New models for implementation of genome-wide evaluation in black poplar breeding program

Marie Pegard

Centre Val de Loire UNIVERSITÉ Spirit of the Loire Valley

UNIVERSITÉ D'ORLÉANS

UNIVERSITE D'ORLEANS

ÉCOLE DOCTORALE SANTÉ, SCIENCES BIOLOGIQUES ET CHIMIE DU VIVANT

Institut National de la Recherche Agronomique

Thèse présentée par :

Marie PEGARD

soutenue le : **19 Décembre 2018**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : **Génétique amélioration variétale**

New models for implementation of genome-wide evaluation in black poplar breeding program

Thèse dirigée par :

Leopoldo SÁNCHEZ	Directeur de recherche, INRA Orléans
Véronique JORGE	Chargée de recherche, INRA Orléans

RAPPORTEURS :

Mathilde CAUSSE	Directrice de recherche, INRA Avignon
Zulma VITEZICA	Directrice de recherche, INRA Toulouse

JURY :

Leopoldo SÁNCHEZ	Directeur de recherche, INRA Orléans
Véronique JORGE	Chargée de recherche, INRA Orléans
Mathilde CAUSSE	Directrice de recherche, INRA Avignon
Zulma VITEZICA	Directrice de recherche, INRA Toulouse
Dominique DE VIENNE	Professeur, Université Paris-Sud
Stéphane Maury	Professeur, Université Orléans
Sophie BOUCHET	Chargée de Recherche, INRA Clermont-Ferrand
Timothée FLUTRE	Chargé de Recherche, INRA Montpellier

Remerciements

Les prochaines lignes ne seront pas les plus difficiles à écrire. Bien que ce soit celles qui rendent la fin de cette aventure bien réelle. Car la thèse ne représente que trois ans d'une vie mais quelles années ! !

Je voudrais commencer comme il est l'usage par remercier M. **Gilles PILATE** pour m'avoir accueilli dans l'unité de recherche Amélioration Génétique et Physiologie Forestière tout d'abord en tant qu'ingénieur puis en tant que doctorante. J'aimerais également remercier M. **Marc Villar** qui a pris la direction début 2018 de l'UMR BioForA, et qui s'est assurer de notre bien être tout au long de cette année mais également pour avoir partagé avec moi ses connaissances et ses livres sur le peuplier noir sauvage "le vrai".

Je remercie également les membres du jury Mme **Mathilde Causse**, Mme **Zulma Vitezica** pour avoir accepté d'être rapporteur de cette thèse ainsi que M. **Dominique de Vienne**, M. **Stephane Maury**, Mme **Sophie Bouchet** et M. **Timothée Flutre** d'avoir accepté de juger ce travail de thèse.

Je souhaite remercier la région Centre Val de Loire ainsi que le méta-programme SELGEN qui ont financé ce travail de thèse.

Je remercie les membres de mes comités de thèse Mme **Catherine Bastien**, Mme **Laurence Moreau**, M. **Vincent Segura** et Mme **Christelle Robert-Granier** pour leur bienveillance, leurs conseils, pour les discussions autour de la sciences (mais pas seulement) au cours de ces comités de thèse.

J'ai lu que la thèse est un travail solitaire, à mes yeux il n'en est rien. La thèse est un travail d'équipe. L'équipe commence par le directeur l'encadrant et le doctorant. Je ne pourrais jamais vous remercier suffisamment **Leopoldo** et **Véronique** pour la

confiance que vous m'avez accordé. D'abord en m'offrant la possibilité de travailler en tant qu'ingénieur en attendant d'avoir les financements et puis au cours de ces trois années de thèse. Vous m'avez permis de travailler en autonomie tout en me conseillant, en m'aiguillant, en me faisant confiance. Leo tu as su diriger cette thèse de façon très humaine et bienveillante, en m'amenant à me questionner pour corriger mes erreurs. Nous avons pu partager notre passion pour les voyages et discuter des pays que nous avions visité. Véro, tu m'as permis d'apprendre beaucoup de par ta rigueur, ton sens du partage et ton sens des priorités. Tu as su nous rappeler les échéances et les priorités pour éviter que Leo et moi ne nous éparpillons de trop. J'ai également adoré partager ma passion du jardinage et notre fou rire devant le téléphone *archaïque* de Leo que je ne savais pas utilisé. Vous avez également toujours pris le temps de corriger mes écris et ne fut pas une mince affaire.

L'équipe commence mais ne se termine pas avec nous trois, bien d'autres personnes méritent leur place sur cette page. Un grand merci à toi **Catherine**, qui a pris le temps pour m'expliquer, m'initier dans le programme d'amélioration du peuplier. Tu m'as emmené sur le terrain pour que je puisse toucher mesuré voir la réalité du terrain. Tu m'as permis également de faire les croisements contrôlé et de suivre mes "bébés" jusqu'au bout. Merci également pour tes conseils judicieux, pour ces conversations enrichissante sur la route en déplacement, pour ton soutien et ta bienveillance même avec ton emploi du temps de ministre.

Vincent S merci pour, les discussions autour de la génétique (mais pas que), des méthodes et, de toutes ces nouvelles idées à tester et ton partage de script (Oui oui j'ai beaucoup copié sur certains) !

Odile, tu m'as initié à la bio-informatique, sans toi toute la partie bio-informatique de la thèse aurait été bien moins fun ! Merci également pour ton sourire quand on entre dans ton bureau même si tu sais que j'ai beaucoup de question à te poser !

Facu merci pour ton aide pour l'ajustement des données et pour ton initiation à linux, je ne suis pas repassée sur windows après ton départ ! !

J'aimerai remercier également les membres de l'Unité Expérimentale et plus particulièrement **Guillaume** qui m'a empêché de faire mes notations de débourrement sur papier et m'a montré Adonis. **Patrick** pour la gestion de la serre lors de l'élevage

des plants et pour nos conversations autours des plantes et du jardinage. **Christophe**, **Benjamin** pour avoir fait avec moi et m'avoir montré les croisements contrôlés du début à la fin pour l'aide lors du repiquage de ces quelque 900 peupliers. Un grand merci à toi **Thomas** qui m'a permis de retrouver mon chemin dans le parc à pied mère marquant le début de notre amitié. Merci pour le travail de conservation que tu as fait avec l'aide d'**Alexis** sur mes bébés comme tu aimes les appeler. Pour tous ces "Coco ! !" dans la pépinière. Je remercie également **Vincent L** pour ce déplacement à Fâtines pour le prélèvement des parents sur cette aire d'autoroute et pour la crève de 15 jours car tu as bu dans ma bouteille. **Mathieu**, **Nadège**, **Hugo**, Hugo merci à vous également petites mains du décotonnage et du rempotage !

Merci à vous **Vanina** et **Corinne** pour toutes ces extractions d'ADN, et le génotypage pour vérifier le pedigree mais également pour ces bons moments que nous avons partagés durant ces trois ans. Merci à **Céline** également, tu m'as montré toute la partie génotypage rouille et le travail que cela représente.

J'aimerai également remercier, les membres de l'EPGV d'Evry pour leur travail sur le génotypage et le séquençage et plus particulièrement Mme **Patricia Faivre-Rampant** et Mme **Marie-Christine Le Paslier** pour votre bonne humeur et votre gentillesse.

Je profite de cette page pour remercier l'ensemble du personnel de la pépinière ONF de Guéméné-Penfao sans qui nous n'aurions pas eu un aussi bon jeu de données. Je remercie plus particulièrement **Philippe Poupart** et **Olivier Forestier** pour m'avoir accueilli, fait visiter, fait participer aux mesures sur mon dispositif et pour avoir partagé vos connaissances. J'aurai aimé venir plus souvent.

Un grand merci à toi **Patricia** pour ton accueil et toute l'aide que tu nous apportes pour faire face au méandre administratif. Merci également à **Brigitte** et **Véronique** pour leur aide en tant que GU.

Merci à **Christopher**, pour ton aide sur les notations de débourrements lors de ton stage.

I would like to make a short change for the English language. I would like to thank **Robert Banks** for welcoming me to his laboratory. To **Bruce Teir** for allowing me to stay in Australia under your supervision but also for allowing me to discover Australia as

a real Australian with the help of your wife **Shirley**. A big thank you to both of you. You have been amazing. I would also like to thank you **Sarita**, **Laura**, **Léa**, **Malou**, **José**, **Juan** for these moments shared on the other side of the world.

La thèse est un travail intense et il est bien de pouvoir décompresser, c'est pourquoi j'aimerais remercier le groupe du midi (ou plutôt la horde) ! **Yves**, **Céline**,**Thibaud**, **Aurélien**, **Vincent**, **Marlène**, **Clément**, **Alexandre**, **Rémy**, pour les rires quasi journaliers. Mais également les adeptes (ils se reconnaitront) du "juste une bière" pour ces soirées à refaire le monde et ces moments de rire !

Mesfins et **Marija**, je souhaiterais vous remercier pour ces discussions sur votre pays et votre culture, cela m'a permis de voyager en restant sur place.

Merci à **Caroline** pour la gestion de cet élément essentiel à la réussite de la thèse ! La pause café mais également pour la sortie orchidée ou chez ton amie apicultrice. Merci également aux membres assidus ou non de la pause café et ces conversations sur des sujets tellement diversifiés.

Je finirais par remercier chaque membre de cette unité qui d'une façon ou d'une autre au contribué à l'aboutissement de cette thèse et même avec un simple sourire au détour d'un couloir !

Pour mes proches : J'aimerais remercier mes amis de longue date **Johanna** pour nos soirées hebdomadaires post piscine où j'ai dû cuisiner chez toi pour le pas manger "pâte-knacchi", **Marine** pour ces heures de discussion, vous avez su me soutenir et m'entendre râler même à plusieurs kilomètres. Mais aussi **Chouchou**, **Allan**, **Mel** , **Ellie**, **Jojo**, **Chloé** et ces collègues rencontré au cours de cette thèse devenu des amis (ou presque ;-)) (Flow, Camille, Thibaud, Marlène, Thomas, Aurélien, Vincent) au fil des années pour ces tous ces moments partagé qui permettent de se changer les idées.

J'aimerais remercier plus particulièrement **Flow** et **Camille** mes deux métalleux avec qui nous avons mangé des tonnes de sushi et participé au Hellfest.

Je dois également un grand merci à **mes parents**, **ma sista**, qui ont toujours cru en moi plus que moi-même et sans qui je ne serais pas arrivée jusqu'ici. Je n'ai pas assez de mots pour décrire ma reconnaissance.

Je ne sais pas si je dois remercier mon compagnon à quatre pâtes (**Locus le chat** pour les intimes) pour avoir dormi et ronflé à mes côté tout au long de l'écriture de ces remerciements et d'un certain nombre de pages de ce manuscrit. Merci quand même Locus de me rappeler que dans la vie il n'y a pas que la thèse qui est importante, il y a la sieste aussi !

Je pense que le dernier remerciement doit aller à la personne de l'ombre celle que l'on ne voit pas dans le processus de la thèse mais qui pourtant fut essentielle au bon déroulement de cette thèse. A toi **David** je dis merci, tu n'as pas hésité à tout quitter pour me suivre dans cette aventure orléanaise, mais également à l'autre bout du monde, à m'aider à maintenir le capte et la motivation dans les moments de doute et de fatigue.

Contents

List of Figures

List of Tables

Preamble

Our study concerns the black poplar (*Populus nigra* L.), belonging to the polar genus, perennial Eucotyledonous Angiosperms from the Salicaceae family. Natural populations of poplar covered almost 54 millions of hectares (FAO, 2016). Poplar wood production implies interspecific hybrid plantations, which occur naturally when the species' ranges overlap. These natural hybrids were first cloned and marketed before being deliberately created to increase wood production. They are used all around the globe and can allow by their hybrid vigor an increase of +17 to +300% in volume in some cases (Marron et al., 2006; Dillen et al., 2009; Stettler et al., 1988). The most commonly used of them is *P.* × *canadensis*. obtained by cross a male *P. nigra* with a female *P. deltoides*. The breeding program is based on a reciprocal recurrent selection of the two parental species (*P. nigra* and *P. deltoides*) followed by an interspecific hybridization (*P* × *canadensis*). France produces in average 1.3 million of m^3 of wood (25% of the hardwood harvest), and have a dedicated breeding program. The black poplar is a dominant and emblematic species of French rivers. It is also of central importance in improvement programmes. It is indeed the male parent of the Euramerican hybrids used for wood production. Black poplar is also known for its purifying role. It improves water quality through a highly developed root system. The root system of black poplar is also an essential factor in maintaining the riverbanks against erosion. The poplar wood is used for many multi-purposes, light packaging (crates, fish basket, cheese boxes, small fruit baskets, cooking packaging, etc.) and the interior design of vehicles. Poplar wood is increasingly used in modern architecture as a framework or as interior design, outdoors, mainly for cladding. In the context of a protective environment policies with the objective of reducing plastic, poplar wood has a bright future as a biodegradable packaging material. This implies a need to produce a sufficient quantity of wood more quickly, an improvement in the quality

of the wood and being able to respond quickly to new demands made by the sector.

Genomic selection (GS) (Meuwissen et al., 2001) is based on genotyping all along the genome and the construction of predictive models using statistical methods. These models used the information provided by all markers to estimate the value of individuals that are candidates for selection. GS started with and revolutionized dairy cattle breeding in many ways. First by increasing the selection accuracy, combined with an early selection and increased the selection intensity. All ingredients were mixed to increase and faster substantially the genetic gain (Schefers and Weigel, 2012). Following the success of GS in dairy cattle, other animal species and plants (Dekkers and Hospital, 2002; Bernardo and Yu, 2007; Goddard and Hayes, 2009; Heffner et al., 2009; Crossa et al., 2011; Heffner et al., 2011; Heslot et al., 2012; Hickey et al., 2014) have implemented GS their improvement programs (Desta and Ortiz, 2014). Forest trees have high expectations of the GS implementation. Currently, the breeding program is operationally optimized but present a lack of precision and efficiency. The transition to genomic selection would allow improving the precision and the efficiency of this program and integrate the genetic diversity management based on the knowledge of markers. The present proof-of-concept study was a first attempt to quantify the feasibility of genomic evaluation under different scenarios of operational interest through three main questions: (i) The first question concerns the extent to which genomic evaluation could be a guarantee of the genetic diversity that is readily available in the breeding program, and for which is known also the species; (ii) A second question concerns the integration of available information to improve the access to quality genotyping, which is ultimately at the basis of quality predictions. (iii) A third question concerns the feasibility of genomic evaluation compared to the pedigree-based counterpart.

Introduction générale

Notre étude porte sur le peuplier noir (*Populus nigra* L.). Cet arbre appartient au genre populus, et est une angiosperme dicotylédone pérenne de la famille des *Salicaceae*. Les populations naturelles de peupliers couvrent près de 54 millions d'hectares (FAO, 2016). La production de bois de peuplier implique la plantation d'hybrides interspécifiques, ces hybrides apparaissent naturellement lorsque les aires de répartition des espèces se chevauchent. Après leur découverte, ces hybrides naturels ont été clonés puis commercialisés. Par la suite, des hybrides interspécifiques ont été délibérément créés et sélectionnés pour augmenter la production de bois. Plantés dans le monde entier, ils peuvent permettre de par leur vigueur hybride d'augmenter de +17 à +300% la production en volume (Marron et al., 2006; Dillen et al., 2009; Stettler et al., 1988). L'hybride le plus communément utilisé est *P.* × *canadensis*, il est obtenu par croisement d'un P. nigra mâle avec un P. deltoides femelle. Le programme de sélection des hybrides est basé sur une sélection récurrente réciproque des deux espèces parentales (*P. nigra* et *P. deltoides*) suivie d'une hybridation interspécifique ($P \times$ *canadensis*).

Le bois de peuplier est une production importante pour la France, elle produit en moyenne 1,3 million de m^3 de bois (soit 25% de la récolte de feuillus), et dispose d'un programme d'amélioration dédié. Le bois de peuplier est utilisé pour de nombreux usages allant de l'emballage léger (caisses, paniers à poisson, boîtes à fromage, paniers à petits fruits, emballages de cuisson, etc...) à la décoration intérieure des véhicules. L'architecture moderne affectionne de plus en plus le bois de peuplier comme cadre, comme décoration intérieure ou à l'extérieur, principalement pour le bardage. Dans le cadre d'une politique de protection de l'environnement visant à réduire les plastiques, le bois de peuplier a un bel avenir en tant que matériau d'emballage biodégradable. Cela implique la nécessité de produire plus rapidement une quantité suffisante de bois, d'améliorer la qualité du bois

et de pouvoir répondre rapidement aux nouvelles demandes du secteur.

Dans ce contexte, nous nous intéressons à l'amélioration variétale du peuplier noir et à l'implémentation de la sélection génomique au sein du programme d'amélioration. Le peuplier noir est une espèce dominante et emblématique des rivières françaises. Il est également d'une importance capitale dans les programmes d'amélioration. Il est le parent mâle des hybrides eur-américains les plus utilisés pour la production de bois. Le peuplier noir est également connu pour son rôle purificateur. Il améliore la qualité de l'eau grâce à un système racinaire très développé qui est également un facteur essentiel dans le maintien des rives contre l'érosion. Cependant comme pour la plupart des arbres forestiers, l'amélioration variétale est longue et coûteuse en ressource. C'est une espèce idéale pour l'implémentation de la sélection génomique.

La sélection génomique (SG) (Meuwissen et al., 2001) utilise l'information génomique contenu tout le long du génome couplé à la construction de modèles prédictifs basés sur des méthodes statistiques pour estimer la valeur des candidats à la sélection. La SG a d'abord été utilisée au sein des programmes d'amélioration des bovins laitiers et s'est avéré être une véritable révolution de bien des façons. D'abord en augmentant la précision de la sélection, combinée à une sélection précoce et en augmentant l'intensité de la sélection. Tous les ingrédients ont été assemblés pour augmenter et accélérer sensiblement le gain génétique (Schefers and Weigel, 2012). Suite au succès de la SG chez les bovins laitiers, d'autres espèces animales et végétales (Dekkers and Hospital, 2002; Bernardo and Yu, 2007; Goddard and Hayes, 2009; Heffner et al., 2009; Crossa et al., 2011; Heffner et al., 2011; Heslot et al., 2012; Hickey et al., 2014) ont cherché à mettre en œuvre la SG au sein de leurs programmes d'amélioration (Desta and Ortiz, 2014). Les différentes études sur le sujet ont montrés que plusieurs paramètres impact la qualité de la prédiction. Les plus cités sont : le déséquilibre de liaison et l'effectif efficace de la population d'entraînement (Solberg et al., 2008; Grattapaglia and Resende, 2011; Nishio and Satoh, 2014). La densité de marqueurs le long du génome impact la qualité de la prédiction génomique (Romero Navarro et al., 2017; Norman et al., 2018; Kainer et al., 2018). Plusieurs études ont montré que la précision de la SG augmente avec la taille de la population d'entraînement (Jannink et al., 2010; Lorenz et al., 2011; Grattapaglia, 2014). De plus, la relation entre la population d'entraînement et la population de validation est cruciale pour des prédictions

précises. Plus les deux populations sont étroitement liées, plus la prévision sera précise (Pszczola et al., 2012; Ly et al., 2013; Daetwyler et al., 2013; Gowda et al., 2014).

Les arbres forestiers ne font pas exceptions et ont de grandes attentes à l'égard de la mise en œuvre du SG. Le programme d'amélioration actuel a été optimisé sur le plan opérationnel mais présente tout de même un manque de précision et d'efficacité. Le passage à la sélection génomique permettrait d'améliorer la précision et l'efficacité de ce programme et d'intégrer la gestion de la diversité génétique à partir des connaissances des marqueurs. La présente étude représente la première tentative de quantifier la faisabilité de l'évaluation génomique chez le peuplier selon différents scénarios d'intérêt opérationnel au moyen de trois questions principales : (i) La première question concerne l'apport de l'évaluation génomique dans la garantie de la conservation de la diversité génétique ; (ii) Une deuxième question concerne l'intégration des informations disponibles pour rationaliser le coût du génotypage tout en garantissant des données de qualité, base d'une évaluation génomique performante. (iii) Une troisième question concerne la faisabilité de l'évaluation génomique par comparaison avec l'évaluation phénotypique de référence.

Chapitre 1

Bibliographic Review

1.1 The Poplar and its breeding

1.1.1 Poplars

Poplars are perennial *Eucotyledonous* Angiosperms belonging to the *Salicaceae* family. The genus *Populus* includes about thirty species of poplars divided into six botanical sections (*Abaso*, *Aigeiros*, *Leucoides*, *Populus*, *Tacamahaca*, and *Turanga*) by morphological and ecological criteria (Dickman and Kuzovkina, 2014). The area covered by natural poplar populations was estimated at 54 million hectares by IPC (International Poplar Commission) in 2016 (FAO, 2016). Poplars usually have a short lifespan compared to other tree species. However, some individuals can reach ages of 200 to 300 years. The root system of some poplars can persist for thousands of years and by suckering have successively several generations of trunk. Among the genus, some species can become enormous trees, exceeding 3 m in circumference and 45 m in height (Kemperman and Barnes, 1976).

