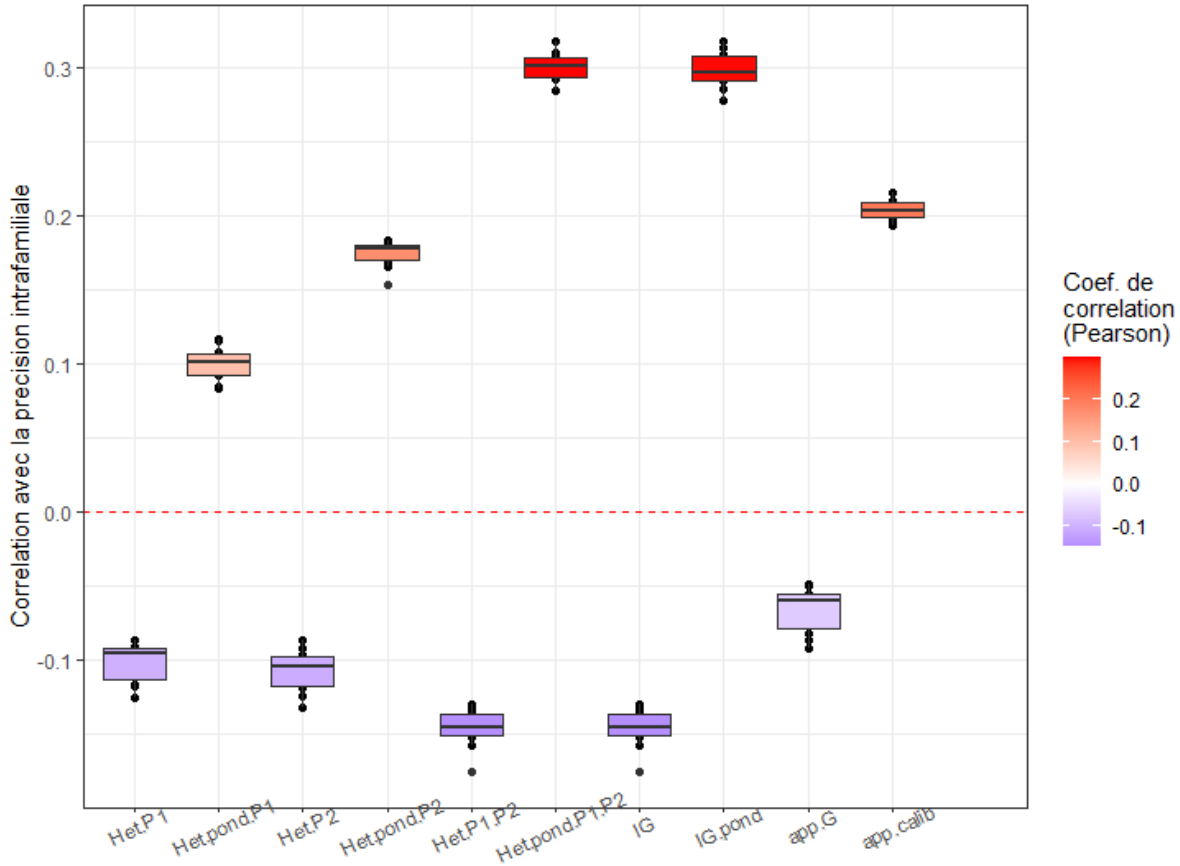



***Figure 3-11: Niveau de corrélation entre les différents paramètres caractérisant les parents d'une famille et la précision de prédiction intrafamiliale.*** ***Het.P1*** *: taux d'hétérozygotie moyen du parent 1 ;* ***Het.P2*** *: taux d'hétérozygotie moyen du parent 2 ;* ***Het.P1.P2*** *: moyenne des taux d'hétérozygotie moyens des parents 1 et 2 ;* ***IG*** *: indice génotypique ;* ***Het.pond.P1, Het.pond.P2, Het.pond.P1.P2, IG.pond :*** *similaires à Het.P1, Het.P2, Het.P1.P2, et IG respectivement, mais calculés en pondérant chaque locus par son effet ;* ***app.G*** *: coefficient d'apparentement génomique entre les deux parents ;* ***app.calib*** *: apparentement moyen des deux parents à la calibration. Par simplicité, les paramètres de polymorphisme et de nombre d'allèles efficaces n'ont pas été représentés : leurs niveaux de corrélation avec la précision intrafamiliale sont très similaires à ceux obtenus avec les paramètres d'hétérozygotie.*

- Son coefficient d'**apparentement génomique** (extrait de la matrice G regroupant l'ensemble des parents)

La corrélation de chacun de ces paramètres avec la précision en intrafamiliale est présentée en **Figure 3-11**. Sans pondération par l'effet des loci, les paramètres d'hétérozygotie et l'indice génotypique apparaissent tous négativement corrélés avec la précision intrafamiliale. A l'inverse, lorsque le calcul implique une pondération des loci, ces corrélations deviennent significativement positives. La variabilité génétique au niveau des loci à effet non-nul expliquent une partie non négligeable du succès de la prédiction. Notons que les meilleures corrélations sont obtenues pour le taux d'hétérozygotie moyen du couple parental et l'indice génotypique (+0.30). L'équivalence entre les corrélations pour ces deux paramètres suggère que la prise en compte de la complémentarité entre parents au niveau allélique n'est nécessairement pas plus explicative de la précision intrafamiliale dans notre cas. Enfin, notons que comme précédemment, l'apparentement à la calibration, calculé comme la moyenne des coefficients d'apparentement génomiques entre le couple parental considéré et l'ensemble des autres géniteurs de $POP_s$, est aussi un facteur explicatif non négligeable avec une corrélation à la précision intrafamiliale de +0.20.

### *Pistes d'analyses complémentaires*

Le taux d'hétérozygotie est un indicateur de la diversité génétique au sein d'une population. Plusieurs approches ont déjà démontré que le maintien de cette diversité permet d'assurer le gain génétique à long-terme. Une approche de référence est l'OCS (« *optimal contribution selection* ») (Meuwissen, 1997; Woolliams et al., 2015; Wray & Goddard, 1994). Allier et al. (2019) ont récemment proposé de prendre en compte la variabilité intra-famille dans ce type de stratégie afin d'améliorer leur performance. L'impact de la diversité sur la précision de la sélection génomique n'a cependant pas été étudié jusque-là.

Nos premiers résultats suggèrent que le taux d'hétérozygotie des individus pourrait être un critère de sélection complémentaire afin, de maintenir la diversité au cours des générations, mais aussi de maximiser la précision intrafamiliale et donc augmenter l'efficacité de la sélection.

Nous envisageons de simuler un processus de sélection simple sur plusieurs générations afin de juger la pertinence de ce nouveau critère complémentaire. Comme dans l'article n°1

(partie simulation), chaque génération de la population d'amélioration serait composée de 40 familles avec 100 individus par famille. La sélection génomique serait appliquée dans une même génération, selon des paramètres qui lui permettent un net avantage de précision par rapport au modèle ABLUP (par exemple $h^2$=0.33, $nT_{set}$=2600 et nSNP =17220). A chaque génération, une sélection de type *forward* serait appliquée. Le choix des futurs géniteurs s'effectuerait d'une part grâce à une procédure de type OCS, et d'autre part en optimisant valeurs génétiques et taux d'hétérozygotie (ces deux paramètres devant être maximisés). Les individus sélectionnés seraient ensuite croisés selon un schéma aléatoire de type « *single pair mating* » pour obtenir la génération suivante. L'évolution de la précision de la sélection génomique, du gain génétique réalisé et du niveau de diversité au cours des générations permettront d'identifier plus clairement les avantages et inconvénients de la nouvelle approche proposée.

Plusieurs points sont encore en réflexion :

- La phase d'optimisation peut s'avérer informatiquement longue et complexe, que ce soit avec l'OCS ou avec notre nouvelle démarche, notamment si le nombre d'individus considéré est grand. Une première étape de pré-sélection des candidats est peut-être nécessaire.
- Egalement, se pose la question du poids accordé à la contrainte de diversité (consanguinité en OCS ou niveau d'hétérozygotie dans notre nouvelle démarche) par rapport au poids accordé à la maximisation du gain génétique.
- Au vu des premiers résultats, l'indice génotypique semble équivalent au taux d'hétérozygotie moyen des parents. La considération de l'indice génotypique impliquerait d'optimiser non plus l'ensemble des parents sélectionnés mais plutôt le choix des différents couples parentaux, ce qui peut être encore plus complexe. L'approche de référence utilisée serait alors plutôt du type optimal cross-selection. (Akdemir & Sánchez, 2016; Gorjanc et al., 2018; Kinghorn, 2011).

## 3.4. Conclusion

L'objectif n°1 de cette thèse visait à évaluer la capacité des modèles génomiques à prédire la variabilité intra-famille, et plus généralement, à définir les conditions favorables pour que ces modèles surpassent l'approche ABLUP.

L'échantillonnage réalisé sur le dispositif A regroupe in fine 833 individus appartenant à 39 familles de pleins-frères. En particulier, 9 familles présentaient des effectifs compris entre 40 à 65 individus. Quel que soit le scénario de cross-validation considéré pour la prédiction génomique, la précision intrafamiliale était en moyenne nulle malgré de fortes hétérogénéités entre familles. De manière générale, les modèles GBLUP ne présentait qu'un léger avantage en terme de précision de prédiction par rapport aux modèles ABLUP intégrant une information de pedigree complète et corrigée.

La construction d'un modèle de simulation cohérent avec nos données réelles a permis de démontrer que le nombre d'individus par famille, et donc la taille globale de la population de calibration, était un paramètre clé pour démontrer la supériorité de la sélection génomique et prédire la variabilité intra-famille. Dans notre design, entre 40 et 65 individus par famille de pleins-frères, soit entre 1600 et 2600 individus en calibration, sont nécessaires pour que la précision intrafamiliale devient significativement positive pour l'ensemble des 40 familles.

Toutefois, cette précision intrafamiliale demeurait très variable entre les différentes familles. Si cette précision résulte sans nul doute d'un processus multifactoriel, nous avons pu identifier que les familles les mieux prédites étaient généralement les familles qui présentaient le plus de variabilité génétique. Cette dernière, représentée par le taux d'hétérozygotie parental, pourrait être utilisé comme critère complémentaire de sélection, afin d'assurer une prédiction génomique efficace au niveau global et intrafamilial.

# 4. Intégration de l'information environnementale en sélection par la construction de normes de réaction

## 4.1. Introduction

En raison de leur immobilité et de leur longévité, les arbres sont particulièrement exposés aux changements d'environnements. Si la considération de caractères de production à un âge avancé a longtemps permis de proposer des variétés performantes pour la production de bois, le changement climatique amène à proposer de nouvelles méthodes de sélection afin d'analyser plus finement la croissance des arbres au regard des conditions environnementales. L'objectif de thèse n°2 vise à développer une démarche d'évaluation par des normes de réaction. L'utilisation de traits-fonction représente un changement de paradigme dans la façon de faire l'évaluation génétique chez les arbres forestiers.

Pour cela, des profils densitométriques ont été acquis pour 650 individus du dispositif B. Ils permettent d'identifier les cernes du bois successives et de tracer rétrospectivement la croissance des arbres au fil des années. Ces données sont assimilables à des mesures répétées au cours du temps, et donc, face aux changements d'environnements. Les modèles de régression aléatoire, classiquement utilisés en génétique animale pour évaluer la productivité au cours du temps, ont ici été exploités pour analyser la croissance annuelle au regard d'indices environnementaux définis sur le même pas de temps. L'intégration de données génétiques dans ce modèle permet d'estimer les paramètres et valeurs génétiques de manière continue au regard du gradient environnemental utilisé. La précision de prédiction du modèle à partir de données génomiques a également été évaluée afin d'en déterminer la pertinence dans un contexte de sélection génomique[10].

Cette démarche est présentée dans la **partie 4.2** et fait l'objet d'un l'article n°2 intitulé « *Integrating environmental gradients into breeding: application of genomic reactions norms in a perennial species* » en cours de révision pour le journal *Heredity*. Différentes perspectives pratiques pour la construction de normes de réaction sont présentées dans la **partie 4.3**.

[10] Au vu des critiques émises dans l'article n°1 concernant la métrique de précision de prédiction (« accuracy »), seule la capacité de prédiction (« predictive ability/performance ») est évaluée ici, considérée comme un estimateur sans biais et sans hypothèse (**cf Article 2 – Material & Methods**).

## 4.2. Article n°2

# Integrating environmental gradients into breeding: application of genomic reactions norms in a perennial species

**Victor Papin[1], Alexandre Bosc[2], Leopoldo Sanchez[3] and Laurent Bouffier[1]**

[1] **INRAE, BIOGECO**, UMR 1202, 33610 Cestas, France.
[2] **INRAE ISPA,** UMR 1391, 33140 Villenave-d'Ornon, France.
[3] **INRAE-ONF, BioForA**, UMR 0588, 45075 Orléans, France.

**Key message**

Growth plasticity of maritime pine was modelled with genomic reaction norms in order to predict genetic values according to environmental gradients facilitating breeding in a climate change context.

**Abstract**

Global warming threatens the productivity of forest plantations. We propose here the integration of environmental information into a genomic evaluation scheme using individual reaction norms, to enable the quantification of resilience in forest tree improvement and conservation strategies in the coming decades. Random regression models were used to fit wood ring series, reflecting the longitudinal phenotypic plasticity of tree growth, according to various environmental gradients. The predictive performance of the models was considered to select the most relevant environmental gradient, namely a gradient derived from an ecophysiological model and combining trunk water potential and temperature. Even if the genotype ranking was preserved over most of the environmental gradient, strong genotype x environment interactions were detected in the extreme unfavorable part of the gradient, which includes environmental conditions that are very likely to increase in the future. Combining genomic information and longitudinal data allowed to predict growth in unobserved environments: considering an equivalent phenotyping effort, the cross-validation scenarios led to predictive performances ranging from 0.25 to 0.59 highlighting the importance of phenotypic data allocation. Genomic reaction norms are useful for the characterization and prediction of the function of genetic parameters and facilitate breeding in a climate change context.

**Keywords** Ecophysiological approach, genomic selection, maritime pine, phenotypic plasticity, random regression, tree rings

**Abbreviations**

**Cir22:** circumference at 22 years old

**CV:** cross-validation

**DM:** annual de Martonne aridity index

**DM':** optimized annual de Martonne aridity index

**GEBV:** genomic estimated breeding values

**$GEBV_{Cir22}$**: genomic estimated breeding values for Cir22

**$GEBV_{pred}$:** GEBV predicted over the entire environmental gradient for individuals in the validation set

**$GEBV_{ref}$:** GEBV calculated over the entire environmental gradient for all individuals, using a RRM integrating all available phenotypic data

**$GG_{max}$ :** maximum genetic gain

**$GG_{true}$ :** true genetic gain

**GP:** annual growth potential index

**GP':** optimized modified annual growth potential index

**GS:** genomic selection

**GxE:** Genotype x Environment

**POP:** trees sampled for this study

**RRM:** random regression model

**RA:** adjusted values of ring area

**$RA_{raw}$:** raw values of ring area

**SNP:** single nucleotide polymorphism

**$T_{set}$:** Training set

**$V_{set}$:** Validation set

# Introduction

Forest trees are keystone species in forest ecosystems supporting biological diversity and providing ecosystem services (Brockerhoff et al., 2017). They also produce wood, which will be a key material for meeting the challenges of the near future, thanks to its multiple uses (construction, paper, furniture, energy, chemistry) and its ability to sequester carbon for long periods of time (Ramachandran Nair et al., 2009; Domke et al., 2020). In this context, forest plantation has been expanding for several decades (FAO, 2010), with the aim of concentrating timber production and relieve pressure on natural forest. However, these benefits of forest plantation will require the adaptation of forest to a new, more challenging climate (Allen et al., 2010; Pawson et al., 2013; Payn et al., 2015) One of the major levers for ensuring sustainable wood productivity for forest plantations will be the deployment of trees capable of maintaining high growth rates even in extreme environments. To meet this goal, the integration of phenotypic plasticity, which is defined as the ability of a genotype to produce different phenotypes in different environmental conditions (Bradshaw, 1965), is becoming a major issue in forest tree breeding programs (Ray et al., 2022). A genotype is considered here as a unique genetic combination found in a single individual, or in several vegetative copies genetically identical. The challenges posed by climate change faced limited scope of traditional genetic analyses of forest trees focusing principally on phenotypic plasticity between experimental sites (Correia et al., 2010; Baltunis et al., 2010; Shalizi & Isik, 2019). These studies highlighting the existence of GxE interactions for conifer trees often consider a limited number of environments, selected so as to avoid high mortality rates. They are, therefore, not designed to be representative of the full range of environments of relevance in a context of rapid climate change. The cost and difficulty of exposing the same genotype to different environmental conditions, particularly for species difficult to propagate vegetatively, are major obstacles to the systematic evaluation of across-site plasticity in the context of tree breeding.

Phenotypic plasticity can be effectively modeled by reactions norms if repeated measurements across ages or clones are available, together with a relevant descriptor of the environment in which the phenotype was expressed (Schlichting & Pigliucci, 1998; Sanchez et al., 2013). A reaction norm is a representation of phenotypic values as a function of an environmental gradient. Various methods for constructing reaction norms have been developed, but the random regression model described by (Kirkpatrick & Heckman, 1989) is particularly relevant in breeding contexts. Through the integration of genetic data, this model can continuously estimate genetic parameters and breeding values according to the gradient. The gradient most

frequently chosen is time (age), and this approach is frequently used in animal breeding (Jamrozik et al., 1997; Schaeffer, 2004; Boligon et al., 2012) and more rarely in plant breeding contexts (Sun et al., 2017; Campbell et al., 2018) including tree breeding (Apiolaza & Garrick, 2001; Wang et al., 2009). However, reaction norms along an environmental gradient have recently been modeled, as a way to meet the challenges of rapid climate change in tree breeding (Alves et al., 2020; Marchal et al., 2019).

In forest tree breeding programs, selection has historically focused on growth traits evaluated at an advanced age, they represent the cumulative reaction of the tree to the environment over a number of years (Mullin et al., 2011; Pâques, 2013). It is not therefore possible to trace back and identify the environmental factors contributing to the final phenotype, as environment can be considered only in a global manner over the whole period. However, yearly growth increments can be correlated with well-characterized environments (Martinez-Meier et al., 2008; Zas et al., 2020). This can be achieved with the use of wood ring series, which define the annual radial growth of each individual in temperate climates. Indeed, the cambial activity of trees depends strongly on environmental conditions, particularly temperature and water availability (Schweingruber, 2007). The variability of annual ring width and wood density characterizes the plastic response of trees to changing environmental conditions. It has been shown to have genetic determinism (Sánchez-Vargas et al., 2007; Dalla-Salda et al., 2009) and could be used as a proxy for the potential reaction of trees to changes in environmental conditions. The analysis of these repeated phenotypes therefore provides an ideal longitudinal dataset for studying phenotypic plasticity at individual level (Marchal et al., 2019). Such analyses can be explanatory in nature, seeking to identify the optimal combination of environmental factors making a significant contribution to annual growth, but they can also be predictive, with the development of functional models for inferring growth in unobserved environments.

The integration of molecular markers into genetic evaluations provides not only more accurate estimates of genetic parameters, but also opportunities to implement genomic selection approaches (R2D2 Consortium et al., 2021). In forest trees, such approaches pave the way for the early selection of important traits, such as wood traits, that would otherwise be evaluable only after many years of cumulative growth. Genomic selection is also particularly valuable in tree breeding, as it allows the integration of traits that are costly and complex to measure (Grattapaglia & Resende, 2011). In many species, the gains provided by the use of genomic data have tended to eclipse the interest in longitudinal data (Oliveira et al., 2019). However,

these two approaches are not antagonistic and their beneficial effects can be combined (Rutkoski et al., 2016; Sun et al., 2017). Genomic reaction norms are rarely used (Lyet al., 2018), but are potentially of great value in this context, as they allow prediction of growth in as yet unobserved environments, thus decreasing the complex and costly evaluation procedures associated with experimentation and phenotyping under different environmental conditions.

We propose here an integration of environmental information into genetic evaluations, using reaction norms in the context of forest tree breeding. A random regression model based on annual ring growth data for maritime pine (*Pinus pinaster* Ait.) and including genomic data was used to fit individual-level reaction norms. The genetic components of these norms were described and the implications of their use in the context of breeding were further investigated with respect to a classical analysis targeting final radial growth. Finally, we investigated the model's ability to predict growth in unobserved environments considering realistic phenotyping conditions for the maritime pine breeding program in a genomic selection context. To our knowledge, this is the first study in a tree breeding context to use a random regression model to combine environmental gradient and genomic information.

***Tableau 4-1: Soil and climate characterization for Site 1 and Site 2***

| Site | Soil characterization[1] | | | | Climate characterization[2] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Organic matter (g/kg of soil)* | *Water table depth (m)* | | | *Cumulative annual rainfall (mm)* | | | | *Mean annual temperature (°C)* | | | |
| | *Mean* | *Min* | *Mean* | *Max* | *2015* | *2016* | *2017* | *2018* | *2015* | *2016* | *2017* | *2018* |
| Site 1 | 37.3 | 0 | 0.9 | 1.8 | 407 | 492 | 582 | 503 | 16.2 | 15.5 | 16.4 | 16.4 |
| Site 2 | 20.5 | 5.8 | 6.9 | 7.8 | 511 | 522 | 647 | 585 | 16.4 | 15.9 | 16.5 | 16.4 |

[1] Soil organic matter content was determined by an analysis of the first 80cm of soil performed in 2015; water table depths were recorded from 2015 to 2016 with soil humidity probes

[2] Climate data are from weather stations located on each site

# Materials and Methods

## Plant material

A maritime pine trial was established at two sites in 1997: Site 1 (Cestas: Lat 44.74, Lng -0.68) and Site 2 (Escource: Lat 44.16, Lng -1.03). Soil characterization revealed greater soil fertility (+16.8 g organic matter/kg of soil) and a shallower water table (mean difference of -6 m) at Site 1 than at Site 2. Climatic measurements showed that there was more rainfall at Site 2 (mean of +15% for total annual rainfall), whereas temperatures were equivalent at the two sites (**Tab. 4-1**). A total of 202 genetic units (35 trees per genetic unit) were studied on both sites. They were planted in a complete block design with single-tree plots (1,250 trees/ha) and consisted of 196 half-sib families obtained from crosses between identified seed parents and two pollen mixtures of identified donors, plus six check lots. Thinning operations were performed at both sites in 2012 and exclusively at Site 1 in 2017, when the trees were 16 and 21 years old, respectively. A subsample (POP) of 25 half-sib families, with 13 individuals per family and per site, was selected as representative of the variability of growth (total of 650 individuals). In this maritime pine experimental context, each genotype is represented by a single individual, so that the notions of "genotype", "individual" and "tree" are considered equivalent in our study.

## Genetic characterization of POP

Genomic DNA was extracted from needles collected from POP, to which we added 186 randomly selected duplicates for repeatability estimates. The concentration and quality of DNA for each sample were determined with a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Genotyping was performed by Thermo Fisher Scientific (Thermo Fisher Scientific, Santa Clara, CA, USA) with the 4TREE Axiom 50K SNP multi-species array (Guilbaud et al., 2020). The preliminary filters recommended by Thermo Fisher Scientific were applied to the genotyping results, at the sample (DishQC ≥ 0.4, CallRate ≥ 97%) and SNP (CallRate ≥ 85, fld-cutoff ≥ 3.2, het-so-cutoff: ≥ -0.3) levels. In addition, sequential filtering was applied, with the removal, in the following order, of SNP with less than 95% repeatability, SNP with more than 5% Mendelian segregation errors and SNP with a minor allele frequency (MAF) below 1%. A genomic relationship matrix (G) was calculated with the VanRaden formula (VanRaden, 2008) using the AGHmatrix package (Amadeu et al., 2016) in R 4.2.2 environment (R Core Team, 2022):

$$G = \frac{(M - P)(M - P)'}{2\Sigma\, p_i(1 - p_i)} \qquad (4.1)$$

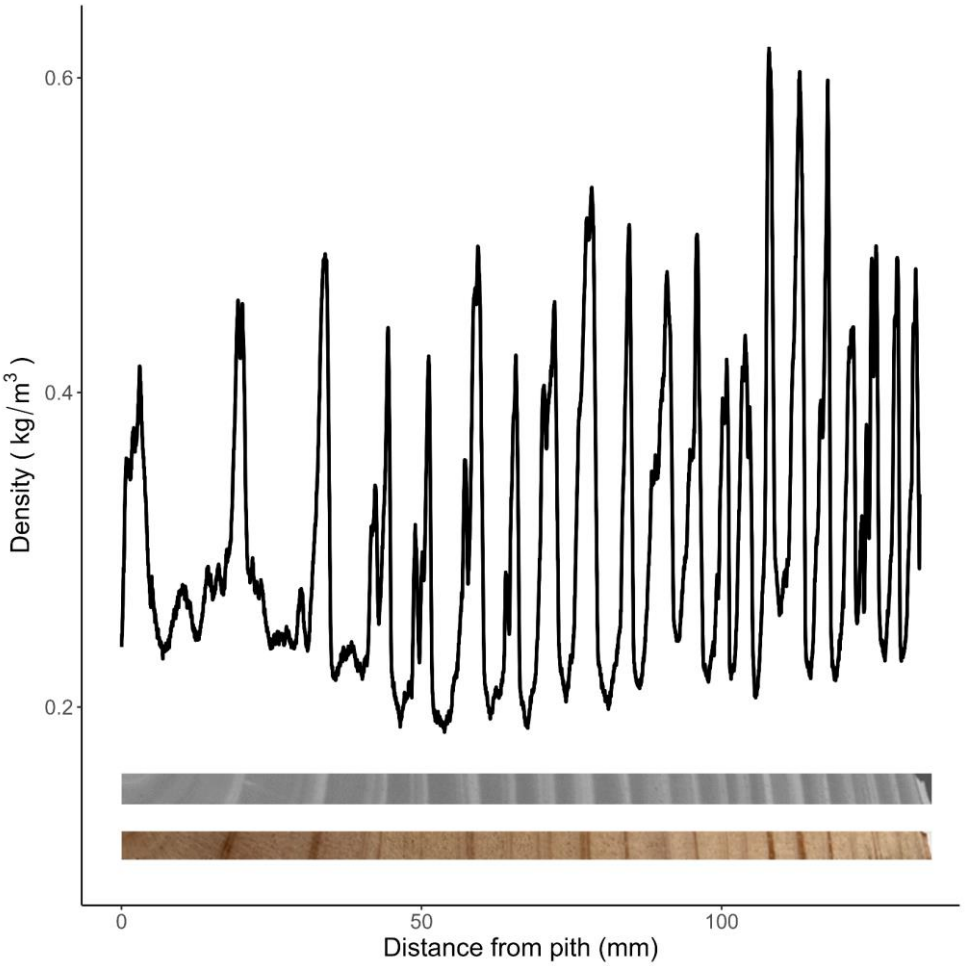



***Figure 4-1 : From wood increment core to wood density profile.*** *From the bottom to the top: the wood increment core picture from one tree, its corresponding radiography, its wood density profile (black line) from pith (position: 0mm) to the bark obtained after processing. Sudden and high drops in wood density mark the end of annual growth and were used to fit each ring limitations.*

where the $M$ matrix ($n$: number of individuals x $m$: number of markers) contains marker information coded as -1 for one of the homozygotes, 0 for heterozygotes and 1 for the other homozygotes; and the $P$ matrix ($n$ x $p$) contains allele frequencies expressed as $2(p_i - 0.5)$, where $p_i$ is the frequency of the second allele at locus $i$ for all individuals.

In addition, pedigree recovery was performed for each tree from POP, with a subset of 161 SNP used to infer the identities of the parents (25 seed parents and 85 pollen parents) and grandparents (69 initial progenitors from the base population of the breeding program) (**Suppl. 4-S1**). The most complete version of the pedigree was used to compute an additive relationship matrix A for further analyses.

## Phenotypic data

Circumference measurements were performed on all the trees in the trial in 2004, 2008, 2012 and 2018, at the ages of 8, 12, 16 and 22 years, respectively. In addition, cores were removed from the trees of POP in December 2019, at breast height, along the same north-south direction for each tree. These cores were cut into 2-mm-thick radial strips for X-ray analysis (Polge, 1966) to obtain wood density profiles (**Fig. 4-1**). The limits between the different rings were identified with Windendro software (Guay et al., 1992) and validated by visual examination. The area of ring $y$ ($RA_{raw_y}$) was calculated at individual level as follows:

$$RA_{raw_y} = \pi \left(L + l_y\right)^2 - \pi L^2 \quad (4.2)$$

where $L$ is the sum of the ring widths from the pith to ring $y$ (ring $y$ excluded) and $l_y$ is the width of ring $y$. $RA_{raw}$ values are a good proxy for biomass produced each year independently of tree age, in contrast to ring widths which tend to decrease progressively over the years due to radial growth of the tree.

We chose to study the 2005-2019 period (15 successive years) here because rings for this period were available for at least 99% of POP and this period excludes the juvenile phase of the trees (**Suppl. Tab. 4-S1 and Fig. 4-S1**). Using the circumference measurements, $RA_{raw}$ values were spatially corrected for each site with spline functions (via the BreedR R package; Muñoz and Sanchez, 2020) and named RA (adjusted ring area). A complete phenotyping series for an individual is thus composed of 15 RA values.

## Characterization of the environment during ring growth

The environmental conditions associated to each ring were characterized with two classes of environmental indices, which depend on both year and site variables. The first class focused on a purely climatic description, with two versions ($DM$ and $DM'$) of the de Martonne aridity index (de Martonne, 1926), whereas the second provided a finer description of the environmental conditions with two indices ($GP$ and $GP'$), extracted from an ecophysiological model combining climatic, silvicultural and soil data (Moreaux et al., 2020).

The de Martonne aridity index was calculated for each ring formed in year $y$ at site $z$ with:

$$DM_{y,z} = \frac{1}{8}\sum_{i=3}^{10}\frac{12P_{i,z}}{T_{i,z}+10} \quad (4.3)$$

where $P_{i,z}$ is the amount of precipitation (in mm) and $T_{i,z}$ is the mean air temperature (in °C), for month $i$ in site $z$. Only the 8 months from March ($i = 3$) to October ($i = 10$) were included here as we considered, as a first approximation, that climatic conditions outside the growth period of maritime pine has no impact on annual RA. In addition, we considered a modified version of the de Martonne index ($DM'$) based on a 30-day sliding window average (instead of calendar months) and considering the impact of the climate of year $(y-1)$ on environmental conditions in year $y$ (inspired by Botzan *et al.*, 1998, **Suppl. 4-2**).

The environmental indices of the second class derived from the ecophysiological model GO+ 3.0 (Moreaux et al., 2020) based on climatic data, silvicultural parameters, soil water properties, soil fertility and reference values for maritime pine growing in the Landes massif (**Suppl. 4-2**). The growth potential index ($GP$) was calculated for each ring, based on mean trunk water potential and temperature estimated daily by the GO+ model. Similarly, to the de Martonne aridity indices, a second index $GP'$ was used to consider a sliding window of 10 days over the course of a year and to take into account the impact of previous year.

## Genetic analysis of radial growth

### *Univariate model*

Radial growth analysis for POP would classically be based on the most recent available circumference measurement (here for 2018, when the trees were 22 years old, denoted Cir22) and the following univariate model:

$$y = Xb + Zu + e \quad (4.4)$$

where $y$ is the vector of Cir22; $b$ and $u$ are the vectors of solutions for fixed site and random genetic additive effects, respectively; $e$ denotes the residuals. $X$ and $Z$ are the corresponding incidence matrices of dimensions $n\,x\,1$ and $n\,x\,n$, respectively, where $n$ is the number of individuals. We assumed that $u \sim N(0, G\sigma_g^2)$ and $e \sim N(0, I\begin{bmatrix}\sigma_{e_1}^2 & 0\\ 0 & \sigma_{e_2}^2\end{bmatrix})$, with $G$ the genomic relationship matrix described above, $\sigma_g^2$ the additive genetic variance, $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$ the residual variances for Site 1 and Site 2 respectively. The genomic estimated breeding values for Cir22 will be denoted $GEBV_{Cir22}$.