Their foliage is deciduous, simple and with opposite phyllotaxis. The leaves can be serrated or lobed (Dickman and Kuzovkina, 2014) (Figure 1.1A). The poplars are mostly diploid ; they contain two sets of 19 (2n = 38) chromosomes. Natural triploids (2n = 57) and tetraploids (2n = 76) have also been identified (Johnsson, 1942; Einspahr et al., 1963; Every and Wiens, 1971). Poplars are mostly pioneers (Rameau et al., 1989) in a variety of ecological habitats both in monospecific stands or in association with other trees

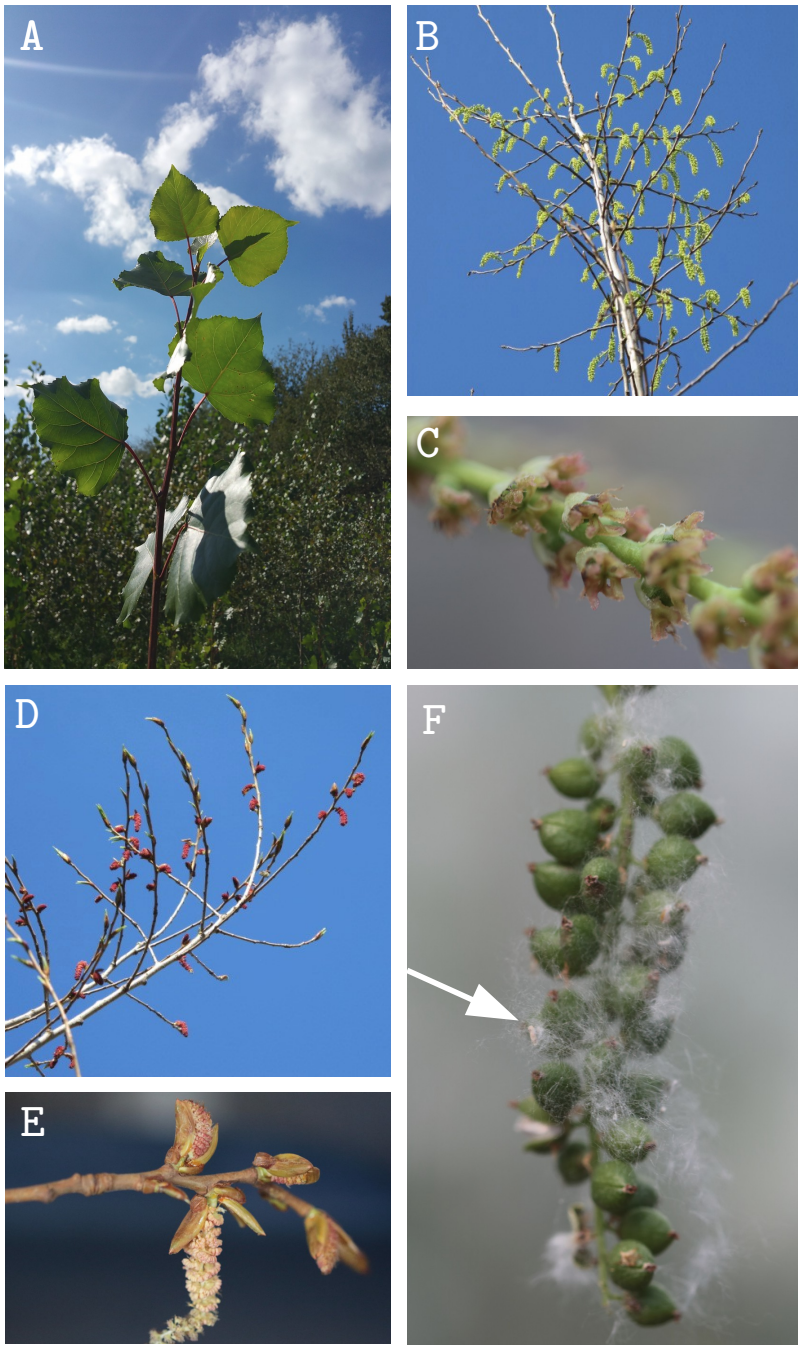


FIGURE 1.1: Flowers, fruits and leaves of Populus nigra. A : P. nigra leaves ; B : Female catkins in a tree ; C : Macro photography of female catkins and flowers ; D : Male catkins in a tree ; E : Macro photography of male catkins and flowers ; F : Ripe P. nigra fruit dropping cotton and seeds (white arrow show a seed). Credits : Marie PEGARD, the photographs were taken in nurseries and greenhouses of INRA Val de Loire.

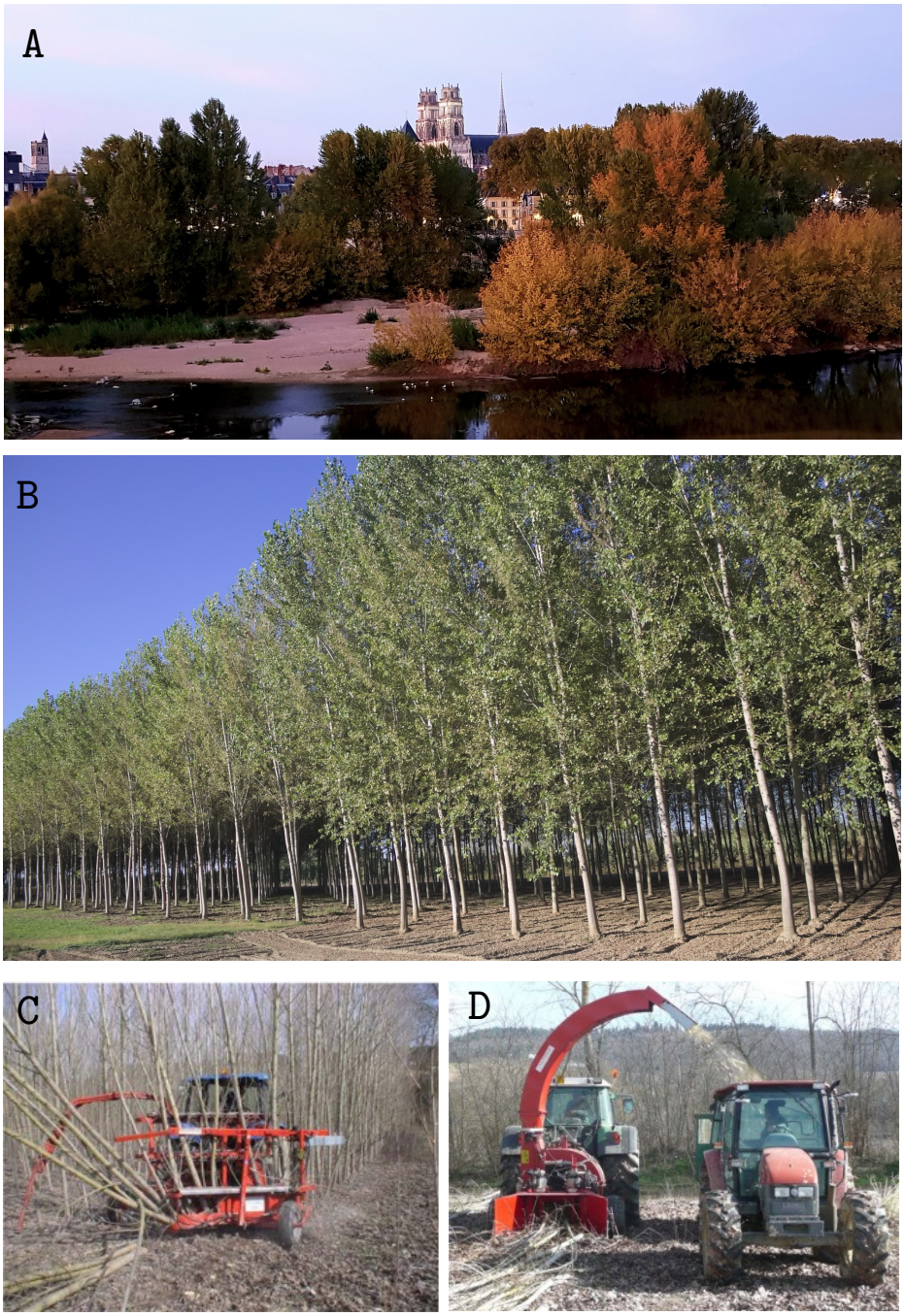


FIGURE 1.2: Poplars in the landscape. A : Populus nigra natural stands on the Loire river at Orléans, France ; B : Cultivated hybrid plantation in the region of Savigliano, Italy ; C : Cut-windrower for two-year old SRC poplar ; D : Pick-up and disk chipper used for the comminution of windrowed poplar. Credits : A and B : Marie Pégard, C and D : Santangelo et al. (2015).

species (Dickman and Kuzovkina, 2014). They are preferably found in wet ecosystems and particularly in riparian ecosystems (Figure 1.2A). They have a wide natural distribution from northern latitudes to the tropics to the latitudinal and altitudinal limits of tree growth. They have been widely introduced in the Southern Hemisphere. Poplars are dioecious with petal-free flowers (Figure 1.1B-E). The flowers, 30-200 in number, are displayed on catkins, they usually appear before the leaves in early spring. Pollination is anemophilous, and the fruit is carried by an elongated bunch of capsules that ripens in a few weeks to a month (Figure 1.1F). Seeds wrapped in cotton are mostly spread by the wind over long distances (10 km or more ; Wyckoff and Zasada (2002); Rathmacher et al. (2010)) and secondary by moving water. Poplar seeds are very small, varying from 300 to 16 000 g^{-1} depending on the species. Old trees can produce between 30 to 50 million seeds in a single season (Wyckoff and Zasada, 2002). Poplar stands can generate a significant amount of cotton, often considered a nuisance in urban areas. Therefore, male trees are preferred as ornamental trees.

Poplar interspecific hybrids

Throughout Populus species, many natural hybrids are common when the species are sympatric, i.e., when their natural or planted distribution intersects. In the *Populus* genus, spontaneous interspecific hybridizations were described. They appeared in the same or between different botanical sections. They are most frequently observed between and inside the *Aigeiros*, *Tacamahaca* and *Leucoides* sections (Cagelli et al., 1995). Nevertheless, some crosses are impossible such as *Populus alba* L. × *Populus nigra* L., and others are unidirectional, i.e. *Populus deltoides* Bart. ex Marsh. (♀) × *P. nigra* (♂). Historically, natural hybrids have been cloned and commercially grown for hundreds of years. Since several decades, tree breeders bred hybrids and deploy them in poplar wood production, motivated by their rapid growth. Some of them, have been used around the world. The most common ones for commercial purpose is *P.* × *canadensis* obtained by crossing a female *P. deltoides* with a male *P. nigra* (Dickman and Kuzovkina, 2014).

In the breeding of poplars, there are several reasons for the use of interspecific hybridization. In poplars, hybrid vigour has been observed more often from intersectional hybridization but not necessarily from intrasectional hybridization (Rood et al., 2017).

Hybrids combine the favourable traits found in both parental species. The hybrid may expressed genetic dominance when the phenotype is similar to that of one parent. The presence of hybrid vigour, which can be defined as the superiority of hybrids over the best of both parents, is also an asset. Finally, hybrids often exhibit more stable behaviour in changing environments (homeostasis) (Stebbins, 1959, 1985; Stettler et al., 1996). Thus the hybrid *P.* × *canadensis* combines a good rooting, a good level of resistance to European leaf rust (*Melampsora larici-populina*) of its parent *P. nigra* with a high juvenile vigour of the parent *P. deltoides*. The productivity gain through the use of hybrid vigor has been investigated in several studies. Different combinations of parent species in different environments agree on growth gains between +17% and +50% and between 44%-176% for volume (Marron et al., 2006; Dillen et al., 2009). Some *P.* × *generosa* (*P. deltoides* × *P. trichocarpa*) hybrids have even shown an individual heterosis effect of +300% (Stettler et al., 1988). Interspecific hybridization is all the more interesting as genetic progress and dominance and epistatic effects are disseminated by clones.

Poplar wood production and valorization

During the 25th Session of the International Poplar Commission, in Berlin in 2016, 31.4 million ha of planted poplars have been reported, of which 58% are managed for multi-purposes, 30% are planted primarily for wood production, 9% for environmental protection and 3% is managed for biomass production for fuelwood. Canada reported the largest area of poplar planted (21.8 million hectares ; 69% of the global area) ahead of China (8.5 million ; 27% of the global area planted with poplars), followed by France (0.2 million hectares), Turkey, Iran, Spain and the United States (0.1 million hectares each).

In France, poplar plantations cover 193,000 ha (Figure 1.3), representing a standing timber (Figure 1.2B) volume of 30 million m3, an annual increase of 2.6 million m^3 and an average annual harvest of about 1.3 million m^3 of wood (25% of the hardwood harvest). Factually, it is an intensive cultivation of trees planted at a final density and wide spacing (densities mainly vary from 155 to 204 stem/ha, i.e., spacing from 8m × 8m to 7m × 7m). Coppice with short or very short rotation (Figure 1.2C and D) is also used for biomass production. Poplar cultivation generally does not require irrigation and very few inputs but needs weeding between rows in early growing stages. Pruning is done between 6m and

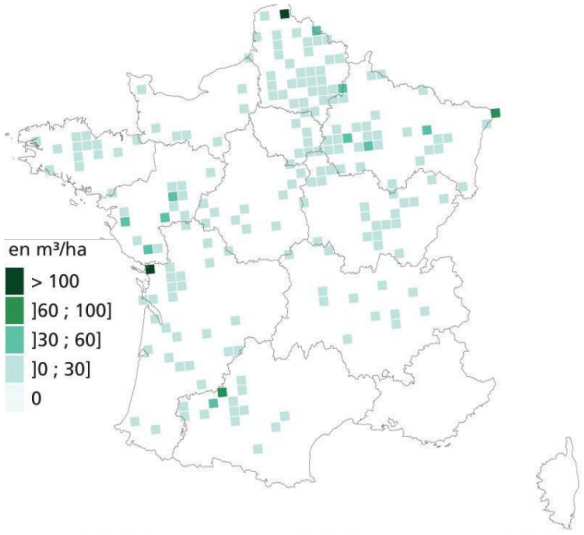


FIGURE 1.3: Map of the volume of poplar trees grown in France, in m^3/ha (IGN 2009-2013)

8m, for an harvesting between 15 and 22 years, depending on the region (Min, 2017).

Poplar wood has particular characteristics combining lightness and mechanical resistance, and its sharp point remains its ability to peel. Annually, approximately 800,000 m^3 of poplars are used for peeling (Figure 1.4B-D) and 500,000 m^3 for sawing in France (Ricodeau et al., 2018). Poplar wood is mainly used for light packaging (crates, fish basket, cheese boxes, small fruit baskets, cooking packaging, etc. Figure 1.4A) and the interior design of vehicles. It is also found in plywood panels and the manufacture of pallets. Poplar wood is increasingly used in modern architecture as a framework or as interior design (Figure 1.4E). In some areas, however, it requires an anti-termite treatment to which it is sensitive. Using preservation processes such as High-Temperature Treatment (HTT), poplar wood can be used outdoors, mainly for cladding (Figure 1.4F). Its by-products are used for mulching, pulp, and paper or biofuel (tec, 2016).

1.1.2 Cultivated genetic resources

The International Poplar Commission manages the international register of poplar cultivars marketed throughout the world. In August 2018, there were 357 accessions (Reg, 2016). Most of them are hybrids (223), the majority (160) are *P.* × *canadensis* cultivars, followed by *P.* × *canescens* (18 : *P. tremula* × *P. alba*), *P.* × *generosa* Henry (13 : *P. deltoides* × *P. trichocarpa*), the rest (30) are more complex hybrids including cultivars

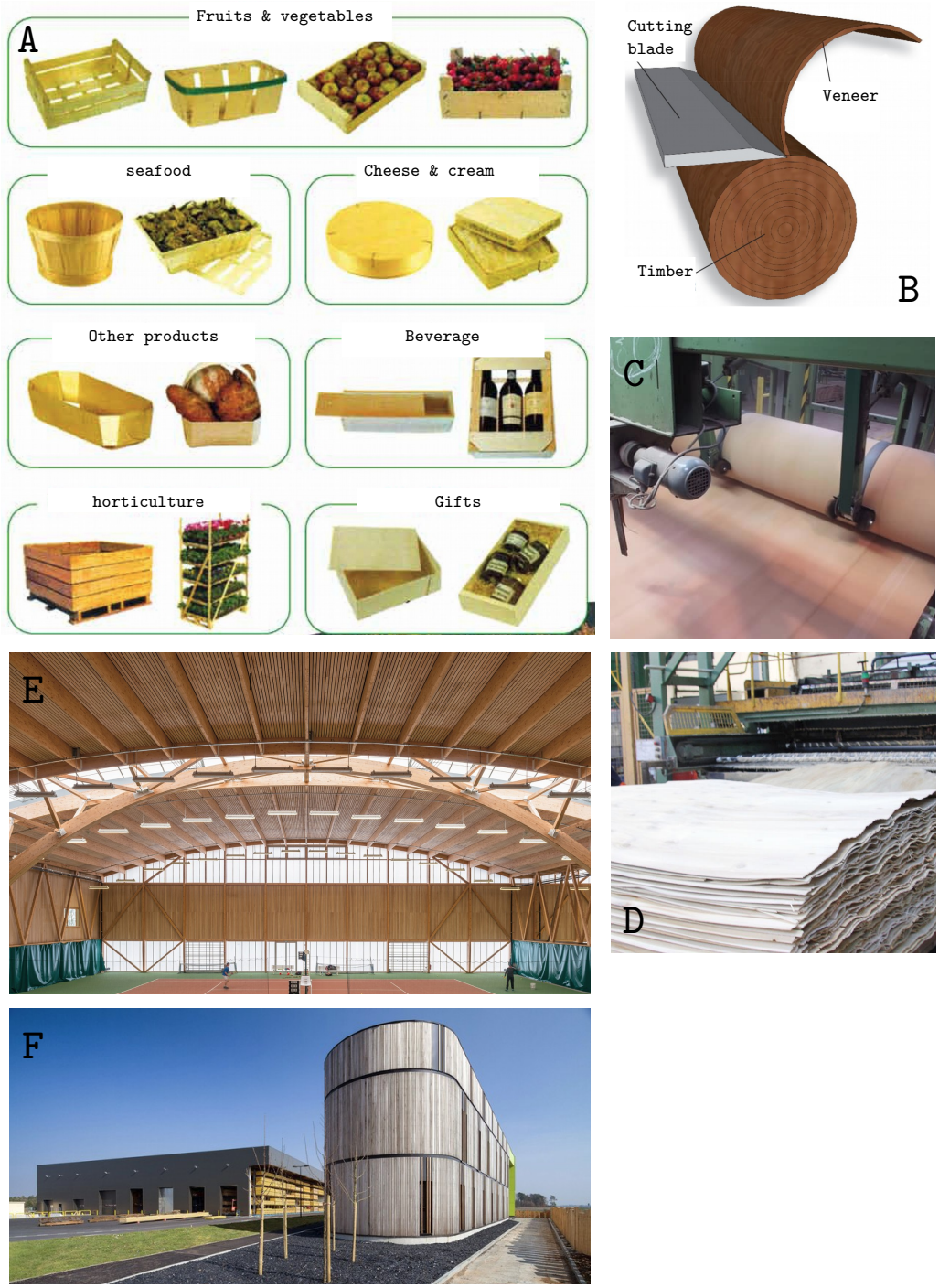


FIGURE 1.4: Poplar wood usages. A : light packaging ; B to D : Veneer production ; E-F : Modern architecture (E : Coopérative Triangle 37, Sublaines / Indre-et-Loire - France and F : New released tennis court Grenoble - France). Credits : A : AR COM ; B : Christophe Février ; C : blb-bois ; D : Philippe Schilde Agence Info ; E : R2K Architectes and F : Sébastien ANDREI.

from three-way hybridization (*P. x canadensis* Mönch × *P. yunnanensis* Dode) and a four-way hybrid (*P.* × *canadensis* Mönch × *P.* × *generosa* Henry). Just over a third (134) of cultivated cultivars are pure species, with *P. deltoides* (57) and *P. nigra* (40) at the top, followed by *P. trichocarpa* (16), *P. alba* (12), *P. tomentosa* (6). Other species (*P. tremula*, *P. tremuloides*, *P. laurifolia* Ledeb, *P. balsamifera* L., and *P. yunnanensis* Dode) appear as minorities within cultivars, they are represented by one or two individuals. Despite the choice of cultivars proposed by the International Register, resources are currently under-exploited. In France, for example, the top 10 clones sold between 2016 and 2017 represented 76.3% of plants sold. They were in decreasing order of popularity : "Koster" (*P.* × *canadensis*), "I45/51" (*P.* × *canadensis*), "Polargo" (*P.* × *canadensis*), "Trichobel" (*P. trichocarpa*), "I-214" (*P.* × *canadensis*), "Soligo" (*P.* × *canadensis*), "Albelo" (*P.* × *canadensis*), "Raspalje" (*P.* × *generosa*), "Vesten" (*P.* × *canadensis*) and the poplar clone "Blanc du Poitou" (*P.* × *canadensis*) (Min, 2017). The French Ministry of Agriculture, in collaboration with the National Poplar Commission (NPC), updates every two years the regionalized lists of poplar clones, defining eligible lists of forest reproductive material according to areas of use. Currently, clones susceptible to leaf rust (*Melampsora* spp.) and woolly poplar aphid (*Phloemyzus passerinii*) are not recommended and are not eligible for state support. Cultivar diversification aims to reduce the threat of biotic and abiotic risks.

European poplar breeding program

European poplar breeding program started during the fifties, and are mostly performed by public research institutions (Table 1.1). Selection criterias are mostly shared by all the european breeding program and concerned yield, disease and pest resistance, phenology. Other traits are selected in others country as phytorémmédiation in serbia or the ability to droughtly and saline soils (Stanton et al., 2014). Each country has its own cultivar catalogues, belgium, dutch and italian breeding programs have selected the most cultivated poplars.

TABLE 1.1: Characteristics of European poplar breeding programmes (except French breeding program) summarized from (Stanton et al., 2014, 152-170)

Country	Start year	Species	Most important commercial cultivars	Private or Public	Selection objectives
Austria	1964	*P. nigra* *P. deltoides* *P. × canadensis* *P. nigra × P. maximowiczii* *P. × canescens*		Public	renewable biomass energy production *Melampsora* rust resistance coppicing ability
Belgium	1948	*P. × generosa* taxon, *P. × canadensis*, *P. nigra*, *P. deltoides*, *P. trichocarpa*, *P. maximowiczii*, *P. trichocarpa × P. maximowiczii* taxon	Beaupre, Unal, Muur, Oudenberg, Vesten, Bakan, Skado	Public	adventitious rooting, growth rate, stem form, phenology, disease resistance (*Melampsora larici-populina*, *Marssonina brunnea* and *Xanthomonas populi*), Wood quality
Finland	1950	*P. tremula × P. tremuloides*		Public	growth, form, wood and fibre quality, *Venturia tremulae* resistance and amenability to micropropagation
Germany	1948	*P. tremula*, *P. tremuloides*, *P. alba*, *P. × canescens*, *P. grandidentata*, *P. maximowiczii*, *P. trichocarpa*, *P. nigra* and *P. deltoides* varietal mixtures of five to ten genotypes		Public	resistance to *Pollaccia elegans*, *Xanthomonas populi*, *Dothichiza* spp., *Marssonina brunnea*, *Melampsora* spp., yield, short-rotation biomass production
Hungary		*P. × canadensis*, *P. alba*, *P. nigra* and *P. deltoides*	Sv2-24, Sv1-64, H758, H425-4	Public	yield, disease and pest resistance, phytoextraction of zinc and copper for the remediation of bauxite mine spoils
Italy	1922	*P. × canadensis*,*P. deltoides*, *P. nigra*, *P. alba*, *P. deltoides × P. maximowiczii*	I214,I45/51,Soligo	Public	Resistance to *Melampsora* spp, *Marssonina brunnea*, *Venturia populina*, *Discosporium populeum*, and the woolly aphid, growth rate and adaptation to local photoperiods
Netherland	1948	*P. × canadensis*, *P. × generosa* and *P. maximowiczii*	Albelo, Polargo, Koster	Public	resistance to *Melampsora larici-populina* and *Marssonina brunnea*, *Xanthomonas populi* resistance, growth rate, stem form, branch architecture and wind tolerance

TABLE 1.1: Characteristics of European poplar breeding programmes (except French breeding program) summarized from (Stanton et al., 2014, 152-170)

Country	Start year	Species	Most important commercial cultivars	Private or Public	Selection objectives
Russia	1930	*P. alba*, *P. nigra*, *P. tremula*, *P. laurifolia*, *P. maximowiczii*, *P. suaveolens*	Sowietica Pyramidalis, Bolleana Kamyshinsky, Bolide, Veduga, Rozier, Thevestina, Pioner, Brabantica, Bachelieri, Gelrica, Marilandica, Regenerata, Robusta, Sacrau-59, Serotina	Public	growth rate, decay resistance (*Fomes igniarius* and *Phellinus tremulae*), wood quality, stem and crown form, phenology
Serbia		*P.* × *canadensis*, *P.* × *canescens*		Public	Disease and insect resistance, adventitious rooting, resistance to aphids (*Aphidoidea* spp.), poplar leaf beetle (*Chrysomela populi*), poplar leaf miner (*Leucoptera sinuella*), black stem disease (*Dothichiza populea*), leaf spot (*Marssonina brunnea*) and leaf rust (*Melampsora* spp.), drought-tolerant genotypes, biomass productivity, wood density, calorific value, several chemical and mechanical wood properties, phytoextraction of heavy metals, nitrates and polycyclic aromatic hydrocarbons
Spain	1950	*P.* × *canadensis*,*P. deltoides*, *P. nigra* and *P. alba*	Siberia Extremena, Platero, Bordils and Poncella	Public	adaptability to droughty and saline soils, calorific value and ash content
Sveden	1930	*P. maximowiczii* × *P. trichocarpa*, *P.* × *generosa*	NE42, Boelare	Public	superior cultivars for short-rotation, renewable energy feedstock production
Turkey	1970s	*P. deltoides*, *P. maximowiczii*, *P. nigra*, *P. trichocarpa* and *P.* × *canadensis*		Public	yield, adventitious rooting, wood quality, disease resistance (*Xanthomonas populi*, *Marssonina brunnea*, poplar mosaic virus) and frost resistance

French Poplar breeding program

Historically, French poplar breeding is manage by three research organizations (INRA, IRSTEA, FCBA). After a very strong attack of foliar rust revealing the high susceptibility of most clones, the implementation of a common program was established (Berthelot et al., 2001). Since 2001, the French breeding program has been managed by a “Groupement d’Intérêt Scientifique” (GIS). It is entitled "GIS Poplar Genetic, Breeding and Protection". It is a cooperation agreement signed and renewed approximately every 7 years by three partner organizations : FCBA, IRSTEA, INRA. Several strategic axes have been defined (BERTHELOT et al., 2005) :

- to design and propose efficient and stable clones in the short, medium and long term
- to develop the selection methodology
- to determine and propose criteria for the evaluation of foreign varieties and selection of future French varieties
- to test progenitors and their progeny and evaluate new French clones and foreign cultivars, in multi-site field tests, in nurseries and laboratories
- to develop and disseminate to operators in the sector the main agronomic and health characteristics of the genotypes tested (French and foreign).

Diverse species and types of hybrids are being selected within the Poplar GIS, including *P. deltoides* and *P. trichocarpa* as parents of hybrids but also for their own value in wood production. Several backcrossing series between American hybrids and parental species are also tested. P. nigra is improved mainly as a parent of the hybrid *P.* × *canadensis*. New hybrids have also been tested by the GIS, namely the cross *P. trichocarpa* × *P. maximowiczii* Henry (Asian origin).

Recurrent semi-reciprocal selection

The breeding scheme for the *P.* × *canadensis* hybrid is based on a semi-reciprocal recurrent selection (Figure 1.5). Classical reciprocal recurrent selection (Comstock et al., 1949) is used to improve two populations with complementary characteristics. Individuals

in population A are randomly crossed with those in population B or with a test individual in the population B. Their progenies are then evaluated phenotypically to determine the cross value of the parents, estimating the additive genetic value of the parents and the dominance value of the evaluated crosses (namely testcross). This makes possible to select the best parents who will be used to produce the most valuable commercial crosses. In the case of poplar, the testcross step is difficult to imagine given the generation time. The individual's own value is indeed strongly correlated to his value at crossbreeding for hybrid production, this is the reason why we speak of recurrent semi-reciprocal selection : the selected individuals serve both as a starting point for further improvement through a new intraspecific cycle and for the production of commercial hybrid material. Semi-reciprocal recurrent selection allows genetic gain by combining additive and dominance effects. Indeed, it improves the additive value of parental populations by concentrating favourable alleles and makes it possible to exploit the heterosis effect within hybrids. The heterosis effect is often related to the genetic distance between the two parental species (Stelkens and Seehausen, 2009).

1.1.3 Selection criterias

The breeding programs in poplars are based on multi-character clonal selection, and estimation of genotype $\times$ environment interaction. Among the selection criteria for European improvement programmes, the potential growth of trees is assessed by measuring height, circumference for conventional valorization, biomass produced and number of sprouts after coppice when considering a coppice valorization with (very) short rotation (SRC or VSRC). Characteristics such as rooting of cuttings and coppice abilities are selected as important factors for clonal diffusion. Vegetative phenology with budburst and budset are also selected traits. Indeed, it determines the duration of seasonal tree growth. Too early vegetative budburst exposes the terminal bud to late frosts damages (Howe et al., 2000) and threatens the individual's recovery or architecture. Budburst is also an important parameter in recommending the variety in production according to the region of plantation. Similarly, early growth cessation will reduce the growth period. Disease resistance is one of the determining criteria in the varietal improvement process, especially when the variety is clonal. Leaf rust disease at *Melampsora larici-populina*

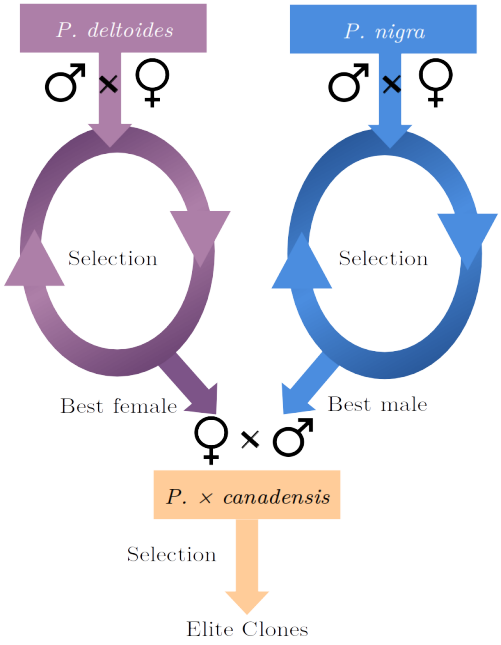


FIGURE 1.5: Recurrent reciprocal selection scheme of *P. × canadensis*. In pink and blue the intra-specific part of the program of respectively *P. deltoides* and *P. nigra*. In yellow, the hybrid part.

is currently the first significant biotic stress and it is almost annually affecting poplar plantations in Europe. Its development is favoured in environments that maintain high humidity on the lower surface of the foliage. In wide-spaced poplar plantations, biomass production losses due to leaf rust attacks can reach 60% in some years for relatively susceptible cultivars (Gastine et al., 2003). The existing resistance sources present in *P. deltoides* and *P. trichocarpa* are easily and quickly overcome by the patogen (Villar et al., 1996; Cervera et al., 1996; Dowkiw and Bastien, 2004; Jorge et al., 2005; Dowkiw and Bastien, 2007; Dowkiw et al., 2010). Research is currently focused on partial resistance in *P. nigra*, which is a reservoir of genetic variability for potential resistance to this pathogen with which it co-evolves. Other resistances are also studied, such as resistance to leaf spot (*Marssonina brunnea* f.sp. *brunnea*) and to bacterial canker (*Xanthomonas populi*). Since the 1990s, a new pest, the woolly aphid (*Phloeomyzus passerinii*), has appeared in poplar stands in Southern France and has been present throughout France since 2011. This pest can cause extremely high losses in some cultivars (80% in cultivar "I-214" (FAO, 2016)).

The tree's architecture is also a selection criteria. Individuals with fastigiate

architecture are prohibited for wood production because their deep knots make them poor quality wood for peeling. Individuals with an open port are favourably selected as well as those with high straightness of their trunk. Individuals efficient for water use are sought. Finally, the quality of the wood is a determining factor in the selection of individuals : physical properties (bark rate, humidity, infradensity, mechanical resistance, percentage of false heart, percentage of tension wood, wood colour, wood shrinkage) and paper/chemical properties (pulp yield, fibre shape, cellulose/lignin ratio, elementary chemical composition) are selection criteria. Other parameters are monitored : sex for cotton production or for genetic effect transmission to the hybrid, in the context of unilateral crosses, resistance to Melampsora allii-populina, wind sensitivity and phototropism.

1.1.4 Different steps of selection

Poplar is a tree with abundant fruiting and controlled greenhouse crosses are well managed. However, due to the number of traits that are evaluated and the age at which the different traits are assessed, not all individuals can be evaluated in the same time and for all traits. The selection of individuals is done in several steps as shown in (Figure 1.6).

The first steps are common for parental species and fr hybrids. After an evaluation in nursery, only hybrids and *P. deltoides* are selected to propose new cultivars. Similarly, characteristics will be assessed at different scales and with different precisions. Plants from controlled crosses are raised in greenhouses, then pruned and gauged to provide cuttings. A first selection is made at this stage on regrowth, cuttings and coppice abilities, with a selection rate of about 50%. In each family, the most promising individuals based on their vigor are also selected to keep a few hundred individuals evaluated in the nursery. The selection rate is again 50%. Characteristics that can be evaluated at a juvenile stage, before 5 years of age, will be evaluated in the nursery. This concerns growth, recovery, shape, phenology, biotic resistance, wood quality (infradensity) for hybrids and *P. deltoides*. This step has the highest selection rate, about 6%, and the phenotypes are well evaluate in a complete random block design in a growing environment. Indeed, a lot of juvenile characters in poplar are very good proxy for the quality of the individual in the adult stage (growth, phenology, resistance). The individuals selected during this stage will be further

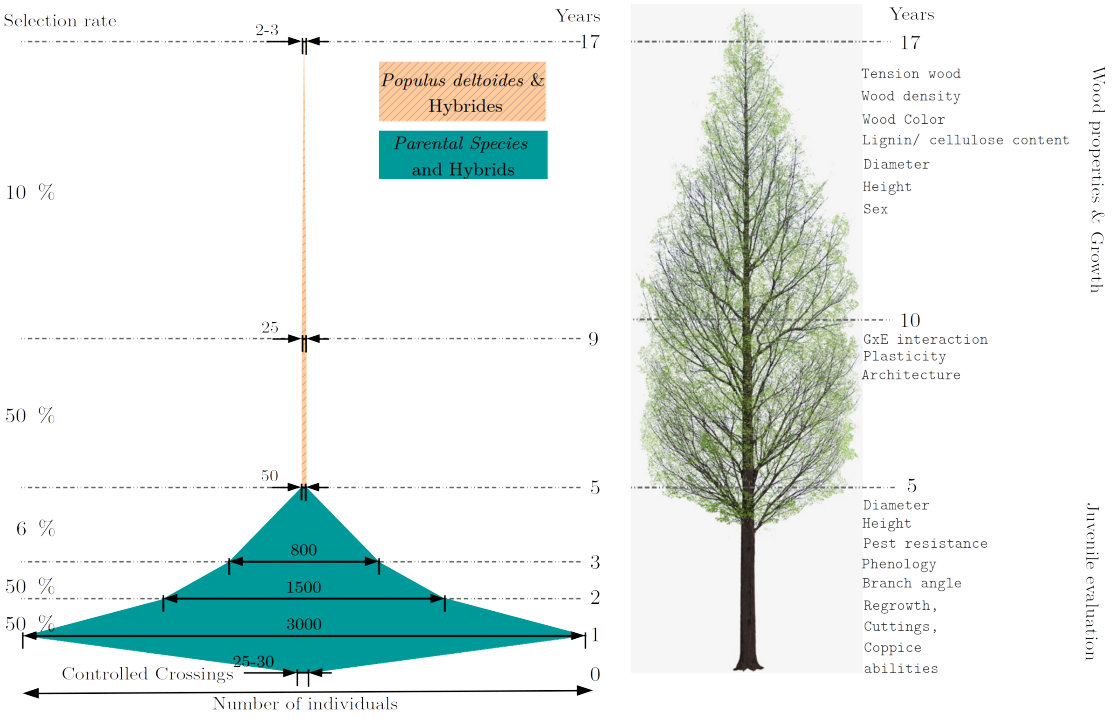


FIGURE 1.6: Representation of the number of individuals variation produced from 25 or 30 controlled crosses at the different steps of selection depending on the timeline. The selection rate indicated in left part for 1 cycle of selection. The selection criteria applied depending on the age are in the right part.

tested. At the laboratory level, further studies on a small number of promising individuals, resistance tests to different strains of pathogens such as leaf rust on leaf discs and *M. brunnea* and on the woolly aphid on potted plants. Finally, the final selection steps are carried out at the plantation level, on a limited number of individuals, varying according to the evaluation system chosen. The planting stages concern species and hybrids used directly as commercial cultivars (not *P. nigra*). Late traits are evaluated on trees more than 10 years old (growth, sex, adaptation to soil or climate and wood quality) and phytosanitary monitoring is being continued. Two types of plantations are used : a single-tree plot (Figure 1.7) system with 10 clonal replicates to evaluate 50 to 60 genotypes and to effectively control environmental effects over a small area. However, this induces a strong competitive effect and does not allow differences to be easily visualized. Multi-tree plots (Figure 1.7) with 9 clones per plot and 3 replicates allow to evaluate the behaviour of individuals in stands with good visibility. However, this type of device is expensive in terms of trial area and only allows to evaluate 12 to 15 clones using reference genotypes (as control) between the different trials. In the end, only 2 to 5 individuals will be selected and may be proposed for registration. The overall selection rate between the controlled crosses and the registered individual is close to 1%. The duration of this selection will be 17 years for pure species (excluding *P. nigra*) and 20-25 years for hybrids.

The strong European and even global interest in *P.* × *canadensis* hybrids resulting from the crossing of a *P. deltoides* female and a *P. nigra* male leads breeders to invest in the genetic resources available within the native *P. nigra* species.

1.2 The black poplar : *Populus nigra* Linnaeus

1.2.1 Species characteristics

In France, poplars are an integral part of the landscape due to its natural or wild stands. There are three pure native species present in the natural state : white poplar (*Populus alba*), aspen (*Populus tremula* L.) and black poplar (*Populus nigra*). White poplar is mainly found along rivers, in the Mediterranean valleys, in the Rhône corridor, and along the Rhine. Aspen is a forest type poplar and is located in all the French forests. The black poplar is a dominant and emblematic species of French rivers. It is also of central

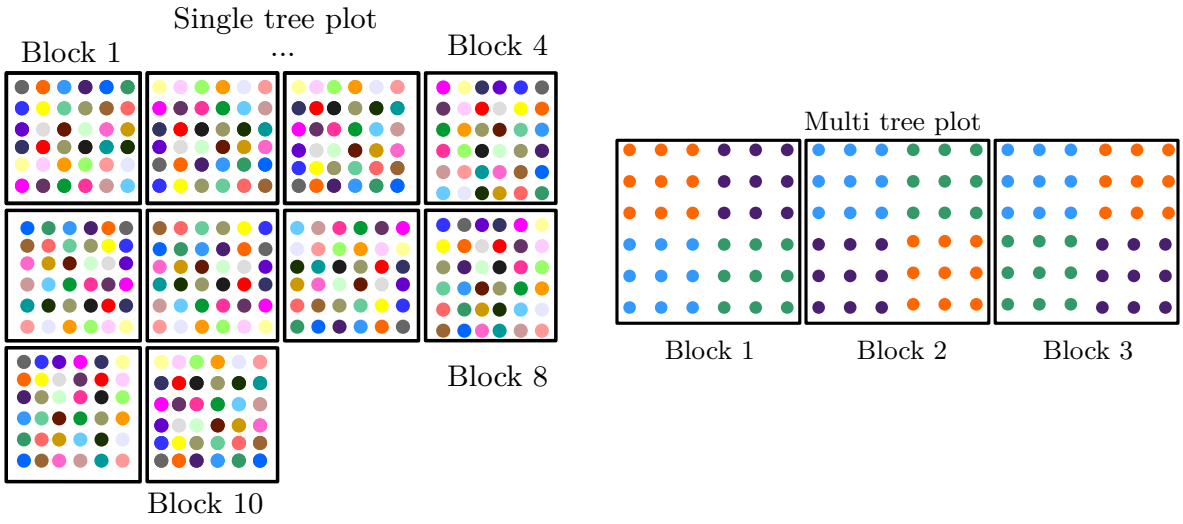


FIGURE 1.7: Example of Single-tree plot plantation (36 clones 10 blocks) and Multiple-tree plot (4 clones in 3 blocks).

importance in improvement programmes. It is indeed the male parent of the Euramerican hybrids used for wood production. The potential natural range of black poplar extends from southern Ireland to the western tip of China through a narrow North African margin (Figure 1.8). It is a pioneering species, with high demands for water and light. Black poplar is the dominant species of riparian forests, and its regeneration depends strictly on the functioning of the river or stream : the seeds produced can only germinate on recent sediments, mobilized by river dynamics and appearing by lowering the groundwater level in spring. The species also has a vegetative reproduction pattern through cuttings (which can be carried by water) or by injured root suckers (natural cloning).

Black poplar is also known for its purifying role. It improves water quality through a highly developed root system. It acts as a natural filter by capturing certain pollutants such as phosphates and nitrates from agricultural or urban sources (Ruffinoni et al., 2003). The root system of black poplar is also an essential factor in maintaining the riverbanks against erosion (Foussadier, 2003; Rodrigues et al., 2007). The high longevity of black poplar is substantial, trees aged 220 years have been found in the Gave d'Oloron. In its final stage, the tree can host a multitude of insects, birds, and bats thanks to its abundance of caches and natural cavities, thus promoting the biodiversity of its natural environment (Villar and Forestier, 2009).

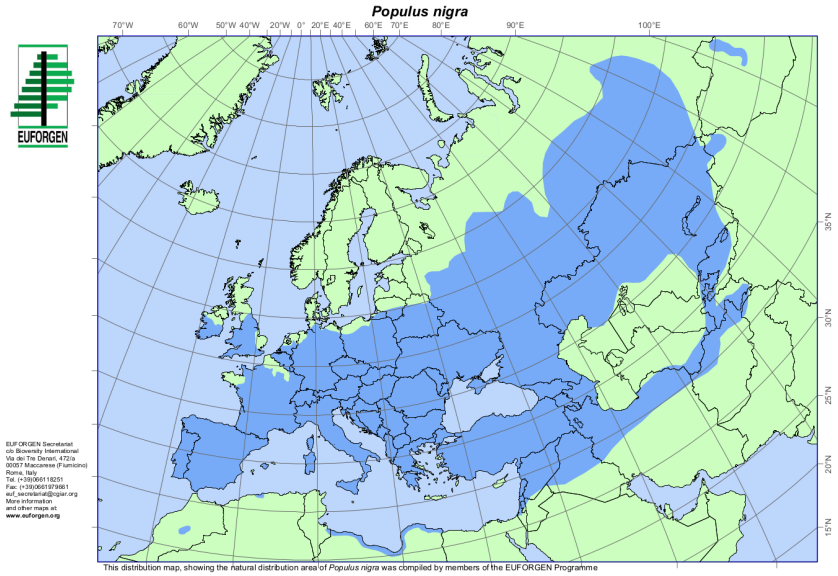


FIGURE 1.8: Distribution map of Black poplar (Populus nigra) EUFORGEN 2015, www.euforgen.org.

In the past, black poplar was a vital source of wood for riverside populations. The straightest black poplars were used as a framework, the rest as firewood. Black poplar foliage could be used as a supplement to feed livestock. Used to mark parcel boundaries or to provide faggots when pruned into pollards, black poplar was widely used by farmers. However, the Euramerican hybrids created in France in 1755 (Dickman and Kuzovkina, 2014), quickly replaced the black poplar crops, because of their high growth rate (Cain and Ormrod, 1984). From then, black poplar was mainly used as a resource for genetic improvement through crossbreeding for its hardiness, its ability to root in soils of variable texture and structure and its resistance to diseases (Cagelli et al., 1995), all these qualities are expressed in the hybrids. Black poplar is, therefore, a crucial genetic resource in forest plantations.

Nowadays, to deal with the various threats to its habitats such as the anthropisation of rivers (development : transport road, irrigation, extraction of aggregates, dams), genetic pollution of natural stands by clones of cultivated poplars can reach 13% (Pospíšková and Šálková, 2006; Ziegenhagen et al., 2008; Rathmacher et al., 2010; Smulders et al., 2008; Bastien et al., 2009; Dowkiw et al., 2014; Paffetti et al., 2018). A national programme for the conservation of genetic resources was implemented by the "Commission des ressources génétiques forestières" (CRGF) in 1992 under the aegis of the French Ministry of

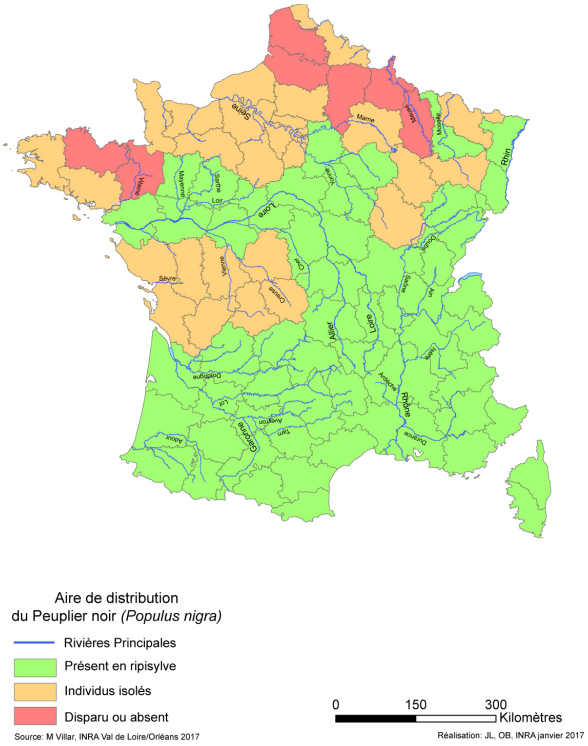


FIGURE 1.9: Black poplar distribution in France (Villar and Forestier, 2017)

Agriculture. The primary objective is to preserve the founding genes of current variability and to preserve local adaptations as well as the natural mechanisms that underlie it. The program build a map of distribution for *P. nigra* in France (Figure 1.9). The programme also seeks to enhance the existing genetic diversity of black poplar for new uses as restoration actions in riparian areas for landscape aspects or bank stabilization (Villar and Forestier, 2017). In recent years, six mixed varieties of black poplar clones (VMC : "Variétés Multi Clonales") structured by watershed have been registered in the national registry of essential materials for forest species (Villar and Forestier, 2017).