### *Random regression model (RRM)*

Individual RA series for POP were modeled as a function of the environmental gradient, using an RRM implemented in Wombat software (Meyer, 2007). The environmental gradients associated with the four indices previously described were modeled independently according to the RRM formulation. Regardless of the environmental index used, the joint analysis of the two sites and year series provided an overall environmental gradient of 30 levels (15 environmental levels per site). Legendre polynomials were used as the base functions (Kirkpatrick et al., 1990) for the following RRM (Mrode & Thompson, 2005):

$$RA_{ijs} = \sum_{k=0}^{k_m} \Phi_{\mathrm{ijk}}\, \mathrm{m_{sk}} + \sum_{k=0}^{k_\alpha} \Phi_{\mathrm{ijk}}\, \alpha_{ik} + \sum_{k=0}^{k_p} \Phi_{\mathrm{ijk}}\, \mathrm{p_{ik}} + r_{ijs} \quad (4.5)$$

where $RA_{ijs}$ is the ring area of individual $i$ for environmental level $j$ at site $s$; $\mathrm{m_{sk}}$ is the $k^{\mathrm{th}}$ fixed regression coefficient used to model the average trajectory at site $s$; $\alpha_{ik}$ and $\mathrm{p_{ki}}$ are the $k^{\mathrm{th}}$ random regression coefficients for the genetic additive and permanent environmental effects, respectively, of individual $i$, the latter effect representing the similarity between repeated records for the same individual of environmental and non-additive genetic origin; $\Phi_{\mathrm{ijk}}$ is the $k^{\mathrm{th}}$ Legendre polynomial for the RA of individual $i$ at environmental level $j$; $k_m, k_\alpha, k_p$

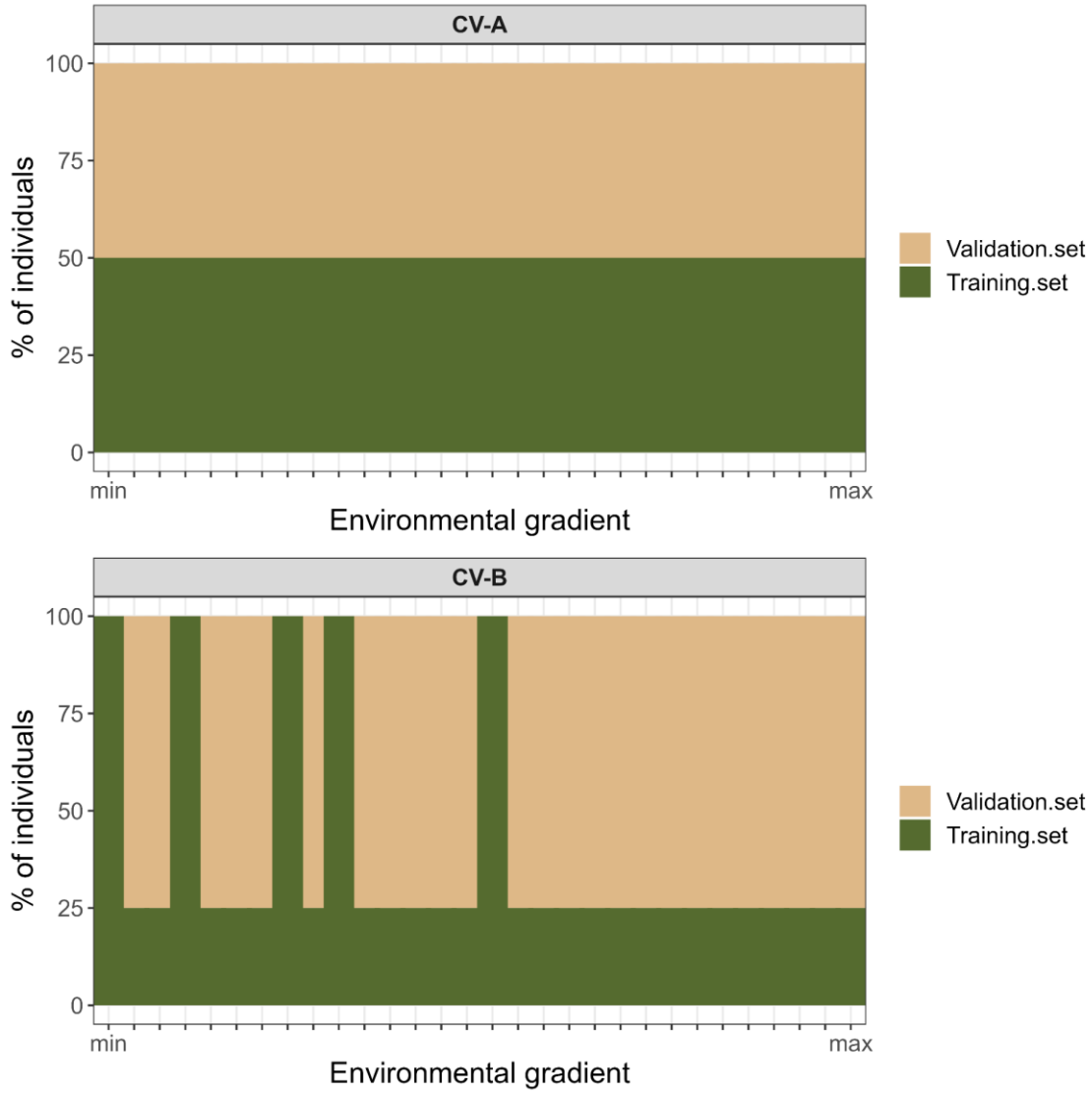



***Figure 4-2 : Cross-validation scenarios CV-A and CV-B performed with a RRM according to the GP' index*** *Both scenarios include the same amount of phenotypic information in the training set (i.e. 50% of total phenotypic data); only the distribution of this information across individuals and environmental levels differ. All families contributed equally to the training set.*

are the order of polynomials for mean trajectory, genetic additive and permanent environmental effects, respectively; and $r_{ijs}$ is a random residual.

Based on the results of a preliminary analysis (**Suppl. Fig. 4-S2**), we decided to model second-order trajectories by setting $k_m = k_\alpha = k_p = 2$.

The equivalent matrix notation for this model is (Mrode & Thompson, 2005):

$$y = Xb + Zu + Qpe + e \quad (4.6)$$

where $y$ is the vector of RA over the environmental levels; $b$ is the vector of solutions for site fixed effect; $u$ and $pe$ are the vectors of the individual genetic additive and permanent environmental random regression coefficients, respectively; $e$ denotes the residuals. For genomic-based RRM, it is assumed that $u \sim N(0, G \otimes \Omega)$, $pe \sim N(0, I \otimes P)$, and $e \sim N(0, I \otimes D)$, where $\otimes$ denotes the Kronecker product, $G$ the relationship matrix described above, $\Omega$ and $P$ the covariance matrices for the RR coefficients for the genetic additive and permanent environmental effects, respectively, and $D$ is a diagonal matrix of heterogeneous residuals for each environmental level. For pedigree-based RRM, $G$ is replaced by $A$.

With a second-order model ($k_m = k_\alpha = k_p = 2$), the RRM estimates three genetic coefficients per individual. From these, individual GEBV were then obtained at all environmental levels as a trajectory, following the formulation of (Mrode & Thompson, 2005). GEBV estimated with an RRM integrating all available phenotypic data and solved at each environmental level are denoted $GEBV_{ref}$. The individual trajectories of $GEBV_{ref}$ as a function of the environmental gradient were clustered with a K-means approach extended to longitudinal data (implemented in the kml R package; Genolini *et al.*, 2015). The Calinski-Harabatz criterion was used to define the best number of clusters (ranging from 2 to 7).

## Genomic selection

### *Cross-validation (CV) scenarios*

The predictive performance of the RRM was assessed over two CV scenarios (**Fig. 4-2**). First, the reference scenario, denoted **CV-A**, where the training set ($T_{set}$) included the complete phenotyping series for 50% of the individuals (randomly selected within sites and families), whilst the remaining 50% of individuals constituting the validation set ($V_{set}$). Second, the CV-B scenario explored the possibility of retaining the same amount of phenotypic information as for the CV-A (i.e. 50% of total phenotypic data) but distributed differently over the individuals. Scenario **CV-B** mimicked the use of a high-throughput phenotyping tool for quick estimation

of the last five RA which, in a context of global warming, would typically correspond to unfavorable years. The $T_{set}$ for **CV-B** included complete phenotypic series (i.e. 15 phenotypic records per individual) for 25% of individuals and only five phenotypic records for the remaining individuals (75% of individuals). We selected the five environmental levels within the eight most unfavorable ones by applying a Kennard-Stone algorithm (Kennard & Stone, 1969) via the prospectr R package (Stevens & Ramirez-Lopez, 2022).

For each CV scenario, the predictive performance of the RRM was estimated as the Pearson correlation coefficient between predicted ($GEBV_{pred}$) and observed RA in $V_{set}$ over the whole environmental gradient and based on 10 independent repetitions. Such performance estimator was used as a criterion for assessing modeling quality (Ly et al., 2018; Arnal et al., 2019; Momen et al., 2019).

### *Genetic gains*

The predictive performance of the RRM for genetic gains in our reference scenario CV-A was assessed over each of the environmental levels. The assessment consisted of calculating the differences in genetic gain between a selection based on $GEBV_{pred}$ obtained in $V_{set}$ and the corresponding maximum that would have been obtained with the same selection intensity based on $GEBV_{ref}$. For this, at each environmental level, the top 5% of individuals selected according to $GEBV_{pred}$ were identified and their corresponding $GEBV_{ref}$ (obtained with all the phenotypic information) used to calculate the true genetic gain ($GG_{true}$) as the $GEBV_{ref}$ average of the selected individuals. This amount was compared for the corresponding environmental level to the maximum gain ($GG_{max}$), which was calculated as the $GEBV_{ref}$ average of the top 5%. Finally, $GG_{true}$ and $GG_{max}$ were centered and reduced to ensure comparability between environmental levels. Any difference between $GG_{true}$ and $GG_{max}$ would indicate a decrease in the correlation between $GEBV_{pred}$ and $GEBV_{ref}$ for the selected percentage.

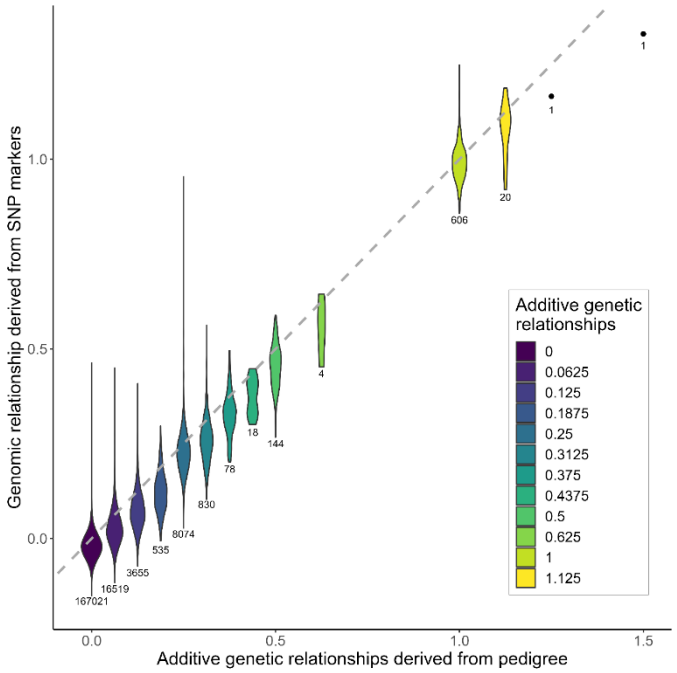



***Figure 4-3 : Comparison between additive genetic relationships derived from pedigree and genomic relationships derived from SNP markers for individuals of POP.*** *For each value of the discrete scale taken by the additive genetic relationships, the corresponding violin plot represents the continuous distribution of genomic relationships. Numbers below each violin plot denote the number of relationship included in the corresponding violin plot. Grey line is the bisector passing through the origin of the graph. The two highest relationships derived from pedigree (1.25 and 1.5) are unique and so represented by single points instead of violin plots.*

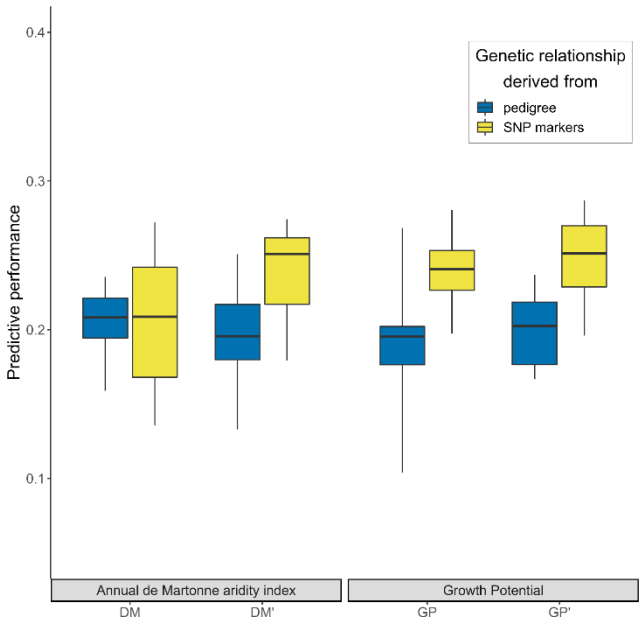



***Figure 4-4 : Predictive performance of the RRM according to the environmental gradient and the genetic information used.*** *Boxplots indicates the Pearson correlation coefficient between observed and predicted RA values over the whole environmental gradient for 10 repetitions of the CV-A scenario. Boxplots are blue when the RRM implemented integrated additive genetic relationship derived from pedigree while boxplots are yellow when the RRM implemented integrated relationships derived from SNP markers. For each kind of genetic information, RRM were run independently with each of the four environmental gradients, respectively derived from DM, DM', GP and GP' indices.*

# Results

## Size and genetic characterization of POP

After phenotype curation (9 wood-density profiles excluded) and genotyping quality control (13 genotypes excluded), POP was finally composed of 628 trees (303 from Site 1 and 325 from Site 2).

Pedigree recovery on POP validated 93% of the pedigree seed parents (monoicous individuals acting as mothers) and allowed the correction of 5%. The remaining 2% of the pedigree seed parents was classified as unknown, as no candidate parent could be validated. Pollen parents (acting as fathers) were successfully recovered for 65% of the individuals. Note that the original design of the study was based on crosses with a mixture of pollen donors, resulting in the fathers initially being unknown in the pedigree. Finally, based on the curated pedigree, a status number ($N_s$; Lindgren *et al.*, 1996) of 21 was obtained for POP, suggesting a high level of relatedness between the families studied.

The genotyping of POP resulted in the characterization of the 628 individuals over 3,832 SNP, with a repeatability of 97% and a total missing data rate of 1%. Genomic relationship coefficients ($g_{xy}$) estimated in $G$ were consistent with the additive relationship coefficients ($a_{xy}$) calculated in $A$ (**Fig. 4-3**). The $a_{xy}$ values were discrete, whereas the $g_{xy}$ values were normally distributed for each level of relatedness. Note that, for most additive coefficient levels, the normal distribution has a long upward-sloping tail (revealing some rare cases of unrecorded relatedness), and a mean slightly below the theoretical value, the latter being represented by the gray line in **Figure 4-3**.

## Quality of model fit

The predictive performance (estimated with CV-A) was used as a criterion for assessing the quality of RRM (**Fig. 4-4**). Mean predictive performances were poor, with correlation coefficients ranging from 0.19 to 0.25. Predictive performance was slightly better (+0.04 better, on average) for genomic-based RRM than for pedigree-based RRM, except for RRM based on the $DM$ environmental index (equivalent mean predictive performance of 0.21). The best predictive performances were obtained for genomic-based RRM with the $DM$' (0.24) and $GP$' (0.25) indices. The optimization of environmental indices improved slightly RRM predictive ability by 16% and 3% relative to the initial $DM$ and $GP$ indices, respectively. Finally, the

genomic-based RRM using the $GP'$ index was selected for the analyses described below, due to its best predictive performance (0.25) for genomic selection.

## Individual reactions norms estimated by genomic-based RRM

Reordering longitudinal data by the annual environmental index, which characterizes the conditions of ring formation, instead of the ordinal year greatly modified the shape of the mean RA curve in a more easily interpretable way (**Fig. 4-5**). When expressed as a function of the environmental index GP', RA increases significantly. The lowest $GP'$ values are associated with the most unfavorable environmental conditions for growth, whereas the highest values are associated with the most favorable conditions for growth. This pattern suggests plasticity at the population level, but hides individual behaviors, which may deviate from this central trajectory. Random individual deviations from the mean trajectories due to additive genetic effects are represented in **Figure 4-6** and were solved over the environmental gradient of $GP'$ ($GEBV_{ref}$). For most genotypes, $GEBV_{ref}$ showed a dependence on $GP'$, highlighting the existence of plasticity for RA. These different behaviors can be characterized thanks to five clusters, depicted in different colors (**Fig. 4-6**). Each individual reaction norm can be characterized by its mean $GEBV_{ref}$ and its slope defined with two coefficients (quadratic and linear regression coefficients as second-order trajectories were modelled). Fifty-seven percent of individual reaction norms were assigned to clusters A and B (**Fig. 4-6**) and were characterized by shallow slopes (**Suppl. Fig. 4-S3**) and mean $GEBV_{ref}$ close to 0 (**Suppl. Fig. 4-S4**). This does not mean that the phenotypic trajectory of these individuals is flat. Instead, it indicates that they have trajectories indistinguishable from the mean trajectory and that this indistinguishability, by its additive genetic origin, would not give extra plasticity to the offspring. The highest and lowest mean $GEBV_{ref}$ were those obtained for individuals from clusters C-E (26%) and cluster D (17%), respectively. These clusters also display reaction norms with the strongest positive and negative slopes (for clusters C-E and D, respectively), leading to a greater range of variation in individual genetic values in favorable than in unfavorable environmental levels. Mean $GEBV_{ref}$ is, thus, strongly correlated with slope regression coefficients (0.52 and 0.95 with quadratic and linear regression coefficients, respectively). There appear to be few intersections between reaction norms, corresponding to changes in individual ranks across environmental levels, over most of the environmental gradient. However, large overlaps occur in the part of the gradient corresponding to unfavorable environmental levels. Moreover, despite the overall similarity of

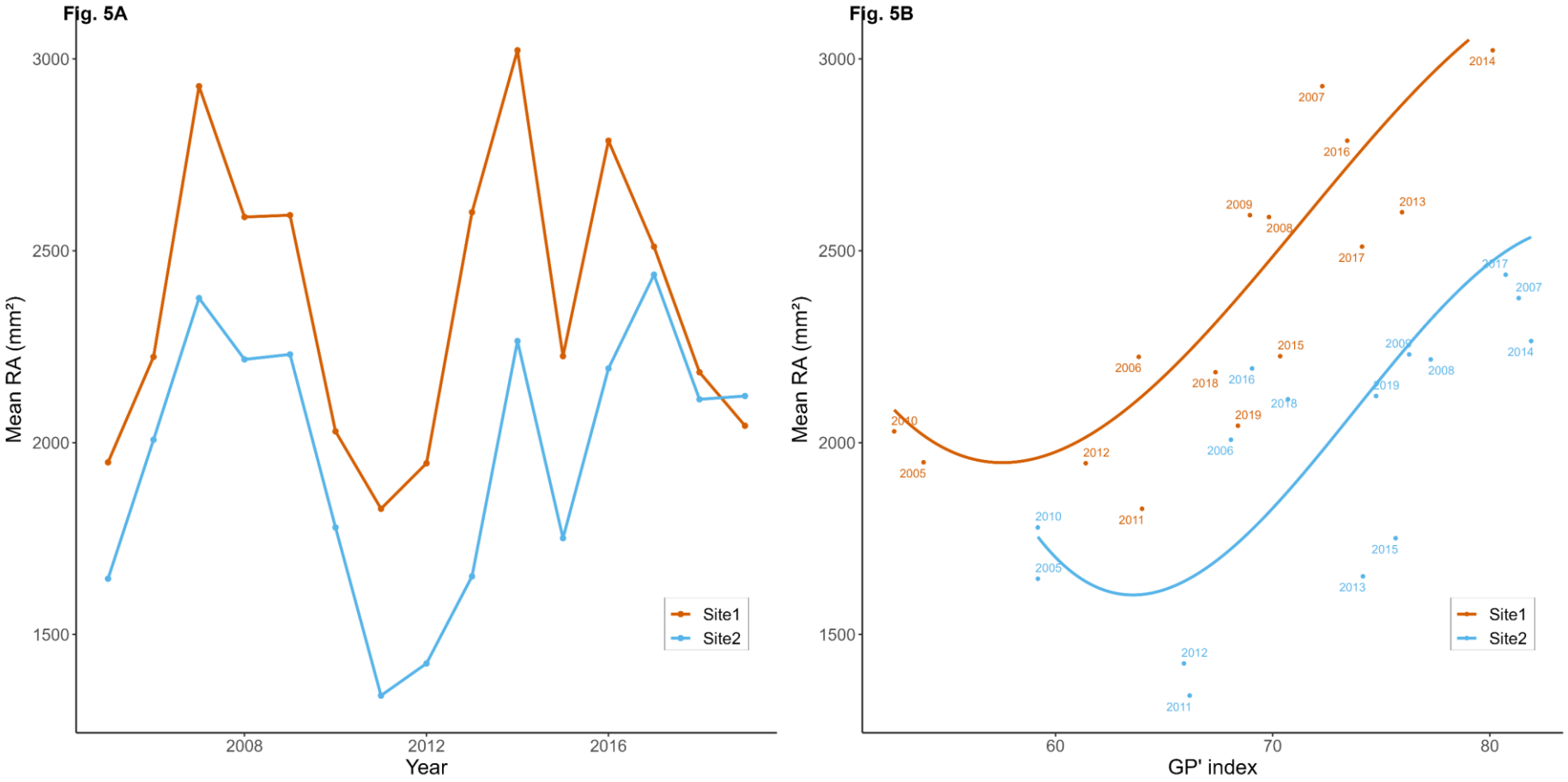



***Figure 4-5 : Evolution of mean RA according to the years for each site (Fig. 5A) or according to the GP' index (Fig. 5B).*** *Figure 5.A presents mean phenotypic trajectories of RA and Figure 5.B presents mean trajectories adjusted by the RRM for each site. Both trajectories are the result of the same model. The significance of the slope parameter for each trajectory in Figure 5.B was assessed with Student's t-test (p-value<0.01)*

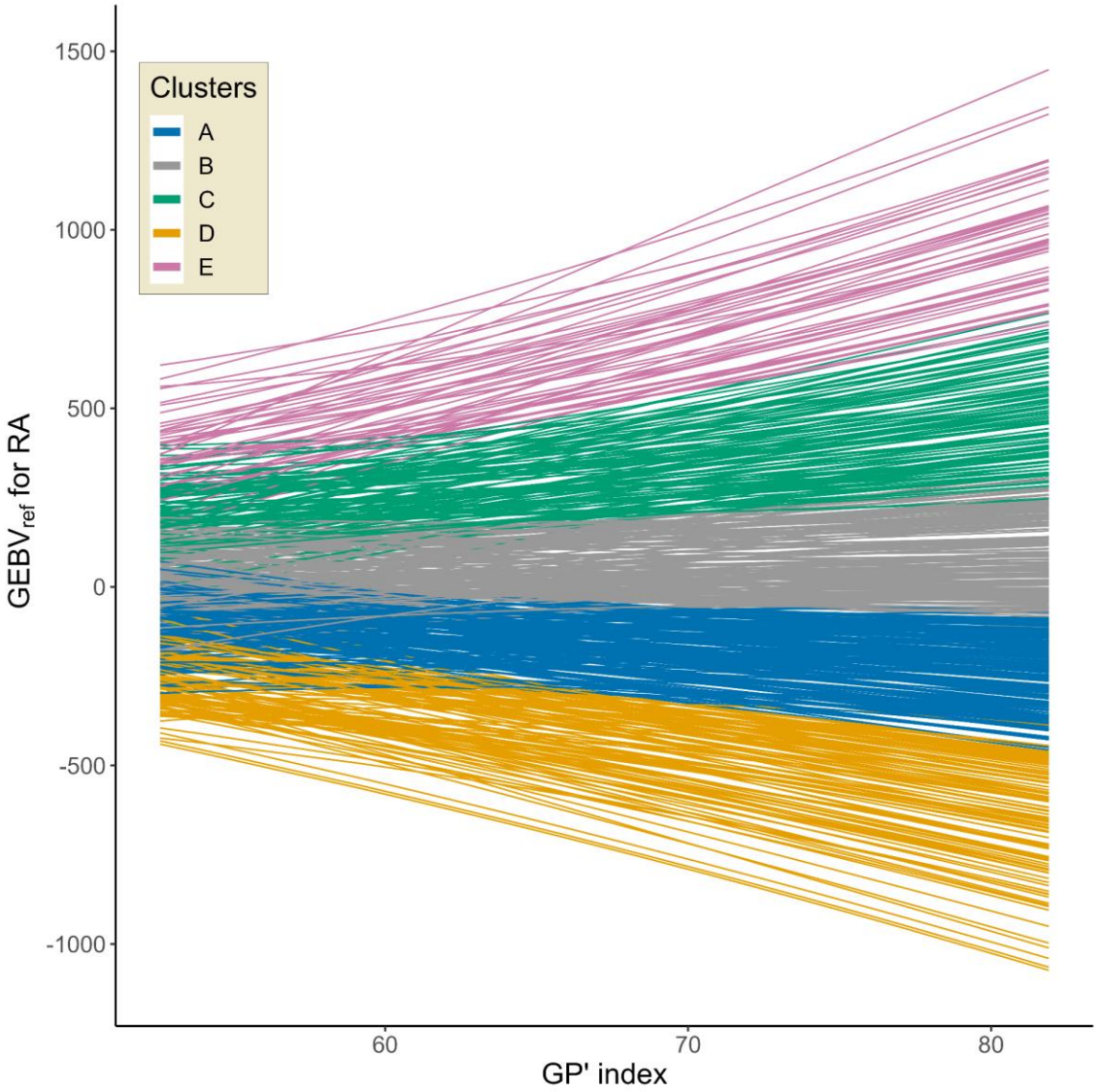



***Figure 4-6 : Individual trajectories of GEBVref associated to RA according to the GP' index.*** *Trajectories correspond to the genetic component of the reaction norms estimated by the genomic based RRM. Trajectories were divided in 5 clusters (A to E) with the following proportions: A-29.6%, B-27.3%, C-18.6%, D-17.2%, E-7.3%.*

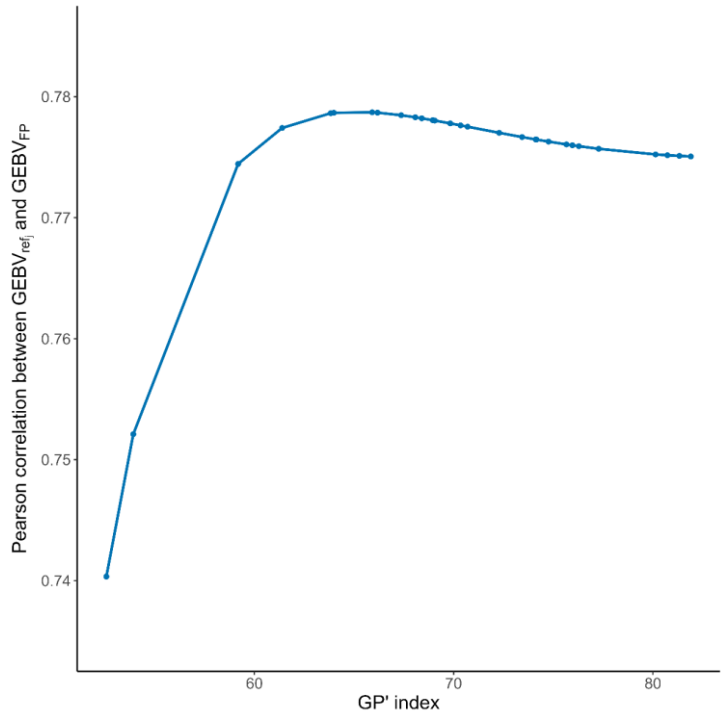



***Figure 4-7 : Pearson correlation coefficients between genomic estimated breeding values obtained from a final-point univariate model ($GEBV_{FP}$) and genomic estimated breeding values obtained from a RRM at each GP' level j ($GEBV_{ref\_j}$)***

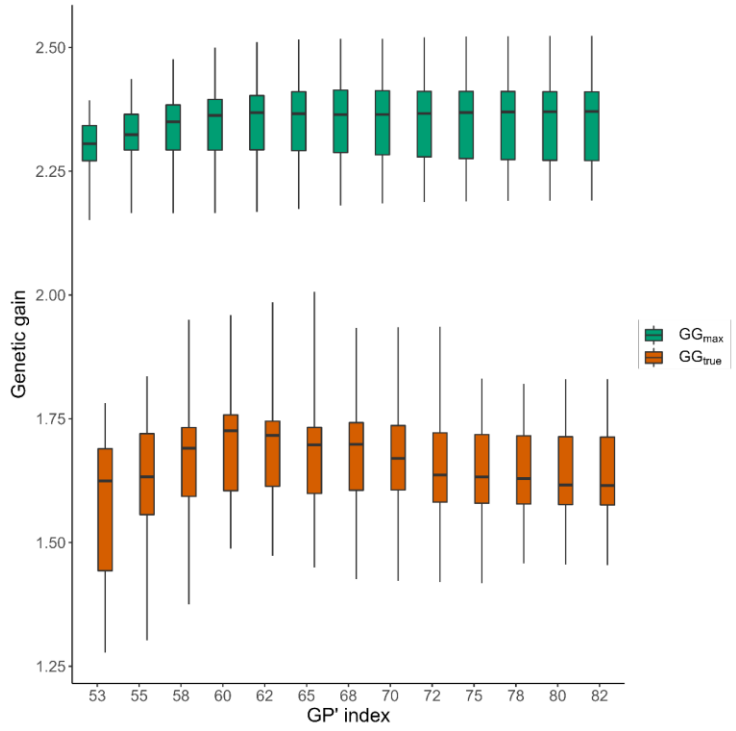



***Figure 4-8 : Maximum genetic gain ($GG_{max}$) and true genetic gain ($GG_{true}$) according to GP' index.*** *The RRM was used with complete phenotypic information for all individuals to estimate $GEBV_{ref}$ over the gradient; and then independently repeated 10 times with the scenario CV-A to predict $GEBV_{pred}$ for individuals in the validation set. $GG_{max}$ was calculated as the mean of the top 5% of $GEBV_{ref}$ and for each iteration $GG_{true}$ was calculated as the mean of $GEBV_{ref}$ associated to the top 5% individuals selected based on $GEBV_{pred}$ for each **GP'** values. $GG_{max}$ and $GG_{true}$ are centered and reduced.*

trajectories within the same cluster, overlaps between norms were detected within clusters, mostly in the unfavorable part of the gradient (**Suppl. Fig. 4-S5**).

Correlations were calculated between the GEBV obtained with the final-point univariate model ($GEBV_{FP}$) and those obtained with the RRM across the successive environmental levels *j* ($GEBV_{ref\_j}$). For GP' environmental levels over 59, correlations remained almost constant ranging from 0.77 to 0.78, highlighting the consistency between $GEBV_{ref\_j}$ and $GEBV_{FP}$ at most environmental levels (**Fig. 4-7**). However, the correlation was clearly weaker for the most unfavorable environmental levels (GP'<59), with the coefficient falling to a minimum of 0.74. These lower values highlight differences in the behavior of some genotypes (and, therefore, genotype ranking) at these environmental levels.

## GS scenarios and cross-validation

### *Genetic gain over the environmental gradient*

The overall predictive performance of the genomic-based RRM (using the GP' environmental index) estimated with the CV-A scenario was 0.25 (**Fig. 4-4**). Breeding efficiency, based on predicted values, was assessed by calculating genetic gains for different environmental levels (**Fig. 4-8**). $GG_{max}$ increased until the environmental value of 62, above which it reached a plateau with maximum value of 2.35. The differences between $GG_{max}$ and $GG_{true}$ was minimal for GP' environmental level 64, increasing to a maximum at the most unfavorable (73) and most favorable (72) environmental levels. The relatively poor predictive performance of the RRM (0.25) necessarily led to a significant loss of genetic gain (no overlap between $GG_{max}$ and $GG_{true}$ boxplots). Nevertheless, depending on the environmental level, $GG_{true}$ accounted for 68% to 73% of $GG_{max}$. $GG_{true}$ was always significantly different from 0 ($p_{value\ T-Test} < 0.001$), indicating a certain efficiency of selection based on predicted values, even in the most extreme environmental levels.