Black poplar grows in metapopulations, as demonstrated by the distribution of existing genetic diversity that have been studied in recent years on a French or European scale. Budset (Rohde et al., 2011), leaf area and water use efficiency (Chamaillard et al., 2011; Guet et al., 2015) have shown low inter-basin differentiation and high intra-basin diversity. A more extensive study using SNP markers, involving 12 populations in Western Europe revealed a geographical structure according to the watersheds. This study also shows the presence of an alpine barrier that has isolated populations in southern Europe (Faivre-

Rampant et al., 2016).

1.2.2 Genetic variability for adaptive and production traits

Several authors studied the genetic variability for adaptive and wood production traits. They concerne unrelated individuals, intraspecific crosses and parent-progeny comparisons in *P. nigra*. They allowed to estimate broad sense and narrow sense heritabilities (Table 1.2). They varied from low to high values depending on the trait and on the genetic background. Growth traits and pest resistance traits tended to be less heritable than bud phenology or branch angle (TEISSIER du CROS, 1977; Pichot and Tessier Du Cros, 1988; Rohde et al., 2011; Fabbrini et al., 2012; Guet et al., 2015; Nepveu et al., 1978; Isik and Toplu, 2004; Guet and Bastien, 2011; Legionnet et al., 1999). Components of wood quality appear to have moderate and less variable heritability (Nepveu et al., 1978; Gebreselassie et al., 2017). All these studies revealed a high genetic variability at an individual level. The high genetic variability identified for several target traits of traditional poplar breeding program, and in particular for rust resistance (Legionnet et al., 1999), explains the increasing interest of these genetic resources in these programs.

1.2.3 Populus genomic resources

To study tree biology, it is necessary to adopt a species as model, as their fundamental biological processes can not be studied with herbaceous models such as Arabidopsis. In facts, perennial habit and long life span, secondary growth from a vascular cambium, wood formation, phenology including winter dormancy and reactivation of growth in spring, and mechanisms of adaptation to local environment over large geo-climatic ranges that are typical of many trees. In early 2000s, poplar was proposed as such a model because of several attributes as ease vegetative and clonal propagation, ease genetic manipulation, rapid growth and a history of breeding and genetics, and use in plantation forestry (Douglas, 2015). Poplar was the third plant genome and the first tree which complete genome sequence to be published (Tuskan et al., 2006). A *P. trichocarpa* female ("Nisqually-1") was selected for the assembly, annotation and interpretation of the poplar genome (Tuskan et al., 2006).

TABLE 1.2: Heritabilities for adaptive and production traits in *Populus nigra*

Trait	Heritabilities		References
	Narrow sense h^2	Broad sense H^2	
Height growth	0,32 - 0,6	0,31 - 0,67	[1]
Circumference	0,15 - 0,55	0,34 - 0,82	[2]
Budburst	0,61 - 0,70	0,72 - 0,94	[3]
Budset	0,32 - 0,49	0,20 - 0,76	[4]
Foliar rust resistance		0,22 - 0,59	[5]
Branch angle	0.27 - 0.73	0.58 - 0.78	[6]
Wood infradensity		0,61 - 0,69	[7]
Wood chemistry		0,6 - 0,8	[8]

[1](TEISSIER du CROS, 1977; Nepveu et al., 1978; Pichot and Tessier Du Cros, 1988; Isik and Toplu, 2004; Sow et al., 2018) ; [2](TEISSIER du CROS, 1977; Nepveu et al., 1978) ; [3](TEISSIER du CROS, 1977; Pichot and Tessier Du Cros, 1988; Guet et al., 2015) ; [4](TEISSIER du CROS, 1977; Rohde et al., 2011; Guet et al., 2015) ; [5](TEISSIER du CROS, 1977; Pichot and Tessier Du Cros, 1988; Legionnet et al., 1999) ; [6](TEISSIER du CROS, 1977; Pichot and Tessier Du Cros, 1988) ; [7](Nepveu et al., 1978) ; [8](Gebreselassie et al., 2017)

However, even if a considerable effort was put into this reference genome, it is still an approximation of the true genome. The poplar genome assembly and annotation was twice revised and the current v3.0 contains 422.9 Mb of assembled sequence (out of a total genome size estimated at 485 Mb), with 41,335 annotated loci with protein-coding transcripts (see *Populus trichocarpa* v3.0 at the JGI plant genome portal Phytozome 11, https://phytozome.jgi.doe.gov/pz/portal.html for details). The assembled genome is a mosaic of haplotypes because of the individual chosen as reference is highly heterozygous. The poplar genome is considered the smallest among all tree genomes and allowed to re-sequence many individuals, including mapping of short next generation sequencing reads to the reference (Douglas, 2015).

Whole genome resequencing in *P. trichocarpa* and *P. nigra* were the supports of the development of genome-wide single nucleotide polymorphism (SNP) studies on evolutionary genetics, population genomics and phenotype-genotype correlation (Slavov et al., 2012; Pinosio et al., 2016).

A 34K Illumina Infinium SNP genotyping array for *P. trichocarpa* was designed for evolutionary studies such as intraspecific genetic differentiation, species assignment and

hybrids detection (Geraldes et al., 2013). After resequencing 51 *Populus nigra* individuals through Western Europe, a 12K Infinium Bead-Chip array was design and 888 individuals were genotypes. The chip was used to study and characterized the genetic structure of *P. nigra*. The SNP repartition follows the genomic genome of known QTLs (Jorge et al., 2005; Fabbrini et al., 2012; Rohde et al., 2011; Rae et al., 2009; Novaes et al., 2009; El Malki, 2013) and was developed to refine QTL location (Faivre-Rampant et al., 2016). This chip is a good candidate for a genomic selection test.

1.3 Genomic selection

1.3.1 Purpose and concept of genomic selection

Genomic selection (GS) is the *direct descendant* of marker-assisted selection (SAM) applied to quantitative traits (Meuwissen et al., 2001). Its application is based on dense genotyping all along the genome and the construction of predictive models using statistical methods, usually a derivative of a mixed model BLUP or a Bayesian method. These models simultaneously rely on the information provided by all markers to estimate the value of individuals that are candidates for selection. With adapted models, GS is able to predict the total genotypic value by estimating the additive (GEBV : genomic breeding value), dominance and epistatic part present in the phenotype of an individual. Markers are generally distributed over the entire genome and for the majority of them there is no *a priori* information indicating whether or not they are related to a QTL. The predictive model is calibrated with a set of individuals that have been phenotyped (or with known estimated additive value) and genotyped. These individuals constitute the calibration (training) population. Candidates for selection are evaluated and selected according to their genotyping information using previously constructed predictive models. Candidates may be progenies of the training population or individuals from other somehow related populations (Figure 1.10).

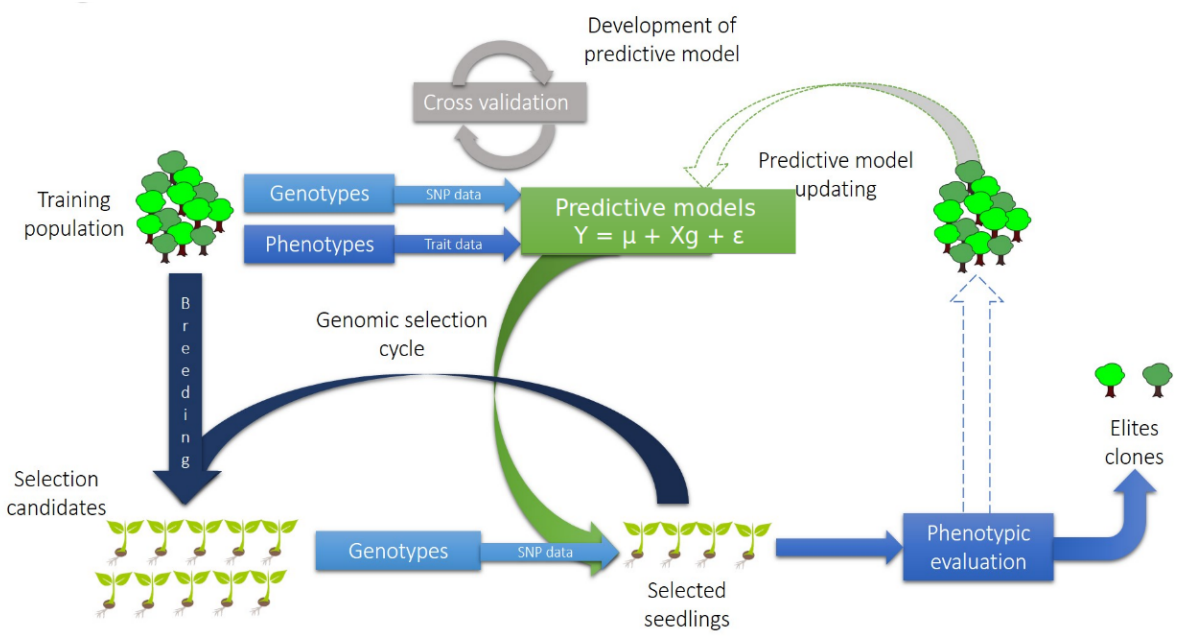


FIGURE 1.10: Concept of genomic selection adapted from Grattapaglia (2014) to the Poplar context.

1.3.2 Advent of the GS

The practical implementation of GS began with Holstein dairy cattle, which benefits from highly efficient data collection, large training populations, and consequently the potential for high prediction accuracies (Guillaume et al., 2008). The first official results of genomic evaluation (GS) in production species were published in 2009 for dairy cattle populations (Hayes et al., 2009a). The use of GS was then extended to other dairy breeds and beef cattle. Today, more than 15 countries use GEBVs in their national improvement programmes and their use has been approved at the international level (Eggen, 2012; Bouquet and Juga, 2013), leading to a globalisation of Holstein genetic progress (Lund et al., 2011). GS has revolutionized dairy cattle breeding by increasing the young animals selection accuracy, making early selection possible and thus shortening the generation interval. Finally, it increased the panel of selected candidates with respect to the traditional selection based on progeny tests. These factors contributes to faster genetic gain (Schefers and Weigel, 2012). Following the success of GS in dairy cattle,

other animal species and plants have launched the implementation of GS in an effort to revolutionize their improvement programs (Dekkers and Hospital, 2002; Bernardo and Yu, 2007; Goddard and Hayes, 2009; Heffner et al., 2009; Crossa et al., 2011; Heffner et al., 2011; Heslot et al., 2012; Hickey et al., 2014; Desta and Ortiz, 2014).

1.3.3 Accuracy of genomic selection

As with conventional selection, the accuracy of genomic prediction is estimated with the correlation between the true breeding value (TBV) and its estimate (GEBV : Genomic Estimated Breeding Value). When using empirical data, the TBV of an individual is not known, and accuracy must be estimated by cross-validation. This consists in splitting the training population into several subsamples. Each subsample is successively used as a validation population, and its GEBVs predicted by a model that is calibrated with the other subsamples. All subsamples participate one single time as validation. The correlation between GEBVs and individual phenotypes is then calculated for each validation population. It is expected that residuals in the phenotypes are uncorrelated to breeding values, thus the correlation of GEBV and phenotypes can be considered as a good proxy of the correlation between TBV and GEBV. The calculated correlation is called predictive ability or prediction accuracy depending on the authors. It measures the ability of genomic selection to predict observed values, rather than the true additive value. If the observed values are phenotypes, the accuracy of prediction can be inferred by dividing the predictive ability with the square root of the heritability of the trait in question. ($r_{GEBV,TBV} = r_{GEBV,Phenotypes}/h^2$) (Falconer, 1981). This assumes that the errors associated with GEBV and phenotype are independent (Lorenz et al., 2011, page=94).

The accuracy of genomic prediction is affected by several factors : the relatedness between training and test populations, the number of individuals in the training population, the linkage disequilibrium between markers and QTLs, the statistical method used to estimate GEBVs, the heritability of the trait, the molecular marker density, the type of markers and the genetic architecture of the trait (number of genes and distribution of their effects) (Hayes et al., 2009a; Jannink et al., 2010; Lorenz et al., 2011; Grattapaglia, 2014).

The predicting ability (or accuracy) is the most common criteria to estimate the

genomic prediction performance (Daetwyler et al., 2013). Other complementary criteria for prediction quality are also proposed, like the slope and intercept of the regression of phenotypes on predictions. This slope should be close to 1 and the intercept close to zero. There are many reasons for deviations in slope and intercept, like model's deficiencies as wrong variance partition or incomplete model, this induced systematic biases or nonrandom choice of individuals for training and validation population (Patry and Ducrocq, 2011a,b; Mäntysaari et al., 2010).

1.3.4 Factor influencing prediction accuracy

Effect of linkage disequilibrium (LD) and effective population size (Ne)

The choice of the population for genomic selection has a strong impact on the prediction accuracy. The two main population parameters important for accuracy are the effective population size (Ne) and the linkage disequilibrium (LD). LD is the non-random association between the alleles of different loci within a population. Several methods can be used to measure LD (Weir, 1979, 1996; Slatkin, 2008; Russell and Fewster, 2009). The rate of decrease in LD depends in particular on the effective population size (Ne). It corresponds to the size of an ideal population that would have resulted in the same random genetic drift (or genetic diversity) as the actual population. A high Ne often implies a reduced LD and conversely, a low Ne can lead to high LD. Several studies have shown that these two parameters are strongly linked with the prediction accuracy (Figure 1.11) and associated with marker density along the genome influence the quality of genomic prediction (Solberg et al., 2008; Grattapaglia and Resende, 2011; Nishio and Satoh, 2014). It is important to have a certain LD in the population to capture the effects of QTLs via nearby markers. The density of molecular markers can represent an adjustment variable to compensate for the constitutive parameters of the population (Ne and LD).

1.3.5 Molecular marker density

Several studies have investigated the impact of molecular marker density on the accuracy of genomic selection (Romero Navarro et al., 2017; Norman et al., 2018; Kainer et al., 2018). They show that accuracy increases with the number of markers before reaching a

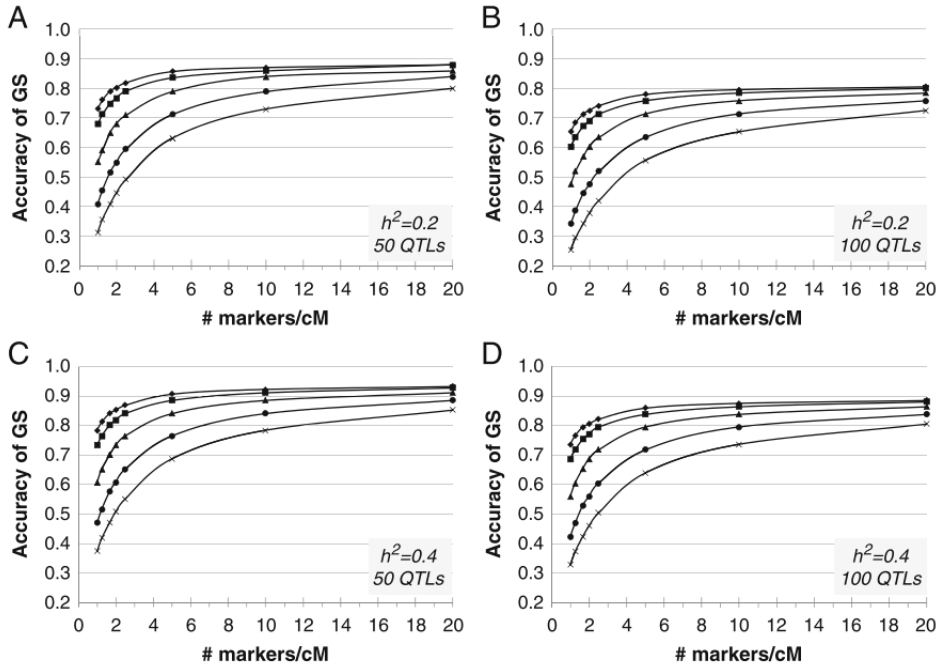


FIGURE 1.11: Accuracy of genomic selection (GS) as a function of marker density (markers/cM) for different combinations of narrow sense heritability (h2) and the total number of QTL controlling the trait(s) with a training set N = 1,000. The plotted curves correspond to effective population sizes Ne = 10 (filled diamond), Ne = 15 (filled square), Ne = 30 (filled triangle), Ne = 60 (filled circle), Ne = 100 (multiplication symbol) (extracted without modification from Grattapaglia and Resende (2011)).

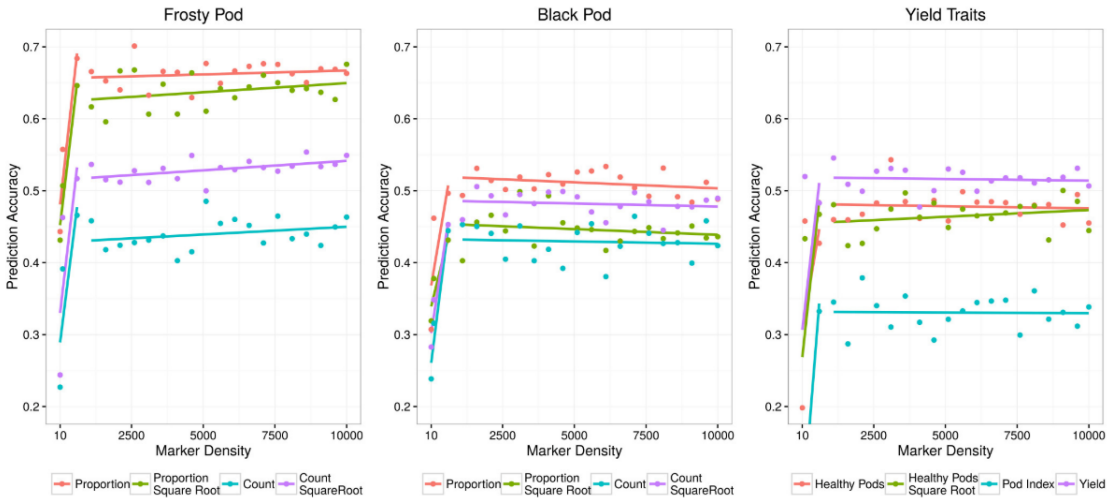


FIGURE 1.12: Genomic selection predictive abilities relative to the number of markers used to estimate each additive relationship matrix (proportion, proportion square root, count, count square root). in Cocoa (Romero Navarro 2017)

plateau (Figure 1.12). This plateau depends typically on traits and population constitutive properties. With increasing access to affordable genomic sequence data, the ability to use the complete genome sequence for prediction is becoming a reality. Several studies have shown that the benefit of densifying genotyping may vary depending on the trait under consideration (Kainer et al., 2018; Zhang et al., 2018). These studies also show that the use of too many markers could eventually reduce the quality of prediction. The relatedness between the training and the validation population is a factor that interacts with marker density to determine the minimum number of markers that are required. Typically, for selection candidates related to the training population, for example descendants of the training population, the minimum marker density will be lower than that required for unrelated individuals (Meuwissen et al., 2009). In addition, sequences and low density genotyping can be combined with the use of imputation to increase the molecular density information in the general population. Some authors, after successfully imputing low and medium densities to high densities, have used the results to improve genomic prediction (Badke et al., 2014; Frischknecht et al., 2014).

Genotypic imputation : Principle and factors influencing the results

The idea of genotypic imputation as additional genotyping data has been described by Burdick et al. (2006), who used the term "in silico genotyping". In this context, imputation refers to the process of predicting genotypic data that is not directly available for an individual. Imputation uses a reference panel composed of individuals with a high marker density genotyping to predict all missing markers affecting another genotyped panel with lower density coverage (Marchini and Howie, 2010). Imputation can be used to correct missing data due to technical or genotyping errors, and to predict unobserved SNPs (Roshyara et al., 2014), or even non genotyped individuals (Berry et al., 2018).

Two basic strategies are used for imputation : either the use of genealogical and Mendelian segregation (Browning and Browning, 2011; Howie et al., 2009; Scheet and Stephens, 2006), or the use of linkage disequilibrium (Daetwyler et al., 2011; Meuwissen and Goddard, 2010). Both strategies can also be combined sequentially (Sargolzaei et al., 2014). The accuracy of imputation depends on similar factors to those relevant to genomic prediction, basically : quality of genotyping, levels of linkage disequilibrium (LD), marker density, and the relationship between the reference population and the imputed population (Hickey and Gorjanc, 2012; Browning and Browning, 2011; Hickey and Gorjanc, 2012; Hayes et al., 2010). Imputation presents several advantages, like reduction of genotyping costs (Huang et al., 2012) and, through the gain in density, improvement of QTL detection and model prediction accuracy (Marchini and Howie, 2010).

1.3.6 Composition of the training population

When implementing genomic selection in a real improvement program, the choice of the training population is essential. The training population usually corresponds to an operational phase of the program, like a parental generation, while the validation set typically corresponds to progenies or collaterals of the training population. Several studies have shown that the accuracy of GS increases with the size of the training population (Jannink et al., 2010; Lorenz et al., 2011; Grattapaglia, 2014). In addition, the relatedness between the training and the validation population is crucial for precise predictions. The more closely the two populations are related, the more accurate the prediction will be

(Pszczola et al., 2012; Ly et al., 2013; Daetwyler et al., 2013; Gowda et al., 2014). The accumulation of several generations into the training has also been shown to be beneficial (Muir, 2007). Many recent studies investigated the way to optimize the construction of the calibration population. These optimization methods are based typically on the choice of individuals that maximize accuracy (Rincent et al., 2012; Isidro et al., 2015; Akdemir et al., 2015; Rincent et al., 2017) or on the use of genetic contributions to design or update the training population Eynard et al. (2017).

1.3.7 The different statistical methods and models

Among the different statistical methods to obtain GEBVs, two main groups can be distinguished : approaches that estimate an additive effect associated with each marker and approaches that directly give the additive value of individuals. More complex models integrating dominance and epistatic effects will be discussed in a separate paragraph.