### *Predictive performance over the CV scenarios*

We considered an alternative cross-validation scenario (CV-B) (**Fig. 4-2**), to improve selection efficiency while preserving phenotyping effort with respect to CV-A. As in CV-A, 50% of the phenotypic data were used to constitute the $T_{set}$ of the CV-B. The key difference between the two is due to a better distribution of phenotypic effort, both between individuals and between environments, in CV-B. This alternative distribution had a considerable impact on improving the predictive performance of the RRM, which increased from 0.25 for the CV-A to 0.59 for

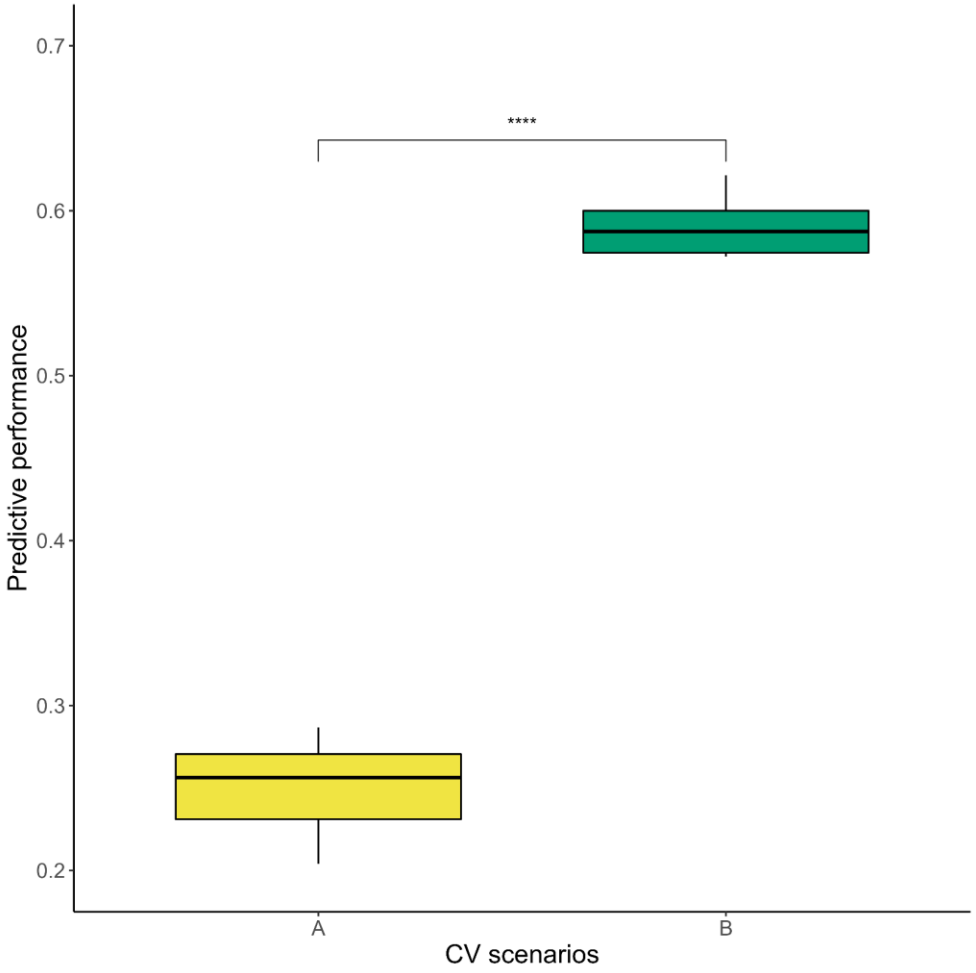



***Figure 4-9 : Predictive performance of the RRM according to the CV-A and CV-B scenarios.*** *Boxplots indicates the Pearson correlation coefficient between observed and predicted RA values over the whole environmental gradient for 10 independent repetitions of the CV scenario. The significance between predictive performances was assessed by a Student's t-test (****: p-value <1e10-4)*

the CV-B (**Fig. 4-9**), with no increase in phenotyping effort. It should be noted that the CV-A scenario is a major challenge for RRM, as it imposes the prediction of entire trajectories for half of the population. This challenge is relaxed in CV-B by including at least partial information for all individuals.

# Discussion

Deciding which genetic material should be planted now to form the forests of tomorrow is becoming increasingly challenging due to the rapidity of climate change (Thomas et al., 2004; Wiens, 2016). Using longitudinal tree-ring data and parallel environmental descriptors, we have successfully modeled genomic individual reaction norms based on random regression. This first example for forest trees provided consistent results for use in the maritime pine breeding program, but may inspire other programs in perennial species.

## Reaction norms in forest trees

Growth measurements at advanced age are generally used for the calculation of breeding values. Such measurements constitute highly integrative phenotypes that can be associated only with a global environmental site index. Using sites with contrasting indices has been a classic strategy to establish comparative trials for genetic x environment evaluation. In this sense, our two sites present strong contrast in terms of fertility and water table depth at the scale of the Landes massif, but even with their differences they are still part of the same breeding area (Jolivet et al., 2007). Wood cores give us access to phenotypic inter-annual variation and can be used to generate longitudinal annual growth data that can be associated with annual environmental variation. Our results showed indeed that the environmental variation between years was much greater than the one between sites (**Fig. 4-5**). Indeed $Cir_{22}$ was associated with a mean environmental index GP' of 68.2 for site 1 and 72.1 for site 2, whereas analysis based on ring measurements covered a larger index range (GP' from 52.6 to 81.9). This much greater annual variation provides an opportunity to infer plasticity at individual level over a large environmental gradient.

In addition to longitudinal data collection, which can be operationally costly, there are other challenges that arise with these data. One is autocorrelation between repeated measurements on the same individual in a time series. Another, not least, is ontogenetic differences between phases of phenotype expression (Sanchez et al., 2013). Finally, a third challenge is the choice of a relevant environmental descriptor. Although we have not shown it for simplicity, we have performed a preliminary RRM for RA with a one-year lag in the climatic index in order to match RA of year *n* with the environmental index of year *n - 1*, and its results pointed to an absence of autocorrelated effect. As for the ontogeny challenge, we have ignored in our longitudinal data series the initial segments corresponding to the juvenile phase, keeping only

the remaining adult phase for which the RA trend was generally flat, despite strong inter-annual oscillations (**Suppl. Fig. 4-S1**).

The third challenge is probably the most difficult to address, the choice of a relevant environmental index (Li et al., 2017). This study was not designed to identify precisely the environmental factors most relevant to tree growth, but we defined two classes of biologically meaningful environmental indices that integrate the key components of temperature and water (Begum et al., 2013; Rathgeber et al., 2016). Both of them depend on the year and the site in which the ring was formed. The first class (aridity indices) is easy to obtain, since it only considers the climatic data (temperature and precipitation over the growing period) of the site and the year associated with to the rings under study. On the other hand, the second class (growth potential indices) requires more complex modeling, including for example the characterization of the daily water status of the trunk. A major difference between the two types of indices is the insensitivity of the former to the intra-annual distribution of precipitation and temperatures. Thus, similar annual aridity values ($DM$ or $DM$') may reflect different climatic realities over the course of the growing season, with temperatures and/or precipitation occurring at different periods and leading to differences in growth. Conversely, by considering the daily environmental status and tree physiology, the growth potential indices ($GP$ and $GP$') allowed a more detailed consideration of within-year environmental variation.

Finally, the predictive performance obtained with DM, DM', GP, GP' (**Fig. 4-4**) confirms the relevance of the proposed environmental indices, but also suggest that they only partially capture the environmental factors influencing radial growth and the differences between individuals' reactions. More specifically, the variability due to site is not fully described by the index, given the remaining high significance value of the corresponding fixed effect in RRM (**Fig. 4-5B**).

## Modelling reaction norms with RRM

Unlike univariate single-point analyses, which are easy to implement but do not integrate longitudinal phenotypic information, or multi-trait models, which can integrate it but are computationally demanding, RRM provides genetic estimates over the chosen continuous environmental gradient with reduced parametrization (Sun et al., 2017). The continuous trajectory of GEBV predicted by the RRM allows a position to be considered at any environmental level, whether it has actually been observed or not. The RRM can model highly complex curves using orthogonal base functions such as Legendre polynomials, which are

widely used and described in the context of breeding (Campbell et al., 2018; Marchal et al., 2019; Schaeffer, 2004). Despite their great flexibility and computational advantages, Legendre polynomials may present numerical problems (Runge's phenomenon) at the extremities for high-order fits (de Boor, 1978; Meyer & Kirkpatrick, 2005). In this study, the adjustment at the extremities of the environmental gradient was particularly important as the unfavorable extreme conditions are likely to increase in frequency in the future (Coumou and Rahmstorf, 2012; Spinoni *et al.*, 2018). The use of low-order polynomials to model RA trajectories overcame this problem. The consistency and quality of the norms fitted with Legendre polynomials were verified by a comparison with norms fitted with B-spline functions, which are considered a more robust alternative to high-order polynomials in terms of extremum fitting, although less advantageous computationally (de Boor, 1978; Meyer & Kirkpatrick, 2005) (Kendall correlations between $GEBV_{ref}$ estimated with Legendre polynomials and those estimated with B-splines yielding coefficients of up to 0.95 over the entire gradient for the final RRM).

## Exploration of individual genetic trajectories

Random individual trajectories (**Fig. 4-6**) highlight the existence of plasticity for genetic values that can be targeted by breeders. It is not easy to discriminate between individual reaction norms that follow a trajectory close to the population average, given their high frequency, the fact that they present shallow slopes and mean $GEBV_{ref}$ close to 0. However, individuals with potentially good growth along the entire gradient are much easier to discriminate from the rest, for which the proposed clustering allows simple and efficient visualization (cluster E), useful for selection purposes.

The distribution of individual GEBV varied between environmental levels, and those more favorable levels enhanced the expression of differences between trajectories relative to less favorable levels, which has already been observed in other biological models (Arnold et al., 2019). Genotype ranking was globally preserved over the trajectories for most of the gradient (van Eeuwijk et al., 2016). This trend was confirmed by strong correlations (**Fig. 4-7**) between the GEBV obtained with this RRM for radial growth and those obtained with the final circumference univariate analysis (final-point model). However, this correlation was weaker for unfavorable environments (environmental index below 59), in agreement with the reranking of genotypes observed for the individual trajectories at the most unfavorable end (**Fig. 4-6**). This precise and localized GxE interaction in our gradient, only possible thanks to

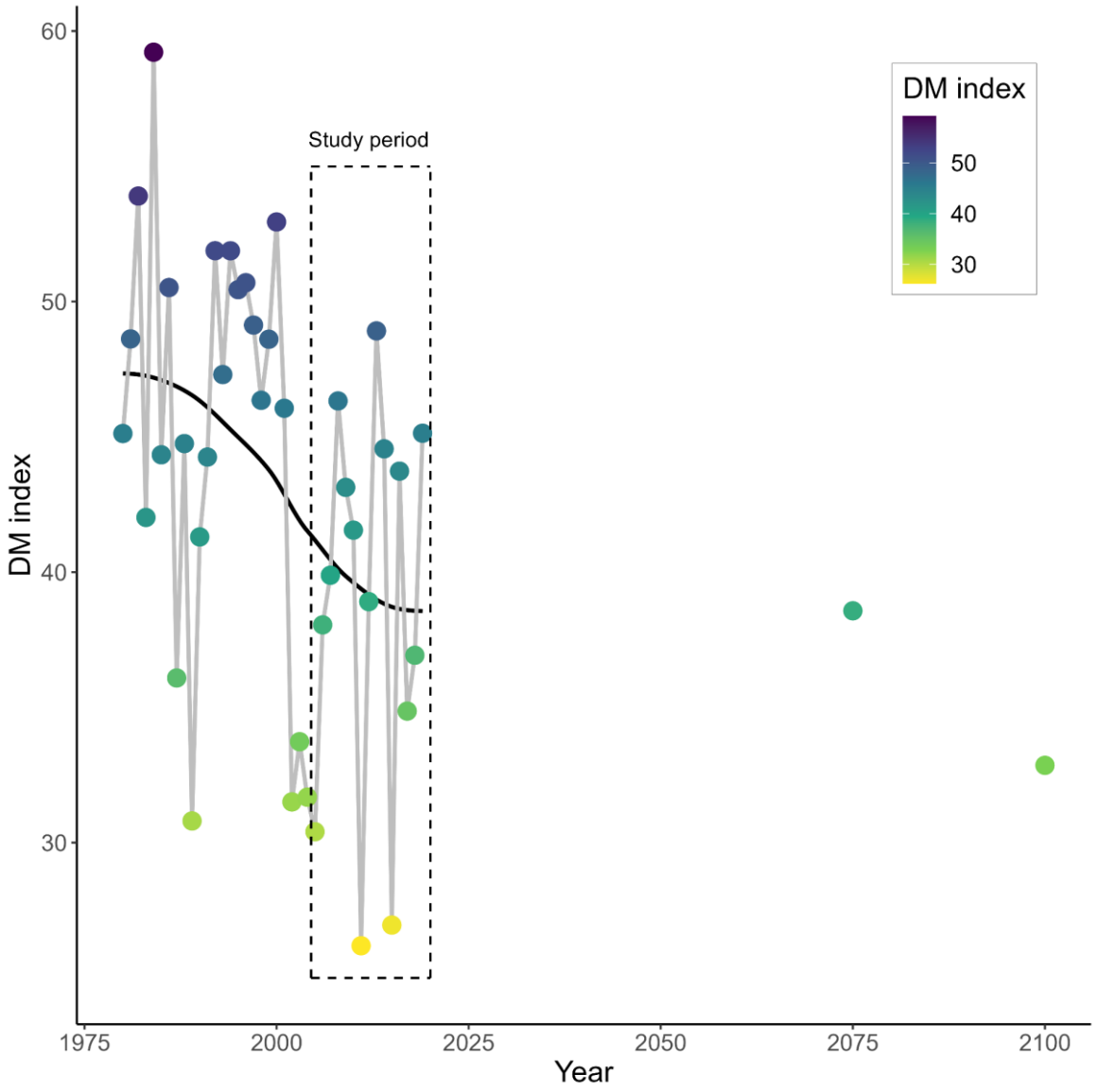



***Figure 4-10 : Evolution of DM index from 1980 to 2019 and prediction of this index for medium and long term horizon.****Evolution from 1980 to 2019 was calculated with historical data from Meteo France weather station close to Site1. Predictions positioned in 2056 and 2086 were calculated using weather predictions available on drias-climat.fr considering the RCP8.5 scenario (Scenario without any climatic politic) for period from 2041 to 2070 (medium horizon) and from 2071 to 2100 (long term horizon). Our study period from 2005 to 2019 is framed in a dotted box.*

the use of the RRM, should not be considered marginal or potentially negligible considering that it affects only one segment of the gradient. In fact, climate projections (**Fig. 4-10**) suggest that such unfavorable environments are likely to become much more common in the future. Even if the expected global level of aridity in 2075 remains close to current levels, according to our de Martonne calculation, aridity in 2100 will be much stronger, with a higher frequency of extreme events as predicted by other studies (Sillmann & Roeckner, 2008; Lehner et al., 2017). Our 15-year study period was already affected by a high global level of aridity and included extreme annual climates that may become frequent in the future. The environmental gradient used for the inference of reaction norms is therefore particularly relevant for identifying genotypes with better potential for growth in the unfavorable years to come.

When GxE interactions must be taken into account in selection decisions, a robust strategy would involve prioritizing the best adapted genotypes across the entire environmental gradient (Li et al., 2017), focusing on the notion of persistence. The definition of persistence may vary according to species and breeding aims (Gengler, 1996; Rocha et al., 2018), but it is generally defined as the capacity of a species to maintain a stable or high level of growth or production over time or in the face of different environmental conditions. For reactions norms, several ways of evaluating persistence and integrating the slope of trajectories in an operational breeding context have been proposed. For example, for feed conversion ratio in large white pigs, Huynh-Tran *et al.*, 2017 suggested combining the EBV estimated by the RRM with the coefficients of eigenvectors estimated from the eigenvalue decomposition of the covariance matrix of additive genetic effects. In their study, two summarized breeding values for each individual were sufficient to describe most of the variation in terms of mean genetic values (first dimension) and the slopes of EBV trajectories (second dimension), and could be used directly in selection. In another example in goat lactation, Arnal *et al.*, 2019 considered "the cumulative deviation in genetic contribution to yield relative to an average animal having the same (initial) yield" for the calculation of persistence-related EBV. Finally, Peixoto *et al.*, 2020 suggested the ranking of cotton genotypes on the basis of area under the reaction norm, the genotype with the highest norm being the most persistent. Another interesting approach would involve calculating the final GEBV for each individual as the mean of the GEBV for each environmental level weighted by the probability of occurrence of the environmental level in the future. Such a strategy would make use of the GxE interaction to maximize genetic gain for individuals performing in environmental conditions close to those predicted for the near future,

while ensuring a certain level of resilience to environmental variation. Any of these proposals could be applied to our data. A possible advantage of the latter strategy could be to take more explicit account of future climate predictions, provided that they have some control over uncertainty.

## Reaction norm in a GS context

The use of genomic reaction norms to predict growth in unobserved environments is a good example of the potential benefits of genomic selection approaches for traits that are complex to evaluate. Wood density profiles provide highly informative longitudinal data on tree growth over the years, but its acquisition via the coring process remains costly and time-consuming at breeding-program scale. This limitation has motivated one of our alternative cross-validation scenarios (CV-B), with a more homogeneous distribution of phenotypic effort, resulting in a training population involving all individuals, 25% of which contribute full time series and the remaining 75% only partial 5-year series. Indeed, relative to our baseline scenario (CV-A), which aimed to predict the full trajectories of 50% of individuals, the CV-B scenario achieved a much higher level of predictive performance (0.59), demonstrating that the allocation of phenotyping effort to constitute the training population is a key optimization to consider. The scenario CV-B would reflect the use of a high-throughput phenotyping tool usable on a large number of individuals at the cost of a smaller number of rings scanned per tree, which is basically what a resistograph does (Bouffier et al., 2008). Resistograph measures the resistance of the wood to penetration with a needle and can estimate RA efficiently for the rings closest to the bark, i.e. the last five rings formed (personal communication). These measurements provide only partial information about plasticity, but when applied to the whole population, they have the advantage of providing information complementary to that obtained by coring. Overall, less phenotyping effort is required, but the benefits are substantial.

The genetic component of reaction norms, the one of greatest interest to breeders, was estimated by integrating pedigree or genomic information in the RRM. Genomic-based RRM had a significantly better predictive performance (with the $GP'$ index) than pedigree-based RRM, suggesting that refining the coefficients of relationships between individuals through their molecular characterization with SNP results in the generation of more suitable models (Bouvet et al., 2016; Gamal El-Dien et al., 2016). The pedigree information tended towards a systematic overestimation of pairing coefficients relative to the genomic information (**Fig. 4-3**). However, some rare pairs of individuals appeared to be much more related on the basis of

genomics than on the basis of pedigree, suggesting that, in some cases, the pedigree may be incomplete, or may contain errors, despite the correction and recovery steps (Tan et al., 2017; Li et al., 2019). The use of genomic data for genomic evaluation is often proposed for forest trees (Grattapaglia & Resende, 2011; Lebedev et al., 2020), but first GS studies for maritime pine (Bartholomé et al., 2016; Isik et al., 2016) highlighted the difficulty of demonstrating a superiority of genomic models over pedigree-based models. In this study, we provide some arguments to go beyond these limitations in the application of the genomic prediction model. The RRM takes greater advantage of genomic information to predict individual trajectories than pedigree information. Indeed, in a context of intense climate change, the importance of integrating environmental information into genetic evaluation may fully justify the additional cost of genotyping (Isik, 2014).

## Supplementary Figures

***Table 4-S1: Number of rings available for POP after filtering according to the year***

| Year | *1998* | *1999* | *2000* | *2001* | *2002* | *2003 to 2017* | *2018* | *2019* |
|---|---|---|---|---|---|---|---|---|
| Number of rings available for POP | 3 | 203 | 530 | 606 | 624 | 628 | 627 | 625 |

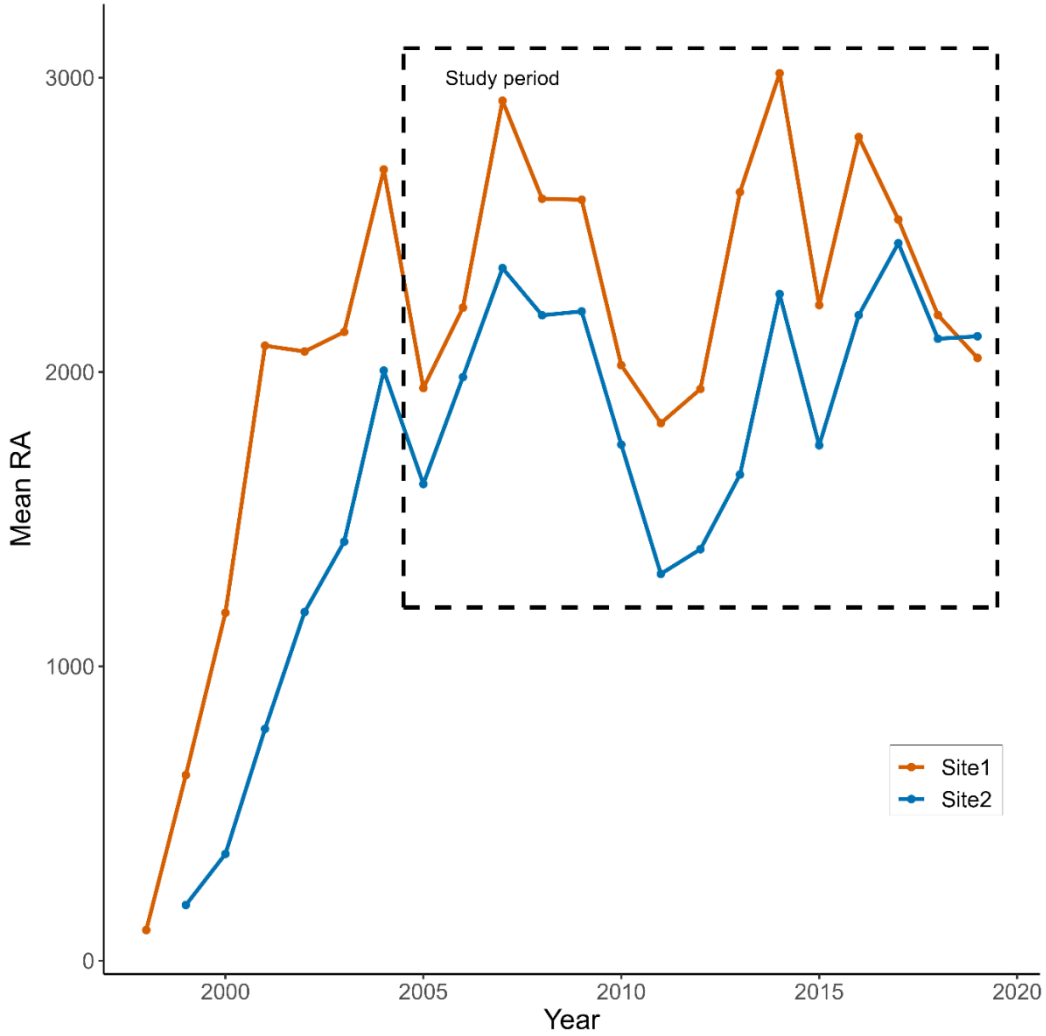



***Figure 4-S1 : Evolution of mean RA over the years for individuals of POP.*** *The orange and blue lines represent the average trajectories of the 303 individuals of Site1 and the 325 individuals of Site2, respectively. Our study period from 2005 to 2019 is framed in a dotted box.*

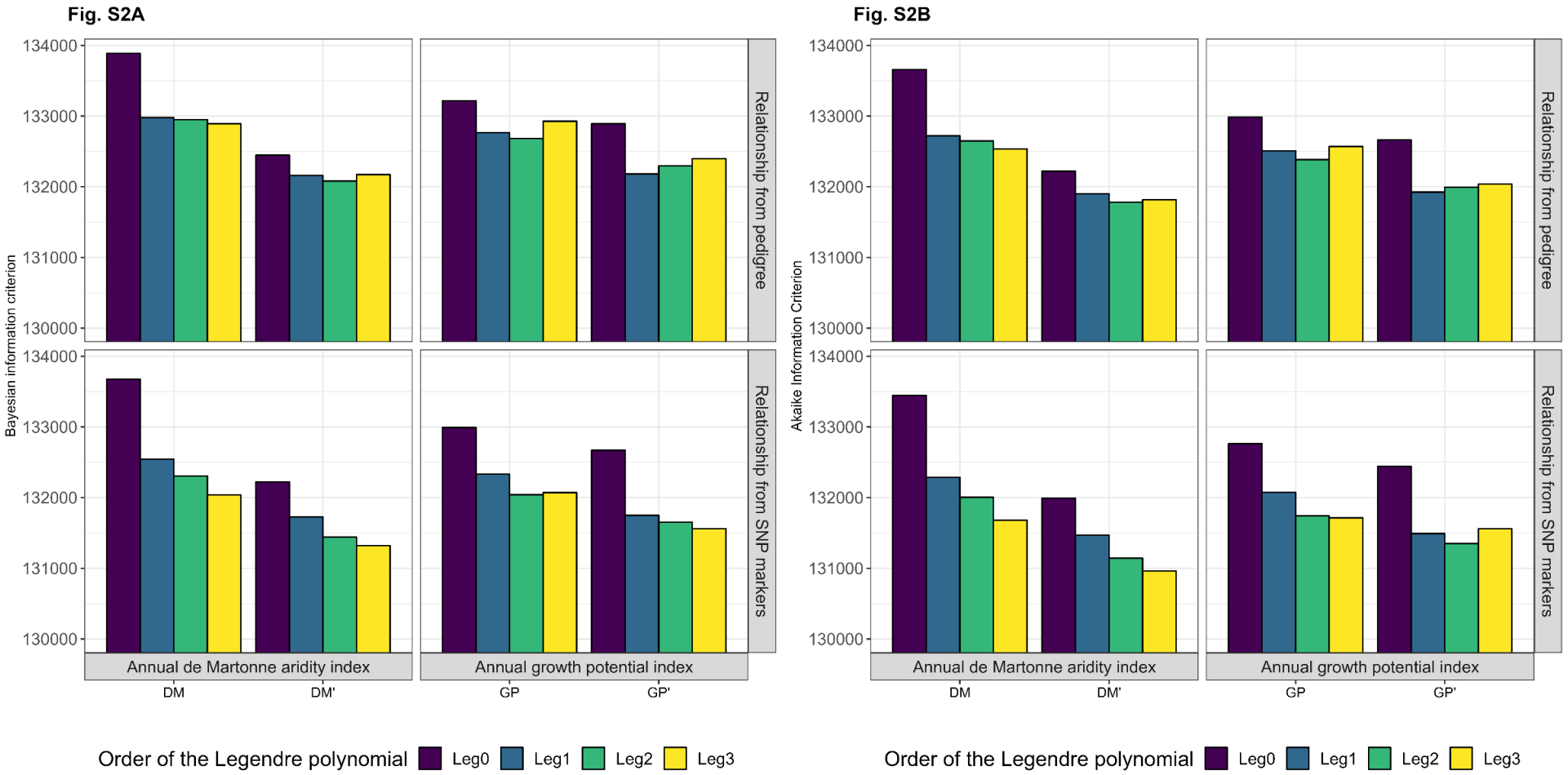



***Figure 4-S2 : Comparison of Bayesian information criterion (BIC) (Fig. S2A) and Aikaike information criterion (AIC) (Fig. S2B) for RRM with different orders of Legendre polynomials.*** *Order 2 RRMs seem more appropriated in our study to model the relatively simple trajectories of RA. Compared to order 3 RRMs, order 2 RRMs display very similar shapes of reaction norms and allow a drastic reduction on computational demand, with a marginal loss of goodness of fit according to AIC and BIC when appearing as second the second best fit.*

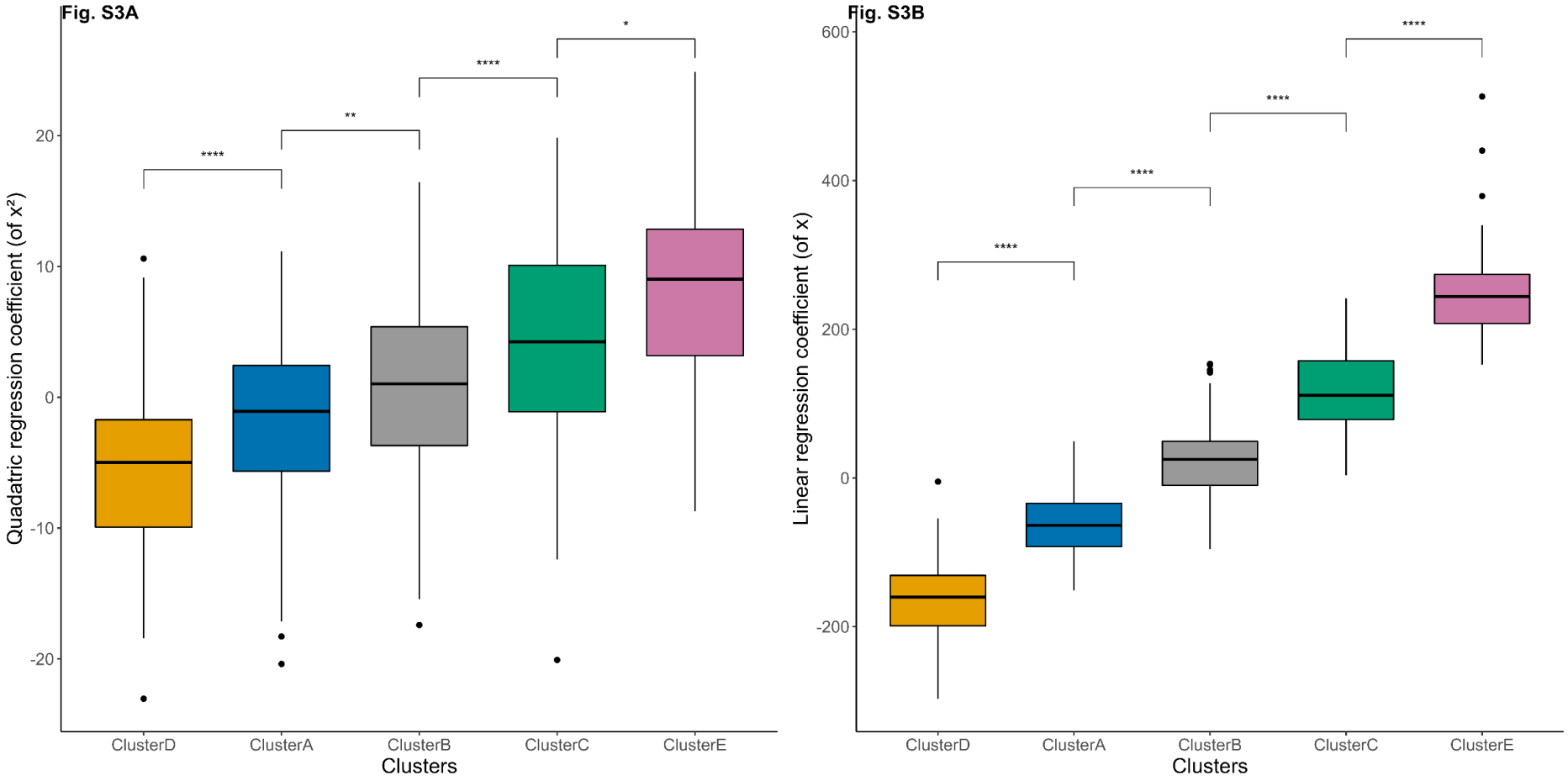



***Figure 4-S3: Quadratic and linear regression coefficients of the trajectories estimated by the RRM depending on the cluster.*** *The differences between boxplots were assessed by a Student's t-test and the significance level associated is indicated above boxplots (*: p-value<0.05, **: p-value<0.01, ****: p-value<0.0001)*

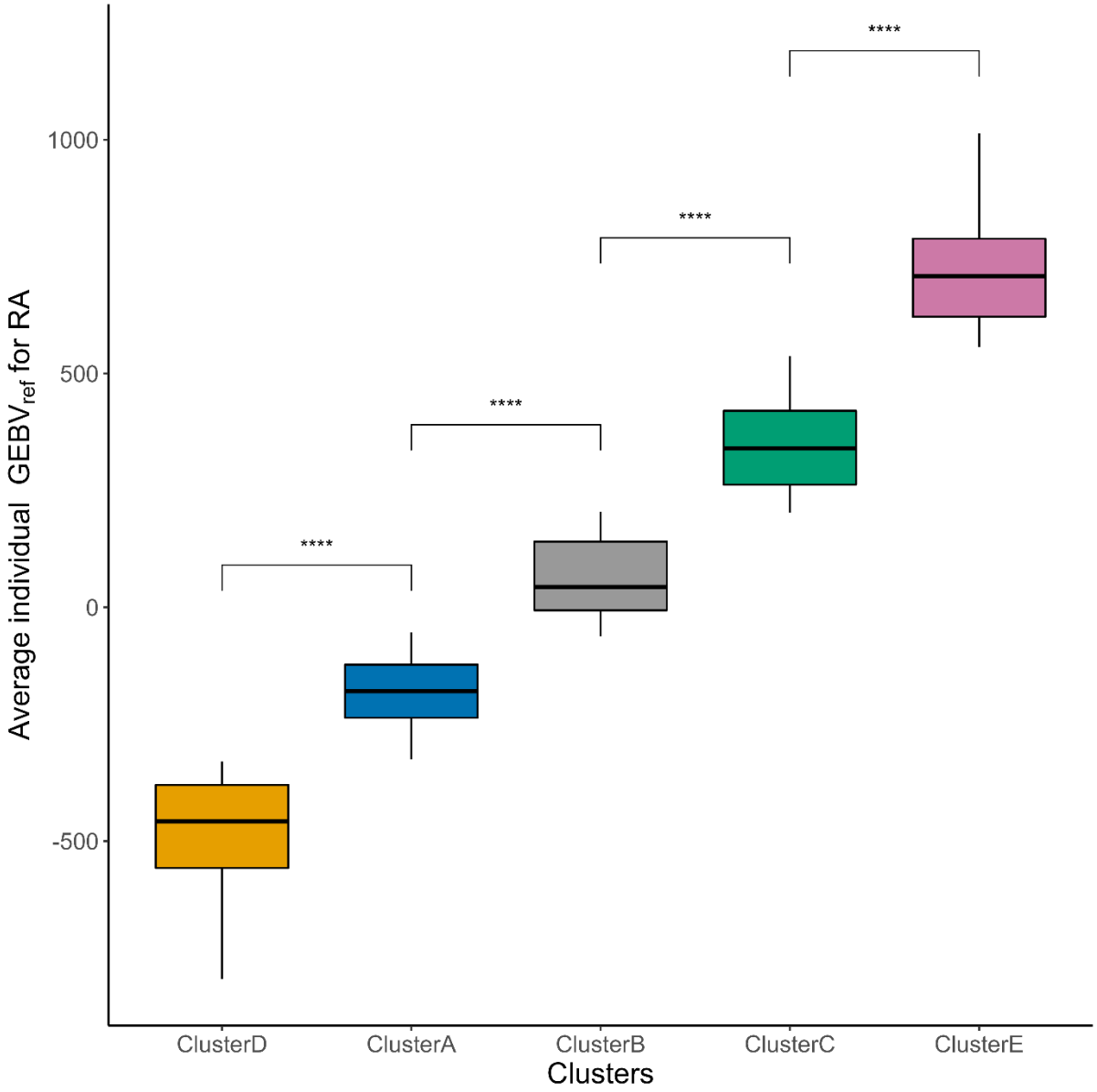



***Figure 4-S4: Mean GEBV$_{ref}$ per individual estimated by the RRM depending on the cluster.***
*The differences between boxplots were assessed by a Student's t-test and the significance level associated is indicated above boxplots (****: p-value<0.0001)*

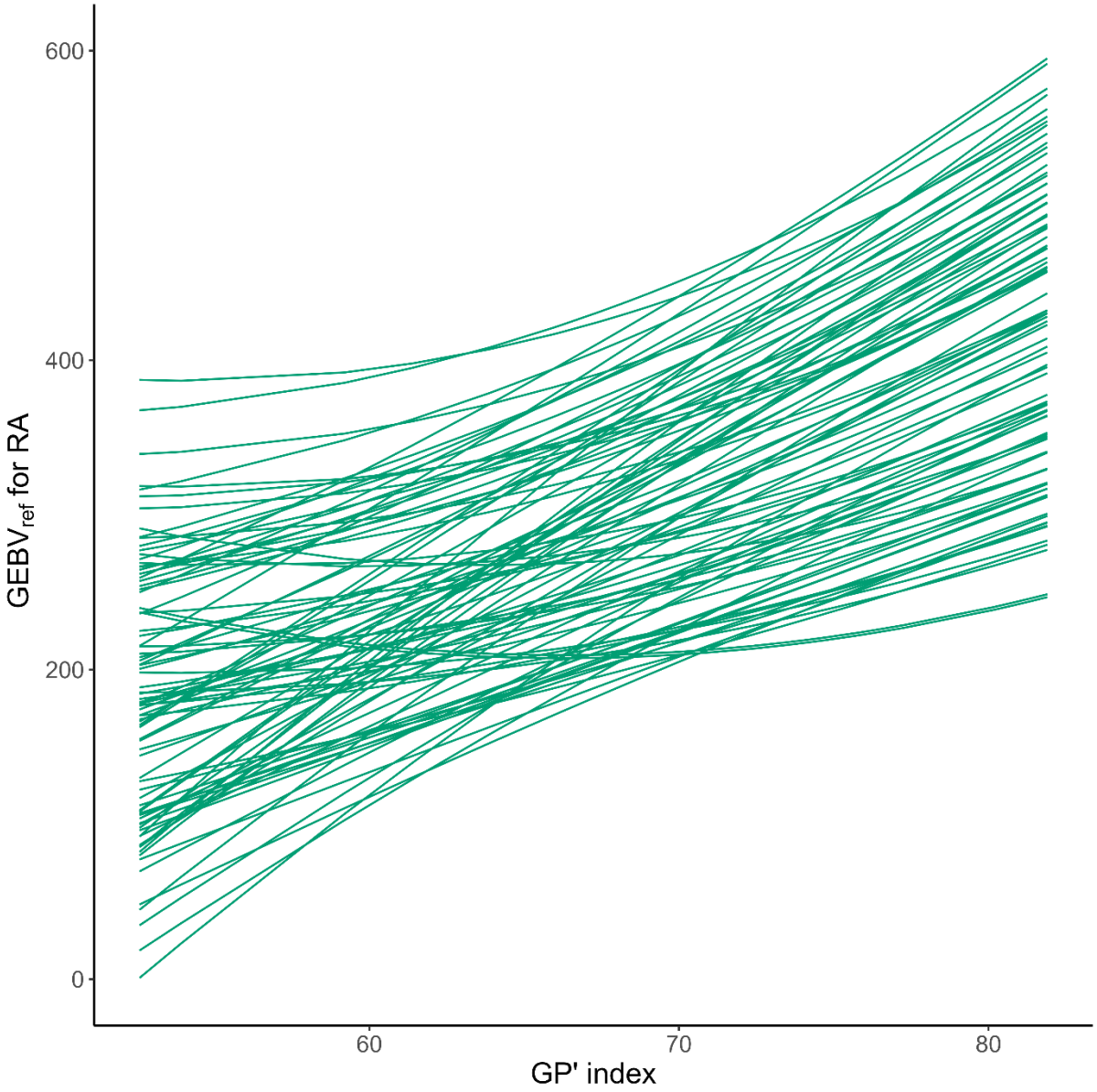



***Figure 4-S5 : Trajectories of individual GEBV$_{ref}$ as a function of annual GP' index estimated by the RRM, only for individuals from cluster C (18.6%)***

# Supplementary 4-1 : Pedigree recovery for POP

## Preliminary stage

The 25 known female parents and the 85 potential male parents (corresponding to two pollen mixtures from 42 and 43 male parents) of POP are grafted in clonal archives and needles were sampled for DNA extraction. Following the same procedure as for POP, an additional genomic relationship matrix (noted $G_{all}$) was computed with the 628 individuals of POP and the 85 parents that successfully passed the quality controls. $G_{all}$ $(713\ x\ 3832)$ was used only for pedigree recovery, as described below.

For the small proportion of parents (1/25 mothers and 24/85 fathers) that did not pass the quality controls, genotyping information was available for a limited number of SNP markers from a previous study (Vidal et al., 2017).

## Criteria for pedigree recovery

A subset of 161 highly discriminant SNP (LD-$r^2$<0.1 and MAF>0.4 in POP) was chosen from the 4TREE array and used for pedigree recovery. The 25 parents not successfully genotyped with the 4TREE array were only characterized for only 21/161 SNP (available from Vidal et al., 2017). For the female and male parents, the two most likely candidates were identified with Cervus 3.0 (Kalinowski et al., 2007; Marshall et al., 1998), as described by Vidal et al. (2017). The candidate considered most likely was examined first, followed by the second most likely candidate if necessary. For validation, a candidate parent had to have (i) fewer than two mismatches with the descendant and (ii) a high level of relatedness ($0.5 \pm 20\%$) to the descendant (relatedness coefficient extracted from $G_{all}$). In the specific case in which information for only 21 SNP was available for the candidate parent, criterion (ii) was replaced with the delta score criterion (estimated with Cervus 3.0 and described by Vidal et al., (2017).

# Supplementary 4-2 : Environmental indices

## Modified de Martonne aridity index

The modified version of the de Martonne index proposed in this study is based on the formula suggested by Botzan et al. (1998):

$$DM'_{y,z} = \alpha\, DM_{y,z} + (1-\alpha) DM_{(y-1),z}$$

Where $DM'_{y,z}$ is the modified de Martonne index for the year $y$ and site $z$; $DM_{y,z}$ and $DM_{(y-1),z}$ are respectively the de Martonne index for the year $y$ and $y-1$; and $\alpha$ is a coefficient defined as below:

$$r = \frac{DM_{y,z} - DM_{(y-1),z}}{DM_{y,z}} \quad , \quad if \begin{cases} r > 0.5 \\ r = 0.2\ to\ 0.5 \\ r < 0.2 \end{cases} \quad then \begin{cases} \alpha = 0.75 \\ \alpha = 0.90 \\ \alpha = 1.00 \end{cases}$$

We tested different threshold values for $r$ classes $\{(0.5-0.2),(0.4-0.2)\}$ and for $\alpha$ classes $\{(0.75-0.90-1.00),(0.50-0.75-1.00),(0.25-0.50-1.00),(0.00-0.25-1.00)\}$. Each time, the gradient of annual de Martonne indices obtained was used in a RRM and the quality of the model was assessed by a cross-validation routine (CV-A). The best model quality was obtained with the annual modified de Martonne index gradient calculated on the sliding window of 30 days with $r$ class = $\{(0.4-0.2)\}$ and $\alpha$ class = $\{(0.25-0.50-1.00)\}$. The final formula used is thus:

$$DM'_{y,z} = \alpha\, DM_{y,z} + (1-\alpha) DM_{(y-1),z} \text{ , with:}$$

$$r = \frac{DM_{y,z} - DM_{(y-1),z}}{DM_{y,z}} \quad , \quad if \begin{cases} r > \mathbf{0.4} \\ r = \mathbf{0.2}\ to\ \mathbf{0.4} \\ r < \mathbf{0.2} \end{cases} \quad then \begin{cases} \alpha = \mathbf{0.25} \\ \alpha = \mathbf{0.50} \\ \alpha = \mathbf{1.00} \end{cases}$$

**Inputs and details for GO+ 3.0 model**

A growth potential index (GP) was calculated for each year $y$ within each site $z$, based on mean trunk water potential ($\varphi_{trunk}$) and temperature ($T_a$) estimated daily by the GO+ v3.0 model (Moreaux et al., 2020):

$$GP_{y,z} = \sum_{i=1}^{365} GP_{\varphi_{tronc_i}} . GP_{T_{a_i}}$$

where $GP_{\varphi_{trunk}}$ and $GP_{T_a}$ are the growth potential components linked to $\varphi_{trunk}$ and $T_a$, respectively. $GP_{\varphi_{trunk}}$ and $GP_{T_a}$ are obtained for each day $i$ with the following response functions:

$$GP_{\varphi_{trunk_i}} = \frac{1}{1 + e^{-\lambda\,(\varphi_{trunk}+c)}} \quad and \quad GP_{T_{a_i}} = Q_{10}^{\frac{T_a - T_{ref}}{10}}$$

with the following parameters for the sigmoid function: $\lambda\ (slope) = 10$ and $c = \frac{\max(\varphi_{trunk}) - \min(\varphi_{trunk})}{2}$ (an additional parameter centering the sigmoid on our range of $\varphi_{trunk}$ values), and the following values for the Q10 function: $Q_{10} = 10$ and $T_{ref} = 29.9\ (°C)$ (maximum temperature $T_a$ over our study period 2005-2019).

The GO+ model was used with maritime pine species parameters (default). Soil parameters used for Site1 and Site2 were respectively typical of wet Landes and dry Landes with low fertility. Real climatic and silvicultural data were added.

The GO+ model produced stand-level average values of microclimate temperature ($T_a$) root water potential ($\varphi_{roots}$) and canopy water potential ($\varphi_{canopy}$), at a daily scale over the period 2005-2019. These last two parameters were combined to estimate a trunk water potential at 1.30m ($\varphi_{trunk}$):

$$\varphi_{trunk} = \varphi_{roots} - \frac{1.3 + 0.7}{Ht_{mean} + 0.7} \left(\frac{\varphi_{roots}}{\varphi_{canpy}}\right)$$

With $Ht_{mean}$ the average stand height for the given day.

Note that the GO+ version used does not incorporate growth data from our stands. The independence between environmental variables and phenotypic trait used in the RRM is therefore guaranteed.

**Modified Growth Potential index**

From the initial GP calculation procedure, different values of slopes $\alpha = \{20, 10, 5, 2\}$ and center of the curve $c = \{c_{ref}\text{-}50\%, c_{ref}\text{-}25\%, c_{ref}, c_{ref}\text{+}25\%, c_{ref}\text{+}50\%\}$ were tested for the sigmoid response function to $\varphi_{trunk}$, as well as different values of $Q_{10}$={2,3,4} were tested for the response function to temperature $T_a$. To get the final growth potential for one year, different types of annual integration of $GP_{\varphi_{tronc_i}} . GP_{T_{a_i}}$ were tested (sum of daily values over a year; average over a sliding window of 2, 5 or 10 days before summing over a year; assignment of the minimum obtained over a sliding window of 2, 5 or 10 days before summing over a year). In addition, modifications to the annual GP values to take into account the impact of previous year were applied in the same way as for the modified de Martonne index. The best model quality (i.e. best predictive performance for CV-A) was obtained using an annual environmental gradient calculated with the initial procedure and initial values (sigmoid function response to $\varphi_{trunk}$ with $\alpha = 10$ and $c = c_{ref}$; Q10 function response to $T_a$ with $Q_{10} = 2$) but using the minimum value of $GP_{\varphi_{tronc_i}} . GP_{T_{a_i}}$ in a sliding window of 10 days before for the annual integration and considering the impact of previous year (modifications with $r\ classes\ \{(0.2 - 0.4)\}\ and\ \alpha\ \{\ (0.25 - 0.50 - 1.00)\}$).

The final formula for the GP' index is thus: $GP'_{y,z} = \alpha\ GP_{y,z} + (1 - \alpha) GP_{(y-1),z}$ , with: (1) and (2)

$$(1): GP_{y,z} = \sum_{i=1}^{365} min\ \left\{ GP_{\varphi_{tronc_{i-10}}} . GP_{T_{a_{i-10}}}, \ldots, GP_{\varphi_{tronc_i}} . GP_{T_{a_i}} \right\}$$

When $i\ \in [1{:}\,10]$, $GP_{\varphi_{tronc_{i-10}}} . GP_{T_{a_{i-10}}}$ are from the previous year (last days of December)

$$(2): r = \frac{GP_{y,z} - GP_{(y-1),z}}{GP_{y,z}}\ ,\quad if\ \begin{cases} r > \mathbf{0.4} \\ r = \mathbf{0.2}\ to\ \mathbf{0.4} \\ r < \mathbf{0.2} \end{cases}\ then\ \begin{cases} \alpha = \mathbf{0.25} \\ \alpha = \mathbf{0.50} \\ \alpha = \mathbf{1.00} \end{cases}$$

## Author contributions

LB and LS: conceptualization, supervision and validation

VP, AB, LB and LS: methodology

VP: data curation, formal analysis, visualization, writing – original draft preparation

VP and AB: software

AB, LB and LS: writing – Review & Editing

## Acknowledgments

The authors would like to thank GIS "Groupe Pin Maritime du Futur" and INRAE - UEFP (https://doi.org/10.15454/1.5483264699193726E12) for the installation of the studied sites, the management of the sites, the help to collect data (circumference measurements) and biological material (needles and increment cores). Authors are thankful to Frederic Lagane (PHENOBOIS Platform) for the cutting and the radiography of the increment cores. Authors thank the Institute of Biosciences and BioResources – Italian National Council of Research (IBBR/CNR), especially Giovanni Giuseppe Vendramin and Sara Pinosio for performing most of DNA extraction and quality monitoring. Part of the experiments (DNA extraction, quantification and manipulation) were also performed at the PGTB (doi:10.15454/1.5572396583599417E12), with the help of Christophe Boury and Céline Lalanne. Raphaël Segura provided soil and climate characterization for the two studied sites. Authors are also thankful to Christophe Plomion for his help during the conceptualization of this study.

## Funding

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme Project under grant agreement n°773383 (B4EST). VP was awarded a doctoral fellowship (N°2020-CK-126) from Ecole Nationale Supérieure Des Sciences Agronomiques de Bordeaux-Aquitaine, 1 cours du Général de Gaulle, CS 40201 33175 Gradignan Cedex.

## Data availability

The data underlying this article are accessible via the private following link: (Data INRAE) https://entrepot.recherche.data.gouv.fr/privateurl.xhtml?token=15f2101e-ebb8-4b7c-838b-9703090cfec4 The corresponding DOI is https://doi.org/10.57745/NUTK1I (Papin, Victor; Bosc, Alexandre; Sanchez-Rodriguez, Leopoldo; Bouffier, Laurent, 2023)

The data will be made public and accessible to all once the article has been accepted.

# References

Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., Kitzberger, T., Rigling, A., Breshears, D. D., Hogg, E. H. (Ted), Gonzalez, P., Fensham, R., Zhang, Z., Castro, J., Demidova, N., Lim, J.-H., Allard, G., Running, S. W., Semerci, A., & Cobb, N. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest Ecology and Management*, *259*(4), 660-684. https://doi.org/10.1016/j.foreco.2009.09.001

Alves, R. S., de Resende, M. D. V., Azevedo, C. F., Silva, F. F. e, Rocha, J. R. do A. S. de C., Nunes, A. C. P., Carneiro, A. P. S., & dos Santos, G. A. (2020). Optimization of Eucalyptus breeding through random regression models allowing for reaction norms in response to environmental gradients. *Tree Genetics & Genomes*, *16*(2), 38. https://doi.org/10.1007/s11295-020-01431-5

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A. F., Resende Jr., M. F. R., & Muñoz, P. R. (2016). AGHmatrix : R package to construct relationship matrices for autotetraploid and diploid species : A blueberry example. *The Plant Genome*, *9*(3), 1-10. https://doi.org/10.3835/plantgenome2016.01.0009

Apiolaza, L. A., & Garrick, D. J. (2001). Analysis of longitudinal data from progeny tests : Some multivariate approaches. *Forest Science*, *47*(2), 129-140. https://doi.org/10.1093/forestscience/47.2.129

Arnal, M., Larroque, H., Leclerc, H., Ducrocq, V., & Robert-Granié, C. (2019). Genetic parameters for first lactation dairy traits in the Alpine and Saanen goat breeds using a random regression test-day model. *Genetics Selection Evolution*, *51*(1), 43. https://doi.org/10.1186/s12711-019-0485-3

Arnold, P. A., Kruuk, L. E. B., & Nicotra, A. B. (2019). How to analyse plant phenotypic plasticity in response to a changing climate. *New Phytologist*, *222*(3), 1235-1241. https://doi.org/10.1111/nph.15656

Arrhenius, S. (1889). Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift für physikalische Chemie*, *4*(1), 226-248.

Baltunis, B. S., Gapare, W. J., & Wu, H. X. (2010). Genetic parameters and genotype by environment interaction in radiata pine for growth and wood quality traits in Australia. *Silvae Genetica*, *59*(1-6), 113-124. https://doi.org/doi:10.1515/sg-2010-0014

Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., & Bouffier, L. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics*, *17*(1), 604. https://doi.org/10.1186/s12864-016-2879-8

Begum, S., Nakaba, S., Yamagishi, Y., Oribe, Y., & Funada, R. (2013). Regulation of cambial activity in relation to environmental conditions : Understanding the role of temperature in wood formation of trees. *Physiologia Plantarum*, *147*(1), 46-54. https://doi.org/10.1111/j.1399-3054.2012.01663.x

Boligon, A. A., Mercadante, M. E. Z., Lôbo, R. B., Baldi, F., & Albuquerque, L. G. (2012). Random regression analyses using B-spline functions to model growth of Nellore cattle. *Animal*, *6*(2), 212-220. https://doi.org/10.1017/S1751731111001534

Botzan, T. M., Mariño, M. A., & Necula, A. I. (1998). Modified de Martonne aridity index : Application to the Napa Basin, California. *Physical Geography*, *19*(1), 55-70. https://doi.org/10.1080/02723646.1998.10642640

Bouffier, L., Charlot, C., Raffin, A., Rozenberg, P., & Kremer, A. (2008). Can wood density be efficiently selected at early stage in maritime pine (*Pinus pinaster* Ait.)? *Annals of Forest Science*, *65*(1), 106-106. https://doi.org/10.1051/forest:2007078

Bouvet, J.-M., Makouanzi, G., Cros, D., & Vigneron, P. (2016). Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping : Prediction accuracy implications. *Heredity*, *116*(2), 146-157. https://doi.org/10.1038/hdy.2015.78

Bradshaw, A. D. (1965). Evolutionary Significance of Phenotypic Plasticity in Plants. In E. W. Caspari & J. M. Thoday (Éds.), *Advances in Genetics* (Vol. 13, p. 115-155). Academic Press. https://doi.org/10.1016/S0065-2660(08)60048-6

Brockerhoff, E. G., Barbaro, L., Castagneyrol, B., Forrester, D. I., Gardiner, B., González-Olabarria, J. R., Lyver, P. O., Meurisse, N., Oxbrough, A., Taki, H., Thompson, I. D., van der Plas, F., & Jactel, H. (2017). Forest biodiversity, ecosystem functioning and the provision of ecosystem services. *Biodiversity and Conservation*, *26*(13), 3005-3035. https://doi.org/10.1007/s10531-017-1453-2

Campbell, M., Walia, H., & Morota, G. (2018). Utilizing random regression models for genomic prediction of a longitudinal trait derived from high-throughput phenotyping. *Plant Direct*, *2*(9), 1-11. https://doi.org/10.1002/pld3.80

Correia, I., Alía, R., Yan, W., David, T., Aguiar, A., & Almeida, M. H. (2010). Genotype × environment interactions in *Pinus pinaster* at age 10 in a multienvironment trial in Portugal : A maximum likelihood approach. *Annals of Forest Science*, *67*(6), 612-612. https://doi.org/10.1051/forest/2010025

Coumou, D., & Rahmstorf, S. (2012). A decade of weather extremes. *Nature Climate Change*, *2*(7), Article 7. https://doi.org/10.1038/nclimate1452

Dalla-Salda, G., Martinez-Meier, A., Cochard, H., & Rozenberg, P. (2009). Variation of wood density and hydraulic properties of Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) clones related to a heat and drought wave in France. *Forest Ecology and Management*, *257*(1), 182-189. https://doi.org/10.1016/j.foreco.2008.08.019

de Boor, C. (1978). *A practical guide to splines* (2nd éd., Vol. 27). New York: Springer Verlag.

de Martonne, E. (1926). Une nouvelle function climatologique : L'indice d'aridité. *Meteorologie*, *2*, 449-459.

Domke, G. M., Oswalt, S. N., Walters, B. F., & Morin, R. S. (2020). Tree planting has the potential to increase carbon sequestration capacity of forests in the United States. *Proceedings of the National Academy of Sciences*, *117*(40), 24649-24651. https://doi.org/10.1073/pnas.2010840117

FAO. (2010). Global forest resources assessment : Main report. *UN Food and Agriculture Organization, Rome*. https://www.fao.org/forest-resources-assessment/past-assessments/fra-2010/en/

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., & El-Kassaby, Y. A. (2016). Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. *G3 Genes|Genomes|Genetics*, *6*(3), 743-753. https://doi.org/10.1534/g3.115.025957

Gengler, N. (1996). Persistency of lactation yields : A review. *Interbull Bulletin*, *12*, 87-96.

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d : R packages to cluster longitudinal data. *Journal of Statistical Software*, *65*(4), 1-34. https://doi.org/10.18637/jss.v065.i04

Grattapaglia, D., & Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genetics & Genomes*, *7*(2), 241-255. https://doi.org/10.1007/s11295-010-0328-4

Guay, R., Gagnon, R., & Morin, H. (1992). A new automatic and interactive tree ring measurement system based on a line scan camera. *The Forestry Chronicle*, *68*(1), 138-141. https://doi.org/10.5558/tfc68138-1

Guilbaud, R., Biselli, C., Buiteveld, J., Cattivelli, L., Copini, P., Dowkiw, A., Esselink, D., Fricano, A., Guerin, V., Jorge, V., & others. (2020). Development of a new tool (4TREE) for adapted genome selection in European tree species. *Proceedings of the Gentree Symposium*. Proceedings of the Gentree Symposium, Avignon, France.

Huynh-Tran, V. H., Gilbert, H., & David, I. (2017). Genetic structured antedependence and random regression models applied to the longitudinal feed conversion ratio in growing Large White pigs. *Journal of Animal Science*, *95*(11), 4752-4763. https://doi.org/10.2527/jas2017.1864

Isik, F. (2014). Genomic selection in forest tree breeding : The concept and an outlook to the future. *New Forests*, *45*(3), 379-401. https://doi.org/10.1007/s11056-014-9422-z

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., Plomion, C., & Bouffier, L. (2016). Genomic selection in maritime pine. *Plant Science*, *242*, 108-119. https://doi.org/10.1016/j.plantsci.2015.08.006

Jamrozik, J., Schaeffer, L. R., & Dekkers, J. C. M. (1997). Genetic evaluation of dairy cattle using test day yields and random regression model. *Journal of Dairy Science*, *80*(6), 1217-1226. https://doi.org/10.3168/jds.S0022-0302(97)76050-8

Jolivet, C., Augusto, L., Trichet, P., & Arrouays, D. (2007). Les sols du massif forestier des Landes de Gascogne : Formation, histoire, propriétés et variabilité spatiale. *Revue forestière française*, *59*(1), 7-30. https://doi.org/10.4267/2042/8480

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, *16*(5), 1099-1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, *11*(1), 137-148. https://doi.org/10.1080/00401706.1969.10490666

Kirkpatrick, M., & Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology*, *27*(4), 429-450. https://doi.org/10.1007/BF00290638

Kirkpatrick, M., Lofsvold, D., & Bulmer, M. (1990). Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*, *124*(4), 979-993. https://doi.org/10.1093/genetics/124.4.979

Lebedev, V. G., Lebedeva, T. N., Chernodubov, A. I., & Shestibratov, K. A. (2020). Genomic selection for forest tree improvement : Methods, achievements and perspectives. *Forests*, *11*(11), 1190. https://doi.org/10.3390/f11111190

Lehner, F., Coats, S., Stocker, T. F., Pendergrass, A. G., Sanderson, B. M., Raible, C. C., & Smerdon, J. E. (2017). Projected drought risk in 1.5°C and 2°C warmer climates. *Geophysical Research Letters*, *44*(14), 7419-7428. https://doi.org/10.1002/2017GL074117

Li, Y., Klápště, J., Telfer, E., Wilcox, P., Graham, N., Macdonald, L., & Dungey, H. S. (2019). Genomic selection for non-key traits in radiata pine when the documented pedigree is corrected using DNA marker information. *BMC Genomics*, *20*(1), 1026. https://doi.org/10.1186/s12864-019-6420-8

Li, Y., Suontama, M., Burdon, R. D., & Dungey, H. S. (2017). Genotype by environment interactions in forest tree breeding : Review of methodology and perspectives on research and application. *Tree Genetics & Genomes*, *13*(3), 60. https://doi.org/10.1007/s11295-017-1144-x

Lindgren, D., Gea, L., & Jefferson, P. A. (1996). Loss of genetic diversity monitored by status number. *Silvae Genetica*, *45*, 52-59.

Ly, D., Huet, S., Gauffreteau, A., Rincent, R., Touzy, G., Mini, A., Jannink, J.-L., Cormier, F., Paux, E., Lafarge, S., Gouis, J. L., & Charmet, G. (2018). Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. *Field Crops Research*, *216*, 32-41. https://doi.org/10.1016/j.fcr.2017.08.020

Marchal, A., Schlichting, C. D., Gobin, R., Balandier, P., Millier, F., Muñoz, F., Pâques, L. E., & Sánchez, L. (2019). Deciphering hybrid larch reaction norms using random regression. *G3 Genes|Genomes|Genetics*, *9*(1), 21-32. https://doi.org/10.1534/g3.118.200697

Marshall, T. C., Slate, J., Kruuk, L. E. B., & Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, *7*(5), 639-655. https://doi.org/10.1046/j.1365-294x.1998.00374.x

Martinez-Meier, A., Sanchez, L., Pastorino, M., Gallo, L., & Rozenberg, P. (2008). What is hot in tree rings ? The wood density of surviving Douglas-firs to the 2003 drought and heat wave. *Forest Ecology and Management*, *256*(4), 837-843. https://doi.org/10.1016/j.foreco.2008.05.041

Meyer, K. (2007). WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *Journal of Zhejiang University SCIENCE B*, *8*(11), 815-821. https://doi.org/10.1631/jzus.2007.B0815

Meyer, K., & Kirkpatrick, M. (2005). Up hill, down dale : Quantitative genetics of curvaceous traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1443-1455. https://doi.org/10.1098/rstb.2005.1681

Momen, M., Campbell, M. T., Walia, H., & Morota, G. (2019). Predicting longitudinal traits derived from high-throughput phenomics in contrasting environments using genomic Legendre polynomials and B-splines. *G3 Genes|Genomes|Genetics*, *9*(10), 3369-3380. https://doi.org/10.1534/g3.119.400346

Moreaux, V., Martel, S., Bosc, A., Picart, D., Achat, D., Moisy, C., Aussenac, R., Chipeaux, C., Bonnefond, J.-M., Figuères, S., Trichet, P., Vezy, R., Badeau, V., Longdoz, B., Granier, A., Roupsard, O., Nicolas, M., Pilegaard, K., Matteucci, G., … Loustau, D. (2020). Energy, water and carbon exchanges in managed forest ecosystems : Description, sensitivity analysis and evaluation of the INRAE GO+ model, version 3.0. *Geoscientific Model Development*, *13*(12), 5973-6009. https://doi.org/10.5194/gmd-13-5973-2020

Mrode, R. A., & Thompson, R. (2005). *Linear models for the prediction of animal breeding values* (2nd ed). CABI Pub.

Mullin, T., Andersson Gull, B., Bastien, J.-C., Beaulieu, J., Burdon, R., Dvorak, W., King, J., Kondo, T., Krakowski, J., Lee, S., Mckeand, S., Pâques, L. E., Russell, J., Skrøppa, T., Stoehr, M., & Yanchuk, A. (2011). Economic Importance, Breeding Objectives and Achievements. In C. Plomion, J. Bousquet, & C. Kole (Éds.), *Genetics, Genomics and Breeding of Conifers* (p. 40-127). Science Publishers and CRC Press: New York.

Muñoz, F., & Sanchez, L. (2020). *breedR: statistical methods for forest genetic resources analysts* [Logiciel]. https://github.com/famuvie/breedR

Oliveira, H. R., Brito, L. F., Lourenco, D. A. L., Silva, F. F., Jamrozik, J., Schaeffer, L. R., & Schenkel, F. S. (2019). Invited review : Advances and applications of random regression models : From quantitative genetics to genomics. *Journal of Dairy Science*, *102*(9), 7664-7683. https://doi.org/10.3168/jds.2019-16265

Pâques, L. E. (Éd.). (2013). *Forest tree breeding in Europe : Current state-of-the-art and perspectives* (Vol. 25). Springer Netherlands. https://doi.org/10.1007/978-94-007-6146-9

Pawson, S. M., Brin, A., Brockerhoff, E. G., Lamb, D., Payn, T. W., Paquette, A., & Parrotta, J. A. (2013). Plantation forests, climate change and biodiversity. *Biodiversity and Conservation*, *22*(5), 1203-1227. https://doi.org/10.1007/s10531-013-0458-8

Payn, T., Carnus, J.-M., Freer-Smith, P., Kimberley, M., Kollert, W., Liu, S., Orazio, C., Rodriguez, L., Silva, L. N., & Wingfield, M. J. (2015). Changes in planted forests and future global implications. *Forest Ecology and Management*, *352*, 57-67. https://doi.org/10.1016/j.foreco.2015.06.021

Peixoto, M. A., Coelho, I. F., Evangelista, J. S. P. C., Alves, R. S., Rocha, J. R. do A. S. de C., Farias, F. J. C., Carvalho, L. P., Teodoro, P. E., & Bhering, L. L. (2020). Reaction norms-based approach applied to optimizing recommendations of cotton genotypes. *Agronomy Journal*, *112*(6), 4613-4623. https://doi.org/10.1002/agj2.20433

Polge, H. (1966). Établissement des courbes de variation de la densité du bois par exploration densitométrique de radiographies d'échantillons prélevés à la tarière sur des arbres vivants : Applications dans les domaines Technologique et Physiologique. *Annales des Sciences Forestières*, *23*(1), 1-206. https://doi.org/10.1051/forest/19660101

R Core Team. (2022). *R: A Language and Environment for Statistical Computing* [Logiciel]. https://www.R-project.org/

R2D2 Consortium, Fugeray-Scarbel, A., Bastien, C., Dupont-Nivet, M., & Lemarié, S. (2021). Why and how to switch to genomic selection : Lessons from plant and animal breeding experience. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.629737

Ramachandran Nair, P. K., Mohan Kumar, B., & Nair, V. D. (2009). Agroforestry as a strategy for carbon sequestration. *Journal of Plant Nutrition and Soil Science*, *172*(1), 10-23. https://doi.org/10.1002/jpln.200800030

Rathgeber, C. B. K., Cuny, H. E., & Fonti, P. (2016). Biological basis of tree-ring formation : A crash course. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.00734

Ray, D., Berlin, M., Alia, R., Sanchez, L., Hynynen, J., González-Martinez, S., & Bastien, C. (2022). Transformative changes in tree breeding for resilient forest restoration. *Frontiers in Forests and Global Change*, *5*. https://doi.org/10.3389/ffgc.2022.1005761

Rocha, J. R. do A. S. de C., Marçal, T. de S., Salvador, F. V., da Silva, A. C., Machado, J. C., & Carneiro, P. C. S. (2018). Genetic insights into elephantgrass persistence for bioenergy purpose. *PLOS ONE*, *13*(9), 1-16. https://doi.org/10.1371/journal.pone.0203818

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3 Genes|Genomes|Genetics*, *6*(9), 2799-2808. https://doi.org/10.1534/g3.116.032888

Sanchez, L., Rozenberg, P., & Bastien, C. (2013). Shifting from growth to adaptive traits and competition : The prospect of improving tree responses to environmental stresses. In *Novel Tree Breeding* (Vol. 24). Instituto Nacional de Investigacion y Tecnologia Agraria y Alimentaria (INIA). https://hal.science/hal-01268435

Sánchez-Vargas, N. M., Sánchez, L., & Rozenberg, P. (2007). Plastic and adaptive response to weather events : A pilot study in a maritime pine tree ring. *Canadian Journal of Forest Research*, *37*(11), 2090-2095. https://doi.org/10.1139/X07-075

Schaeffer, L. R. (2004). Application of random regression models in animal breeding. *Livestock Production Science*, *86*(1), 35-45. https://doi.org/10.1016/S0301-6226(03)00151-9

Schlichting, C., & Pigliucci, M. (1998). *Phenotypic Evolution : A Reaction Norm Perspective*. Sinauer associates.