Since the number of markers (p) is often greater than the number of individuals (n) in the calibration model, the known p>>n problem, estimating the effect of markers by multiple regression using ordinary least squares is not possible. Some kind of variable selection or shrinkage estimation procedure is then required to make such problem solvable (De Los Campos et al., 2009). In recent years, several methods have been developed : derivatives of BLUP methods (Henderson, 1975), Bayesian methods and non-parametric methods (Gianola and van Kaam, 2008; Neves et al., 2012; González-Camacho et al., 2012; Ornella et al., 2014; Howard et al., 2014). One of the simplest formulas for the estimation of GEBV is the following :

$$y = \mu + Za + \epsilon$$

where y is the vector of observations ($n \times 1$), μ the mean of the observations, a the vector of markers' effects ($p \times 1$), Z an incidence matrix ($n \times p$) containing the copy number (0, 1 or 2) of the most frequent allele and ϵ the vector of residuals ($n \times 1$), with n individuals (observations) and p markers (unknowns or parameters). GEBV of individuals can then be obtained as a result of summing up across all markers their respective additive effects in a.

For approaches giving GEBV directly (GBLUP), a basic formula is :

$$y = \mu + Xg + \epsilon$$

where g is a vector $(p \times 1)$ of GEBV, and X a design matrix linking observations to individuals. GEBVs follow a normal distribution, with a covariance between effects modelled through an additive relationship matrix that is derived from markers for GEBV or from pedigree in classical pedigree-based BLUPs.

A wide variety of statistical methods have been developed to handle the problem of greater number of parameters than observations by following different strategies. Roughly, the different strategies are thought to accommodate implicitly different underlying genetic architectures of quantitative traits, from simpler infinitesimal-like architectures to those more complex with a variety of heterogeneous effects across genes. Thus, some statistical methods for GS are based on a constraint that imposses the same variance for all markers (RR-BLUP, GBLUP, etc.), and that corresponds to pseudo-infinitesimal genetic architectures (very large number of genes with very small effects). Other methods impose variable selection and/or shrinkage strategies to reduce the number of relevant parameters to be estimated, and this can be done for instance by using *a priori* information in a Bayesian framework that forces most effects to be close to zero with just a few with larger values. With the Bayes family of methods (A, B, etc) there is the possibility to assume variances that are specific for each marker. This allows greater emphasis to be placed on certain markers with respect to others that will end up with negligible effects. Another somehow simpler alternative is to use an iterative weighted GBLUP (Legarra et al., 2009). Each marker is weighted depending on its own estimated effect, obtained in a previous iteration of GBLUP and the process can be repeated several cycles to narrow the set of selected markers. Some implementations showed similar results to Bayesian methods, with the advantage of being much faster and easier (WANG et al., 2012).

The different statistical methods used in this thesis are described in the material and methods in Chapter 2. Several review studies compared and discussed some of the most common statistical methods used in GE (Jannink et al., 2010; Lorenz et al., 2011; Heslot et al., 2012; de los Campos et al., 2013). Often, the ranking amongst methods depend on the genetic architecture of the trait (e.g.,Hayes et al. (2009b); VanRaden et al. (2009); Daetwyler et al. (2010); Clark et al. (2012). Due to this dependency, no single method emerges as a global solution. Therefore, it is necessary to try at least two methods.

Commonly, authors use one where loci are equally weighted versus another one where markers can have larger differences in effects.

1.3.8 Towards more complex models

Non-additive genetic effects

Non-additive effects in quantitative genetics are well known by breeders and geneticists since the beginning of the 20th century. The infinitesimal model for quantitative traits based on relatives' similarities was first described as an additive model (Fisher, 1919), and then extended to the case with dominance (Fisher, 1919; Wright, 1921). Finally, the model was extended to accommodate epistatic effects (Cockerham, 1954; Kempthorne, 1954). In the context of GS, the interest in non-additive models, according to the number of publications treating the subject (see Figure 1.13), raised significantly from 2016 onwards. First authors proposed to include dominance in GS models by extending the basic model to estimate a dominance effect associated to each SNP marker (Toro and Varona, 2010; Su et al., 2012). However, this method used "observed" heterozygotes, and predicted the "biological" additive genotypic effect instead of predicting individual breeding values (Varona et al., 2018). Huang and Mackay (2016) pointed out that the estimated proportion of variance due to additivity, dominance, and epistasis does not reflect necessarily the "biological" (or "functional") effect of the genes, even if the exercise can be useful for prediction and selection purposes. Vitezica et al. (2013) proposed a reformulation of the model to estimate breeding values and dominance deviations. Methods to take into account epistatic effects were proposed and shown to produce promising results (Vitezica et al., 2017; Martini et al., 2017).

Several studies integrated dominance or epistatic effects in the GS. The results on real datasets showed either no improvement in terms of accuracy (Jiang et al., 2017; Heidaritabar et al., 2014; Gamal El-Dien et al., 2016) even if a non-additive proportion of variance was observed for the traits, or a small improvement in prediction accuracy (Moghaddar and van der Werf, 2017; Aliloo et al., 2016; Tan et al., 2018). This so far limited success may be due to the fact that the populations under study were not big enough nor with an optimal design to reveal the benefits of adding non-additive effects in

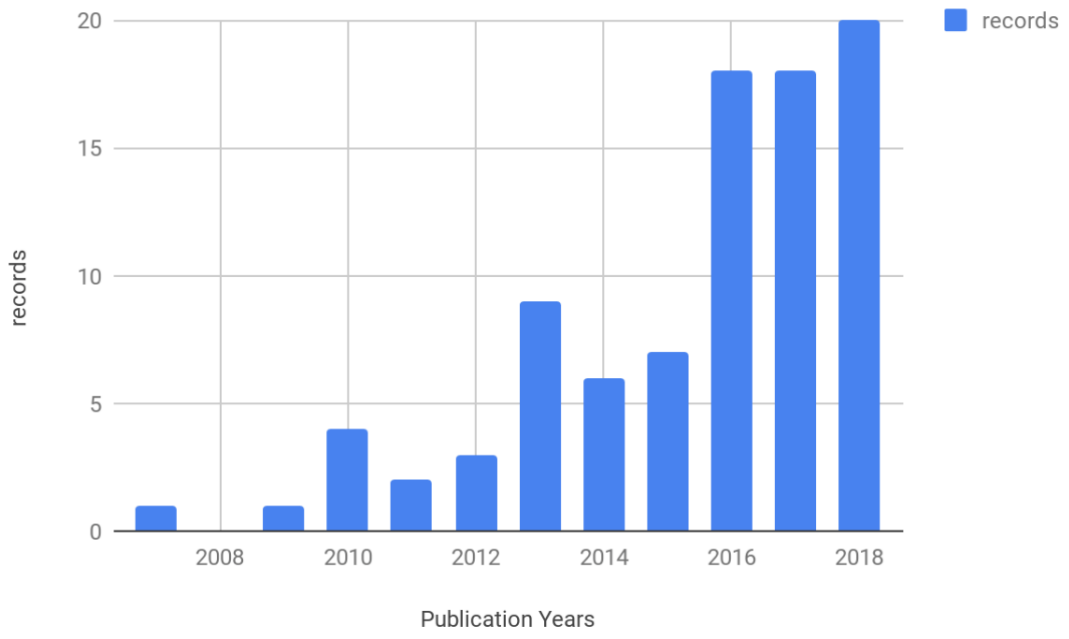


FIGURE 1.13: Number of publications per year referring to non-additive effects in genomic selection. This report reflects citations to source items indexed within Web of Science Core Collection in October 2018.

genomic prediction.

Multi-trait

The majority of genomic selection studies involve a single-trait analysis, while operational selection in a majority of cases comprises several traits. Few have so far highligted the potential of genomic selection with a multi-trait method. This idea has been first addressed by simulation (Calus and Veerkamp, 2011; Guo et al., 2014) and with real data (Jia and Jannink, 2012). Considering several traits jointly with a multivariate analysis using the mixed model framework is expected to benefit from existing genetic correlations between traits to increase prediction accuracy (Gilmour et al., 2008). In other words, the prediction accuracy of a trait with low heritability can beneficiate from the information of a correlated trait with high heritability and a good accuracy. The stronger is the genetic correlation, the greater will be the benefit of using multi-trait approaches (Calus and Veerkamp, 2011; Jia and Jannink, 2012).

However, empirical studies showed contrasting results, with some showing a benefit

of this approach (Marchal et al., 2016; Schulthess et al., 2016), while others resulted in no apparent advantage over a single-trait method (Jia and Jannink, 2012; Dos Santos et al., 2016; Rambolarimanana et al., 2018). These studies were focused on the eventual advantages in terms of prediction accuracy and did not addressed explicitly the operational implications when traits were antagonist. Despite this fact, the multi-trait GS approach may help to shorten selection cycles or combine several selection steps whenever they are done at independent steps.

1.4 Objectives, Opportunities and Challenges of the genomic selection for forest trees

Forest trees are usually known to have high levels of natural heterozygosity, which represents important reservoirs for adaptation and breeding. Trees, compared to other domesticated species, are able to do asexual propagation which is operationally interesting for breeding and evaluation (Bisognin, 2011). On the other hand, trees present typically long juvenile phases (Miller and Gross, 2011), and field evaluations that are costly and time consuming. Two aspects of trees in general and poplars in particular make the advent of genomic selection especially desirable. The first is the fact that, as many other large perennials, poplar breeding relies in lengthy, costly and complex phenotypic evaluation before selection and mating can take place. One example comes from resistance to rust that usually needs to be evaluated in two phases. In a growth chamber phase, resistance is evaluated against different races of the pathogen to identify specific and general resistance factors ; and in a field phase, rust tolerance as the impact of rust attacks on growth is evaluated in specific costly experiments. Another example of a difficult-to-evaluate trait, common to several forest species, is wood quality. This undermines the screening capacity and the possibility of high selection intensities to take benefit of the large diversity that is available. The second is the fact that improved poplar varieties take the form of cloned genotypes, whose selection relies on lengthy crosses and large segregation families in order to benefit from non-additive genetic effects at the extreme genotypes. This again undermines the capacity to evaluate a large panel of breeding stock. Genome-based selection is expected to bring higher precision at earlier stages, thus to allow for

shortenings in selection cycles. Breeding programs like that of poplar are based on multiple-trait selection, where genomic selection could allow for selection at the seedling stage of quality traits at once, which are otherwise evaluated at 10 years or more at different steps.

Recent studies of GS in forest trees were conducted on several species, like eucalypts, spruces and pines (Resende et al., 2017; Tan et al., 2017, 2018; Gamal El-Dien et al., 2016; Lenz et al., 2017; de Almeida Filho et al., 2016). The results synthesized and discussed by Grattapaglia (2017), reflect a large range of prediction accuracies (Table 1.3). Even if the numerous proof-of-concepts so far of genomic selection in forest trees pinpoint to promising perspectives, some caveats remain. First, most of the proposed models in GS for trees assume additive effects, with still few attempts with clear-cut results of non-additive models. The implementation of non-additive models have seen so far rather small benefits (Muñoz et al., 2014; Gamal El-Dien et al., 2016; Tan et al., 2018). However, for clonally propagated crops as poplar, dominance and epistatic effects are to be considered as potentially relevant for selecting elite genotypes. It is important, therefore, to gain understanding on the factors behind the limiting benefits of non-additive approaches so far, whether constitutive to the underlying architecture, linked to the modelling approaches or even due to design and size issues in the experiments. Secondly, we have seen that linkage disequilibrium is important to ensure a good prediction accuracy. In most cases, forest trees are highly heterozygous, with LD patterns that break down at really short distances, i.e. between 50 and 300 base pairs for *P. nigra* (Chu et al., 2009; Marroni et al., 2011). This means that, to compensate for the short LD patterns in our species, more markers those usually at hand are needed in order to increase the likelihood of genomic prediction performances. Thirdly, as for all species, the genetic architecture of the trait of interest affects prediction accuracy (Nakaya and Isobe, 2012). For complex highly polygenic traits, typically large training populations are required to cover well all genotypic combinations, with a proper design to be able to capture interactive effects (Jannink et al., 2010). However, most of the training populations in use so far could well be rather limited by operational constraints, undermining the generality and power of the genomic selection experiments that are available.

This thesis was developed in the framework of a project aiming at studying the feasibility of genomic selection within the black poplar improvement program taking into

consideration the simultaneous management of the genetic diversity. This thesis more particularly focused on some issues for which GS could *a priori* bring an efficient solution, notably for optimizing the selection of best individuals from crosses, for integrating more efficiently multiple-trait selection into a more simple step, and for gaining in cost-benefit efficiency while incorporating new schemes of phenotypic evaluation and genotyping. Black poplar is one of the parents of *P.* × *canadensis* which is one of the most economically important hybrids for plantations worldwide. Moreover, phenotypic and genomic resources together with knowledge have been accumulated in the last decades as basis for fundamental science and praxis for support of the breeding program. The expectations for genomic selection are first and foremost a significant reduction in the evaluation cycle. Currently, between 15 and 20 years are required for a cultivar to be available for production. Genomic selection appears as an appropriate tool to shorten our selection cycle by several years or even by halving the evaluation cycle. Time savings could be achieved in several ways : (i) by combining genomic and phenotypic selection cycles without creating varieties to accelerate intraspecific recombination and prospects of gain ; (ii) by centralizing evaluation efforts on a single easily managed training population in order to mobilize promptly ressources with each new demand, either economic or adaptive.

One of the main weaknesses of the current breeding program is the lack of precision and efficiency in the selection of individuals at the very early stage of sprouted plants. It is also at this stage where most of the genetic variation is available. Selecting with low efficiency at such early stages can easily weed out useful variation. The possibility to select on several traits at once at early stages would limit the losses in selection intensity that are expected from selections conducted subsequently and at independent levels. The transition to genomic selection would also allow better prediction of low heritability traits by integrating non-additive effects and multiple-trait approaches. Ultimately, the improvement program could integrate genetic diversity management based on the knowledge of markers. Although some of these forecited prospects would require a specific experimental approach for careful assessments, the present proof-of-concept study was a first attempt to quantify the feasibility of GS under different scenarios of operational interest. This multifaceted aim can be further developed into three basic questions :

The first question concerns the extent to which GS could be a guarantee of the genetic

diversity that is readily available in the breeding program, and for which is known also the species. Assuming that the genotyping that is required for GS could also serve to monitor explicitly the genetic diversity, the implementation of GS would readily facilitate the incorporation of genetic diversity management schemes. Moreover, if the prospect of speeding up genetic progress with GS is a reality for our species, such a rapid turnover of generations could easily accelerate the erosion of diversity per unit of time. The need to take into account short and long term gains is then essential. Despite of the interest, this point could not be fully developed for the thesis due to lack of time. It was taken as an important point for the discussion and perspectives at the end of the thesis, in the Discussion chapter. A second question concerns the integration of available information to improve the access to quality genotyping, which is ultimately at the basis of quality predictions. We studied the benefits of imputation from low density to sequences, and in order to increase the number of available markers and improve also the homogeneity of the coverage across genomes. Such a system would allow to increase the efficiency of genotyping at low costs, making GS eventually more competitive with respect to pedigree-based evaluation. The first article of this thesis (Chapter 3) presents the feasibility of imputation under the extreme demand of reconstructing sequences from low density genotyping, assesses the different factors affecting imputation quality and the eventual consequences in the imputed genotypes in terms of linkage disequilibrium and annotation profiles.

A third question concerns the feasibility of GS compared to the pedigree-based counterpart. Such feasibility was studied under different aspects, in order to identify the conditions for which the new methodology could be competitive, whether with the help of multiple-trait approaches, with non-additive modelling, assuming different genetic architectures across different traits, through marker densification, or using phenotypes with varying repeatabilities. We devised different scenarios to challenge genomic selection that mimicked operational situations, and used different criteria to scrutinize the advantages of GS. A second paper (Chapter 4) compiles most of these aspects, taking advantage of the available data from the black poplar population. Some aspects like the feasibility of genomic selection for early stages were not covered explicitly, although some of the tested scenarios could shed some light on the issue. In light of the results presented

in Chapters 3 and 4, Chapter 5 will discuss how and where to implement GS in the French black poplar breeding program in order to increase efficiency. Also at this chapter, the management of genetic diversity will be addressed. More generally, the prospects of GS in poplar will be discussed, giving also a few elements of cost-benefits.

TABLE 1.3: Genomic evaluations (GS) in forest trees.

Species or Genus	Population size	Number and type of markers used	Trait Heritability variations	Predictive ability	References
Eucalypts (*E. grandis*, *E. urophylla*, *E. globulus*, Hybrids)	738 - 1126	3K DArT fixed array 40K SNP from EuCHIP60K fixed array	0.22-0.93	0.05-0.72	[1]
Loblolly pine (*Pinus taeda*)	165-951	3K-5K SNP Infinium chip	0.11-0.95	0.17-0.86	[2]
White spruce (*Picea glauca*)	1694 - 1748	6K SNPs Infinium chip	0.04-0.57	0.33-0.79	[3]
Intereior spruce (*Picea glauca* × *Picea engelmannii*)	769-1126	50K-60K SNPs by GBS after imputation	0.29-0.98	0.47-0.77	[4]
Maritime pine (Pinus pinaster)	661-818	2.5K - 4K SNPs Infinium chip	0.17-0.30	0.38-0.82	

[1] (Resende et al., 2012a; Lima, 2014; Tan et al., 2017, 2018) ; [2] (Resende et al., 2012b; Resende, 2012; Zapata-Valenzuela et al., 2012, 2013; Muñoz et al., 2014) ; [3] (Beaulieu et al., 2014a,b) ; [4] (Gamal El-Dien et al., 2015; Ratcliffe et al., 2015) ; [5] (Isik et al., 2015; Bartholomé et al., 2016)

Chapitre 2

Materials and Methods

2.1 Résumé du Chapitre

Cette thèse concerne le matériel du programme d'amélioration du peuplier noir français. La population utilisée se compose de vingt-trois parents comprenant onze mâles et douze femelles avec 1000 descendants répartis en trente-cinq familles. L'étude de la faisabilité de la densification du génotypage puis de la sélection génomique chez le peuplier noir a d'abord été réalisée sur une partie du jeux de données pour permettre de mettre en place des pipelines d'analyse de données, de l'imputation à la sélection génomique. Au fur et à mesure que les données de génotypage et de séquençage sont devenues disponibles, d'abord en novembre 2016 (deuxième essai), puis en septembre 2017 (jeu de données complet), les modèles ont été de nouveau testés pour affiner notre choix de méthodes et logiciels.

Dans un premier temps, une étape d'imputation a été menée afin de densifier le génotypage tout le long du génome. L'ensemble des individus ont été génotypés avec une puce 12K d'Illumina (Faivre-Rampant et al., 2016). Quarante-trois des individus nodaux ont été entièrement séquencés (référence), tandis que la majorité restante (cible) a été imputée de 8K à 1,4 million de SNPs en utilisant le logiciel FImpute (Sargolzaei et al., 2014). Chaque SNP et chaque individu ont été évalués pour les erreurs d'imputation au moyen d'une validation croisée. Certaines statistiques telles que la "p-value" exacte du test d'équilibre de Hardy Weinberg (Wigginton et al., 2005), la qualité du séquençage, la profondeur du séquençage par site et par individu, la fréquence des allèles mineurs, le rapport de densité des marqueurs ou la redondance des informations SNP (Speed et al.,

2017) ont été calculés. Des analyses en composantes principales et des analyses Boruta ont été utilisées sur tous ces paramètres pour classer les facteurs affectant la qualité de l'imputation. De plus, nous caractérisons l'impact de la relation entre la population de référence et la population cible.

Suite à cette densification, un test de sélection génomique a été mis en place. Dans cette étude, nous avons tenté de comparer l'évaluation génomique à l'évaluation traditionnelle fondée sur le pedigree et d'évaluer dans quelles conditions l'évaluation génomique surpasse le pedigree classique. Plusieurs conditions ont été testées comme la constitution de la population formatrice par validation croisée, la mise en œuvre de modèles multi-traits, mono-caractères, additifs et non-additifs avec différentes méthodes d'estimation (G-BLUP, G-BLUP pondéré ou BayesCπ), enfin l'impact de la densification par imputation a été testé à travers quatre jeux de données avec différentes densité de marqueur (7K, 50K, 100K et 250K SNP).

Nous avons utilisé sept caractères évalués au sein de quatres séries d'évaluations à la même localisation sur différentes années. La qualité de la prédiction est évaluée avec le calcul de la précision, de la corrélation de rang du Spearman et du biais de prédiction. Ces valeurs sont évalués à travers une stratégie de validation croisée et testée avec une partie du jeux de données, un jeu de données indépendants à également été utilisé.

2.2 Plant material, Phenotypic and Genotypic Data

2.2.1 Breeding designs and families selection

This thesis involved French black poplar breeding program material. It started with a four by four factorial mating design, created between 1990 and 1995 by Marc Villar (UMR BioForA), representing fourteen families for a total of 468 progenies. Twenty additional families from a double pair mating design from the GIS Peuplier (Groupement d'Intérêt Scientifique) were available. Ten additional families were obtained in spring 2015 to reinforce the connection between both mating designs. A total of thirty-one parents involving fifteen males and sixteen females with 1700 progenies were at our disposal for this thesis (Table 2.1). The aim here was firstly to increase the population at our disposal, secondly to be the closests to the breeding population by including new relevant crosses,

	SRZ	BDG	71077-308	92510-1	72145-007	72131-017	73182-009	73193-056	72131-036	3824-3	71034-2-406	72146-11	H487	72144-009	72159-004	72156-003
VGN-CZB25	(10-11)10B 55	(10-11)11B 57	(10-11)12B 54	(10-11)13B 34+36	1014E 65	1015E 98	1016E 66	1017E 65								
71041-3-402		1211B 28	1212B 11	1213B 21	1214E 40		1216E 36									
71072-501	1310B 25	1311B 28 + Famille carto	1312B 29													
SSC	1410B 15	1411B 20	1412B 20	1413B 34												
71040					GIS88 24	GIS87 20										
662200037					GIS86 25	1515E 116	1516E 138	1517E 86								
73193-089						GIS96 25	GIS79 32	GIS89 20								
662200216							GIS84 110		GIS85 20							
71069-914										GIS70 22						
73193-091							GIS90 62			GIS69 21		GIS68 30				
H480										GIS61 13	GIS60 20					
71036-2-401								GIS67 16								GIS63 84
72131-001							GIS66 20									
72143													GIS62 44			
72160														GIS65 40		
72149-029															GIS82 20	

Table 2.1: Mating design *P. nigra* × *P. nigra* : factorial design tested in 1999 in Guemene-Penfao in green, GIS double pair matings tests in blue, and new families created in of 2015 in orange. Underlined families are phenotypically unevaluated families. The color gradient represents the priorities in the selection of individuals for genotyping, knowing that green and orange are essential and that blue is extra, dark blue a priority over light blue, and families in white are optional families.

and thirdly the have a good balance between a sufficient number of families and families of relatively large size.