Schweingruber, F. H. (2007). *Wood Structure and Environment*. Springer.

Shalizi, M. N., & Isik, F. (2019). Genetic parameter estimates and GxE interaction in a large cloned population of *Pinus taeda* L. *Tree Genetics & Genomes*, *15*(3), 46. https://doi.org/10.1007/s11295-019-1352-7

Sillmann, J., & Roeckner, E. (2008). Indices for extreme events in projections of anthropogenic climate change. *Climatic Change*, *86*(1), 83-104. https://doi.org/10.1007/s10584-007-9308-6

Spinoni, J., Vogt, J. V., Naumann, G., Barbosa, P., & Dosio, A. (2018). Will drought events become more frequent and severe in Europe? *International Journal of Climatology*, *38*(4), 1718-1736. https://doi.org/10.1002/joc.5291

Stevens, A., & Ramirez-Lopez, L. (2022). *An introduction to the prospectr package*.

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J.-L., & Sorrells, M. E. (2017). Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *The Plant Genome*, *10*(2), 1-12. https://doi.org/10.3835/plantgenome2016.11.0111

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., & Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biology*, *17*(1), 110. https://doi.org/10.1186/s12870-017-1059-6

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., de Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Townsend Peterson, A., Phillips, O. L., & Williams, S. E. (2004). Extinction risk from climate change. *Nature*, *427*(6970), 145-148. https://doi.org/10.1038/nature02121

van Eeuwijk, F. A., Bustos-Korts, D. V., & Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype × environment interactions? *Crop Science*, *56*(5), 2119-2140. https://doi.org/10.2135/cropsci2015.06.0375

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414-4423. https://doi.org/10.3168/jds.2007-0980

Vidal, M., Plomion, C., Raffin, A., Harvengt, L., & Bouffier, L. (2017). Forward selection in a maritime pine polycross progeny trial using pedigree reconstruction. *Annals of Forest Science*, *74*(1), 21. https://doi.org/10.1007/s13595-016-0596-8

Wang, C., Andersson, B., & Waldmann, P. (2009). Genetic analysis of longitudinal height data using random regression. *Canadian Journal of Forest Research*, *39*(10), 1939-1948. https://doi.org/10.1139/X09-111

Wiens, J. J. (2016). Climate-related local extinctions are already widespread among plant and animal species. *PLOS Biology*, *14*(12), 1-18. https://doi.org/10.1371/journal.pbio.2001104

Zas, R., Sampedro, L., Solla, A., Vivas, M., Lombardero, M. J., Alía, R., & Rozas, V. (2020). Dendroecology in common gardens : Population differentiation and plasticity in resistance, recovery and resilience to extreme drought events in *Pinus pinaster*. *Agricultural and Forest Meteorology*, *291*, 108060. https://doi.org/10.1016/j.agrformet.2020.108060

***Figure 4-11 : Mesure au résistographe***

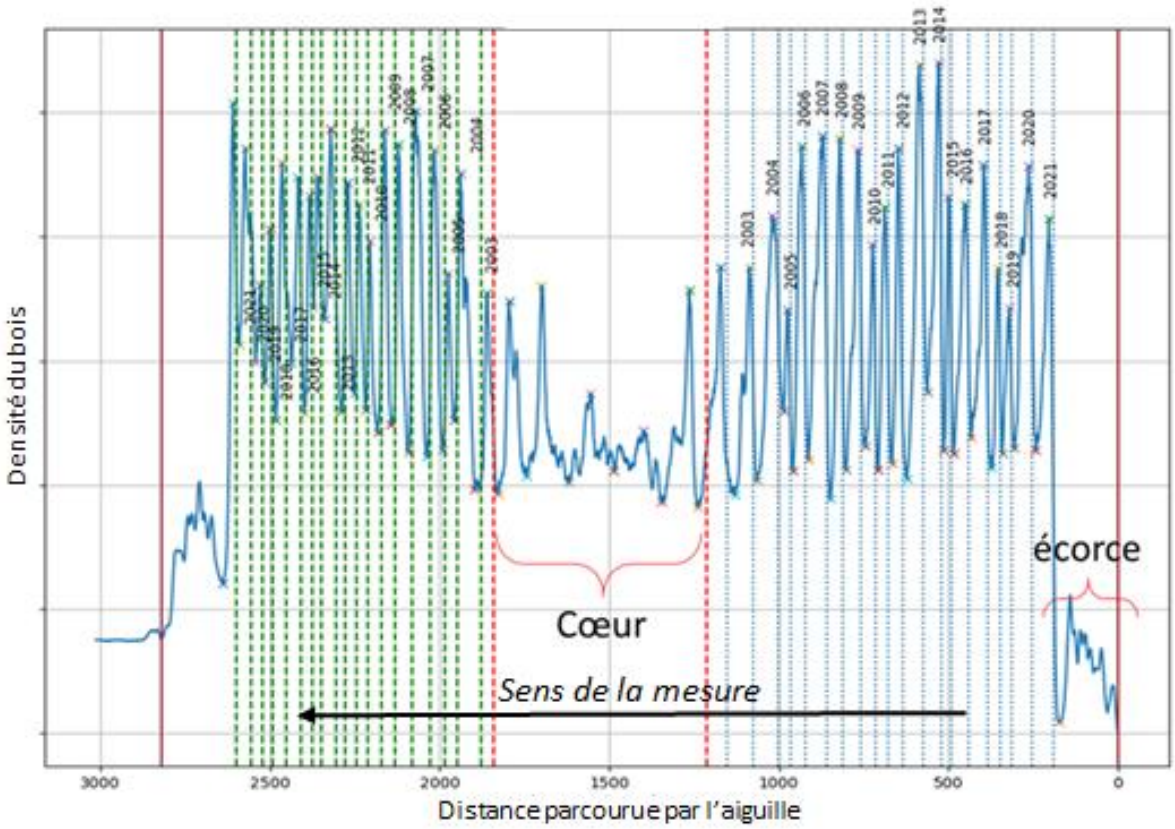



***Figure 4-12 : Profil de densité obtenu avec le résistographe.*** *L'aiguille pénètre le tronc de la droite vers la gauche. La partie située après le cœur a été considéré comme non exploitable.*

## 4.3. Perspectives pratiques pour la construction de normes de réaction

L'approche de densitométrie est considérée à l'heure actuelle comme la méthode de référence pour la caractérisation précise d'un très grand nombre de cernes. Il s'agit cependant d'une procédure lente et couteuse, et donc peu envisageable en routine sur un grand nombre d'individus. Dans un contexte de sélection génomique, d'autres alternatives moins onéreuses et plus rapides sont envisagées dans cette partie afin de faciliter l'intégration des normes de réaction pour l'évaluation génétique.

### 4.3.1. Prédiction génomique à partir de données de résistographe

En plus des données de densitométrie, le scénario de cross-validation CV-B proposé dans l'article 2 s'appuie sur mesures phénotypiques théoriquement réalisables avec un résistographe. Nous allons brièvement expliquer ici les potentialités et les limites de cet outil.

Le résistographe, proposé par Rinn et al.(1996), est classiquement utilisé chez les arbres forestiers pour estimer la densité moyenne du bois du tronc. Cet appareil mesure la résistance à la pénétration d'une aiguille qui traverse le bois à la manière d'une perceuse (**Fig. 4-11**). Les versions les plus récentes de cet outil fournissent désormais un profil précis de l'évolution de la densité au cours de la mesure (**Fig. 4-12**). Ces profils peuvent être analysés de manière similaire à des profils obtenus par densitométrie.
25 à 30 mesures par heure sont réalisables avec le résistographe qui fournit de plus des profils directement exploitables. Par opposition, le prélèvement des carottes de bois se fait à un rythme d'environ 20 carottes par heure. A cela s'ajoute les phases de découpe, de séchage, d'extraction des résines et de radiographie qui s'étalent généralement sur plusieurs jours (**Fig. 2-7**). Si la mesure par résistographe séduit donc par son aspect haut-débit, des études préliminaires sont nécessaires pour évaluer la précision des profils obtenus par cette approche.

En ce sens, des premières analyses ont été réalisées dans le cadre d'un stage de Master 2 effectué par Olivier Le Bourdellès. Des profils de densité ont été obtenus avec le résistographe pour 325 individus du site de Cestas. Ces individus, inclus dans l'échantillonnage réalisé sur le dispositif B, possèdent comme référence un profil obtenu par densitométrie. La corrélation entre surfaces de cernes estimées par résistographe et surfaces de cernes estimées par densitométrie est très variable selon l'année considérée (**Fig. 4-13**). Ces corrélations sont particulièrement

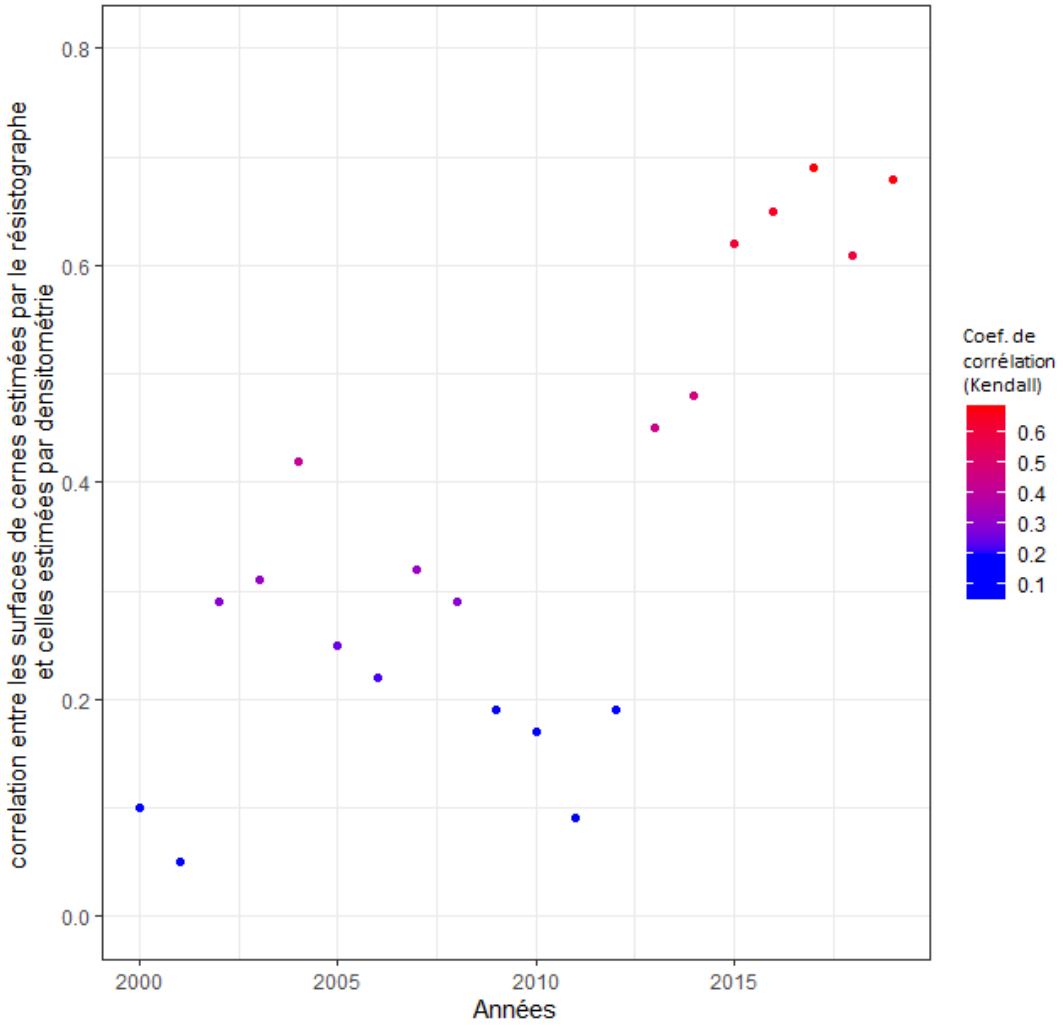



***Figure 4-13 : Corrélation par année entre les surfaces de cernes estimées par résistographe et celles estimées par densitométrie.***

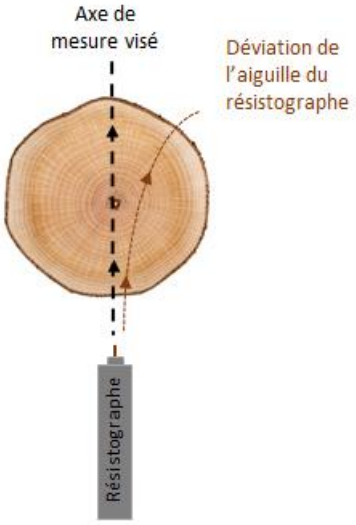



***Figure 4-14 : Déviation de l'aiguille lors de la mesure au résistographe***

faibles pour la période 2000-2014 où elles ne dépassent jamais 0.5. Au-delà d'une routine d'analyse à perfectionner, cela peut s'expliquer par un problème mécanique au niveau de l'aiguille du résistographe. Cette dernière reste difficilement parallèle à l'axe de mesure et a tendance à s'écarter de l'axe de mesure lorsqu'elle se déplace vers l'intérieur, en raison de la grande flexibilité de l'aiguille en acier de l'appareil. Cette déviation affecte la caractérisation des cernes les plus éloignés de l'écorce (**Fig. 4-14**). Les corrélations sont nettement plus élevées pour les 5 dernières années (2015-2019) avec une moyenne à 0.65.

En raison de cette précision limitée, l'utilisation du résistographe pour estimer des surfaces de cernes puis construire des normes de réaction n'est pas encore envisagée. La caractérisation des 5 derniers cernes de croissance par résistographe est plutôt proposée comme un complément des données de densitométrie, comme c'est le cas dans le scénario de cross-validation CV-B présenté dans l'article 2. La distribution alternative de l'effort de phénotypage en intégrant ces mesures de résistographe augmente drastiquement les précisions de prédiction.
Des scénarios dérivés de CV-B ont été proposés afin d'évaluer l'impact de la distribution des environnements associés aux 5 derniers cernes de croissance au sein de la gamme environnementale globale. Dans tous les cas, les niveaux moyens de précision de prédiction étaient équivalents, confirmant que les scénarios proposés sont robustes et applicables quel que soit le contexte environnemental des dernières années passées.

### 4.3.2. Prédiction génomique en intégrant une mesure de circonférence

La circonférence à un âge avancé est mesurée en routine dans les programmes d'amélioration forestiers. Elle est facilement convertible en surface transversale du tronc et apparait alors comme l'intégrale des surfaces de cernes du bois. Cette information a été valorisée dans un nouveau scénario de cross-validation basé sur une démarche en trois étapes :

- Les surfaces de cerne ont été estimées par densitométrie pour les individus de la population de calibration. Pour cette même population, les mesures de circonférence à 12 ans sont converties en surface totale (notées $RA_{tot}$). Ces deux sources d'information ont été utilisées pour estimer les coefficients $a$ et $b$ de la régression linéaire suivante :

$$RA_{y=7} = a * RA_{tot} + b \quad (4.7)$$

où $RA_{y=7}$ désigne la surface du cerne « central », c'est-à-dire le cerne associé au 7$^{e}$ niveau du gradient environnemental.

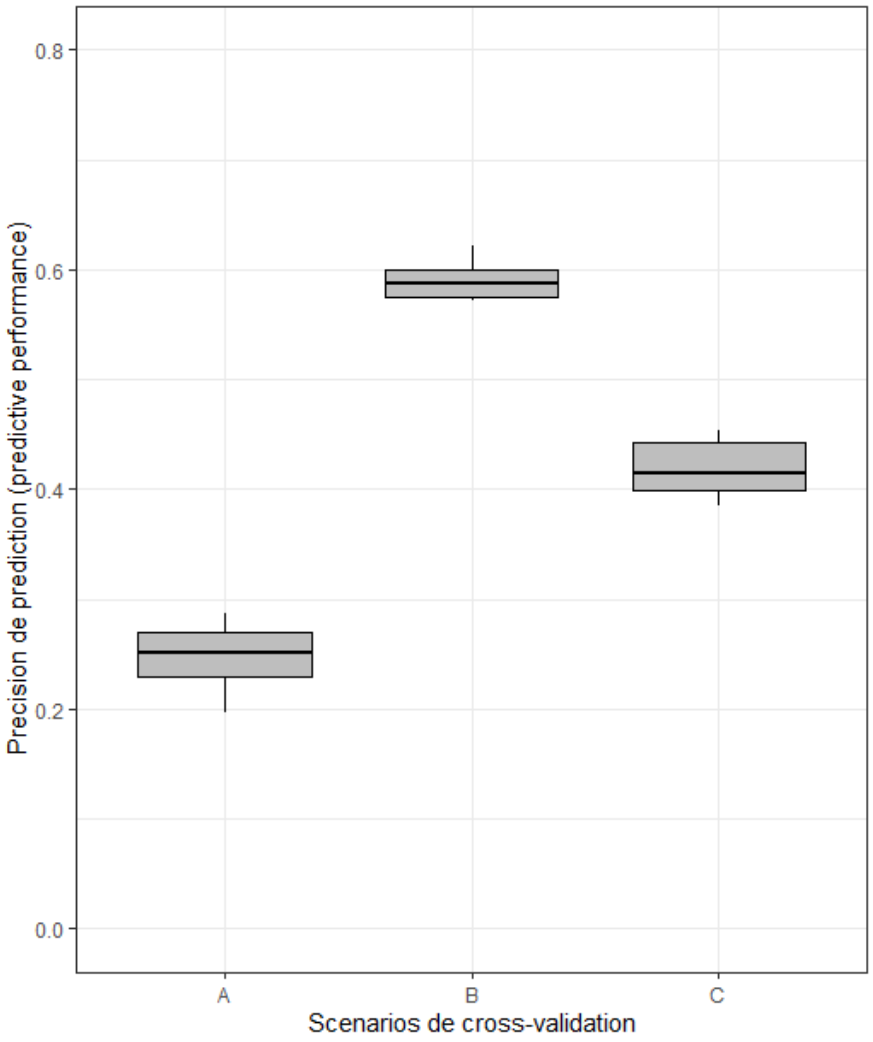



***Figure 4-15 : Précision de prédiction (predictive performance) pour les différents scenarios de cross-validation (10 répétitions par scenario)***

- Les coefficients $a$ et $b$ sont ensuite utilisés pour estimer les valeurs $RA_{y=7}$ des individus de la population de validation. Comme précédemment, cette estimation se base sur la mesure de circonférence convertie en surface totale pour ces individus ($RA_{tot}$). Les individus en validation possèdent désormais une observation de surface cerne.

- Enfin, nous définissions le scénario de cross-validation CV-C suivant :
    - 49% des individus sont inclus dans la calibration avec une information phénotypique complète (15 mesures de surface de cerne obtenues par densitométrie)
    - 51% des individus sont inclus dans la validation avec une seule observation phénotypique (correspondant à la surface de cerne estimée à partir de la circonférence à 12 ans)

  La quantité d'information phénotypique inclus en calibration est identique à celles des scénarios CV-A et CV-B. Par rapport au scénario CV-A, le nombre d'individu intégré dans la calibration avec la totalité de leur information phénotypique est abaissée de 1% car nous ajoutons parallèlement 1% d'information pour les individus en validation.

  La précision de prédiction obtenue avec le scénario CV-C est presque doublée par rapport à celle obtenue avec le scenario CV-A (**Fig. 4-15**). Même si l'estimation d'une surface de cerne est réalisée de manière très grossière pour les individus en validation, elle permet de positionner la croissance des individus au regard de la population. Cela signifie également que la performance des génotypes est relativement facile à prédire car elle est globalement stable dans les différents environnements. De manière générale, la comparaison des scénarios CV-A, CV-B et CV-C montrent que l'intégration de quelques informations phénotypiques pour les individus en validation augmente significativement les précisions de prédiction. Nous avons montré que ces observations « partielles » de surface de cerne peuvent être obtenues à partir de la mesure de circonférence ou grâce au résistographe, ces deux méthodes étant beaucoup moins coûteuses que celles obtenues à partir de données de densitométrie classiques.

## 4.4. Conclusion

L'objectif n°2 visait à démontrer l'intérêt d'une évaluation génétique par des normes de réaction chez les arbres forestiers. Un modèle de régression aléatoire a été utilisé pour analyser des séries de cernes de bois en fonction de différents gradients environnementaux. La précision de prédiction des modèles a été utilisée comme critère pour sélectionner le gradient environnemental le plus pertinent. Il s'agit dans notre étude d'un gradient dérivé d'un modèle éco physiologique, combinant le potentiel hydrique du tronc et la température du microclimat.

Cette évaluation par des normes de réaction a permis de démontrer que même si le classement des génotypes est préservé sur la majeure partie du gradient environnemental, des interactions GxE dites qualitatives ont été détectées dans la partie du gradient la plus défavorable à la croissance. Cette partie du gradient correspond des conditions environnementales susceptibles de devenir de plus en plus fréquentes dans le futur.

La combinaison des informations génomiques et des données longitudinales a également permis de prédire la croissance dans différents environnements. En considérant un effort de phénotypage équivalent, les scénarios cross-validation ont conduit à des précisions de prédiction allant de 0.25 à 0.59, soulignant l'importance de l'allocation des données phénotypiques. Proposée dans cette partie, l'intégration des données de circonférence, obtenues en routine, ou des mesures par résistographe, en plein développement, sont de véritables perspectives faciles à mettre en œuvre dans le cadre de l'amélioration forestière.

Notons enfin que dans ce contexte, les précisions de prédiction obtenues avec les données génomiques étaient systématiquement supérieures à celles obtenues à partir des données de pedigree. La modélisation de données longitudinales permet de démontrer plus facilement le potentiel de la sélection génomique, malgré une structure de population de calibration assez éloignée des recommandations faites dans le chapitre n°1. La combinaison en sélection génomique, d'une part d'une structure de calibration captant la ségrégation mendélienne, et d'autre part de données longitudinales mises au regard d'une information environnementale, laisse présager des perspectives attrayantes pour l'amélioration forestière.

# 5. Discussion et perspectives

## 5.1. Préambule : Comment la sélection génomique peut répondre aux enjeux de l'amélioration forestière ?

L'argument le plus fréquemment mis en avant pour justifier l'intérêt de la sélection génomique chez les arbres forestiers est la réduction de la durée des cycles de sélection (Grattapaglia, 2017; Isik, 2014). En comparaison à l'évaluation conventionnelle réalisée sur des caractères mesurés entre 5 et 15 ans après la plantation, la prédiction précoce des valeurs génétiques de la nouvelle génération (via la génomique ou le pedigree) doit en effet permettre d'identifier les meilleurs individus avec plusieurs années d'avance dans le processus de sélection. Alors que les programmes forestiers les plus avancés atteignent aujourd'hui leur quatrième génération d'amélioration, la grande majorité n'en sont actuellement qu'à leur premier ou deuxième cycle de sélection. Par conséquent, la précision de la sélection génomique n'a été évaluée que sur une seule génération dans la plupart des études forestières (Lebedev et al., 2020), de sorte qu'il n'existe pas encore de vision opérationnelle claire sur la manière de mettre en œuvre un tel changement méthodologique. En outre, la capacité de prédiction au fil des générations devra être maintenue dans le cas de la sélection génomique. Le maintien de cette capacité peut être entravé par les processus de recombinaison génétique, de sélection et de dérive, qui modifient le déséquilibre de liaison et l'apparentement entre la population de calibration et la population en sélection (Grattapaglia, 2022). Si la prédiction multi-génération semble possible chez les arbres forestiers (Bartholomé et al., 2016; Haristoy et al., 2023; Thistlethwaite et al., 2019, Simiqueli et al., 2023), une stratégie de « mise à jour » du modèle de sélection génomique est recommandée en intégrant dans la calibration des individus de chaque génération (Grattapaglia, 2022; Iwata et al., 2011). Ainsi, l'amélioration forestière ne pourra pas supprimer complètement les tests de descendance qui ralentissent le processus de sélection, du moins à moyen terme. La réalisation de nouveaux croisements reste dans tous les cas tributaire de la durée nécessaire à la maturité sexuelle des arbres, soit entre 3 et 10 ans selon les espèces, et des moyens humains et financiers requis pour réaliser ces opérations complexes chez les arbres forestiers.

Au-delà de l'augmentation du gain génétique par unité de temps pour des critères de productivité et de qualité du bois, c'est aujourd'hui la réponse des arbres aux conditions environnementales changeantes qui apparait comme le principal enjeu (Rivers et al., 2023). Celle-ci est pourtant difficile à évaluer avec les critères de sélection actuels. Le système

d'évaluation a été développé sur la base d'une certaine stabilité des conditions environnementales au cours des décennies précédentes, ce qui peut justifier l'utilisation de phénotypes hautement intégratifs qui sont évalués une ou quelques fois seulement, comme la croissance (Li et al., 2017). Par opposition, c'est une caractérisation fine de la réponse des arbres à des variables environnementales précises qui semble nécessaire aujourd'hui. Cette réponse des arbres peut être évaluée en termes de croissance, de qualité du bois ou encore de résistance par rapport aux pathogènes. En concentrant son effort de phénotypage sur une population de calibration de taille limitée, la sélection génomique peut permettre ce changement de paradigme en favorisant l'intégration de nouveaux critères de sélection, plus complexes mais plus adaptés aux nouvelles conditions environnementales. Deux prérequis sont indispensables pour assurer le plein potentiel de la sélection génomique : une précision de prédiction suffisante et une meilleure prise en compte des variables environnementales dans l'évaluation génétique. Cette thèse propose d'explorer ces deux aspects qui sont indispensables pour adapter les méthodologies de la sélection génomique aux nouveaux enjeux. En premier lieu, la prise en compte de la ségrégation mendélienne devrait permettre la pleine efficacité de la sélection génomique, à la fois pour accélérer les gains génétiques et pour permettre une gestion plus fine de la diversité génétique. Puis, ce n'est que sur des modèles génomiques performants où le phénotypage est limitée à une population de calibration que des critères de sélection complexes intégrant une dimension environnementale pourront être valorisés.

## 5.2. La précision de prédiction de la sélection génomique

Le succès d'un programme de sélection génomique repose en grande partie sur la précision des prédictions (Grattapaglia, 2017). Les sélectionneurs forestiers, déçus par l'échec de la SAM (Muranty et al., 2014), affichent une certaine réticence envers l'implémentation cette approche génomique. Pour les en convaincre, les études évaluant sa précision chez les arbres forestiers se sont multipliées ces dernières années amenant à des conclusions très séduisantes. Ma (courte) expérience dans ce domaine contraste cependant quelque peu avec ces messages positifs.

### 5.2.1. Regard critique sur la mise en exergue de la sélection génomique chez les arbres forestiers

#### *Précision de prédiction équivalente au modèle ABLUP*

La précision de la sélection génomique est difficile à appréhender en valeur absolue (Durán et al., 2017; Resende et al., 2012). C'est plutôt une comparaison avec la précision d'une approche de référence qui peut réellement permettre de juger les gains apportés. Le modèle ABLUP est la référence pour l'amélioration forestière, car il s'agit du système actuellement utilisé, qui intègre les données généalogiques disponibles dans les programmes forestiers et permet de prédire les valeurs génétiques d'individus non phénotypés.

Une métrique de référence de la littérature forestière est l'efficacité relative par unité de temps (RE) de la sélection génomique (Chen et al., 2018; Ratcliffe et al., 2015). Elle est définie par :

$$RE = \frac{r(GEBV_{GS}, EBV)}{r(EBV_{TS}, EBV)} \times \frac{T_{TS}}{T_{GS}} \quad (5.1)$$

où $r(GEBV_{GS}, EBV)$ et $r(EBV_{TS}, EBV)$ sont les précisions des modèles génomiques et ABLUP respectivement, et $T_{TS}$ et $T_{GS}$ les durées estimées d'un cycle de sélection conventionnelle (sélection réalisée après phénotypage de tous les candidats) et génomique (sélection réalisée après génotypage des candidats mais sans phénotypage). Par principe, un grand nombre d'études assument une réduction par deux de la durée des cycles avec la sélection génomique, i.e. $T_{GS} = 0.5T_{TS}$ (Lebedev et al., 2020). Selon cet indicateur, la sélection génomique reste avantageuse même avec une précision jusqu'à deux fois inférieure à celle du modèle ABLUP. Comme évoqué en Introduction, si la sélection génomique permet en effet de réduire la durée des cycles par rapport à une sélection conventionnelle basée sur l'observation des phénotypes dans les tests de descendance, le modèle ABLUP permet également de prédire des valeurs

génétiques pour des individus non-phénotypés, ce qui amènerait à une réduction de la durée des cycles de même ampleur.

Dans ces conditions, la précision de la sélection génomique pour l'objectif n°1 a été évaluée en comparaison avec celle d'un modèle ABLUP intégrant une information de pedigree complète et corrigée. Au vu de la littérature forestière, ces conditions peuvent paraitre « défavorables » pour montrer l'avantage de la sélection génomique, mais elles nous semblent nécessaires pour évaluer sans ambiguïté son potentiel. Malgré une densité de marquage plus que doublée par rapport aux précédentes études de sélection génomique chez le pin maritime (Bartholomé et al., 2016; Isik et al., 2016) et un design original de la population de calibration favorisant la dimension intrafamiliale, les résultats de notre étude expérimentale démontrent une équivalence entre les modèles GBLUP et ABLUP en terme de précision de prédiction. De manière générale, en remontant aux estimateurs de la précision de prédiction $r(GEBV_{GS}, EBV)$ et $r(EBV_{TS}, EBV)$, ces deux types de modèles apparaissent équivalents dans une grande majorité d'études chez les arbres forestiers, ce qui à priori ne justifierait pas le passage à la génomique (Beaulieu et al., 2014; Calleja-Rodriguez et al., 2019; Thistlethwaite et al., 2017).

### *Le cas particulier du genre Eucalyptus*

La précision des modèles génomiques apparait supérieure à celles des modèles ABLUP dans quelques études sur les espèces du genre *Eucalyptus*, ce qui légitime dans ces conditions la faisabilité de la sélection génomique. Au sein des arbres forestiers, ces espèces se démarquent par des tailles de génomes très réduites, autour de 640Mb (Myburg et al., 2014), tandis que celles des conifères dépassent par exemple les 20Gb (Chagné et al., 2002). Les densités de marquage apparaissent ainsi très contrastées, autour de 1 SNP/12-20kb pour le genre *Eucalyptus* contre 1 SNP/2000kb pour les conifères. Le déploiement clonal du genre *Eucalyptus* facilite également la qualité du phénotypage et permet d'estimer plus efficacement la composante génétique additive (Grattapaglia, 2022). Si ces éléments peuvent expliquer l'avantage de la sélection génomique pour ces espèces, notons toutefois que seules 4 études sur les 29 recensées en 2023 pour le genre *Eucalyptus* ont réalisé une comparaison claire avec le modèle ABLUP et présentent une précision de prédiction supérieure (Estopa et al., 2023; Kainer et al., 2018; Suontama et al., 2019; Thavamanikumar et al., 2020).

## *Pourquoi ne pas utiliser l'ABLUP dans une configuration similaire à celle du GBLUP ?*

Le modèle ABLUP est utilisé en routine en amélioration forestière pour estimer les valeurs génétiques de l'ensemble des individus en sélection, à partir d'observations de leurs performances propres ou celles de leurs apparentés. Ce modèle n'est cependant pas utilisé en routine selon le principe de la sélection génomique, c'est-à-dire pour prédire les valeurs génétiques d'une nouvelle génération ou d'une partie entière de la population volontairement non-phénotypée. Mais si les précisions du modèle ABLUP sont similaires à celles des modèles génomiques, alors pourquoi ne pas l'utiliser ainsi ? En cause principalement, l'incapacité de ces modèles à prédire la variabilité intra-famille. Rappelons que l'intérêt d'un croisement repose sur la valeur génétique moyenne attendue dans la descendance au vu des valeurs génétiques parentales mais aussi sur la déviation de la descendance par rapport à cette moyenne afin d'identifier des individus « supérieurs » (Zhong & Jannink, 2007). Sans observation phénotypique dans une famille de pleins-frères, le modèle prédit une valeur génétique identique pour l'ensemble des individus et offre uniquement la possibilité d'une sélection aléatoire en intra-famille.

## *Un biais de publication sur la précision intrafamiliale ?*

Notre étude expérimentale de sélection génomique a permis de mettre en évidence une précision intrafamiliale nulle en moyenne avec le modèle GBLUP. La précision globale de ce modèle est donc principalement due à la prédiction des moyennes familiales (Beaulieu et al., 2014). Il s'agit d'un élément clé qui appuie l'idée que l'implémentation de la sélection génomique dans ces conditions n'est pas envisageable car elle conduirait aux mêmes limitations que celles décrites précédemment avec l'utilisation en prédiction d'un modèle ABLUP. Au vu du nombre conséquent d'études de sélection génomique publiées chez les arbres forestiers, il semble surprenant que cette précision nulle en intra-famille n'ait pas été mise en lumière plus tôt, surtout dans un contexte d'équivalence généralisée entre modèles ABLUP et GBLUP. Si la structure de la population étudiée ne permet pas toujours d'évaluer la prédiction intrafamiliale, nous pouvons tout de même envisager un biais de publication à ce niveau. Une précision nulle en intra-famille n'est certes pas le résultat le plus plaisant à publier, mais nous pensons qu'il est nécessaire de le mettre en évidence.