The priorities of genotypic information acquisition were represented by a color gradient in table 2.1. 15 families have been considered as essential whereas 14 families were extra (light blue and white) , dark blue being the highest priority and white the lowest priority due to the fact that one or both parents were already well represented or because they were less connected with the others families. Among families with a large number of progenies, one was chosen to predict mendelian segregation within families from relatives information (family 1516E).

Not all families provided phenotypic and genotypic information. In table 2.1, available genotypic and phenotypic data concerned reduced factorial mating design (8 families and 294 individuals), genotyping data remains to be acquired concerning all the other families (27 families and 740 individuals). Phenotypic data were available for most families except for 12 families and 907 individuals (underlined ones). Given that families had in general more members than those finally used in the study, a sample was done to represent the intrafamilial variability at phenotypic level (see next section).

Individual sampling for the study

Individuals were primarily chosen to valorize existing genotyping and phenotyping data on *P. nigra* × *P. nigra* factorial mating design. A limited part of the factorial mating design (dark green Table 2.1) had already a large set of phenotypic and genotypic data available and were selected by default. This subsample comprised 268 individuals who were insufficient to evaluate GS feasibility. To improve the reliability of our results, the number of individuals in our study was increased. Additionally, new crosses within the factorial mating design were selected, which were not genotyped initially, because of their high rust sensibility. Their inclusion was assumed to be interesting to increase discrimination for the trait of rust sensibility.

Some extra families were chosen from outside the factorial depending on their availability of phenotypes and their relatedness with already selected families, and in order to have a minimum of 2 Full-Sibs (FS) families per parent. Twenty individuals were sampled within each family. For each family with more than twenty individuals, 10 000 random samplings were performed among all available individuals in the cohorts and based on phenotypic data. For each sample, the difference between the variance for the sample and the variance within the whole family was calculated by trait. The same was calculated for covariances between traits. All components were summed up to get a criterion for the choice of representative sampling, as shown by the equation 2.2.1 :

$$criteria = (var_{si} - var_{pi})^2 + (covar_{sij} - covar_{pij})^2 \tag{2.1}$$

where var_{si} is the variance within subsample for the trait i,var_{pi} the variance within the family for the trait i, $covar_{sij}$ the covariance between the trait i and the trait j within the subsample, and $covarp_{ij}$ the covariance between the trait i and the trait j within the family.

The samplings were ranked in ascending order, and the first ten were selected. Finally, a graphical visualization of the dispersions (figure 2.1) was used to choose which sample best represented the dispersion within the families.

In summary, the designed population used in this thesis corresponded to a pedigree

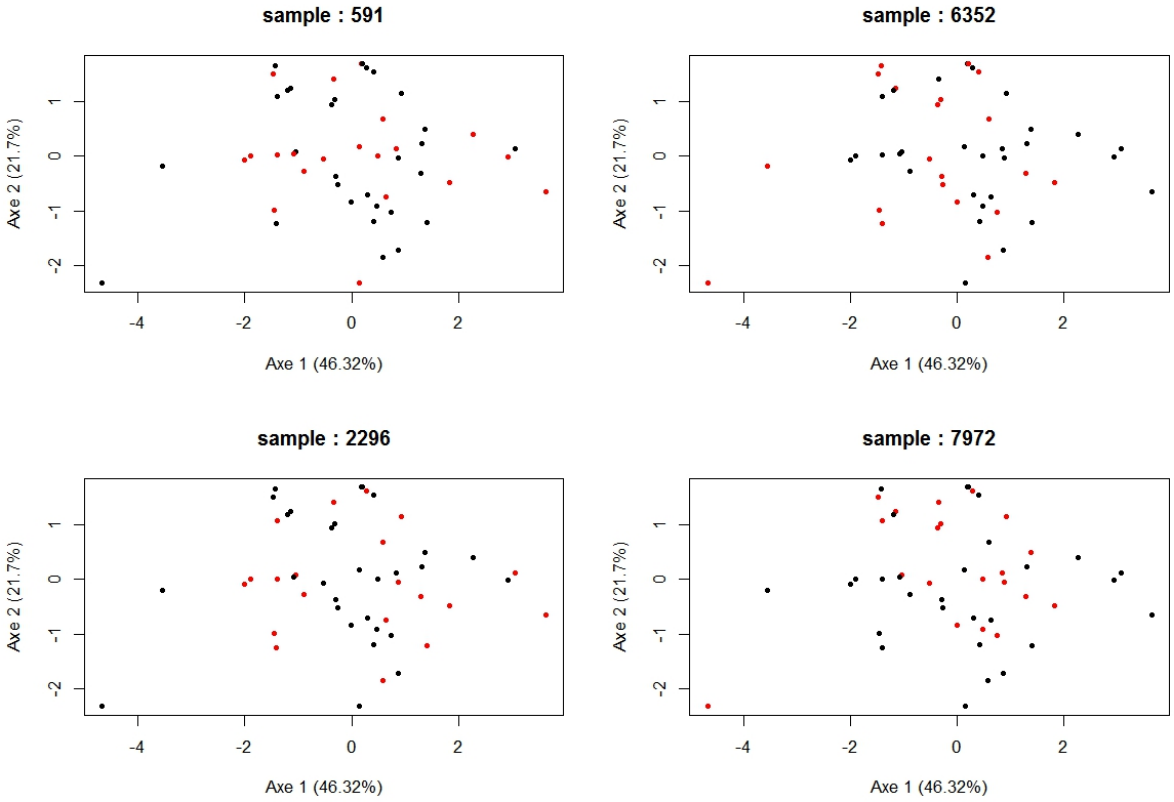


Figure 2.1: Four of the best sampling graphical visualization of the individual dispersions used to choose the best sample within the family GIS68.

	SRZ	BDG	71077-308	92510-1	72145-007	72131-017	73182-009	73193-056	72131-036	3824-3	71034-2-406	72146-11
VGN-CZB25	(10-11)10B 55	(10-11)11B 57	(10-11)12B 54	(10-11)13B 32		1015E 34	1016E 14	1017E 30				
71041-3-402		1211B 28	1212B 11	1213B 17	1214E 30		1216E 25					
71072-501	1310B 25	1311B 28	1312B 29									
SSC	1410B 15	1411B 20	1412B 20	1413B 22								
71040					GIS88 24	GIS87 20						
662200037					GIS86 25	1515E 32	1516E 118	1517E 31				
73193-089						GIS96 22	GIS79 20	GIS89 18				
662200216							GIS84 31		GIS85 19			
71069-914										GIS70 22		
73193-091										GIS69 21		GIS68 30
H480										GIS61 13	GIS60 19	

Table 2.2: Selected individuals among crossing design *P. nigra* x *P. nigra* : compound factorial design tested in 1999 in Guemene-Penfao in green, GIS2012 tests in pink and GIS2014 in blue and crossovers of 2015 in orange.

of one grand-parent, 23 parents and 1011 progenies. Individuals were structured into 35 full-sib cohorts, 14 from the "4 by 4" factorial mating design and 21 from a series of multiple pair-mating designs. Family size ranged from 10 to 118, with an average of 26 individuals per family. The details of the final set are presented in the table 2.2.

Controlled crosses and installation of field evaluation

The great aptitude for vegetative propagation in poplar greatly facilitates controlled crosses, notably in *P. nigra*. Male and female flowering shoots are collected during winter with a pole or a rifle. The branches obtained are placed in plastic bags and stored in a cold chamber to ensure the cold requirements for the break-up of dormancy. During March, the male flowering shoots are placed in jars with water and in cages (Figure 2.2A). On average about thirty male flowers were used for the pollen collection at a rate of five flowers for large circumference branches and one to two flowers for the thinnest ones. Flowering budburst occurred and polled was collected on a sheet from male flowers. The pollen was then stored in a filter paper placed on silica gel for drying up. The pollen was either stored in the refrigerator in this way for use within the year or stored in a tube and placed in a freezer at -20°C for later use. A pollen germination test was carried out before pollination by placing it on agar medium at 25°C for 24 hours and examining the number of pollen tubes under a binocular magnifier.

The female flowering shoots were also taken out of the cold chamber and placed in pots with holes to let in water input (Figure 2.2B). As the catkins develop, and before the

flowers were fertile, a number of 30 were kept in cages to avoid genetic pollution. Around the fertility peak, pollination took place. To do this, the pollen was placed in a petri dish placed in another petri dish where a fine stream of water is present in order to slowly re-wet it with the help of ambient humidity. The pollen was collected and deposited on female flowers with a brush. The area around the female shoots was humidified to pin down the flying extra pollen and the cage put back in place. The manual fertilization step was done once or twice per day around the fertility peak (when the catkins are open and when the flowers move away from each other). Once fertilization was successful, the capsules develop on the catkins. They ripened after 15 days to 1.5 months. Once mature, the capsules opened to release the cotton and seeds. Before the release was complete, paper cups were placed on the catkins to collect the seeds and cotton. The seeds and cottons are harvested and then separated. Between 100 and 150 seeds were mixed in boxes containing water-soaked blotting paper to allow germination.

Once germination was complete, the seedlings were collected and planted in mini-moots of compost, and then raised in mini-greenhouses to maintain a humidity level propitious to their growth. They were then repotted following several stages before being raised in greenhouses for the production of cuttings in Orléans nursery, as part of the additional families for the thesis, or in the nursery of Guéméné-Penfao for the GIS families. The aim was to have at least 6 cuttings of good circumference. The slicing of cutting took place around the second half of January. The cuttings were stored in a cold room in plastic bags hermetically sealed at 2°C. The cuttings were soaked 24 hours before being placed on the ground at the end of March or at the beginning of April, in a six randomized complete block design.

2.2.2 Phenotypic data

Because of their availability, their faster and cheaper acquisition, only juvenile traits were used for this thesis. Field evaluations corresponded to four different measurement campaigns. Recorded traits were height (at 1 and 2 years), circumference (at 2 years, as proxy for growth), budburst, bud set, rust resistance (at 1 and 2 years), defoliation, and the branch angle as an architecture assessment. The study used traits assessed on all plants, being already of interest for breeding, and for which we expected to have different

Figure 2.2: Steps of controlled crossings at the implementation of the mating design. 1. Male flowering branches in cages to collect pollen; 2. Macro-photography of a male black poplar flower; 3. Pollen in a petri dish; 4-5. Female flower shoots placed in pots to induce floral budburst; 6. Macro-photography of a female black poplar flower receptive to pollen; 7. Female containment cage; 8. brush fertilization of a catkins with female flowers; 9. Female flowers after fecundation that have passed the pollen receptivity period; 10. Capsule catkins after successful fertilization; 11. release of cotton and seeds; 12-13. De-cottoning steps; 14. Black poplar seeds in a petri dish with water-soaked absorbent paper; 15. Post-germination seedling cotyledonary stage; 16. Seedling after transplanting at the cotyledonary stage; 15. 2nd to 3rd leaf stage seedlings after transplanting; 18. Seedlings of a dozen leaves; 19. Last step of potting for greenhouse growing; 20. Greenhouse plants after 3 months of growth; 21 Evaluation batches 2 months after the rooting of cuttings and 22. Evaluation batches after two years of growth. Credits : Marie Pegard except for 15 : Marlène Lefèbvre (INRA) and 22 : Phillipe Poupart (ONF).

	Years 1999-2000	Years 2012-2013	Years 2014-2015	Years 2017-2018	Number of records
Rust year 1	1	1	1	1	4
Rust year 2	1	1	1	1	4
Defoliation	1	0	0	0	1
Height year 1	1	1	1	1	4
Height year 2	1	1	1	1	4
Diameter year 1	1	1	1	0	3
Circumference year 2	1	1	1	1	4
Budburst	0	1	1	1	3
Budset	0	1	1	1	3
branch angle	1	1	1	1	4

Table 2.3: Summary of available phenotypic data for each evaluation batches.

genetic architectures and heritabilities. Traits were recorded in the same geographical location at Guéméné-penfao nursery but at different years and in different plots. The table 2.3 summarised the availability and the year of measurements. The first phenotyping campaign during 1999 and 2000 involved the factorial mating design with a total of 14 families and 413 offspring phenotyped. From the second campaign during 2012 and 2013, 126 phenotyped individuals in 6 families were evaluated. From the third campaigns in 2014 and in 2015, 105 in 5 families were evaluated. Lastly, the 10 additional full-sib families were phenotyped between 2017 and 2018. In total, 367 individuals were phenotyped in this last period. To complete phenotypic data, a budburst evaluation of factorial design individuals at the INRA clonal park (Orléans) were performed during spring 2016.

Data acquisition protocol

The protocol for the data acquisition was the same for all the evaluations. Measurements took place in the first and second year of growth. The notation of the rust took place in early September when the rust pressure was maximal. Rust resistance was assessed with scale notations from 1 (no symptom) to 9 (generalized symptoms) at year one (rust1) and year two (rust2) (Figure 2.3). The growth was measured during the winters of the first and second year of growth. Growth was assessed for stem circumference and height. Stem circumference was considered at 1m for the second year (circ2). Height was assessed with a graduated rod after one (heigth1) and two years of growth (heigth2). During the

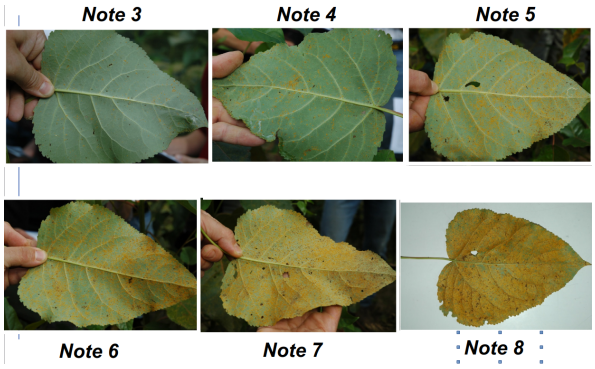


Figure 2.3: Melampsora foliar rust notation in nursery. 1 = no sore ; 2 = 1 to 10 sores on less than 50% of the leaves; 3 = 1 to 10 sores on more than 50% of the leaves; 4 = 11 sores with 50% coverage on less than 50% of the leaves; 5 = 11 sores with 50% coverage on more than 50% of the leaves; 6 = 50 to 75% of the leaf area covered on less than 50% of the leaves; 7 = 50 to 75% of the leaf area covered on more than 50% of the leaves; 8 = more than 75% of the leaf area covered on less than 50% of the leaves; and 9 = more than 75% of the leaf area covered on more than 50% of the leaves. Credits : Poplar GIS.

spring of the second growth year, budburst phenology of the main stem terminal bud was evaluated by measuring its kinetics (every 3 or 5 days from March to April) from a six-class rating scale rated from 0 to 5, where stage 0 corresponded to a completely closed bud and stage 5 when stem internode elongation started (Figure 2.4 ; Castellani et al. (1967)). A local polynomial regression model was used to predict a mean date for each stage from 1 to 4 stages, even if not all were observed for the same given clone. Estimates were in Julian days and given for stage 3 to assess individual susceptibility to late frosts (Howe et al., 2000). Mean branch angle was scored on proleptic branches at two years old with a 1 to 4 scoring scale (angbranch), where score 1 was given to the narrowest angle between the branch and the trunk and score 4 to the widest angle (Figure 2.5).

Phenotype adjustments

All seven phenotypes were independently adjusted to field micro-environmental heterogeneity with the breedR package (Muñoz and Sanchez (2018), implemented in R3.3.1 platform). We used an individual-tree mixed model over all 4 evaluation batches with random effects to fit bi-splines surfaces (Cappa and Cantet, 2007; Cappa et al., 2015), which were nested to each evaluation batch (field experiment) (Supplementary material). The model comprised all phenotyped individuals in the trials, involving

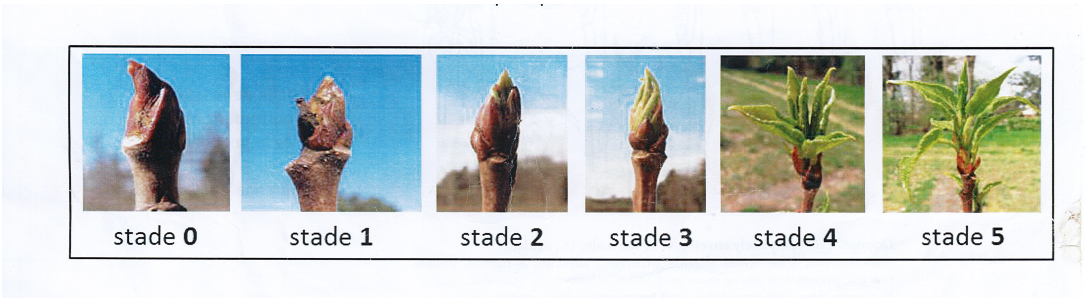


Figure 2.4: Illustration of the six-classes measurement scale for vegetative budburst in black poplar. Measurements of young plants aged 1 or 2 years in nurseries on the terminal bud of the main axis. (Source : Catherine Bastien)

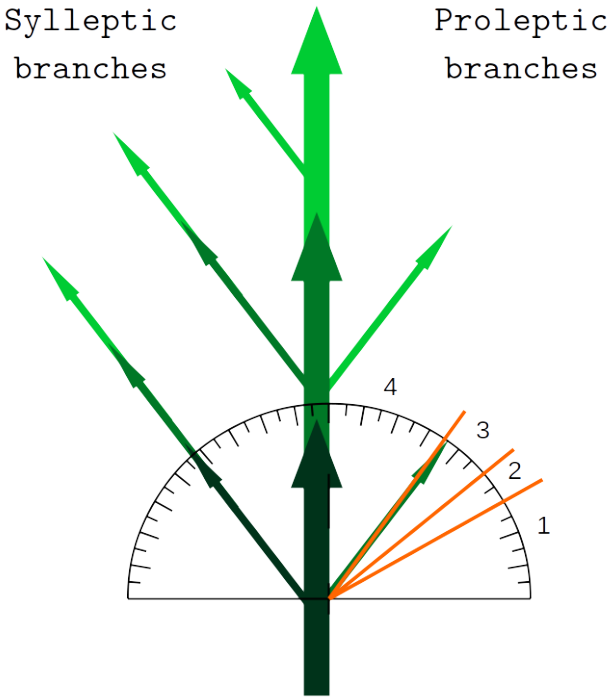


Figure 2.5: Young tree architecture with sylleptic branches (growth in the same vegetative cycle than trunk) at the left and proleptic at right. The branch angle is measured on proleptic branches according to a protractor's scale and a score is assigned between defined radiations. 1: between 0°and 30°; 2: between 30°and 40°; 3: Between 40°and 55°; 4: and Between 55°and 90°.

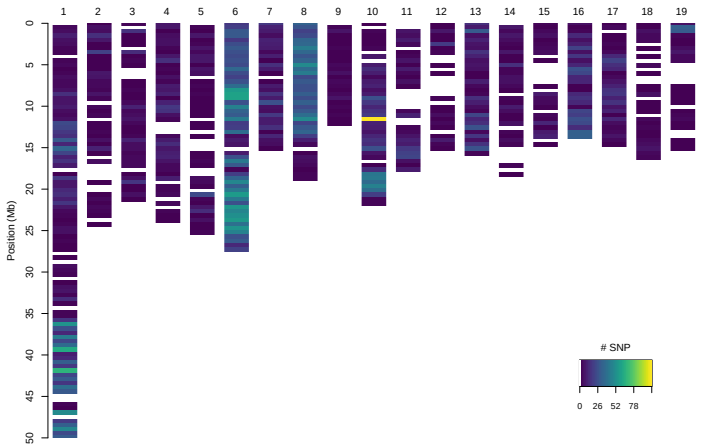


Figure 2.6: Marker density of 7540 SNPs from the chip in 500kb non-overlapping windows.

genotyped and non-genotyped individuals, and according to a single-step formulation (Legarra et al., 2009). The use of all information in field trials with minimum gaps allowed to predict with a maximum of precision the micro-environmental individual effect. This effect was subtracted from the observed phenotype, and the resulting spatially adjusted phenotype was used as raw phenotype for the rest of the study.

2.2.3 Genotypic datasets

Whole genome sequences

Available data : In the population, fourteen parents were already sequenced by Genome Analyzer IIx from Illumina. The SNP detected in this latter paper were used to design the Populus nigra 12K custom Infinium Bead-Chip (Illumina, San Diego, CA) (Faivre-Rampant et al., 2016). The SNPs are not evenly spaced (Figure 2.6), the range of density varied from 5 SNP/Mb to 80/Mb.

Sequences acquisition : For the others parents (22), 1 grandparent, 14 progenies and 6 unrelated, Illumina paired-end shotgun indexed libraries were prepared from one µg of DNA per accession, using Illumina TruSeq®DNA PCR-Free Sample Preparation kit. Briefly, indexed library preparation was performed with DNA fragmentation by AFA (Adaptive Focused Acoustics™) technology on Covaris focused-ultrasonicator, all enzymatic steps and clean up were realized according to manufacturer's instructions.

Single or dual indexes were used. Final libraries were quantified by using qPCR using KAPA Library Quantification Kit and Life Technologies QuantStudio™Real-Time PCR system. Fragment size distribution of libraries was assessed by High Sensitivity DNA assay either on Agilent 2100 Bioanalyzer or on Caliper LabChip®GX nucleic acid analyzer. Equimolar pools of multiplexed samples, up to 11, were engaged in sequencing using 4 lanes. After clusters generation on CBot, paired-end sequencing 2 × 150 sequencing by synthesis (SBS) cycles was performed either on an Illumina HiSeq®2000/2500 running in high output mode (one lane) or on Illumina HiSeq®4000 (three lanes).