### 5.2.2. Conditions de la supériorité de la sélection génomique

#### *Identifier les facteurs clés de la précision de la sélection génomique par simulation*

Les facteurs clés impactant la précision de la sélection génomique ont été largement décrits dans la littérature et sont rappelés en introduction (Daetwyler et al., 2008; Hayes et al., 2009). Les formules déterministes constituent un moyen efficace pour déterminer cette précision sans avoir à passer par des études expérimentales longues et couteuses à mettre en place. L'étude de Grattapaglia et Resende (2011) est considérée comme fondatrice dans le domaine forestier. Les auteurs prennent en compte les spécificités de l'amélioration forestière pour évaluer la précision de la sélection génomique en fonction de la taille de la population de calibration, du nombre de QTL contrôlant le caractère, de son héritabilité, et de l'étendue du DL entre marqueurs et QTL représentée par différentes densités de marqueurs pour différentes tailles efficaces de population. Toutefois, il est difficile de représenter avec ces formules l'effet de la structure de la population de calibration ou de l'apparentement entre les populations de calibrations et de validation. En outre, la précision de la sélection génomique n'est prédite qu'au niveau global, la précision intra-famille étant encore complexe à prédire (Schopp et al., 2017). En général, il existe encore une certaine réticence à l'égard de leurs prédictions générales, qui contrastent parfois avec les résultats obtenus dans des cas réels. Cette situation a récemment donné lieu à de nouveaux développements qui tentent d'améliorer les performances de ces prédictions générales (Elsen, 2017). Ces différents éléments justifient notre passage à des simulations stochastiques de type allélique afin d'identifier les déterminants de la précision globale et intrafamiliale en sélection génomique dans le cadre de notre design chez le pin maritime.

En accord avec de précédents résultats (Grattapaglia, 2017), la taille de la population de calibration apparait dans notre étude comme le facteur limitant de la précision de la sélection génomique. Les simulations montrent que ce n'est qu'à partir de 40-65 individus par famille et 1600 individus au total dans la calibration que la précision globale du modèle GBLUP devient significativement supérieure à celle du modèle ABLUP. A partir de ces effectifs seuils, la précision intrafamiliale devient positive pour l'ensemble des familles étudiées. De plus, il apparait que les effets de la ségrégation mendélienne sont mieux prédits dans les familles qui présentent le plus de variabilité génétique. En pratique ces effectifs globaux et par famille ne sont encore jamais atteints dans la plupart des études chez les arbres forestiers ce qui peut expliquer de manière générale les précisions limitées des modèles génomiques

(Lebedev et al., 2020). A l'inverse, la précision de la sélection génomique n'augmente que marginalement avec le nombre de marqueurs et semble plutôt avoir atteint un plateau autour de 10 marqueurs/cM. En résumé, notons que bien que les conditions actuelles de l'amélioration forestière ne semblent pas favorables à une implémentation efficace de la sélection génomique, son potentiel pourrait être plus grandement valorisé au moyen d'un effort supplémentaire pour constituer une population de calibration de taille suffisante et d'un regard particulier sur la variabilité génétique dans les croisements réalisés.

### *Pourquoi ne pas avoir fait des simulations en amont de notre étude expérimentale ?*

Si les données échantillonnées sur le dispositif A ont permis de valider les premiers résultats obtenus avec le modèle de simulation, il peut paraitre regrettable de ne pas avoir réaliser des simulations en amont de notre étude expérimentale pour déterminer préalablement les conditions du succès de la sélection génomique. En pratique, des simulations ont bel et bien été réalisées dès le début de la démarche afin de définir un échantillonnage pertinent. Avec cet échantillonnage, les simulations indiquaient une précision supérieure des modèles GBLUP rapport au ABLUP ainsi que des précisions intrafamiliales significativement positives. Mais plusieurs éléments peuvent expliquer pourquoi nous ne retrouvons pas les résultats espérés avec les données réelles. Tout d'abord, les simulations ont été réalisées pour un caractère associé à une héritabilité de 0.33, l'héritabilité moyenne de la hauteur mesurée à 12 ans dans le programme d'amélioration du pin maritime. Cependant, l'héritabilité finalement estimée avec les données échantillonnées sur le dispositif A est de 0.13. Cette forte baisse peut s'expliquer par un phénotypage de mauvaise qualité ou par forte variance environnementale sur le site. Également, effectifs et nombre de marqueurs attendus ont été revus à la baisse suite à une qualité de génotypage insatisfaisante. Contrairement à la caractérisation initialement prévue de 1000 individus avec une densité 9 SNP/cM (potentiel théorique de la puce 4TREE), nous n'avons finalement pu exploiter qu'une densité de 5.7 SNP/cM pour 833 individus. Si à l'avenir les procédures de génotypage et de traitement des résultats peuvent s'affiner, la priorité est plutôt d'augmenter le nombre d'individus génotypés pour assurer la potentialité de la sélection génomique.

## 5.3. Caractérisation phénotypique fine et lien avec des variables environnementales

La sélection génomique apparait particulièrement valorisable quand elle offre la possibilité de travailler sur des caractères plus fins ou plus complexes grâce à un phénotypage limité à la population de calibration. Outre une précision de prédiction suffisante, c'est une intégration de l'information environnementale dans les procédés d'évaluation qui fait défaut aujourd'hui.

### 5.3.1. L'opportunité de la dendroplasticité

La stabilité des génotypes dans différentes conditions environnementales est classiquement déterminée chez les arbres forestiers par l'évaluation de la croissance à un âge avancé dans des dispositifs multi-sites (Li et al., 2017). Si la sélection génomique peut permettre de réduire les contraintes financières et logistiques liées à ce type de dispositif, l'évaluation tardive empêche toute mise en relation avec une information environnementale précise. La performance d'un individu ne peut être associée qu'à une valeur environnementale moyenne caractérisant le site sur toute la période de croissance. Les deux sites inclus dans le dispositif B sont parmi les plus contrastés du massif. Ils présentent toutefois des conditions climatiques similaires au vu de leur faible éloignement géographique (70km). Les différences entre les deux sites résident davantage dans l'humidité et la fertilité de leurs sols, ces derniers étant difficiles à caractériser de manière simple et a posteriori. Ainsi, l'intégration d'une information environnementale dans la procédure d'évaluation multi-sites est complexe, les deux sites étant plutôt analysés comme des répétitions.

Plutôt que d'exploiter des contrastes entre sites, l'évaluation d'une croissance annuelle grâce à la dendroplasticité permet de valoriser les contrastes environnementaux entre années. La gamme environnementale ainsi explorée est nettement plus importante, et la caractérisation fine de la croissance permet une mise en relation avec une information environnementale plus pertinente. Pour l'indice GP', qui intègre par l'intermédiaire du modèle GO+ les différences d'humidité des sols entre les deux sites, les contrastes entre années génèrent une gamme environnementale 8 fois plus étendue que celle obtenue avec ces deux sites pourtant considérés comme « extrêmes ».

Dans notre étude, la caractérisation par densitométrie a été réalisée sur des arbres âgés de 23 ans. Si cela a permis d'évaluer les arbres sur une grande plage d'années, notons que le carottage peut être réalisé plus tôt. Après la période juvénile, une succession de quelques années

contrastées du point de vue environnemental peut suffire pour caractériser la croissance des arbres en intégrant une information environnementale pertinente.

### 5.3.2. Challenge du choix de la variable environnementale

Un challenge majeur pour la modélisation de normes de réaction réside dans le choix de l'indice environnemental. Dans la littérature, les modélisations par régression aléatoire se font principalement au regard d'une échelle temporelle (Apiolaza & Garrick, 2001; Wang et al., 2009). Un indice environnemental pertinent caractérisé par un large gradient peut non seulement répondre à de nouveaux critères d'évaluation, mais aussi fournir de nouvelles informations sur le fonctionnement de l'arbre, combinant ainsi un intérêt prédictif et explicatif. S'il n'existe pas de critère consensus pour juger de la pertinence des indices intégrés et de la modélisation qui en découle, l'évaluation de la précision de prédiction par cross-validation apparait comme un critère de comparaison particulièrement pertinent dans un contexte de sélection génomique (Arnal et al., 2019; Ly et al., 2018; Momen et al., 2019).

Les différents types d'indices proposés dans notre étude intègrent une composante hydrique et une composante de température, deux facteurs clés expliquant la croissance des arbres (Begum et al., 2013; Larson, 1962; Richter, 1976; Schweingruber, 2007). Pour chaque indice, l'augmentation de la disponibilité en eau est perçue comme un élément favorisant la croissance des arbres. La différence entre indices réside dans la façon dont est caractérisée cette ressource en eau, par le relevé des précipitations dans les indices DM et DM', et par la caractérisation du potentiel hydrique du tronc dans les indices GP et GP'.

A l'inverse, la température n'est pas appréhendée de la même manière entre les deux types d'indices. Au regard des indices DM et DM', une augmentation de la température est appréhendée comme un facteur d'aridification du milieu, ce qui est donc perçu comme défavorable à la croissance des arbres (Bréda et al., 2006; Rennenberg et al., 2006). A l'inverse, l'augmentation de la température prise en compte dans les indices de potentiel de croissance (GP et GP') est considérée comme favorable à la croissance car elle accélère de nombreux processus physico-chimiques (Arrhenius, 1889; Farquhar et al., 1980), tels la division et l'élongation cellulaire (Begum et al., 2013).

La caractérisation de la croissance à une échelle annuelle et sa mise en relation avec des indices environnementaux définis sur le même pas de temps présente deux limitations majeures. Tout d'abord, notre modélisation repose sur l'hypothèse que deux années présentant les mêmes aridités moyennes (DM ou DM') seront associées à des niveaux de croissance équivalents. Or une même valeur d'aridité annuelle peut refléter différentes distributions des températures et des précipitations au cours de l'année. Ces différentes distributions peuvent quant à elles induire des niveaux de croissance différents, ce qui invalide notre hypothèse initiale. Les indices GP et GP' ont été proposés pour aller au-delà de cette première limitation. La caractérisation, non plus des variables environnementales, mais plutôt du potentiel de croissance des arbres en fonction de ces variables environnementales permet d'obtenir des indices annuels qui divergent en fonction des variations intra-annuelles. Par exemple, une même période de fortes températures, favorable à la croissance, n'aura pas le même impact sur le potentiel de croissance annuel si elle intervient au printemps, lorsque la ressource en eau est généralement suffisante pour permettre une croissance rapide, que si elle intervient en été alors que le manque d'eau est déjà un facteur limitant de la croissance.

Deuxièmement, la croissance d'une année peut aussi être influencée par les conditions environnementales des années précédentes. La forte aridité de l'année 2012 peut par exemple expliquer la faible surface de cerne moyenne observée en 2013, malgré que 2013 ait été une année peu aride (**Fig. 4-5**). Pour pallier à cette problématique, l'impact interannuel a partiellement été pris en compte par les changements appliqués à l'indice DM pour obtenir l'indice DM'. Le gain obtenu en termes de précision de prédiction avec cet indice DM' confirme l'importance de cette considération interannuelle. Plus précisément, pour cet indice caractérisant l'aridité de l'environnement, ce changement peut capturer l'impact du climat passé sur la teneur en eau du sol les années suivantes. Alors que le modèle GO+ tient compte de manière continue de la capacité hydrique du sol, l'indice GP qui en découle intègre donc déjà cet effet. Cela peut expliquer que seul un gain marginal (en termes de précision de prédiction du modèle) est obtenu par les changements appliqués pour passer de l'indice GP à l'indice GP'. L'existence d'un gain, même faible, entre GP et GP' suggère que les changements appliqués à l'indice GP peuvent capturer assez partiellement un effet du climat passé sur la structure fonctionnelle de l'arbre, par exemple une perte de surface foliaire chargée de capturer la ressource carbone.

Malgré des conceptions éloignées, les différents indices environnementaux donnent des résultats assez proches dans les modélisations de normes de réaction. L'existence d'un effet

« site » significatif suggère cependant que ces indices ne capturent que partiellement les facteurs environnementaux impactant la croissance des arbres. Si les composantes ciblées (température et eau) jouent en effet un rôle important dans la plasticité des cernes, notons que cette dernière peut par exemple aussi être impactée par le niveau de la fertilité des sols, le rayonnement ou encore le niveau de compétition avec les arbres voisins. Cependant, notre étude s'est davantage concentrée sur la nouveauté et la pertinence de la sélection génomique dans le contexte de la prédiction longitudinale des caractères que sur la recherche de l'indice climatique sous-jacent à la variation phénotypique. Notre approche de validation croisée pourrait par la suite être utilisée pour évaluer plus systématiquement divers indices environnementaux.

### 5.3.3. Une sélection complexe dans un contexte dynamique

La modélisation de normes de réaction permet d'estimer les valeurs génétiques de façon continue le long d'un gradient d'environnemental. Dans notre étude, les interactions GxE détectées sont principalement quantitatives, c'est-à-dire que les différences relatives entre individus varient en fonction des environnements mais entrainent peu de reclassement des individus. Dans ces conditions, on peut se questionner sur l'intérêt d'une évaluation basée sur des normes de réaction, qui à première vue aurait induit les mêmes choix de sélection qu'une analyse simple site. Toutefois, la présence d'interactions GxE qualitatives dans les environnements les plus secs (premier tiers du gradient environnemental (indice GP')) ne doit pas être marginalisée. Ces environnements défavorables à la croissance qui amènent un reclassement partiel des individus sont à mêmes de gagner en fréquence et en intensité dans les prochaines décennies (Allen et al., 2015; Pawson et al., 2013). La projection sur les conditions environnementales futures est particulièrement importante pour des espèces pérennes dont les cycles de sélection sont longs. L'objectif général ici était de fournir une méthodologie facilement applicable chez les arbres forestiers pour qu'une information environnementale puisse être intégrée en routine lors des processus d'évaluation.

Lors d'un processus d'évaluation classique, les meilleurs individus sont sélectionnés sur la base de leur valeur génétique estimée dans une « breeding zone » dont les conditions environnementales sont homogènes. Lorsque des normes de réaction sont estimées, le choix des meilleurs individus doit alors se faire à partir des trajectoires des valeurs génétiques qui s'entrecroisent et ne présentent pas la même disparité le long du gradient environnemental. Cette approche dynamique ouvre donc une nouvelle discussion concernant le critère de

sélection à employer. Une première stratégie souvent suggérée pour gérer ces interactions GxE consiste à sélectionner les individus présentant une bonne performance dans un grand nombre d'environnements (Li et al., 2017). Dans un contexte de normes de réaction, les meilleurs individus seraient ceux dont l'aire sous la courbe est maximale (Alves et al., 2020; Peixoto et al., 2020). Souvent associé à la notion de stabilité, ce type de critère s'assimile plutôt à une performance moyenne dans à un environnement global, ce qui ne semble pas pertinent dans un contexte de changement climatique. Une seconde stratégie plus appropriée consisterait à calculer un index de sélection en pondérant la performance des individus en fonction de la probabilité d'occurrence des différents environnements du gradient dans le futur. En ce sens, la gamme environnementale utilisée pour l'évaluation est clé (**Fig. 4-10**). Elle doit inclure, dans la mesure du possible, les conditions environnementales probables dans le futur afin de sélectionner des individus adaptés à ces conditions considérées comme défavorables. Mais aussi, il est nécessaire de conserver une représentation des conditions actuelles, afin d'assurer encore un niveau de croissance des plus élevés dans des conditions considérées comme plus favorables. Ce type d'index de sélection permettrait également de prendre en compte les variances génétiques hétérogènes en fonction des différents environnements qui apparaissent clairement dans notre étude (**Fig. 4-6**).

De manière générale, la prise en compte des prévisions climatiques est essentielle pour l'amélioration des espèces à longue durée de vie. Si la qualité de ces prévisions à court terme est généralement bonne, beaucoup d'incertitudes demeurent sur le moyen et long terme. La définition de l'environnement du futur est d'autant plus complexe qu'il se caractérisera par une augmentation de la fréquence des évènements climatiques extrêmes (Coumou & Rahmstorf, 2012) et que cet environnement évoluera au cours de la vie des peuplements forestiers. Ainsi, le critère de sélection pourrait inclure des aspects de performance mais aussi de stabilité, par exemple au moyen d'un indice combinant le niveau de différenciation de la trajectoire par rapport à la trajectoire moyenne de la population et la variance des points de la trajectoire. Ce 'double' critère pourrait permettre de sélectionner les individus les plus adaptés aux conditions futures et démontrant une certaine résilience face aux variations environnementales.

Une alternative à la quête du génotype idéal réside peut-être dans l'utilisation d'un mélange de génotypes. La combinaison au sein d'une même plantation de génotypes aux potentiels d'adaptation contrastés en fonction des différents types d'environnements pourrait garantir un certain niveau de production finale. Au vu de la longue croissance des arbres, la rusticité des

mélanges pourrait être particulièrement valorisable dans un contexte environnemental très changeant.

### 5.3.4. Normes de réaction et sélection génomique

Pour être pertinente, cette évaluation fine de la croissance et cette mise en relation avec des variables environnementales doit être faite pour l'ensemble des individus en sélection. Cela est rendu possible dans le cadre de la sélection génomique où le phénotypage est ciblé sur la population de calibration. Alors que la précision prédictive du modèle pour le cas de la validation croisée avec 50% des individus sans aucune observation phénotypique était plutôt limitée, la possibilité de distribuer l'effort de phénotypage de manière plus homogène pourrait permettre des améliorations considérables. Dans notre exemple, la réduction du nombre d'individus avec phénotypage complet dans la calibration (i.e. 25%), compensée par le phénotypage à haut débit de quelques cernes pour le reste de la population avec un résistographe, pourrait augmenter la précision prédictive d'un facteur 2.5 sans qu'il soit nécessaire d'augmenter le nombre de points de phénotypage. La réflexion à avoir n'est donc pas uniquement en termes de structure de la population de calibration mais aussi dans la répartition de l'information phénotypique entre ses membres et à la combinaison de méthodes intensives (carottage) versus peu coûteuses (résistographe) qui permettront d'obtenir un rendement maximal.

Notons que l'intégration des normes de réaction dans un contexte de sélection génomique pour l'objectif n°2 n'a pas respecté les recommandations faites dans l'objectif n°1, en termes de taille et de composition de la population de calibration. Une des difficultés en amélioration forestière est de travailler avec les dispositifs existants. Ces derniers, mis en place il y a plusieurs décennies, ne sont pas nécessairement optimisés pour réponde aux questions actuelles. La construction des normes de réaction a été préférablement réalisée sur un dispositif âgé ayant expérimenté un grand nombre d'environnements au cours de son existence, bien que ses effectifs et sa structuration en familles demi-frères ne permettent pas de s'approcher des conditions suggérées dans l'objectif n°1. Si la façon de repenser les dispositifs à l'avenir est discutée plus loin, notons que les précisions de sélection génomique obtenues avec les données longitudinales étaient légèrement supérieures à celles obtenus avec le modèle ABLUP. Nous pouvons faire l'hypothèse que ces gains auraient été encore plus significatifs avec une population de calibration adaptée.

## 5.4. Perspectives dans les programmes d'amélioration forestiers

Les travaux de cette thèse ont permis, d'un côté, de clarifier les conditions de mise en œuvre de la sélection génomique et de la capture de l'effet mendélien (objectif n°1), et d'un autre côté, de proposer une approche d'intégration des données environnementales dans le processus d'évaluation (objectif n°2). L'objectif n°2 était initialement imaginé comme une extension de l'objectif n°1 : la prédiction de la ségrégation mendélienne est un préalable essentiel pour l'implémentation d'une sélection génomique efficace et pour envisager dans un second temps d'ajuster des modèles plus complexes compilant phénotypes longitudinaux, données environnementales et données génomiques. Toutefois, l'intégration de ces données longitudinales, essentielle pour assurer le gain génétique dans un contexte de changement climatique, contribue également à démontrer plus facilement l'avantage de la sélection génomique par rapport aux approches de pedigree. Autrement dit, ces deux objectifs convergent pour justifier la mise en œuvre de la sélection génomique et le coût supplémentaire lié à son génotypage. Si aujourd'hui le passage à la sélection génomique des arbres forestiers semble une perspective certaine, l'implémentation en routine nécessite tout de même une réflexion sur l'organisation des programmes d'amélioration.

### 5.4.1. Faire évoluer la structure des dispositifs

Les études de sélection génomique chez les arbres forestiers s'appuient sur des dispositifs existants et initialement conçus pour l'amélioration conventionnelle. Dans le cadre de la sélection récurrente du pin maritime, deux types de dispositifs sont couramment déployés par le GIS PMF[11]. Les tests de descendance biparentaux, support de l'approche *forward*, permettent de sélectionner les individus au sein du dispositif qui constitueront la nouvelle génération de la population d'amélioration. Ces croisements biparentaux sont privilégiés afin d'assurer le suivi du pedigree au fil de générations. Ils sont toutefois très contraignants d'un point de vu logistique ce qui ne permet de tester qu'un nombre limité de combinaisons parentales chaque année. Les tests de descendance *polycross*, support de l'approche *backward*, permettent d'évaluer l'AGC des candidats à la sélection via la performance de leurs descendants. Les meilleurs géniteurs sont ensuite installés dans les vergers à graines. Ces croisements *polycross* ne sont pas valorisés dans une sélection *forward*, principalement parce que le pedigree des individus évalués est incomplet, le père étant inconnu.

[11] Groupement d'intérêt scientifique Pin Maritime du Futur (cf Encadré 1-1)

L'introduction des marqueurs moléculaires dans les programmes d'amélioration permet d'envisager de nouvelles approches. La stratégie de polymix breeding (Lambeth et al., 2001) se base uniquement sur des tests de descendance *polycross* en incluant une étape supplémentaire de reconstruction du pedigree grâce aux marqueurs. Au sein de ce dispositif, les meilleurs individus sont sélectionnés à la fois pour constituer la nouvelle génération d'amélioration mais aussi pour former la population de production en verges à graines. Cette approche simplifie les étapes de croisement et raccourcit les cycles de sélection, tout en assurant un gain génétique dans les verges à graines équivalent à celui d'une approche de sélection classique *backward* (Vidal et al., 2017).

L'utilisation de dispositifs de *polycross* avec une étape de reconstruction du pedigree permet de tester un grand nombre de combinaisons parentales sans augmenter l'effort en croisements contrôlés, avec un surcoût en temps et économique. Toutefois, le nombre d'individus évalués in fine par combinaison dans ce type de dispositif (i.e. le nombre d'individus par famille de pleins-frères) est généralement très faible. Dans le dispositif B, par exemple, nous avons identifié après reconstruction du pedigree que les 25 familles de demi-frères échantillonnées (i.e. 25 croisements *polycross*) correspondaient à 295 familles de pleins-frères avec en moyenne 1.3 individus par famille. Ainsi, la probabilité de sélectionner dans la descendance d'un croisement un individu supérieur à chacun de ses deux parents est limitée. A l'inverse, les tests de descendance biparentaux permettent d'évaluer un grand nombre d'individus par famille de pleins-frères, mais un nombre plus limité de combinaisons parentales. Sur le dispositif A, 90 familles de pleins-frères étaient évaluées avec 48 individus par famille à la plantation. D'une manière générale, dans un système où le nombre d'individus en évaluation est restreint, l'objectif est de trouver un équilibre entre le nombre de combinaisons parentales testées et le nombre d'individus évalués par combinaison.

Les simulations suggèrent qu'un dispositif idéal pour la sélection génomique regrouperait au moins 40-65 individus par famille de pleins-frères. En se basant par exemple sur le dimensionnement du site de Cestas (6440 arbres au total), cela permettrait de tester entre 161 et 215 combinaisons parentales. Toutefois, une prédiction de la valeur génétique est ici possible pour un grand nombre de pleins-frères sans évaluation phénotypique, conservés par exemple en pépinière. Cela permettrait d'augmenter drastiquement l'intensité de sélection au niveau intrafamilial par rapport à une approche conventionnelle (cf **5.4.4**). Le choix des parents est clé afin de maximiser la variance dans la descendance et donc la probabilité d'obtenir des individus supérieurs, mais aussi pour assurer une bonne précision de prédiction en intra-famille

(cf **partie 3.2**). L'amélioration des performances du modèle de sélection génomique pourrait permettre de diminuer le nombre d'individus évalué par famille au bénéfice d'un nombre plus important de combinaisons parentales testées.
L'équilibre entre le nombre de combinaisons parentales et la taille de la famille peut également dépendre d'autres facteurs. Une factorisation plus élevée pourrait être utile s'il est nécessaire d'évaluer les composantes non-additives plus efficacement. Inversement, un système moins factorisé, par exemple biparental, peut être intéressant s'il est nécessaire d'imposer un régime explicite de croisement, par exemple pour contrôler plus efficacement la diversité dans la descendance. Dans ce sens, les approches de simulation telles que celles utilisées dans cette thèse seront très utiles pour évaluer les avantages d'un système factoriel par rapport à un système biparental.

L'évaluation multi-sites est également remise en question par nos analyses. La variabilité environnementale entre années apparait nettement supérieure à celle observée entre deux sites contrastés au sein du massif landais. Si des contrastes au niveau de la fertilité du sol peuvent par exemple exister, la difficulté à caractériser chaque site et à maintenir les dispositifs au fil des années plaide désormais en faveur d'une évaluation simple site.

### 5.4.2. Vers un phénotypage de précision et haut-débit

Changement climatique et sélection génomique vont bouleverser la façon d'appréhender le phénotypage. Jusqu'ici, les caractères évalués en routine à un âge avancé étaient simples, mesurables facilement sur un grand nombre d'individus. Désormais, l'intégration de l'information environnementale dans les processus d'évaluation impose de caractériser la croissance à une échelle beaucoup plus fine. L'amélioration génétique pourrait évoluer vers un phénotypage de précision pour la population de calibration sur des sites expérimentaux hautement instrumentés, par exemple en permettant des mesures fines et répétées dans le temps. Dans un contexte de changement climatique, nous avons souligné l'importance des normes de réaction, pouvant être construites avec un modèle de régression aléatoire comprenant une matrice G et permettant d'appliquer un critère de sélection combinant performance et stabilité. Outre la caractérisation de la dendroplasticité par densitométrie ou résistographe, le développement du phénotypage haut-débit par drones peut par exemple permettre de suivre finement la croissance en hauteur des arbres chaque année ou le développement de maladies (Liao et al., 2022; Solvin et al., 2020). Dans le cas de la résistance au nématode, ce phénotypage

pourrait être réalisé dans un espace confiné sous serre, ce qui valorise encore plus l'augmentation de l'intensité de sélection permise par la prédiction génomique. En parallèle de la révolution de la sélection génomique, c'est une révolution du phénotypage haut-débit qui s'opère également et qui peut être encore plus valorisable en foresterie au vu de la dimension des arbres et des dispositifs (Bian et al., 2022).

Toutes ces approches génèrent un gros volume de données longitudinales, reconnues pour être plastiques aux conditions environnementales. Les travaux de cette thèse contribuent à proposer une approche pour intégrer les données phénotypiques, environnementales et génomiques qui apparait aujourd'hui comme un challenge majeur pour la sélection.

### 5.4.3. Le génotypage en routine

Si cela a longtemps été un facteur limitant, le génotypage s'est désormais démocratisé pour la plupart des espèces forestières. Les puces à SNP constituent aujourd'hui la meilleure alternative pour générer des données génomiques de bonne qualité, facilement utilisable par les sélectionneurs, et pour un coût faible généralement compris en 20€ et 30€ par échantillon (Grattapaglia, 2022). Développées par des projets publics de grande envergure, ces puces sont aujourd'hui largement accessibles via les entreprises de génotypage ce qui constitue un avantage majeur pour promouvoir l'application pratique de la sélection génomique, à l'instar de ce qui s'est passé pour les espèces animales.
La puce 4TREE intègre 13407 SNP désignés pour le pin maritime, soit une densité théorique d'environ 9 SNP/cM. Au vu des simulations proposées dans l'objectif n°1, cette densité apparait suffisante pour assurer des capacités de prédiction satisfaisantes et un avantage net par rapport aux approches basées sur le pedigree. Comme observé dans d'autres études, l'augmentation du nombre de marqueurs n'apporterait qu'un gain de précision marginal. La précision de prédiction semble atteindre un plateau à partir de 10000-15000 marqueurs pour la plupart des espèces.

Si les outils du génotypage sont aujourd'hui opérationnels, de nouvelles problématiques apparaissent en amont dans les procédures de préparation des échantillons. Au vu de la grande taille des arbres et des dispositifs, le prélèvement de feuilles ou aiguilles s'avère long et complexe. Une extraction systématique d'ADN pourrait être envisagée dès le stade plantule afin limiter ces contraintes de prélèvement par la suite. Toutefois, les procédures de mise en

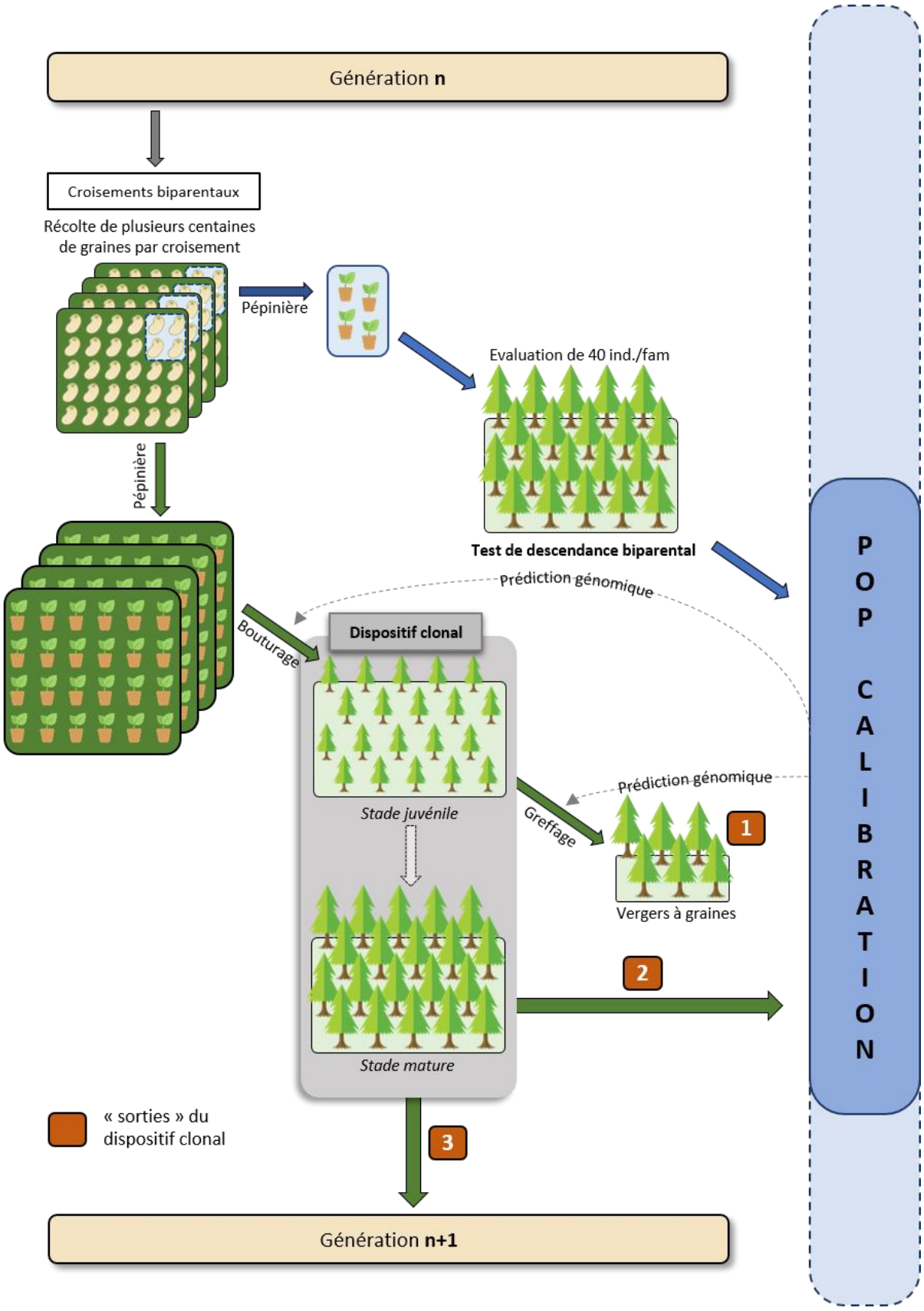



***Figure 5-1 : Exemple de schéma de sélection intégrant la prédiction génomique***

plaque et d'extraction d'ADN sont également fastidieuses et nécessitent du personnel qualifié qui tend à manquer. Le sous traitement de ces étapes par des organismes ou entreprises extérieurs n'est pas encore disponible.

Le génotypage de l'ensemble des individus en sélection et le passage à la sélection génomique chez les arbres forestiers ne se fera pas du jour au lendemain. La population de calibration devra se constituer progressivement avant d'atteindre sa taille et sa structure optimale. La mise à jour de cette population par l'intégration d'individus de chaque génération est de plus essentielle pour assurer le maintien de la capacité prédictive au fil des générations. L'approche single-step GBLUP (ssGBLUP) offre la possibilité d'inclure simultanément les individus génotypés et non-génotypés dans le modèle d'évaluation (Legarra et al., 2009; Misztal et al., 2009). La matrice d'apparentement A ou G peut être remplacée par une matrice H qui intègre à la fois les relations de parenté dérivées du pedigree mais aussi les apparentements génomiques. Cette approche simplifie la procédure d'évaluation tout en améliorant la précision des valeurs génétiques prédites pour la population en sélection. Largement utilisée chez les espèces animales, cette approche pourrait être particulièrement intéressante chez les arbres forestiers dont les tailles de population d'amélioration sont conséquentes.

### 5.4.4. Intégration pratique de la prédiction génomique

La sélection génomique pourrait être valorisée dans un premier temps chez les arbres forestiers afin d'augmenter très significativement l'intensité de sélection (i), notamment en intra-famille. La capacité du pin maritime à produire un très grand nombre de graines par croisement, et sa facilité au bouturage au stade plantule sont deux éléments clés du schéma de sélection génomique proposé ici (**Fig. 5-1**).

Actuellement, seules quelques dizaines d'individus par famille sont installés dans les dispositifs d'évaluation et sont sujets à la sélection. Or, avec environ 100 graines récoltables par cône femelle, il est envisageable de générer des familles avec plusieurs centaines de pleins-frères. Si l'installation en dispositif d'évaluation et le phénotypage de l'ensemble de ces individus n'est pas réalisable au vu des contraintes financières et logistiques imposées par de tels effectifs, la sélection génomique permettrait de prédire une valeur génétique pour l'ensemble des pleins-frères de chaque famille à partir du phénotypage d'un nombre réduit d'entre eux. Les résultats

de cette thèse suggèrent que le phénotypage d'un nombre minimal de 40 individus par famille est tout de même nécessaire pour assurer un niveau de précision suffisant. Si le choix de familles de pleins-frères bien connectées peut faciliter encore plus les prédictions, il convient tout de même de brasser une large diversité dans un contexte de sélection. Le génotypage et le phénotypage de ces familles de plein-frères dans le cadre de dispositifs d'évaluation, alimenteraient la population de calibration. Celle-ci pourrait ensuite être complétée à chaque génération afin de « mettre à jour » le modèle de prédiction génomique.

Conservées jusqu'ici (les graines de conifères se conversent plusieurs années), des graines additionnelles des familles installées dans les dispositifs d'évaluation seraient élevées au stade plantule avant d'être génotypées. La prédiction de leur valeurs génétiques grâce au modèle prédictif construit avec la population de calibration permettrait de réaliser une première étape de pré-sélection avec une intensité très forte au vu de nombre de plantules générées (par exemple, seuls 5% des individus par famille pourraient être retenus). Comme nous l'avons vu lors de l'étude des normes de réaction, si la prédiction est possible pour des individus non-phénotypés, la précision est nettement améliorée lorsqu'une information phénotypique, même partielle, est incluse pour ces individus. Ainsi, les plantules pré-sélectionnées seraient multipliées par bouturage avant d'être installées dans des dispositifs clonaux, regroupant une dizaine de ramets par génotype. A un stade encore juvénile (par exemple 4 à 6 ans), une mesure de circonférence ou une mesure au résistographe permettrait de réaliser une deuxième étape de prédiction génomique, avec une plus grande précision puisque chaque individu à prédire disposerait d'un phénotype propre. Les meilleurs génotypes seraient alors sélectionnés pour être déployés dans les vergers à graines par greffage. Par la suite, le suivi du dispositif clonal permettrait de réaliser une sélection plus large, valorisant la diversité afin de choisir les individus qui constitueront la nouvelle génération d'amélioration mais aussi ceux qui pourraient intégrer la population de calibration avec des mesures au stade mature.

La caractérisation précise des individus au niveau génomique devrait également permettre de gérer plus explicitement la diversité et de mieux orienter les croisements qu'avec une information uniquement basée sur le pedigree. Le choix des parents, réalisé par exemple à partir de leur taux d'hétérozygotie moyen, pourrait permettre de maximiser la ségrégation mendélienne au sein des familles et d'améliorer ainsi la précision de la sélection génomique.

Déjà évoqué précédemment, le dispositif servant à évaluer la population de calibration pourrait prendre la forme d'un site d'évaluation ultra-suivi, tant au niveau de la fréquence et de la précision des mesures phénotypiques que de l'enregistrement des variables environnementales.

L'observation des performances à chaque génération est importante pour valider la précision de la sélection génomique et alimenter régulièrement le modèle de prédiction avec de nouveaux phénotypes. A terme, l'efficacité de la prédiction multi-génération pourrait permettre de réduire le nombre d'individus par famille évalués en dispositif, afin par exemple d'évaluer la descendance d'un nombre de croisements encore plus important.

Bien que cela demande une réflexion sur les compétences techniques nécessaires ainsi que sur l'organisation des opérations, ces propositions sont largement envisageables pour le programme d'amélioration du pin maritime qui disposent de moyens financiers et opérationnels relativement importants. Mais face à la demande en diversification des espèces forestières, on peut s'interroger sur la réalité du déploiement de la sélection génomique dans des programmes d'amélioration aux ressources plus limitées. Dans ces programmes, la sélection génomique pourrait être particulièrement bénéfique afin d'augmenter l'intensité de sélection tout en maintenant un coût faible de phénotypage. La population de calibration pourrait s'apparenter à une 'core collection' de taille limitée mais décrivant largement la diversité de l'espèce. Toutefois, les importants effectifs à génotyper en routine pour obtenir des précisions de prédiction satisfaisantes impose un coût d'entrée élevé qui peut aujourd'hui encore être une limite pour des programmes aux ressources financières limitées. Également, le coût de mise en place de dispositifs adaptés ainsi que le manque d'infrastructures et de personnel pour la préparation des échantillons d'ADN constituent également des entraves importantes. Enfin, ces programmes ne sont pas nécessairement habitués à gérer de gros volumes de données et à gérer des systèmes d'évaluation complexes.

# 6. Conclusion générale

L'objectif de cette thèse était d'étudier la potentialité de la sélection génomique chez les arbres forestiers afin d'intégrer des phénotypes plus complexes en réponse à l'évolution des conditions environnementales. Dans les deux objectifs de la thèse, nous avons souligné les conditions dans lesquelles la sélection génomique pourrait offrir des avantages substantiels par rapport à la sélection conventionnelle, dans la prédiction de traits complexes tels que les normes de réaction et dans un contexte où la sélection intrafamiliale est plus efficace. Bien que cette thèse n'ait pas permis de combiner les deux objectifs de manière empirique, faute de dispositifs adaptés et de temps, nous pouvons émettre l'hypothèse que leur combinaison permettrait de révéler de manière plus claire les avantages potentiels de la sélection génomique. Ainsi, la plus grande précision générée dans un design favorisant l'expression de la variation intrafamiliale faciliterait grandement la prédiction dans des modèles complexes tels que ceux utilisés ici avec les normes de réaction.

Le succès d'un programme de sélection génomique est en premier lieu déterminé par la précision des prédictions. Nous avons montré dans un test de descendance biparental que l'équivalence des précisions entre modèles GBLUP et ABLUP, commune dans les études forestières, se traduisait par une précision intrafamiliale nulle. D'après nos résultats de simulation, l'avantage de la génomique et sa capacité à prédire la variabilité intra-famille se concrétisent avec des populations de calibration intégrant plus de 40 individus par famille de pleins-frères et près de 2000 individus au total. La variabilité génétique dans les familles apparait comme un élément clé conditionnant la précision intrafamiliale.

En plus d'un effort sur l'optimisation de la taille et de la structure des populations de calibration, l'amélioration forestière doit aussi adapter ses critères de sélection. Nous avons montré que l'estimation des croissances annuelles par l'étude de la dendroplasticité pouvait être mise en relation avec une information environnementale pertinente dans les procédés d'évaluation. La modélisation de normes de réaction génomiques permet de comprendre et de valoriser la plasticité individuelle dans un contexte environnemental changeant.

Les caractères intégrateurs, assimilables à des « boites noires », ont permis de réaliser des gains génétiques importants dans un contexte environnemental stable au cours des dernières décennies. Désormais, l'évolution rapide des conditions environnementales impose de revoir les conditions de l'évaluation génétique. Face à ce défi, de nouveaux outils sont aujourd'hui disponibles pour le phénotypage haut-débit, pour la caractérisation de l'environnement ou

encore pour le génotypage. Ils constituent des clés pour mieux comprendre le fonctionnement de l'arbre, sa réponse à différents paramètres environnementaux, et pour proposer des méthodes et critères de sélection adaptés aux enjeux actuels. Le besoin de compréhension et d'intégration de ces données multiples positionne aujourd'hui le sélectionneur à la croisée de nombreuses disciplines. Nous avons par exemple montré que la mise en relation entre les phénotypes étudiés et les variables environnementales pouvait s'appuyer sur des connaissances en éco-physiologie lors de la construction des normes de réaction. L'amélioration génétique peut ainsi profiter des apports de disciplines extérieures dans des contextes variés. Les données de pédologie et la caractérisation du système racinaire peuvent aider à comprendre les potentiels de captation des ressources. Les connaissances en entomologie, pathologie et même chimie doivent permettre de caractériser finement les interactions avec les bio-agresseurs et d'identifier des mécanismes de défense. La génétique des populations informe sur la structuration des espèces et facilite l'introduction de nouvelles provenances dans les populations d'amélioration. En somme, il devient impératif que l'amélioration génétique, qui a longtemps évolué de manière isolée, s'intègre dans une approche multidisciplinaire afin d'assurer la pérennité et la productivité des forêts de plantation dans un contexte environnemental changeant.

# Références

AAF (2014a) Académie d'Agriculture de France. (Y. Birot, Ed). La Forêt et le Bois en France en 100 Questions : Vol. Quelles sont les principales industries liées à la forêt ? https://www.academie-foret-bois.fr/

AAF (2014b) Académie d'Agriculture de France. (Y. Birot & B. Romand-Amat, Éds.) La Forêt et le Bois en France en 100 Questions : Vol. Quelle place pour le secteur forêt-bois dans l'économie nationale ? https://www.academie-foret-bois.fr/

ADS (2023) Académie des sciences. Rapport du Comité des sciences de l'environnement de l'Académie des sciences et points de vue d'Académiciens de l'Académie d'Agriculture de France - juin 2023. https://www.academie-sciences.fr/pdf/rapport/060623_foret.pdf

Agreste (2021) Forêt-Bois DRAAF Nouvelle-Aquitaine. La filière forêt-bois Nouvelle-Aquitaine – Mémento 2021. https://draaf.nouvelle-aquitaine.agriculture.gouv.fr/foret-bois-r247.html

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Akdemir, D., & Sánchez, J. I. (2016). Efficient Breeding by Genomic Mating. *Frontiers in Genetics*, *7*. https://www.frontiersin.org/articles/10.3389/fgene.2016.00210

Allen, C. D., Breshears, D. D., & McDowell, N. G. (2015). On underestimation of global vulnerability to tree mortality and forest die-off from hotter drought in the Anthropocene. *Ecosphere*, *6*(8), art129. https://doi.org/10.1890/ES15-00203.1

Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., & Teyssèdre, S. (2019). Improving short- and long-term genetic gain by accounting for within-family variance in optimal cross-selection. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01006

Alves, R. S., Resende, M. D. V. de, Rocha, J. R. do A. S. de C., Peixoto, M. A., Teodoro, P. E., Silva, F. F. e, Bhering, L. L., & Santos, G. A. dos. (2020). Quantifying individual variation in reaction norms using random regression models fitted through Legendre polynomials : Application in eucalyptus breeding. *Bragantia*, *79*, 485-501. https://doi.org/10.1590/1678-4499.20200125

Apiolaza, L. A., & Garrick, D. J. (2001). Analysis of longitudinal data from progeny tests : Some multivariate approaches. *Forest Science*, *47*(2), 129-140. https://doi.org/10.1093/forestscience/47.2.129

Arnal, M., Larroque, H., Leclerc, H., Ducrocq, V., & Robert-Granié, C. (2019). Genetic parameters for first lactation dairy traits in the Alpine and Saanen goat breeds using a random regression test-day model. *Genetics Selection Evolution*, *51*(1), 43. https://doi.org/10.1186/s12711-019-0485-3

Arrhenius, S. (1889). Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift für physikalische Chemie*, *4*(1), 226-248.

Aspinwall, M. J., Loik, M. E., Resco De Dios, V., Tjoelker, M. G., Payton, P. R., & Tissue, D. T. (2015). Utilizing intraspecific variation in phenotypic plasticity to bolster agricultural and forest productivity under climate change. *Plant, Cell & Environment*, *38*(9), 1752-1764. https://doi.org/10.1111/pce.12424

Ballesta, P., Maldonado, C., Pérez-Rodríguez, P., & Mora, F. (2019). SNP and Haplotype-Based Genomic Selection of Quantitative Traits in Eucalyptus globulus. *Plants*, *8*(9). https://doi.org/10.3390/plants8090331

Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., & Bouffier, L. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics*, *17*(1), 604. https://doi.org/10.1186/s12864-016-2879-8

Beaulieu, J., Doerksen, T., Clément, S., MacKay, J., & Bousquet, J. (2014). Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity*, *113*(4), Article 4. https://doi.org/10.1038/hdy.2014.36

Beaulieu, J., Doerksen, T. K., MacKay, J., Rainville, A., & Bousquet, J. (2014). Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genomics*, *15*(1), 1048. https://doi.org/10.1186/1471-2164-15-1048

Beaulieu, J., Lenz, P., & Bousquet, J. (2022). Metadata analysis indicates biased estimation of genetic parameters and gains using conventional pedigree information instead of genomic-based approaches in tree breeding. *Scientific Reports*, *12*, 3933. https://doi.org/10.1038/s41598-022-06681-y

Beaulieu, J., Nadeau, S., Ding, C., Celedon, J. M., Azaiez, A., Ritland, C., Laverdière, J.-P., Deslauriers, M., Adams, G., Fullarton, M., Bohlmann, J., Lenz, P., & Bousquet, J. (2020). Genomic selection for resistance to spruce budworm in white spruce and relationships with growth and wood quality traits. *Evolutionary Applications*, *13*(10), 2704-2722. https://doi.org/10.1111/eva.13076

Begum, S., Nakaba, S., Yamagishi, Y., Oribe, Y., & Funada, R. (2013). Regulation of cambial activity in relation to environmental conditions : Understanding the role of temperature in wood formation of trees. *Physiologia Plantarum*, *147*(1), 46-54. https://doi.org/10.1111/j.1399-3054.2012.01663.x

Bian, L., Zhang, H., Ge, Y., Čepl, J., Stejskal, J., & EL-Kassaby, Y. A. (2022). Closing the gap between phenotyping and genotyping : Review of advanced, image-based phenotyping technologies in forestry. *Annals of Forest Science*, *79*(1), 22. https://doi.org/10.1186/s13595-022-01143-x

Bosc, A. (2011). Climator - Simuler les impacts du changement climatique pour mieux prévoir la vulnérabilité des forêts au changement climatique. Que nous apprend la Recherche sur la vulnérabilité des forêts au changement climatique ?, ffhal-02802128f

Bouffier, L., Klápště, J., Suontama, M., Dungey, H. S., & Mullin, T. J. (2019). Evaluation of forest tree breeding strategies based on partial pedigree reconstruction through simulations : Pinus pinaster and Eucalyptus nitens as case studies. *Canadian Journal of Forest Research*, *49*(12), 1504-1515. https://doi.org/10.1139/cjfr-2019-0145

Bouffier, L., Raffin, A., & Kremer, A. (2008). Evolution of genetic variation for selected traits in successive breeding populations of maritime pine. *Heredity*, *101*(2), Article 2. https://doi.org/10.1038/hdy.2008.41

Bouvet, J.-M., Couteau, N., & Vigneron, P. (1992). *Premiers éléments de l'analyse des plans factoriels du schéma de sélection récurrente réciproque de l'Eucalyptus au Congo* (Congo) [Conference_item]. Production de variétés génétiquement améliorées d'espèces forestières à croissance rapide : Actes, Bordeaux, France, 14-18 septembre 1992; AFOCEL. https://agritrop.cirad.fr/465610/

Bradshaw, A. D. (1965). Evolutionary Significance of Phenotypic Plasticity in Plants. In E. W. Caspari & J. M. Thoday (Éds.), *Advances in Genetics* (Vol. 13, p. 115-155). Academic Press. https://doi.org/10.1016/S0065-2660(08)60048-6

Bradshaw, H. D., Jr, & Stettler, R. F. (1995). Molecular genetics of growth and development in populus. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. Genetics, 139(2), 963-973. https://doi.org/10.1093/genetics/139.2.963

Bréda, N., Huc, R., Granier, A., & Dreyer, E. (2006). Temperate forest trees and stands under severe drought : A review of ecophysiological responses, adaptation processes and long-term consequences. *Annals of Forest Science*, *63*(6), 625-644. https://doi.org/10.1051/forest:2006042

Calleja-Rodriguez, A., Pan, J., Funda, T., Chen, Z.-Q., Baison, J., Isik, F., Abrahamsson, S., & Wu, H. X. (2019). *Genomic prediction accuracies and abilities for growth and wood quality traits of Scots pine, using genotyping-by-sequencing (GBS) data* (p. 607648). bioRxiv. https://doi.org/10.1101/607648

Campbell, M., Walia, H., & Morota, G. (2018). Utilizing random regression models for genomic prediction of a longitudinal trait derived from high-throughput phenotyping. *Plant Direct*, *2*(9), 1-11. https://doi.org/10.1002/pld3.80

Campoy, J. A., Ruiz, D., & Egea, J. (2011). Dormancy in temperate fruit trees in a global warming context : A review. *Scientia Horticulturae*, *130*(2), 357-372. https://doi.org/10.1016/j.scienta.2011.07.011

Candau, J. N. (2008). Impacts du changement climatique sur les insectes ravageurs des forêts méditerranéennes. *Forêt Méditerranéenne*, *XXIX*(2), 145-154.

Caudullo, G., Welk, E., & San-Miguel-Ayanz, J. (2017). Chorological maps for the main European woody species. Data in Brief, 12, 662-666. https://doi.org/10.1016/j.dib.2017.05.007

Chagné, D., Lalanne, C., Madur, D., Kumar, S., Frigério, J.-M., Krier, C., Decroocq, S., Savouré, A., Bou-Dagher-Kharrat, M., Bertocchi, E., Brach, J., & Plomion, C. (2002). A high density genetic map of maritime pine based on AFLPs. *Annals of Forest Science*, *59*(5-6), 627-636. https://doi.org/10.1051/forest:2002048

Chancerel, E., Lamy, J.-B., Lesur, I., Noirot, C., Klopp, C., Ehrenmann, F., Boury, C., Provost, G. L., Label, P., Lalanne, C., Léger, V., Salin, F., Gion, J.-M., & Plomion, C. (2013). High-density linkage mapping in a pine tree reveals a genomic region associated with

inbreeding depression and provides clues to the extent and distribution of meiotic recombination. *BMC Biology*, *11*(1), 50. https://doi.org/10.1186/1741-7007-11-50

Chen, Z.-Q., Baison, J., Pan, J., Karlsson, B., Andersson, B., Westin, J., García-Gil, M. R., & Wu, H. X. (2018). Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. *BMC Genomics*, *19*(1), 946. https://doi.org/10.1186/s12864-018-5256-y

Chen, Z.-Q., Baison, J., Pan, J., Westin, J., Gil, M. R. G., & Wu, H. X. (2019). Increased Prediction Ability in Norway Spruce Trials Using a Marker X Environment Interaction and Non-Additive Genomic Selection Model. *Journal of Heredity*, *110*(7), 830-843. https://doi.org/10.1093/jhered/esz061

Citepa, (2023). Gaz à effet de serre et polluants atmosphériques. Bilan des émissions en France de 1990 à 2022. Rapport Secten éd. 2023. https://www.citepa.org/wp-content/uploads/publications/secten/2023/Citepa_Secten_ed2023_v1.pdf

Comstock RE, Moll RH (1963) Genotype x Environment Interactions. Stat Genet Plant Breed

Correia, I., Alía, R., Yan, W., David, T., Aguiar, A., & Almeida, M. H. (2010). Genotype × environment interactions in *Pinus pinaster* at age 10 in a multienvironment trial in Portugal : A maximum likelihood approach. *Annals of Forest Science*, *67*(6), 612-612. https://doi.org/10.1051/forest/2010025

Cotterill, P. P., Dean, C. A., & Van Wyk, G. (1987). *Additive and dominance genetic effects in Pinus pinaster, P. radiata and P. elliottii and some implications for breeding strategy*. https://publications.csiro.au/rpr/pub?list=BRO&pid=procite:7bf244fa-e4a1-4f7a-84eb-aff5f1777dd8

Coumou, D., & Rahmstorf, S. (2012). A decade of weather extremes. *Nature Climate Change*, *2*(7), 491-496. https://doi.org/10.1038/nclimate1452

Cros, D., Mbo-Nkoulou, L., Bell, J. M., Oum, J., Masson, A., Soumahoro, M., Tran, D. M., Achour, Z., Le Guen, V., & Clement-Demange, A. (2019). Within-family genomic selection in rubber tree (Hevea brasiliensis) increases genetic gain for rubber production. *Industrial Crops and Products*, *138*, 111464. https://doi.org/10.1016/j.indcrop.2019.111464

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., Campos, G. de los, Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic Selection in Plant Breeding : Methods, Models, and Perspectives. *Trends in Plant Science*, *22*(11), 961-975. https://doi.org/10.1016/j.tplants.2017.08.011

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PLOS ONE*, *3*(10), e3395. https://doi.org/10.1371/journal.pone.0003395

Dalla-Salda, G., Martinez-Meier, A., Cochard, H., & Rozenberg, P. (2009). Variation of wood density and hydraulic properties of Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco)

clones related to a heat and drought wave in France. *Forest Ecology and Management*, *257*(1), 182-189. https://doi.org/10.1016/j.foreco.2008.08.019

Dehli Vigeland, M. (2022). *pedtools : Creating and Working with Pedigrees and Marker Data* (R package version 1.3.0) [Logiciel]. https://github.com/magnusdv/pedtools

de Martonne, E. (1926). Une nouvelle function climatologique : L'indice d'aridité. *Meteorologie*, *2*, 449-459.

Devey, M. E., Carson, S. D., Nolan, M. F., Matheson, A. C., Te Riini, C., & Hohepa, J. (2004). QTL associations for density and diameter in Pinus radiata and the potential for marker-aided selection. Theoretical and Applied Genetics, 108(3), 516-524. https://doi.org/10.1007/s00122-003-1446-2

Doerksen, T. K., & Herbinger, C. M. (2010). Impact of reconstructed pedigrees on progeny-test breeding values in red spruce. *Tree Genetics & Genomes*, *6*(4), 591-600. https://doi.org/10.1007/s11295-010-0274-1

Domec, J., & Gartner, B. L. (2002). How do water transport and water storage differ in coniferous earlywood and latewood? *Journal of Experimental Botany*, *53*(379), 2369-2379. https://doi.org/10.1093/jxb/erf100

Durán, R., Isik, F., Zapata-Valenzuela, J., Balocchi, C., & Valenzuela, S. (2017). Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile. *Tree Genetics & Genomes*, *13*(4), 74. https://doi.org/10.1007/s11295-017-1158-4

Durel, C.-E. (1992). Gains génétiques attendus après sélection sur index en seconde génération d'amélioration du Pin maritime. *Revue forestière française*, *44*(4), 341-355. https://doi.org/10.4267/2042/26331

DSF Aquitaine (2010). Avertissement santé des forêts n°10-05. Suite de la tempête de janvier 2009. Exploitation curative des peuplements scolytés durant l'hiver 2010-2011. DRAAF Aquitaine. Pôle santé des forêts Aquitaine - Midi-Pyrénées. 27 octobre 2010.

Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C., & Neale, D. B. (2010). Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (Pinus taeda L., Pinaceae). *Genetics*, *185*(3), 969-982. https://doi.org/10.1534/genetics.110.115543

El-Dien, O. G., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., & El-Kassaby, Y. A. (2018). Multienvironment genomic variance decomposition analysis of open-pollinated Interior spruce (Picea glauca x engelmannii). *Molecular Breeding*, *38*(3), 26. https://doi.org/10.1007/s11032-018-0784-3

Elsen, J.-M. (2017). An analytical framework to derive the expected precision of genomic selection. *Genetics Selection Evolution*, *49*(1), 95. https://doi.org/10.1186/s12711-017-0366-6

Estopa, R. A., Paludeto, J. G. Z., Müller, B. S. F., de Oliveira, R. A., Azevedo, C. F., de Resende, M. D. V., Tambarussi, E. V., & Grattapaglia, D. (2023). Genomic prediction of growth and wood quality traits in Eucalyptus benthamii using different

genomic models and variable SNP genotyping density. *New Forests, 54*(2), 343-362. https://doi.org/10.1007/s11056-022-09924-y

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Longman, Harlow.

FAO (2020) Food and Agriculture Organization of the United Nations. Évaluation des ressources forestières mondiales 2020 (FRA 2020) https://www.fao.org/3/CA8753FR/CA8753FR.pdf#page=7

FAO (2022) Food and Agriculture Organization of the United Nations. FAOSTAT : Forestry Production and Trade. https://www.fao.org/faostat/en/#data/FO

Farquhar, G. D., von Caemmerer, S., & Berry, J. A. (1980). A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species. *Planta, 149*(1), 78-90. https://doi.org/10.1007/BF00386231

FBF (2019) France Bois Forêts. La forêt des Landes de Gascogne. https://franceboisforet.fr/2019/01/28/la-foret-des-landes-de-gascogne/

FIBOIS (2022) FIBOIS Nouvelle-Aquitaine. Le pin maritime : Ressource, renouvellement et problématiques de l'essence. https://fibois-na.fr/wp-content/uploads/2021/09/Fiche-essence-Pin-maritime-Fibois-Landes-de-Gascogne-2021.pdf

Fisher, R. A. (1918). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh, 52*(2), 399-433. https://doi.org/10.1017/S0080456800012163

Fritz, S., Druet, T., Guillaume, F., Malafosse, A., Boscher, M. Y., Eggen, A., Gautier, M., Colleau, J. J., & Boichard, D. (2007). Assessment of Marker-Assisted Selection in the three main French breeds of dairy cattle and future developments.

Fuentes-Utrilla, P., Goswami, C., Cottrell, J. E., Pong-Wong, R., Law, A., A'Hara, S. W., Lee, S. J., & Woolliams, J. A. (2017). QTL analysis and genomic selection using RADseq derived markers in Sitka spruce : The potential utility of within family data. *Tree Genetics & Genomes, 13*(2), 33. https://doi.org/10.1007/s11295-017-1118-z

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., & El-Kassaby, Y. A. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics, 16*(1), 370. https://doi.org/10.1186/s12864-015-1597-y

Geraldes, A., DiFazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., Gunter, L. E., Wymore, A. M., Grassa, C. J., Farzaneh, N., Porth, I., McKown, A. D., Skyba, O., Li, E., Fujita, M., Klápště, J., Martin, J., Schackwitz, W., Pennacchio, C., … Tuskan, G. A. (2013). A 34K SNP genotyping array for Populus trichocarpa : Design, application to the study of natural populations and transferability to other Populus species. *Molecular Ecology Resources, 13*(2), 306-323. https://doi.org/10.1111/1755-0998.12056

GIS PMF 2014 : GIS PMF (2014) GIS Groupe Pin maritime du futur. Les cahiers de la reconstitution n°4: matériel végétal de reboisement. http://www.onf.fr/outils/medias/20130708-143100-661300/++files++/4.

GIS PMF 2014 : GIS PMF (2016) GIS Groupe Pin maritime du futur. Les cahiers de la reconstitution n°5: 20 ans de progrès et d'innovation https://www.gisgpmf.fr/IMG/pdf/-2.pdf

Goddard, M. e., Hayes, B. j., & Meuwissen, T. h. e. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, *128*(6), 409-421. https://doi.org/10.1111/j.1439-0388.2011.00964.x

Gorjanc, G., Gaynor, R. C., & Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, *131*(9), 1953-1966. https://doi.org/10.1007/s00122-018-3125-3

Grattapaglia, D. (2017). Status and Perspectives of Genomic Selection in Forest Tree Breeding. In R. K. Varshney, M. Roorkiwal, & M. E. Sorrells (Éds.), *Genomic Selection for Crop Improvement : New Molecular Breeding Strategies for Crop Improvement* (p. 199-249). Springer International Publishing. https://doi.org/10.1007/978-3-319-63170-7_9

Grattapaglia, D. (2022). Twelve Years into Genomic Selection in Forest Trees : Climbing the Slope of Enlightenment of Marker Assisted Tree Breeding. *Forests*, *13*(10), Article 10. https://doi.org/10.3390/f13101554

Grattapaglia, D., & Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genetics & Genomes*, *7*(2), 241-255. https://doi.org/10.1007/s11295-010-0328-4

Grattapaglia, D., Vilela Resende, M. D., Resende, M. R., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Takahashi, E. K., Zamprogno, K. C., & Kilian, A. (2011). Genomic Selection for growth traits in Eucalyptus : Accuracy within and across breeding populations. *BMC Proceedings*, *5*(7), O16. https://doi.org/10.1186/1753-6561-5-S7-O16

Greenwood, M. S., Adams, G. W., & Gillespie, M. (1991). Stimulation of flowering by grafted black spruce and white spruce : A comparative study of the effects of gibberellin A4/7, cultural treatments, and environment. *Canadian Journal of Forest Research*, *21*(3), 395-400. https://doi.org/10.1139/x91-049

Guilbaud, R., Biselli, C., Buiteveld, J., Cattivelli, L., Copini, P., Dowkiw, A., Esselink, D., Fricano, A., Guerin, V., Jorge, V., & others. (2020). Development of a new tool (4TREE) for adapted genome selection in European tree species. *Proceedings of the Gentree Symposium*. Proceedings of the Gentree Symposium, Avignon, France.

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics*, *177*(4), 2389-2397. https://doi.org/10.1534/genetics.107.081190

Hardner, C., Dieters, M., DeLacy, I., Neal, J., Fletcher, S., Dale, G., & Basford, K. (2011). Identifying deployment zones for Eucalyptus camaldulensis x E. globulus and x E. grandis hybrids using factor analytic modelling of genotype by environment interaction. *Australian Forestry*, *74*(1), 30-35. https://doi.org/10.1080/00049158.2011.10676343

Haristoy, G., Bouffier, L., Fontes, L., Leal, L., Paiva, J. A. P., Pina, J.-P., & Gion, J.-M. (2023). Genomic prediction in a multi-generation Eucalyptus globulus breeding population. *Tree Genetics & Genomes*, *19*(1), 8. https://doi.org/10.1007/s11295-022-01579-2

Hasan, O., & Reid, J. B. (1995). Reduction of generation time in Eucalyptus globulus. *Plant Growth Regulation*, *17*(1), 53-60. https://doi.org/10.1007/BF00024495

Hayes, B., & Goddard, M. (2010). Genome-wide association and genomic selection in animal breedingThis article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed Brassicas : A model for genetic improvement of major OECD crops for sustainable farming". *Genome*, *53*(11), 876-883. https://doi.org/10.1139/G10-076

Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, *91*(1), 47-60. https://doi.org/10.1017/S0016672308009981

Heffner, E. L., Sorrells, M. E., & Jannink, J.-L. (2009). Genomic Selection for Crop Improvement. Crop Science, 49(1), 1-12. https://doi.org/10.2135/cropsci2008.08.0512

Heffner, E. L., Lorenz, A. J., Jannink, J.-L., & Sorrells, M. E. (2010). Plant Breeding with Genomic Selection : Gain per Unit Time and Cost. *Crop Science*, *50*(5), 1681-1690. https://doi.org/10.2135/cropsci2009.11.0662

Henderson CR (1963) Selection index and the expected genetic advance. In: Hanson WD, Robinson HF (Eds). Statistical genetics and plant breeding, pp. 141-163.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, *31*(2), 423-447. https://doi.org/10.2307/2529430

Heslot, N., Yang, H.-P., Sorrells, M. E., & Jannink, J.-L. (2012). Genomic Selection in Plant Breeding : A Comparison of Models. *Crop Science*, *52*(1), 146-160. https://doi.org/10.2135/cropsci2011.06.0297

Hill, W. G., & Weir, B. S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics Research, 93(1), 47-64. https://doi.org/10.1017/S0016672310000480

IFN (2003) Inventaire forestier national n°2. Les tempêtes de décembre 1999 - Bilan national et enseignements. https://inventaire-forestier.ign.fr/IMG/pdf/L_IF_no02_tempetes.pdf

IFN (2009) Inventaire forestier national n°21. Tempête Klaus du 24 janvier 2009. http://inventaireforestier.ign.fr/spip/IMG/pdf/IF21_internet.pdf.