Polymorphism detection from sequences : The same protocol was used on the available and new sequences. Reads were trimmed with Trimmomatic (v. 0.32) (Bolger et al., 2014), and mapped to the *P. trichocarpa* version 3.1 genome (Tuskan et al., 2006) using BWA-MEM 0.7.12- with default parameters (Li, 2013). Picard Tools (v. 2.0.1) (http://broadinstitute.github.io/picard) were used to remove duplicated reads. Local and Indel realignments were performed using a Genome Analysis Toolkit (GATK v. 3.5) (DePristo et al., 2011; McKenna et al., 2010). The variant detection was performed on all individuals by two variant callers: (1) in parallel with Freebayes (V1.0.0) (Lajoie et al., 2013) and (2) by each individual separately with GATK HaplotypeCaller, to be subsequently assembled using GenotypeGVCFs (called later gVCF-GATK). The VCFtools 0.1.15 (Danecek et al., 2011) was used to filter variants with no missing data, with a minimum quality score of 30 and a min depth of 2. Among selected SNPs, three alleles were allowed, because mapping was done on another *Populus* species reference genome, so it was possible to have two alternative alleles and no reference allele in the aligned sequences. Finally, only SNPs and Indels that were detected by both callers and whhich were consistent with Mendelian segregation were kept. To simplify, SNPs and Indels (Insertion/ Deletion were both called SNPs hereafter. In average, 91.7% of reads were mapped, 76.5% were paired and only 2.2% were singletons. The genome coverage was calculated by individual, and it varied between 4X and 52X, with a mean coverage of 13X (Supplementary data).

Low-density genotypic data

For the low-density genotypic data acquisition, two strategies were considered. Either extra individuals could be genotyped by the same chip, which is best in terms of compatibility with previous data but expensive, or genotyping by sequencing (GBS). This second option was cheaper but required time costing bioinformatics analysis and protocol adjustments to ensure a good quality of sequencing. A total of 54 individuals and 64 sequences (some repeated individuals) were genotyped by GBS with a Miseq Illumina platform. It was found that few SNPs were in common between the Chip and the GBS: by aggregating individuals, only 1072 common SNPs with many missing values between individuals were obtained. Because of lack of time for further refinement and somehow poor results, chip genotyping was used instead of genotypinge the extra individuals. Therefore, all individuals were genotyped using the Populus nigra 12K custom Infinium Bead-Chip (Illumina, San Diego, CA) (Faivre-Rampant et al., 2016). The original already published run of genotyping comprised 6 parents, 1 grand-parent and 261 progenies in the factorial mating design. The rest of the parents and progenies were genotyped with the same chip in a second new run, with data availability in two batches: november 2016 and september 2017.

Genotype Data filtering and cleaning

Genotyping chip data cleaning and verification : The concordance between the two genotyping runs was checked with a common control. This indicator showed that the difference between the two chips was 2.18 % with 7806 common SNPs. The presence of duplicate individuals allowed for verification between batches by estimating the proportion of similarity between each pair of samples (equation 2.2.3) :

$$k = 1m(2 - |g_{ik} - g_{jk}|)/2m \tag{2.2}$$

where m is the number of SNPs coded 0, 1, 2, i and j are the sample and k the SNP. We used Plink software (Purcell et al., 2007) to estimate IBS. 43 individuals were identified as replicated. Some verifications have been done for all of these replicates and

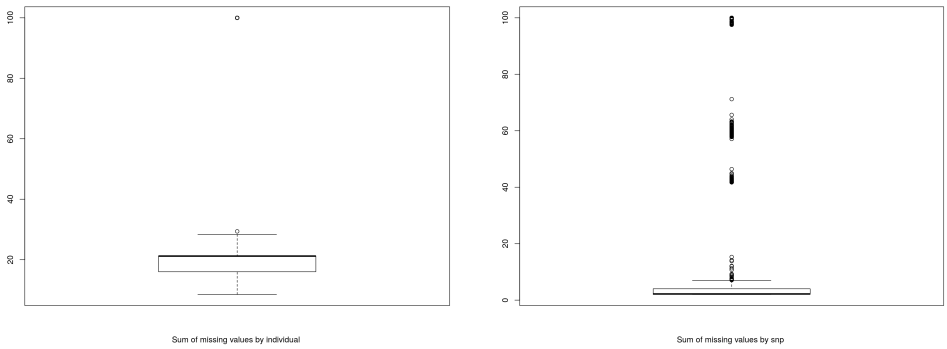


Figure 2.7: Sum of missing values in the chip data by individual (left) and by SNP (right)

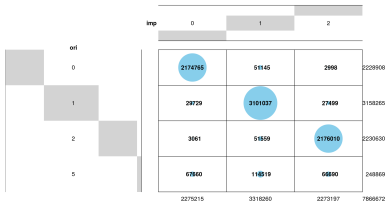


Figure 2.8: FImpute genotypes allelic dose (0, 1, 2) corrections after imputation, missing values are coded 5.

decisions have been taken according to these analyses: for each replicate either the sample was kept and merged or dropped. 10 of them were dropped because it was impossible to know which one was the right one. The proportion of missing data by individuals and by SNP was represented in figure 2.7.

The individuals and SNPs with more than 80% of missing values were removed. Figure 2.7 showed the proportion of missing data by individual, ranging from 0 to 17% with an average of 7%. Whereas for SNP, the proportion was mostly around 0, with some SNPs reaching 70% of missing data. The genotype matrix included 4.81 % of missing values and they were imputed with FImpute (Sargolzaei et al., 2014). FImpute has done 2% of genotypes corrections when the genotype was inconsistent between parents and progenies (figure 2.8).

Sequence and chip information comparison : In total 9432 positions on 10223 expected positions from the chip were detected on the 43 individual sequences. From

those, there were 9260 SNP completely identical at the genotype level with respect to the chip, and remaining 963 SNPs were not identical due to different kind of errors. The most common error affected 800 positions, for which there were too many missing values. The second kind of errors came because the reference genome used for alignment belongs to another species, which could lead in some situations to have two different alleles within our panel being different to the one in the reference sequence. This occured in 115 positions, and all were reintegrated into the data set. The last source of errors was when too many alleles were detected, and 48 positions were concerned. At the end, 9375 SNPs were detected without errors on our panel of 43 individuals.

The 43 individuals sequenced allowed to estimate the similarity between the SNPs resulting from the sequences and those from the chip. A similarity matrix was estimated between sequenced individuals and all genotyped individuals. The similarity between two individuals was made with a comparison by identity, for each position the value 1 is assigned if the two genotypes were identical, 0 otherwise. The similarity was then calculated as the percentage of identical markers in relation to the total number of markers that were in common between the chip and the sequence. We defined a threshold at 0.85 of similarity between the chip and the sequences to consider they come from the same individual, a lower similarity can result in wrong DNA or a pollution during the process. The Parents-progenies trios were used to check the Mendelian segregation for each sequence position. Some parents were not genotyped with the chip: *72131-036*, *71034-2-406*, *73193-091*, *3824-3*, *H480*, *72146-11* and *71069-914*. They were imputed for the chip markers by using information from their descendants and the complementary parent. They were considered like the other individuals genotyped with the chip. As the positions were detected with two different SNP-callers, the similarities were estimated independently for each of them. Both were represented in figure 2.9.

The similarities between the genotype extracted from the sequence and the best hit from the chip were around 95% for the majority of individuals. We compared the name of the individuals from the sequences with the name of the most similar individual in the chip (Best-hit). The individuals with a similarity under 0.8 were non-genotyped individuals and they could not be imputed because they were unrelated to the other genotyped individuals. The results were similar regardless the SNP-caller used. For the non-genotypes individuals,

the chip information will be replaced with the sequence information at the same locations.

Sequence position filtering : A total of 27,475,756 SNPs and Indels were detected by gVCF-GATK, whereas 26,489,941 SNPs were detected by Freebayes (table 2.4). After scoring the SNPs on a quality criterion (Phred score >30), the number of trimmed positions were twice as many with gVCF-GATK than with Freebayes (table 2.4). Among the remaining positions, some were monomorphic between *P. nigra* individuals but different from the reference sequence: about 1 million for gVCF-GATK and twice as much for Freebayes. A total of 2,488,736 positions were common between the two callers at that point of the filtering. Among these positions, 17% were Indels and 83% SNPs. To have the best quality in genotype calling, we kept the positions where the genotype calling was at least 95% similar between the two callers for all individuals. Mendelian segregation was checked on available trios, and 142,974 positions for which the progeny were inconsistent with parents were removed . We used the 7,540 SNPs from the chip to imputed 1,466,586 SNPs from sequences along the 19 Chromosomes (Figure 2.10).

Filtering step	Freebayes	gVCF-GATK
No filter	26,489,941	27,475,756
vcftools (max allele=3, min allele=2, minQ=30)	5,011,303	10,474,367
Monomorphic within P. nigra individuals	1,246,546	2,504,973
Common positions between the two callers	2,488,736 (375,566)	
Homology between two callers more than 95%	1,612,432	
Consistent Mendelian Segregation	1,466,586 (208,217)	

Table 2.4: Number of variants detected in the 43 sequenced individuals using two callers with no filter and after filtering with different parameters to obtain the input dataset used for imputation. In brackets, the number of Indels out of the total number of variants.

2.3 Genotype imputation

The details of this part are given in the Chapter 3 corresponding to the first article of this thesis. We give here a quick overview of the method. We used the FImpute software (v 2.2) (Sargolzaei et al., 2014), as many studies have already pinpointed its good performance for imputation when compared to many other alternatives (Chud et al., 2015; Johnston et al., 2011; Toghiani et al., 2016; Ye et al., 2018). FImpute can use different sizes of

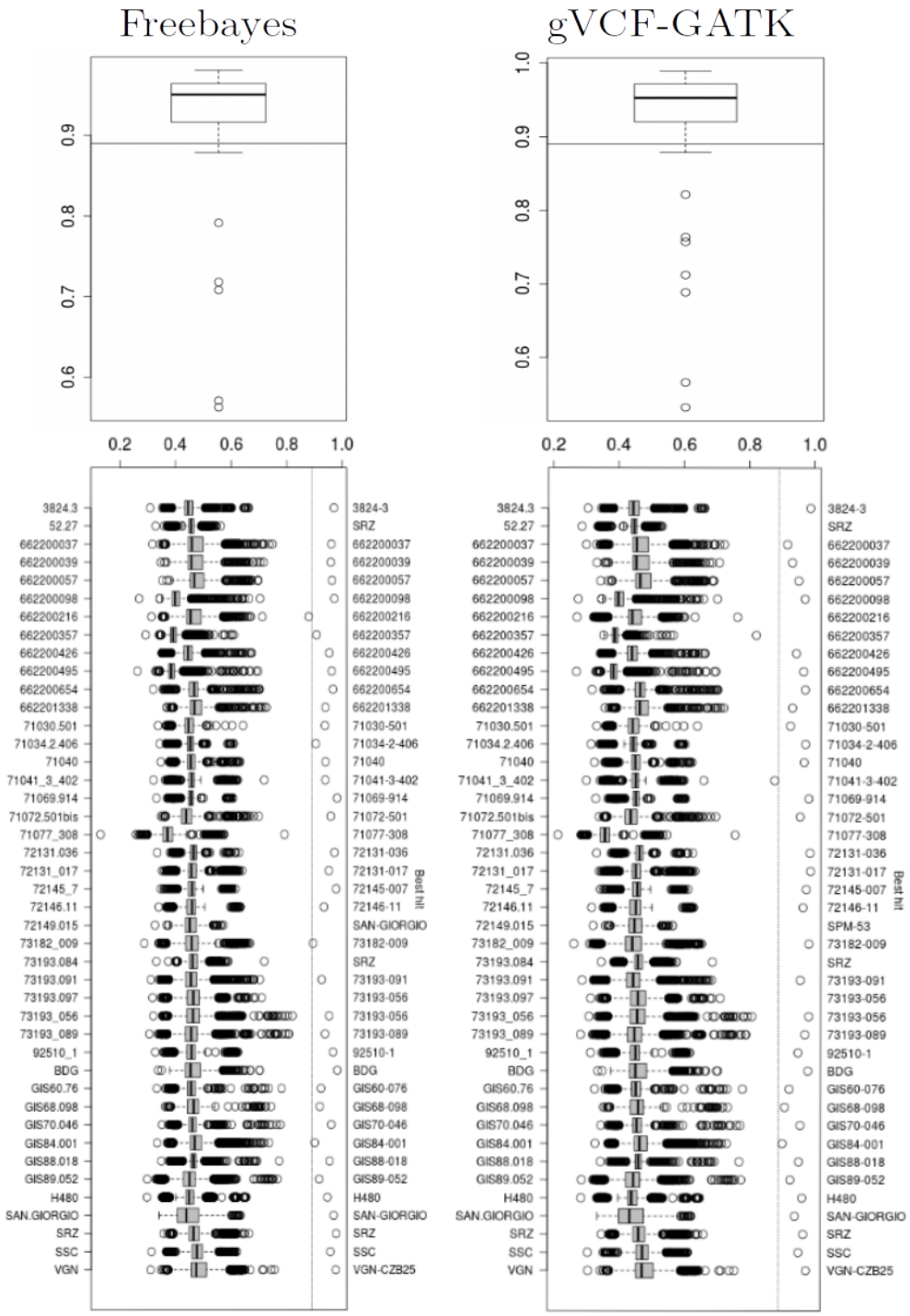


Figure 2.9: Genotype comparison between the sequence and the chip. Results obtained with the two callers are shown, Freebayes on left and gVCF-GATK on right. In the top, the best hits similarities are represented in boxplot. The lower part showed the similarities of each sequenced individual with all the genotyped individuals in box plot. The sequenced individual names were on left and the best hits individual name from the chip on right.

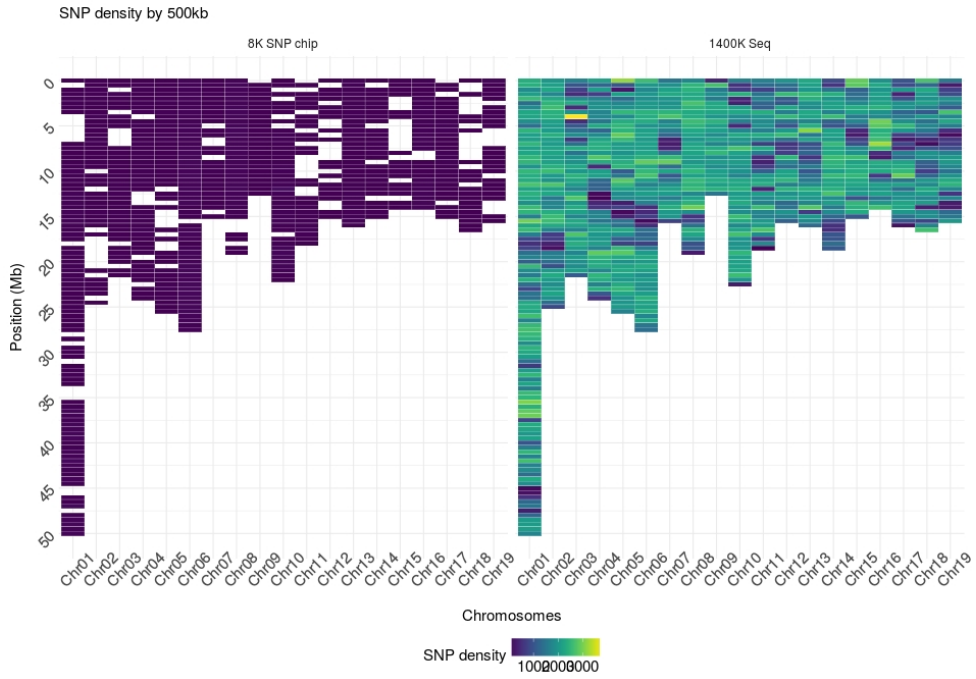


Figure 2.10: Density marker comparison. Marker density of 7540 SNPs from the chip (left) and from SNP calling (right) in 500kb non-overlapping windows.

rolling windows with a given overlap to scan the genomes of target and reference datasets. The pedigree information is used to increase imputation accuracy. Therefore, FImpute combines two sources of information for imputation: the pedigree and the LD.

2.3.1 Preliminary tests

Imputation software selection

A first test was designed to compare two imputation software based on the best results found in the litterature. It was performed on the reduced factorial mating design (dark green in table 2.2), composed by 294 individuals. These individuals were genotyped with the 12K custom illumina Bead Chip (Faivre-Rampant et al., 2016). Markers with more than 80% of missing values, SNPs with faulty segregation between parent and progenies, and monomorphic SNPs were removed. A cross-validation scheme with a masked proportion (10% or 50%) of genotypes for 75% of the progenies was set, with all the parents with complete genotypes and the remaining 25% of the progenies. The masked individuals and positions were randomly selected for each cross-validation run.

The imputation was performed by the FImpute software (v 2.2) (Sargolzaei et al., 2014) and BEAGLE 4.0 (Browning and Browning, 2007) software with the pedigree information provided. The imputation error rate was estimated as the proportion of imputed alleles differing from the true genotypes.

Preliminary sequence imputation in the factorial subsample

This test was the first full-scale test using available data in November 2016. We used the complete Factorial mating design with 392 individuals (383 progrenies, 8 parents and 1 grand-parent). Genotype imputation was performed with FImpute software (Sargolzaei et al., 2014), to impute 1% of missing data on the SNP chip panel. FImpute was also used to phase and impute the genotypic data from 8K SNP (SNPchip) to 2.4 millions of SNP (SNPSeq). To assess the imputation accuracy, a percentage of SNP correctly imputed was calculated by cross validation. For this, each SNPSeq of sequenced individual was masked and imputed in a leave-one-out scheme. Imputation was also tested in a more challenging scheme with 6 sequenced individuals with no known relatives in the dataset. The SNPs with more than 5% of error in the imputation were removed, and the most frequent genotype outcome from all cross-validations was kept for the next steps. The high density imputed dataset resulting at the end comprised 1 million of SNP, or 350K with MAF <0.05.

2.3.2 Final imputation

A most substantial imputation scheme was between the sequence data (1,466,586) on 43 individuals and the genotypic data from the SNP chip (7540) on 1039 individuals. To assess imputation accuracy, a leave-one-out cross validation scheme was performed among the 43 sequenced individuals. The SNPseq were masked for one individual at a time, and this individual with only SNPchip data was subsequently imputed with the rest of individuals. There were several cases depending on the relationship between individuals. The imputation software was able to use several types of information to perform the imputation: (from most advantageous to least advantageous) 1. Information from Brothers/Sisters, parents and half-brothers/sisters was available; 2. Information from parents was available; 3. Information from one of the parents and the Brothers/Sisters

or half-brothers/sisters was available; 4. Information from one of the parents was available 5. Information from Brothers/Sisters or half-brothers/sisters was available 6. There was no related individual bringing information, the software used the global information of the population. To have enough individuals in each case we divided the individuals in : Factorial mating parents (they have more information and more sequenced progenies); Multiple-pair mating parents (they are less related to each other); Factorial mating progenies (they have more sequenced sull-sibs and half-sibs); Multiple-pair mating progenies; unrelated individuals.

Differents statistics allow to assess the imputation quality or accuracy. One was the proportion of alleles correctly imputed by individual (eq. 2.3), and by positions estimate (eq. 2.4).

$$IA_j = 1 - \frac{\sum_{i=1}^{M_j} |g_{ij} - \hat{g_{ij}}|}{2 \times M_j} \tag{2.3}$$

$$IA_i = 1 - \frac{\sum_{j=1}^{N_i} |g_{ij} - \hat{g_{ij}}|}{2 \times N_i} \tag{2.4}$$

where gij is the observed allelic dosage (0,1,2) of the SNP i in individual j, gij is the imputed allelic dosage (0,1,2) from FImpute, M is the total number of SNP and Niis the number of individuals with called genotypes for SNP i.

The proportion of alleles correctly imputed by SNP might be subjected to frequency-dependent bias, in the sense that imputation could be correct more often when the imputed allele is already highly frequent. To overcome this, some authors (Badke et al., 2013; Calus et al., 2014) have proposed alternatives statistic. The Pearson's correlation coefficient between true and imputed individuals or between true and imputed positions (Calus et al., 2014). A correction of the proportion of alleles correctly imputed by the probability of correct imputation by chance (Badke et al. (2013) : eq. 2.5).

$$IA_{freq} = p(AA)_{ref} \times p(AA)_{val} + p(AB)_{ref} \times p(AB)_{val} + p(BB)_{ref} \times p(BB)_{val} \tag{2.5}$$

where $p(AA)_{refi}$, $p(AB)_{refi}$, and $p(BB)_{refi}$ are the observed frequencies for genotypes AA, AB, and BB for SNP i in the reference and $p(AA)_{vali}$, $p(AB)_{vali}$, and $p(BB)_{vali}$

are the predicted genotypic frequencies in the testing population for SNP i. IAfreq can be interpreted as the expected probability of correctly imputing a genotype in the testing population by assigning a randomly sampled genotype from the reference panel. We estimated the proportion of correctly imputed alleles adjusted for MAF using the equation 2.6 :

$$IA_{MAF} = \frac{IA - IA_{freq}}{1 - IA_{freq}} \tag{2.6}$$

where IA is computed as described in equation 2.4 and IA_{freq} in equation 2.5. FImpute offers an imputation mode based on allelic frequency (option "random_fill"), which gives us a lower bound for imputation accuracy and allow to illustrate the pedigree contribution in the imputation process.

Factors affecting SNP imputation

Different factors were used to describe the heterogeneity between individuals and between markers in terms of imputation quality. At the individual level: the sequence depth and the level of relatedness defined according to the following categories : parent of factorial, parent of multiple pair mating design, progeny of factorial, progeny of multiple pair mating design and French wild population. At SNP level, the following factors were considered: sequencing depth across individuals, per-site SNP quality from the SNP calling, the minor allele frequency in the 43 sequenced individuals, the ratio between the chip density and sequences density in non-overlapping 500kb windows, the p-value of an exact Hardy-Weinberg Equilibrium test for each site as defined by Wigginton et al. (2005) and the level of unique information contributed by each SNP given the level of LD with neighbouring SNPs, and calculated as the weight obtained by the LDAK5 software (Speed et al., 2017). The responses of the different variables to the factors considered were analysed by a principal component analysis and a feature selection algorithm called Boruta.

Genotype imputation impact

Linkage Disequilibrium : We were interested on the differences in linkage disequilibrium before and after imputation. We used the Plink software (Chang et al.,

2015; Purcell et al., 2007) to estimate the linkage disequilibrium parameter D' (Gaunt et al., 2007) in the chip dataset and after imputation in the sequence dataset.

Annotation analysis : We were interested in assessing to what extent imputation could change the annotation profile of covered SNPs, notably given the fact that the process involved a substantial change in density. Changes in annotation profiles from enriched to non-enriched but denser genotypes could be of relevance when using the resulting genotypes to fit prediction models for a large spectrum of traits. To get an annotation profile, a gene annotation analysis was performed. The tool Annovar (v. 2017Jul16) (Wang et al., 2010) was used with the command "–geneanno -buildver" in the Populus trichocarpa v3.1 gene set.

The results are described and discussed in a first article. This article has been submitted and the draft is available in Chapter 3.