IGN (2017) Institut national de l'information géographique et forestière. La feuille de l'inventaire forestier - Mai 2017 (40). https://inventaire-forestier.ign.fr/IMG/pdf/if40_plantations.pdf

IGN (2022) Institut national de l'information géographique et forestière. Inventaire Forestier National - Mémento 2022. https://inventaire-forestier.ign.fr/IMG/pdf/memento_2022.pdf

Illy, G. (1966). Recherches sur l'amélioration génétique du Pin maritime. *Annales des sciences forestières*, *23*(4), 765-948. https://doi.org/10.1051/forest/19660401

IPCC (2023) The Intergovernmental Panel On Climate Change. Synthesis Report of the IPCC Sixth Assessment Report. https://www.ipcc.ch/report/sixth-assessment-report-cycle/

Isik, F. (2014). Genomic selection in forest tree breeding : The concept and an outlook to the future. *New Forests*, *45*(3), 379-401. https://doi.org/10.1007/s11056-014-9422-z

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., Plomion, C., & Bouffier, L. (2016). Genomic selection in maritime pine. *Plant Science*, *242*, 108-119. https://doi.org/10.1016/j.plantsci.2015.08.006

Iwata, H., Hayashi, T., & Tsumura, Y. (2011). Prospects for genomic selection in conifer breeding : A simulation study of Cryptomeria japonica. *Tree Genetics & Genomes*, *7*(4), 747-758. https://doi.org/10.1007/s11295-011-0371-9

Jena, K. K., & Mackill, D. J. (2008). Molecular Markers and Their Use in Marker-Assisted Selection in Rice. Crop Science, 48(4), 1266-1276. https://doi.org/10.2135/cropsci2008.02.0082

Jensen, J. (2001). Genetic Evaluation of Dairy Cattle Using Test-Day Models1. *Journal of Dairy Science*, *84*(12), 2803-2812. https://doi.org/10.3168/jds.S0022-0302(01)74736-4

Kainer, D., Stone, E. A., Padovan, A., Foley, W. J., & Külheim, C. (2018). Accuracy of Genomic Prediction for Foliar Terpene Traits in Eucalyptus polybractea. *G3 Genes|Genomes|Genetics*, *8*(8), 2573-2583. https://doi.org/10.1534/g3.118.200443

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, *16*(5), 1099-1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Kimura, M., & Crow, J. F. (1964). THE NUMBER OF ALLELES THAT CAN BE MAINTAINED IN A FINITE POPULATION. *Genetics*, *49*(4), 725-738. https://doi.org/10.1093/genetics/49.4.725

Kinghorn, B. P. (2011). An algorithm for efficient constrained mate selection. *Genetics Selection Evolution*, *43*(1), 4. https://doi.org/10.1186/1297-9686-43-4

Kremer, A. (1992). Predictions of age-age correlations of total height based on serial correlations between height increments in Maritime pine (Pinus pinaster Ait.). *Theoretical and Applied Genetics*, *85*(2), 152-158. https://doi.org/10.1007/BF00222853

Kremer, A., Potts, B. M., & Delzon, S. (2014). Genetic divergence in forest trees : Understanding the consequences of climate change. *Functional Ecology*, *28*(1), 22-36. https://doi.org/10.1111/1365-2435.12169

Lambeth, C., Lee, B.-C., O'Malley, D., & Wheeler, N. (2001). Polymix breeding with parental analysis of progeny : An alternative to full-sib breeding and testing. *Theoretical and Applied Genetics*, *103*(6), 930-943. https://doi.org/10.1007/s001220100627

Larson, P. R. (1962). *A Biological Approach to Wood Quality*. https://nefismembers.org/documents/a-biological-approach-to-wood-quality-1962/

Lebedev, V. G., Lebedeva, T. N., Chernodubov, A. I., & Shestibratov, K. A. (2020). Genomic selection for forest tree improvement : Methods, achievements and perspectives. *Forests*, *11*(11), 1190. https://doi.org/10.3390/f11111190

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, *92*(9), 4656-4663. https://doi.org/10.3168/jds.2009-2061

Legarra, A., Robert-Granié, C., Manfredi, E., & Elsen, J.-M. (2008). Performance of Genomic Selection in Mice. *Genetics*, *180*(1), 611-618. https://doi.org/10.1534/genetics.108.088575

Lenz, P. R. N., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M., & Bousquet, J. (2017). Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). *BMC Genomics*, *18*(1), 335. https://doi.org/10.1186/s12864-017-3715-5

Lenz, P. R. N., Nadeau, S., Azaiez, A., Gérardi, S., Deslauriers, M., Perron, M., Isabel, N., Beaulieu, J., & Bousquet, J. (2020). Genomic prediction for hastening and improving efficiency of forward selection in conifer polycross mating designs : An example from white spruce. *Heredity*, *124*(4), Article 4. https://doi.org/10.1038/s41437-019-0290-3

Lenz, P. R. N., Nadeau, S., Mottet, M.-J., Perron, M., Isabel, N., Beaulieu, J., & Bousquet, J. (2020). Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. *Evolutionary Applications*, *13*(1), 76-94. https://doi.org/10.1111/eva.12823

Li, Y., Klápště, J., Telfer, E., Wilcox, P., Graham, N., Macdonald, L., & Dungey, H. S. (2019). Genomic selection for non-key traits in radiata pine when the documented pedigree is corrected using DNA marker information. *BMC Genomics*, *20*(1), 1026. https://doi.org/10.1186/s12864-019-6420-8

Li, Y., Suontama, M., Burdon, R. D., & Dungey, H. S. (2017). Genotype by environment interactions in forest tree breeding : Review of methodology and perspectives on research and application. *Tree Genetics & Genomes*, *13*(3), 60. https://doi.org/10.1007/s11295-017-1144-x

Liao, L., Cao, L., Xie, Y., Luo, J., & Wang, G. (2022). Phenotypic Traits Extraction and Genetic Characteristics Assessment of Eucalyptus Trials Based on UAV-Borne LiDAR and RGB Images. *Remote Sensing*, *14*(3), Article 3. https://doi.org/10.3390/rs14030765

Lindgren, D., Gea, L., & Jefferson, P. A. (1996). Loss of genetic diversity monitored by status number. *Silvae Genetica*, *45*, 52-59.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E., & Jannink, J.-L. (2011). Chapter Two - Genomic Selection in Plant Breeding : Knowledge and Prospects. In D. L. Sparks (Éd.), *Advances in Agronomy* (Vol. 110, p. 77-123). Academic Press. https://doi.org/10.1016/B978-0-12-385531-2.00002-5

Ly, D., Huet, S., Gauffreteau, A., Rincent, R., Touzy, G., Mini, A., Jannink, J.-L., Cormier, F., Paux, E., Lafarge, S., Gouis, J. L., & Charmet, G. (2018). Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. *Field Crops Research, 216*, 32-41. https://doi.org/10.1016/j.fcr.2017.08.020

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), Article 7265. https://doi.org/10.1038/nature08494

Markussen, T., Fladung, M., Achere, V., Favre, J. M., Faivre-Rampant, P., Aragones, A., Perez, D. D. S., Harvengt, L., Espinel, S., & Ritter, E. (2002). Identification of QTLs Controlling Growth, Chemical and Physical Wood Property Traits in Pinus pinaster (Ait.).

Marshall, T. C., Slate, J., Kruuk, L. E. B., & Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology, 7*(5), 639-655. https://doi.org/10.1046/j.1365-294x.1998.00374.x

Martinez-Meier, A., Sanchez, L., Pastorino, M., Gallo, L., & Rozenberg, P. (2008). What is hot in tree rings ? The wood density of surviving Douglas-firs to the 2003 drought and heat wave. *Forest Ecology and Management, 256*(4), 837-843. https://doi.org/10.1016/j.foreco.2008.05.041

McEwan, A., Marchi, E., Spinelli, R., & Brink, M. (2020). Past, present and future of industrial plantation forestry and implication on future timber harvesting technology. *Journal of Forestry Research, 31*(2), 339-351. https://doi.org/10.1007/s11676-019-01019-3

Meuwissen, T. H. E. (1997). Maximizing the response of selection with a predefined rate of inbreeding. *Journal of Animal Science, 75*(4), 934-940. https://doi.org/10.2527/1997.754934x

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics, 157*(4), 1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Meyer, K. (2000). Random regressions to model phenotypic variation in monthly weights of Australian beef cows. *Livestock Production Science, 65*(1), 19-38. https://doi.org/10.1016/S0301-6226(99)00183-9

Meyer, K., & Hill, W. G. (1997). Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated' records by restricted maximum likelihood. *Livestock Production Science, 47*(3), 185-200. https://doi.org/10.1016/S0301-6226(96)01414-5

Misztal, I., Legarra, A., & Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science, 92*(9), 4648-4655. https://doi.org/10.3168/jds.2009-2064

Momen, M., Campbell, M. T., Walia, H., & Morota, G. (2019). Predicting longitudinal traits derived from high-throughput phenomics in contrasting environments using genomic Legendre polynomials and B-splines. *G3 Genes|Genomes|Genetics*, *9*(10), 3369-3380. https://doi.org/10.1534/g3.119.400346

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., & Myles, S. (2015). LinkImpute : Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3 Genes|Genomes|Genetics, 5(11), 2383-2390. https://doi.org/10.1534/g3.115.021667

Mora, O., Banos, V., Jean-Michel, C., & Regolini, M. (2012). *Le massif des Landes de Gascogne à l'horizon 2050* (Rapport de l'étude prospective, Conseil régional d'Aquitaine-INRA).

Moreau, L., Charcosset, A., & Gallais, A. (2001). SELECTION Etude de l'efficacité de la sélection assistée par marqueurs par rapport à la sélection classique. Oléagineux, Corps gras, Lipides, 8(5), Article 5. https://doi.org/10.1051/ocl.2001.0496

Moreau, L. (2017, octobre 9). La sélection génomique, une innovation pour l'amélioration génétique. *Jardins de France*. https://www.jardinsdefrance.org/la-selection-genomique-une-innovation-pour-lamelioration-genetique/

Moreaux, V., Martel, S., Bosc, A., Picart, D., Achat, D., Moisy, C., Aussenac, R., Chipeaux, C., Bonnefond, J.-M., Figuères, S., Trichet, P., Vezy, R., Badeau, V., Longdoz, B., Granier, A., Roupsard, O., Nicolas, M., Pilegaard, K., Matteucci, G., … Loustau, D. (2020). Energy, water and carbon exchanges in managed forest ecosystems : Description, sensitivity analysis and evaluation of the INRAE GO+ model, version 3.0. *Geoscientific Model Development*, *13*(12), 5973-6009. https://doi.org/10.5194/gmd-13-5973-2020

Mphahlele, M. M., Isik, F., Mostert-O'Neill, M. M., Reynolds, S. M., Hodge, G. R., & Myburg, A. A. (2020). Expected benefits of genomic selection for growth and wood quality traits in Eucalyptus grandis. *Tree Genetics & Genomes*, *16*(4), 49. https://doi.org/10.1007/s11295-020-01443-1

Mrode, R. A., & Thompson, R. (2005). *Linear models for the prediction of animal breeding values* (2nd ed). CABI Pub.

Munoz, P. R., Resende Jr., M. F. R., Huber, D. A., Quesada, T., Resende, M. D. V., Neale, D. B., Wegrzyn, J. L., Kirst, M., & Peter, G. F. (2014). Genomic Relationship Matrix for Correcting Pedigree Errors in Breeding Populations : Impact on Genetic Parameters and Genomic Selection Accuracy. *Crop Science*, *54*(3), 1115-1123. https://doi.org/10.2135/cropsci2012.12.0673

Muranty, H., Jorge, V., Bastien, C., Lepoittevin, C., Bouffier, L., & Sanchez, L. (2014). Potential for marker-assisted selection for forest tree breeding : Lessons from 20 years of MAS in crops. Tree Genetics & Genomes, 10(6), 1491-1510. https://doi.org/10.1007/s11295-014-0790-5

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., Goodstein, D. M., Dubchak, I., Poliakov, A., Mizrachi, E., Kullan, A. R. K., Hussey, S. G., Pinard, D., van der Merwe, K., Singh,

P., … Schmutz, J. (2014). The genome of Eucalyptus grandis. *Nature*, *510*(7505), Article 7505. https://doi.org/10.1038/nature13308

Nagano, S., Hirao, T., Takashima, Y., Matsushita, M., Mishima, K., Takahashi, M., Iki, T., Ishiguri, F., & Hiraoka, Y. (2020). SNP Genotyping with Target Amplicon Sequencing Using a Multiplexed Primer Panel and Its Application to Genomic Prediction in Japanese Cedar, Cryptomeria japonica (L.f.) D.Don. *Forests*, *11*(9), Article 9. https://doi.org/10.3390/f11090898

Nageleisen, L.-M. (2018). Effets du changement climatique sur les insectes forestiers. *Revue forestière française*, *70*(6), 653-660. https://doi.org/10.4267/2042/70317

Namkoong, G., Kang, H. C., & Brouard, J. S. (1988). Basic Concepts in Recurrent Selection. In G. Namkoong, H. C. Kang, & J. S. Brouard (Éds.), *Tree Breeding : Principles and Strategies* (p. 37-55). Springer. https://doi.org/10.1007/978-1-4612-3892-8_3

Neale, D. B., & Savolainen, O. (2004). Association genetics of complex traits in conifers. Trends in Plant Science, 9(7), 325-330. https://doi.org/10.1016/j.tplants.2004.05.006

Nicotra, A. B., Atkin, O. K., Bonser, S. P., Davidson, A. M., Finnegan, E. J., Mathesius, U., Poot, P., Purugganan, M. D., Richards, C. L., Valladares, F., & van Kleunen, M. (2010). Plant phenotypic plasticity in a changing climate. Trends in Plant Science, 15(12), 684-692. https://doi.org/10.1016/j.tplants.2010.09.008

Nicotra, A., & Davidson, A. (2010). Adaptive phenotypic plasticity and plant water use. *Functional Plant Biology - FUNCT PLANT BIOL*, *37*. https://doi.org/10.1071/FP09139

Nielsen, R., Tarpy, D. R., & Reeve, H. K. (2003). Estimating effective paternity number in social insects and the effective number of alleles in a population. *Molecular Ecology*, *12*(11), 3157-3164. https://doi.org/10.1046/j.1365-294X.2003.01994.x

Oliveira, R. de S., Ribeiro, C. V. G., Neres, D. F., Porto, A. C. da M., Ribeiro, D., Siqueira, L. de, Zauza, E. Â. V., Coelho, A. S. G., Reis, C. A. F., Alfenas, A. C., & Novaes, E. (2020). Evaluation of genetic parameters and clonal selection of *Eucalyptus* in the Cerrado region. *Crop Breeding and Applied Biotechnology*, *20*, e29982031. https://doi.org/10.1590/1984-70332020v20n3a35

Osorio, L. F., White, T. L., & Huber, D. A. (s. d.). *Age Trends of Heritabilities and Genotype-by-Environment Interactions for Growth Traits and Wood Density from Clonal Trials of Eucalyptus grandis HILL ex MAIDEN*.

Pâques, L. E. (Éd.). (2013). *Forest tree breeding in Europe : Current state-of-the-art and perspectives* (Vol. 25). Springer Netherlands. https://doi.org/10.1007/978-94-007-6146-9

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545-554. https://doi.org/10.1093/biomet/58.3.545

Pawson, S. M., Brin, A., Brockerhoff, E. G., Lamb, D., Payn, T. W., Paquette, A., & Parrotta, J. A. (2013). Plantation forests, climate change and biodiversity. *Biodiversity and Conservation*, *22*(5), 1203-1227. https://doi.org/10.1007/s10531-013-0458-8

Payn, T., Carnus, J.-M., Freer-Smith, P., Kimberley, M., Kollert, W., Liu, S., Orazio, C., Rodriguez, L., Silva, L. N., & Wingfield, M. J. (2015). Changes in planted forests and future global implications. *Forest Ecology and Management*, *352*, 57-67. https://doi.org/10.1016/j.foreco.2015.06.021

Peixoto, M. A., Coelho, I. F., Evangelista, J. S. P. C., Alves, R. S., Rocha, J. R. do A. S. de C., Farias, F. J. C., Carvalho, L. P., Teodoro, P. E., & Bhering, L. L. (2020). Reaction norms-based approach applied to optimizing recommendations of cotton genotypes. *Agronomy Journal*, *112*(6), 4613-4623. https://doi.org/10.1002/agj2.20433

Pichot, Ch., & Teissier Du Cros, E. (1988). Estimation of genetic parameters in the European black poplar (Populus nigra L.). Consequence on the breeding strategy. *Annales des Sciences Forestières*, *45*(3), 223-238. https://doi.org/10.1051/forest:19880304

Pugh, T. A. M., Lindeskog, M., Smith, B., Poulter, B., Arneth, A., Haverd, V., & Calle, L. (2019). Role of forest regrowth in global carbon sink dynamics. *Proceedings of the National Academy of Sciences*, *116*(10), 4382-4387. https://doi.org/10.1073/pnas.1810512116

Ratcliffe, B., El-Dien, O. G., Klápště, J., Porth, I., Chen, C., Jaquish, B., & El-Kassaby, Y. A. (2015). A comparison of genomic selection models across time in interior spruce (Picea engelmannii × glauca) using unordered SNP imputation methods. *Heredity*, *115*(6), 547-555. https://doi.org/10.1038/hdy.2015.57

Ray, D., Berlin, M., Alia, R., Sanchez, L., Hynynen, J., González-Martinez, S., & Bastien, C. (2022). Transformative changes in tree breeding for resilient forest restoration. *Frontiers in Forests and Global Change*, *5*. https://doi.org/10.3389/ffgc.2022.1005761

Région Nouvelle-Aquitaine (2022). Incendies été 2022 - Girande et Landes - Retour d'expérience. https://www.gironde.gouv.fr/

Rennenberg, H., Loreto, F., Polle, A., Brilli, F., Fares, S., Beniwal, R. S., & Gessler, A. (2006). Physiological Responses of Forest Trees to Heat and Drought. *Plant Biology*, *8*(5), 556-571. https://doi.org/10.1055/s-2006-924084

Resende, J., M. F. R., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., & Kirst, M. (2012). Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). *Genetics*, *190*(4), 1503-1510. https://doi.org/10.1534/genetics.111.137026

Resende Jr, M. F. R., Muñoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., Resende, M. D. V., & Kirst, M. (2012). Accelerating the domestication of trees using genomic selection : Accuracy of prediction models across ages and environments. *New Phytologist*, *193*(3), 617-624. https://doi.org/10.1111/j.1469-8137.2011.03895.x

Resende, M. D. V., Resende Jr, M. F. R., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., Abad, J. M., Takahashi, E. K., Rosado, A. M., Faria, D. A., Pappas Jr., G. J., Kilian, A., & Grattapaglia, D. (2012). Genomic selection for growth and wood quality in Eucalyptus : Capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist*, *194*(1), 116-128. https://doi.org/10.1111/j.1469-8137.2011.04038.x

Resende, R. T., Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., & Grattapaglia, D. (2017). Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. *Heredity*, *119*(4), Article 4. https://doi.org/10.1038/hdy.2017.37

Richter, H. (1976). The Water Status in the Plant Experimental Evidence. In O. L. Lange, L. Kappen, & E.-D. Schulze (Éds.), *Water and Plant Life : Problems and Modern Approaches* (p. 42-58). Springer. https://doi.org/10.1007/978-3-642-66429-8_4

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V. M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C.-C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., & Moreau, L. (2012). Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals : Comparison of Methods in Two Diverse Groups of Maize Inbreds (Zea mays L.). *Genetics*, *192*(2), 715-728. https://doi.org/10.1534/genetics.112.141473

Rincent, R., Salon, C., Gouis, J. L., Roumet, P. P., Lafarge, S. S., & Beauchene, K. (2019). *ArchiRac : Caractérisation de la diversité génétique de l'architecture racinaire chez le blé tendre et le blé dur*.

Rinn, F., Schweingruber, F.-H., & Schär, E. (1996). *RESISTOGRAPH and X-Ray Density Charts of Wood. Comparative Evaluation of Drill Resistance Profiles and X-ray Density Charts of Different Wood Species*. *50*(4), 303-311. https://doi.org/10.1515/hfsg.1996.50.4.303

Rivers, M., Newton, A. C., Oldfield, S., & Contributors, G. T. A. (2023). Scientists' warning to humanity on tree extinctions. *PLANTS, PEOPLE, PLANET*, *5*(4), 466-482. https://doi.org/10.1002/ppp3.10314

Rutkoski, J., Benson, J., Jia, Y., Brown-Guedira, G., Jannink, J.-L., & Sorrells, M. (2012). Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. *The Plant Genome*, *5*(2). https://doi.org/10.3835/plantgenome2012.02.0001

Sambatti, J. B. M., & Caylor, K. K. (2007). When is breeding for drought tolerance optimal if drought is random? *New Phytologist*, *175*(1), 70-80. https://doi.org/10.1111/j.1469-8137.2007.02067.x

Sánchez-Vargas, N. M., Sánchez, L., & Rozenberg, P. (2007). Plastic and adaptive response to weather events : A pilot study in a maritime pine tree ring. *Canadian Journal of Forest Research*, *37*(11), 2090-2095. https://doi.org/10.1139/X07-075

Schlichting, C. D., & Smith, H. (2002). Phenotypic plasticity : Linking molecular mechanisms with evolutionary outcomes. *Evolutionary Ecology*, *16*(3), 189-211. https://doi.org/10.1023/A:1019624425971

Schopp, P., Müller, D., Wientjes, Y. C. J., & Melchinger, A. E. (2017). Genomic Prediction Within and Across Biparental Families : Means and Variances of Prediction Accuracy and Usefulness of Deterministic Equations. *G3 Genes|Genomes|Genetics*, *7*(11), 3571-3586. https://doi.org/10.1534/g3.117.300076

Schweingruber, F. H. (2007). *Wood Structure and Environment*. Springer.

Sierra-Lucero, V., Huber, D., Mckeand, S., White, T., & Rockwood, D. (2003). Genotype-by-environment interaction and deployment considerations for families from florida provenances of loblolly pine. *Forest Genetics, 10*, 85-92.

Silva, V. E., Buzzetti, S., Montanari, R., Panosso, A. R., Moreira, S. C. D., & Silva, J. F. da. (2022). Influence of the climate on productivity and the eucalyptus drought response and a proposal for maximizing wood productivity in function of soil attributes in Brazil. *Ciência Florestal, 32*, 523-547. https://doi.org/10.5902/1980509832690

La mise à jour automatique des citations est désactivée. Pour voir la bibliographie, cliquez sur Actualiser dans l'onglet Zotero.Solvin, T. M., Puliti, S., & Steffenrem, A. (2020). Use of UAV photogrammetric data in forest genetic trials : Measuring tree height, growth, and phenology in Norway spruce (Picea abies L. Karst.). Scandinavian Journal of Forest Research, 35(7), 322-333. https://doi.org/10.1080/02827581.2020.1806350

Sousa, E., JM, R., Bonifacio, L., Naves, P., & Rodrigues, A. (2011). Management and Control of the Pine Wood Nematode, Bursaphelenchus Xylophilus, in Portugal. In *Nematodes : Morphology, Functions and Management Strategies* (p. 21pp).

Souza, L. M., Francisco, F. R., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., Fritsche-Neto, R., & Souza, A. P. (2019). Genomic Selection in Rubber Tree Breeding : A Comparison of Models and Methods for Managing G×E Interactions. *Frontiers in Plant Science, 10*. https://www.frontiersin.org/articles/10.3389/fpls.2019.01353

Stearns, S. C. (1989). The Evolutionary Significance of Phenotypic Plasticity : Phenotypic sources of variation among organisms can be described by developmental switches and reaction norms. *BioScience, 39*(7), 436-445. https://doi.org/10.2307/1311135

Stocks, J. J., Metheringham, C. L., Plumb, W. J., Lee, S. J., Kelly, L. J., Nichols, R. A., & Buggs, R. J. A. (2019). Genomic basis of European ash tree resistance to ash dieback fungus. *Nature Ecology & Evolution, 3*(12), Article 12. https://doi.org/10.1038/s41559-019-1036-6

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J.-L., & Sorrells, M. E. (2017). Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *The Plant Genome, 10*(2), 1-12. https://doi.org/10.3835/plantgenome2016.11.0111

Suontama, M., Klápště, J., Telfer, E., Graham, N., Stovold, T., Low, C., McKinley, R., & Dungey, H. (2019). Efficiency of genomic prediction across two Eucalyptus nitens seed orchards with different selection histories. *Heredity, 122*(3), Article 3. https://doi.org/10.1038/s41437-018-0119-5

Sykes, R., Li, B., Isik, F., Kadla, J., & Chang, H.-M. (2006). Genetic variation and genotype by environment interactions of juvenile wood chemical properties in *Pinus taeda* L. *Annals of Forest Science, 63*(8), 897-904. https://doi.org/10.1051/forest:2006073

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., & Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biology, 17*(1), 110. https://doi.org/10.1186/s12870-017-1059-6

Terray, L., Pagé, C., Déqué, M., & Flecher, C. (2010). *L'évolution du climat en France au travers de quelques indicateurs agroclimatiques*.

Thavamanikumar, S., Arnold, R. J., Luo, J., & Thumma, B. R. (2020). Genomic Studies Reveal Substantial Dominant Effects and Improved Genomic Predictions in an Open-Pollinated Breeding Population of Eucalyptus pellita. *G3 Genes|Genomes|Genetics*, *10*(10), 3751-3763. https://doi.org/10.1534/g3.120.401601

Thistlethwaite, F. R., Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., Stoehr, M. U., Ingvarsson, P. K., & El-Kassaby, Y. A. (2020). Linkage disequilibrium vs. pedigree : Genomic selection prediction accuracy in conifer species. *PLOS ONE*, *15*(6), e0232201. https://doi.org/10.1371/journal.pone.0232201

Thistlethwaite, F. R., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., Stoehr, M. U., & El-Kassaby, Y. A. (2017). Genomic prediction accuracies in space and time for height and wood density of Douglas-fir using exome capture as the genotyping platform. *BMC Genomics*, *18*(1), 930. https://doi.org/10.1186/s12864-017-4258-5

Thistlethwaite, F. R., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., Stoehr, M. U., & El-Kassaby, Y. A. (2019). Genomic selection of juvenile height across a single-generational gap in Douglas-fir. *Heredity*, *122*(6), Article 6. https://doi.org/10.1038/s41437-018-0172-0

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., de Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Townsend Peterson, A., Phillips, O. L., & Williams, S. E. (2004). Extinction risk from climate change. *Nature*, *427*(6970), 145-148. https://doi.org/10.1038/nature02121

Traimond, B. (1980). Le feu est dans la lande ou l'incendie comme fait social. *Revue forestière française*, *32*(S), 333-343. https://doi.org/10.4267/2042/21474

Ukrainetz, N. K., & Mansfield, S. D. (2019). Assessing the sensitivities of genomic selection for growth and wood quality traits in lodgepole pine using Bayesian models. *Tree Genetics & Genomes*, *16*(1), 14. https://doi.org/10.1007/s11295-019-1404-z

Valladares, F., Sanchez-Gomez, D., & Zavala, M. A. (2006). Quantitative estimation of phenotypic plasticity : Bridging the gap between the evolutionary concept and its ecological applications. *Journal of Ecology*, *94*(6), 1103-1116. https://doi.org/10.1111/j.1365-2745.2006.01176.x

van Eeuwijk, F. A., Bustos-Korts, D. V., & Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype × environment interactions? *Crop Science*, *56*(5), 2119-2140. https://doi.org/10.2135/cropsci2015.06.0375

VanRaden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull Bulletin*, *37*, Article 37.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414-4423. https://doi.org/10.3168/jds.2007-0980

Via, S. (1993). Adaptive Phenotypic Plasticity : Target or By-Product of Selection in a Variable Environment? *The American Naturalist*, *142*(2), 352-365. https://doi.org/10.1086/285542

Vidal, M., Plomion, C., Harvengt, L., Raffin, A., Boury, C., & Bouffier, L. (2015). Paternity recovery in two maritime pine polycross mating designs and consequences for breeding. *Tree Genetics & Genomes*, *11*(5), 105. https://doi.org/10.1007/s11295-015-0932-4

Vidal, M., Plomion, C., Raffin, A., Harvengt, L., & Bouffier, L. (2017). Forward selection in a maritime pine polycross progeny trial using pedigree reconstruction. *Annals of Forest Science*, *74*(1), 21. https://doi.org/10.1007/s13595-016-0596-8

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., & Martin, N. G. (2006). Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLOS Genetics*, *2*(3), e41. https://doi.org/10.1371/journal.pgen.0020041

Wang, C., Andersson, B., & Waldmann, P. (2009). Genetic analysis of longitudinal height data using random regression. *Canadian Journal of Forest Research*, *39*(10), 1939-1948. https://doi.org/10.1139/X09-111

White, T. L., Adams, W. T., & Neale, D. B. (2007). Advanced-generation breeding strategies—Breeding population size, structure and management. *Forest genetics*, 479-522. https://doi.org/10.1079/9781845932855.0479

Wiggans, G. R., Cole, J. B., Hubbard, S. M., & Sonstegard, T. S. (2017). Genomic Selection in Dairy Cattle : The USDA Experience. *Annual Review of Animal Biosciences*, *5*(1), 309-327. https://doi.org/10.1146/annurev-animal-021815-111422

Windig, J. J., De Kovel, C. G., & De Jong, G. (2004). Genetics and mechanics of plasticity. *Phenotypic plasticity: functional and conceptual approaches*, 31-49.

Woolliams, J. a., Berg, P., Dagnachew, B. s., & Meuwissen, T. h. e. (2015). Genetic contributions and their optimization. *Journal of Animal Breeding and Genetics*, *132*(2), 89-99. https://doi.org/10.1111/jbg.12148

Wray, N., & Goddard, M. (1994). Increasing long-term response to selection. *Genetics Selection Evolution*, *26*(5), 431. https://doi.org/10.1186/1297-9686-26-5-431

Xie, C.-Y. (2003). Genotype by environment interaction and its implications for genetic improvement of interior spruce in British Columbia. *Canadian Journal of Forest Research*, *33*(9), 1635-1643. https://doi.org/10.1139/x03-082

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565-569. https://doi.org/10.1038/ng.608

Zapata-Valenzuela, J., Isik, F., Maltecca, C., Wegrzyn, J., Neale, D., McKeand, S., & Whetten, R. (2012). SNP markers trace familial linkages in a cloned population of Pinus taeda—Prospects for genomic selection. *Tree Genetics & Genomes*, *8*(6), 1307-1318. https://doi.org/10.1007/s11295-012-0516-5

Zapata-Valenzuela, J., Whetten, R. W., Neale, D., McKeand, S., & Isik, F. (2013). Genomic Estimated Breeding Values Using Genomic Relationship Matrices in a Cloned Population of Loblolly Pine. *G3 Genes|Genomes|Genetics*, *3*(5), 909-916. https://doi.org/10.1534/g3.113.005975

Zas, R., Merlo, E., & Fernández-López, J. (2004). Juvenile – Mature Genetic Correlations in Pinus pinaster Ait. Under Different Nutrient x Water Regimes. *Silvae Genetica*, *53*(1-6), 124-129. https://doi.org/10.1515/sg-2004-0022

Zas, R., Sampedro, L., Solla, A., Vivas, M., Lombardero, M. J., Alía, R., & Rozas, V. (2020). Dendroecology in common gardens : Population differentiation and plasticity in resistance, recovery and resilience to extreme drought events in *Pinus pinaster*. *Agricultural and Forest Meteorology*, *291*, 108060. https://doi.org/10.1016/j.agrformet.2020.108060

Zhong, S., & Jannink, J.-L. (2007). Using Quantitative Trait Loci Results to Discriminate Among Crosses on the Basis of Their Progeny Mean and Variance. *Genetics*, *177*(1), 567-576. https://doi.org/10.1534/genetics.107.075358

Zhou, L., Chen, Z., Olsson, L., Grahn, T., Karlsson, B., Wu, H. X., Lundqvist, S.-O., & García-Gil, M. R. (2020). Effect of number of annual rings and tree ages on genomic predictive ability for solid wood properties of Norway spruce. *BMC Genomics*, *21*(1), 323. https://doi.org/10.1186/s12864-020-6737-3

# Annexe 1 : Publications et communications au cours de la thèse

## Publications:

Victor Papin, Alexandre Bosc, Leopoldo Sanchez and Laurent Bouffier (2023) Integrating environmental gradients into breeding: application of genomic reactions norms in a perennial species. *Journal of Experimental Botany*
Soumis le 5 avril 2023. Décision « Major Revision » le 26 juillet 2023. Resoumis après corrections le 7 septembre 2023. En attente.

Victor Papin, Gregor Gorjanc, Ivan Pocrnic, Laurent Bouffier and Leopoldo Sanchez (2023) Unlocking genomic selection potential: within-family prediction in conifers. *Annals of Forest Sciences*

En préparation, soumission prévue en novembre 2023

## Communications orales :

Victor Papin, Laurent Bouffier, Leopoldo Sanchez, « Case study 3: Norm of reaction for maritime pine », B4EST training course, Online, 7 octobre 2021.

Victor Papin, Laurent Bouffier, Leopoldo Sanchez, « Sélection génomique et construction de normes de réaction chez le pin maritime », séminaire R2D2, Saint Germain au Mont d'or, 8 novembre 2021.

Victor Papin « Sélection génomique chez le pin maritime », Journée de l'Ecole Doctorale, Bordeaux, 8 avril 2022.

Victor Papin, Laurent Bouffier, Leopoldo Sanchez, « Genomic selection and reactions norms for maritime pine », Final B4EST Conference, Lisbonne, Portugal, 20 juin 2022.

Victor Papin, Laurent Bouffier, Leopoldo Sanchez, Workshop « Construction of norms of reaction – Case study : Maritime pine (Pinus pinaster) », Final B4EST Conference, Lisbonne, Portugal, 23 juin 2022.