2.4 Genomic Prediction

This exploratory study of the feasibility of genomic selection in black poplar was initially conducted on the available factorial data (First test). This has allowed us to set up data analysis pipelines from imputation to genomic selection. As genotyping and sequencing data became available, firstly in November 2016 (Second test), then in September 2017 (All dataset), the models were tested again to refine our choice of methods and software. The final result are presented in the draft of Article 2 in Chapter 4. In this section, I retrace the different tests we have performed, while results will be presented in a section of Chapter 4. After the genotype imputation, we extracted three marker sets of 50K, 100K and 250K to be used to test the impact of genotype information densification with respect to our chip data set (7K SNP). The details of how we composed each marker set are in the chapter 4.

2.4.1 Cross-validation strategy

Concerning cross-validation for assessing predicting abilities under the different evaluation methods, we chose three proportions of training and validation sets (TS/VS) : 75/25,

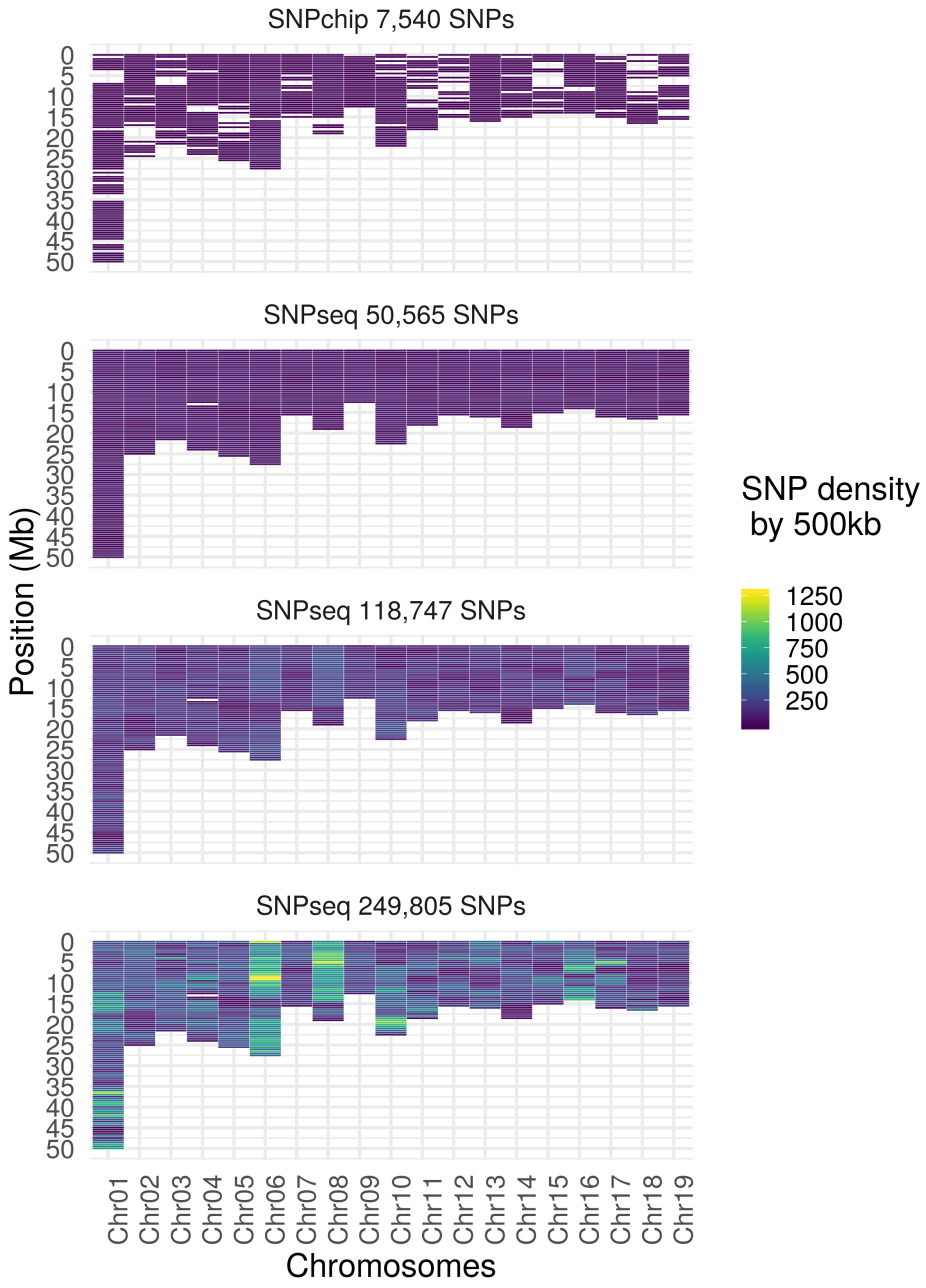


Figure 2.11: Density marker comparison. Marker density of 7540 SNPs from the chip; 50K , 100K and 250K from SNP calling after imputation and filtering the best positions p in 500kb non-overlapping windows.

50/50 and 25/75. We composed each training and validation set following several methods. Training sets were composed with all the parents and completed with randomly selected offspring from the global set. Between 4 (all dataset) and 12 (First and second test) repetitions were sampled for cross-validation. We tried two cross-validations based on the family sampling. Because parents were crossed with several other parents in the factorial design, we expected to be able to predict any of their unobserved (or masked) crosses from remaining crosses, in what constituted our training set for family sampling in the first and second tests. The results were not satisfying, and we chose another scenario for cross-validating families for the whole dataset. With all the data set, we selected randomly a number of families to be present or absent in the training (absent or present in the validation) in such a way that the resulting percentages of training versus validation (75/25 and 25/75) were met. To avoid selecting the same families for training or validation, we compared different random sets of families and took the ones with the most complementary family composition. Finally, we tried to optimize the individual sampling with a CDmean (Rincent et al., 2012) approach. We used CDmean inside families and for the global data set.

2.4.2 Main evaluation methodologies

GBLUP

We used the genomic BLUP or GBLUP based on the classic mixed model approach for genetic evaluations, replacing the relationship matrix A with a molecular relationship matrix G (VanRaden, 2007; Habier et al., 2007). Matrix A gives an expected relationship, ignoring the random sampling of parental alleles at each locus at the time of meiosis (Mendelian segregation), and thus resulting in siblings having the same relationship value. On the contrary, matrix G estimates the relatedness achieved by taking into account Mendelian segregation. Thanks to this additional information on the segregational variance, GBLUP can potentially achieve more accurate predictions of genetic values than matrix A. As with pedigree-based BLUP, GBLUP has the benefit of using equations with dimensions that remain reasonable, as it is linked to the number of pedigree records rather than to the number of markers.

Several methods for calculating G matrix have been proposed in the literature. The Lynch method improved by Li (Lynch and Ritland, 1999; Li et al., 1993) uses a similarity index applied to each locus and from which the relationship coefficient is calculated. This method assumes that all alleles Identical by State (IBS) are IBD (Identical by Descent), i. e. that each allele was originally present in a single copy in the founder population (Eding and Meuwissen, 2001). Otherwise, the relatedness is overestimated. It is possible to correct this by taking into account the probability that at a locus an allele is IBS but not IBD. However, this correction is difficult to implement because it requires to know the genotype of founders, which are often unavailable. An alternative method consists in using allele frequencies to correct relatedness coefficients (VanRaden, 2007; Habier et al., 2007). Ideally, the allelic frequencies must be those in the founder population. In practice, these are often unknown and the allele frequencies of the population under study are used instead. Forni et al. (2011) propose a method to "normalize" the previous relatedness matrix so that the average value of its diagonal is 1. This method aims at obtaining molecular relationships that are optimally compatible with pedigree-based relatedness, especially in cases where the two types of relatedness are to be combined, for example when jointly analyzing genotyped and non-genotyped individuals. Powell et al. (2010) and Yang et al. (2010) proposed derivatives to the weighting of the G matrix. These alternatives comprise a second adjustment using the variance associated to each marker as weighting factor for the construction of G, in the sense that highly variable markers are penalized. We tried the method proposed by Forni et al. (2011) together with that of Powell et al. (2010) for our GBLUP. Given that results in a preliminary set were similar, we kept the method proposed by Forni et al. (2011) based on VanRaden (2007) and Habier et al. (2007) for the final data set analysis.

Marker weighting : Studies conducted in the zebra finch (Lopes et al., 2013) and pork (Santure et al., 2010) showed that the best way to estimate relationships between individuals was with markers in linkage equilibrium. We tried differents methods to estimate relatedness from markers in linkage equilibrium or to weight markers' contribution to relatedness according to their own contribution to linkage disequilibrium. The first method is based on VIF (variance inflation factor) and is calculated with PLINK

(Purcell et al., 2007). We defined the size of the sliding SNP window to be tested, the number of SNPs for the offset of the window at each stage and VIF threshold. VIF is calculated using the following equation 2.7.

$$VIF = \frac{1}{1 - R^2} \tag{2.7}$$

where R^2 is a multiple correlation coefficient for a SNP regressed on all SNPs simultaneously. This factor considers the correlations between SNPs and between linear combinations of SNPs. A VIF of say 10 is often used to represent collinearity problems in multiple regression analysis (e.g. implies a R^2 of 0.9). A VIF of 1 would imply that the SNP is completely independent of all other SNPs. The second method was proposed by Speed et al. (2012), and implemented in the LDAK software. This method assesses patterns of local LD by estimating pairwise correlation matrix between SNPs. LDAK would then determine a weight per marker (i.e. 0 for redundant markers, and 1 for irreplaceable markers) according to the degree of LD as a proxy of the degree of independence of the information provided by markers.

RR-BLUP, BayesCπ and weighted GBLUP

Under the assumption of normal marker effects, the GBLUP is equivalent to the RR-BLUP. Ridge regression (RR-BLUP), random regression BLUP or SNP BLUP (Meuwissen et al., 2001) are different ways to denote the same application that uses mixed models to predict the (random) effect of markers. Marker effects are summed up to obtain individual breeding values. It is assumed that the variances of marker effects are the same across markers. RR-BLUP and GBLUP are actually equivalent models, in the sense that both give the same variance and individual predictions. The first datasets were analysed with the GS3 software to estimate the markers effects and the individuals GEBV (Bayesian equivalent to RR-BLUP). However, it usually took substantial amounts of computing time, notably with high densities like 250K SNP. We chose to use GBLUP instead, and to derive markers effects (u) with the following equation 2.8 (Strandén and Garrick, 2009):

$$u = WX'G^{-1}g \tag{2.8}$$

where W was a diagonal matrix of weights, either an identity matrix (GBLUP) or a diagonal of weights (wGBLUP), X was the genotyping (n $\times$ p) matrix , g the genomic estimated breeding values (GEBV).

The BayesCπ method Habier et al. (2011) is an extension of BRR (Bayesian random regression (Pérez et al., 2010), which is the Bayesian version of RR-BLUP. The BayesCπ therefore also considers that all markers have the same variance but assuming that a proportion π of markers, estimated by the model, have no effect. For the remaining $1-\pi$ markers, the BRR rules apply. The a priori beta distribution is assumed for π. With π=0, BayesCπ is identical to BRR. We used GS3 software to assess BayesCπ. Unlike previous definition of π, the BayesCπ in GS3 assumes π to be the proportion of markers having an effect. The π was either estimated directly by GS3 or given as a fixed value from the start (5%, 1% and 0.05% for the two first tests). Then with all the data, and given the computational time that was required for BayesCπ, we changed for the weighted GBLUP (WANG et al., 2012). Zhang et al. (2016) shows that weighted GBLUP is a good proxy for BayesCπ, giving similar results but within a much shorter computing time as there is no Gibbs sampling. For comparison purposes, BayesCπ was also tested with the R package BGLR, and compared to that of GS3 for two traits and three different genotyping densities.

We used one of the procedures described in (WANG et al., 2012) for the weighted GBLUP (denoted wGBLUP hereafter). Unlike GBLUP, where all markers are assumed to have the same variance and the same weight a priori, the derivative wGBLUP uses transformed G according to marker weights to select SNP. Weights were obtained by normalized the squared marker effects, and those were obtained with a first iteration of GBLUP and subsequent derivation from Strandén and Garrick (2009) formula. Once the weighting done, a second round of GBLUP with the new weighted G matrix was obtained. The whole process was repeated three times, thus is three subsequent weightings.

Additional models with dominance and multiple-traits

The range of evaluation methodologies that were used in the study (BLUP, GBLUP, wGBLUP and BayesCπ) are shown in the table 2.5, together with the information of the evaluation dataset, the different genotyping datasets and the alternative models (additive,

Table 2.5: Summary of statistical methods and models used depending on the dataset

		Factorial mating design			Test November 2016			Last Choice		
	SNP set	ADD	ADD + DOM	MultiTrait	ADD	ADD + DOM	MultiTrait	ADD	ADD + DOM	MultiTrait
P-BLUP	none	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-BLUP corrected	none							Yes	Yes	Yes
GBLUP	7K	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	50K							Yes		
	100K							Yes		
	250K	Yes			Yes			Yes		
wGBLUP	7K							Yes	Yes	
	50K							Yes		
	100K							Yes		
	250K							Yes		
BayesCπ ; π estimated by the model *	7K	Yes	Yes		Yes	Yes		Yes		
	50K							Yes		
	100K							Yes		
	250K	Yes	Yes		Yes	Yes		Yes		
BayesCπ ; π fixed at 10%	7K	Yes	Yes		Yes	Yes				
	250K	Yes	Yes		Yes	Yes				
BayesCπ ; π fixed at 1%	7K	Yes	Yes		Yes	Yes				
	250K	Yes	Yes		Yes	Yes				
BayesCπ ; π at 0.05%	7K	Yes	Yes		Yes	Yes				
	250K	Yes	Yes		Yes	Yes				
		* GS3 was used for the two first test and the package R BGLR with all the data set								
		Converge problem								
		With GS3 the convergence criteriaea still not converge								
		The results were similar between GBLUP and RR-BLUP and were changed for the weigthed GBLUP								
		Not tested								

additive+dominance and multiple-trait). Not all methods were used in combination with dominance effects as shown in table 2.5, the analyses were performed with the R package breedR and sommer.

For dominance models with GBLUP, the dominance relationship matrix was calculated with a modified method based on Vitezica et al. (2013), and normalized following Forni et al. (2011). We also tried dominance with wGBLUP, by weighting the markers in the corresponding dominance relationship matrix with the estimated dominance effect from a previous iteration (see details in the Chapter 4). Multiple-trait models were obtained with R package breedR, and involving the following traits : rust resistance, height and cicumference in first and second year of growth, budburst and branch angle.

A method based on haplotypes

During my stay in the AGBU research unit (University of New England, NSW 2351, Australia) under the supervision of Bruce Tier, we implemented a method to build haplotype-based relationship matrixes. A haplotype is a group of alleles of different loci

located on the same segment of a given parental chromosome and that segregate together during meiosis. Using phased haplotypes and tracking their transmision from parents to offspring would facilitate the distinction whether IBS and IBD between any two alleles in a locus.

The aim of building a haplotype-based relationship matrix is to calculate a genomic relationship matrix between the gametes in the population and subsequently use this information to compute a relationship matrix of individuals (Tier and Sölkner, 1993). The algorithm uses two matrices, of size $n \times p$ with n the number of gametes (twice the number of individuals) and p the number of loci (or SNP). The first matrix contains the phased gametes (see table 2.6 for a toy example).

Table 2.6: Example of an individual phasing by gamete for 10 SNP, 4 parents (P) and 2 progenies (D)

individu	gametes	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
P1	1	1	1	1	1	1	1	1	1	1	1
	2	2	2	2	2	2	2	2	2	2	2
P2	3	3	3	3	3	3	3	3	3	3	3
	4	4	4	4	4	4	4	4	4	4	4
P3	5	1	1	1	1	1	2	2	2	2	2
	6	3	3	3	3	3	4	4	4	3	3
P4	7	2	2	2	2	1	1	1	1	1	1
	8	4	4	4	4	4	4	4	3	3	3
D1	9	5	5	5	6	6	6	6	6	6	6
	10	7	7	7	7	7	7	7	7	7	7
D2	11	8	8	8	8	8	8	8	8	8	8
	12	4	4	4	4	4	5	5	5	5	5

The parents are considered as founders, and each homologous chromosome has its own identification number to allow tracking down to progeny. The progenies are therefore composed of a combination of two gametes derived from the recombination of each parental chromosome (lines 5 and 6 matrix 1). The second matrix (table 2.7 for a toy example) contains the genotyping coded 0 for allele a and 1 for allele A.

The algorithm compares genotyping and haplotypes to build the relationship matrix between gametes. If the genotyping is the same between two gametes and both alleles come from the same haplotype, the corresponding value in the gametic relationship matrix is worth 1. If the genotyping is identical but does not come from the same haplotype, then that relationship is worth p, with p being the haplotypic confidence interval between 0 and 1. Finally, if the genotyping is different, zero is added. Once all the gametes have

Table 2.7: Individual genotyping by gamete for 10 SNP

individu	gametes	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
P1	1	0	1	0	1	0	0	1	1	0	1
	2	0	1	0	0	1	1	0	1	0	1
P2	3	1	0	1	0	1	0	1	0	0	0
	4	1	0	1	1	1	0	1	0	0	1
P3	5	0	1	0	1	0	1	0	1	0	1
	6	1	0	1	0	1	0	1	0	0	0
P4	7	0	1	0	0	0	0	1	1	0	1
	8	1	0	1	1	1	0	1	0	0	0
D1	9	0	1	0	0	1	0	1	0	0	0
	10	0	1	0	0	0	0	1	1	0	1
D2	11	1	0	1	1	1	0	1	0	0	0
	12	1	0	1	1	1	1	0	1	0	1

been compared on all the loci, we divide by the number of loci (table 2.8 example with p=0.5).

Table 2.8: Gametic relationship matrix

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.0	0.3	0.0	0.1	1.0	0.0	0.5	0.1	0.3	0.4	0.1	0.1
2	0.3	1.0	0.2	0.1	0.3	0.2	0.8	0.1	0.5	0.4	0.1	0.1
3	0.0	0.2	1.0	0.4	0.0	1.0	0.1	0.4	0.2	0.1	0.4	0.4
4	0.1	0.1	0.4	1.0	0.1	0.4	0.0	1.0	0.1	0.0	0.5	1.0
5	1.0	0.3	0.0	0.1	1.0	0.0	0.5	0.1	0.3	0.4	0.1	0.1
6	0.0	0.2	1.0	0.4	0.0	1.0	0.1	0.4	0.2	0.1	0.4	0.4
7	0.5	0.8	0.1	0.0	0.5	0.1	1.0	0.0	0.4	0.5	0.0	0.0
8	0.1	0.1	0.4	1.0	0.1	0.4	0.0	1.0	0.1	0.0	0.5	1.0
9	0.3	0.5	0.2	0.1	0.3	0.2	0.4	0.1	1.0	0.4	0.1	0.1
10	0.4	0.4	0.1	0.0	0.4	0.1	0.5	0.0	0.4	1.0	0.0	0.0
11	0.1	0.1	0.4	0.5	0.1	0.4	0.0	0.5	0.1	0.0	1.0	0.5
12	0.1	0.1	0.4	1.0	0.1	0.4	0.0	1.0	0.1	0.0	0.5	1.0

The relationship between individuals (A*) (table 2.9) is calculated with the following equation 2.9 :

$$A* = 0.5 \times (g_{ii} + g_{jj} + g_{ij} + g_{ji}) \tag{2.9}$$

where g_{ii} and g_{jj} are the diagonal elements equal to 1, gij and gji the relationships between the gamete i and the gamete j of an individual. When $p = 0$, the matrix A* is equivalent to the matrix A, and when $p = 1$, A* is equivalent to G (VanRaden, 2007; Habier et al., 2007), with allelic frequencies equal to 0.5 for all loci. We masked randomly 25% of the genotyped individuals to compare A* with A , H (Legarra et al., 2009) and G (VanRaden, 2007; Habier et al., 2007). We compute A* with $p = 0$ to $p = 1$ by implementing p by 0.1 and used in a BLUP model to estimate heritabilities and AIC

criteria.

Table 2.9: Individual relationship matrix A*

	P1	P2	P3	P4	D1	D2
P1	1.30	0.20	0.75	0.75	0.8	0.20
P2	0.20	1.40	0.75	0.75	0.2	1.15
P3	0.75	0.75	1.00	0.55	0.5	0.50
P4	0.75	0.75	0.55	1.00	0.5	0.75
D1	0.80	0.20	0.50	0.50	1.4	0.10
D2	0.20	1.15	0.50	0.75	0.1	1.50

Quality of prediction

We used cross-validation to assess the prediction quality for the different models and evaluation methodologies. The basic parameter was the predicting ability, obtained as the Pearson correlation between the phenotypes and the GEBV for each validation population. For comparative purposes, the prediction accuracy was obtained as the ratio between the predicting ability and the square root of the narrow sense heritability. The heritability being the one obtained from the pedigree-based model for each trait. Other quality parameters were also studied: the Spearman correlation between phenotypes and GEBV and the slope of the regression of phenotypes on the GEBV. Additionally to the cross-validation sets, we used an independent set of individuals, representing the next generation, to assess the genomic prediction performance. The GEBV of the individuals in such testing set were predicted with each of the calibration sets in the cross-validation, and the same quality parameters were obtained as for the cross-validation.

Chapter 3

Densification of the genotyping information by imputation

3.1 Résumé du Chapitre

Ce chapitre se place dans la deuxième question de la thèse et cherche à répondre à la question suivante : Sommes-nous capables d'améliorer les performance de la prédiction génomique chez le peuplier en intégrant un plus grand nombre d'informations ? Plus précisément, nous avons étudié les avantages de la densification génotypique par l'imputation.

Tous les résultats présentés dans ce chapitre convergent vers le fait qu'une imputation génotypique de bonne qualité est possible dans le contexte des populations d'amélioration utilisées dans l'étude. Une stratégie de séquençage du génome relativement restreinte impliquant quelques dizaines d'individus nodaux combinée à un processus d'imputation a permis de multiplier par 185 le nombre de marqueurs disponibles pour plus de mille individus, pour lesquels seules de faibles densités étaient disponibles. Toutefois, la qualité de l'imputation est dans une certaine mesure hétérogène entre les marqueurs et les individus. Le nombre relativement important d'individus séquencés par rapport à la population à imputer, ainsi que les niveaux de parenté dans cette population impliquant les parents, la progéniture et les frères et sœurs ont eu un impact positif sur la qualité des résultats de l'imputation.

Nous montrons également dans ce chapitre que tous les facteurs de pré-imputation

peuvent expliquer en partie les différences de qualité d'imputation, mais qu'aucun d'entre eux ne peut être utilisé de manière opérationnelle pour filtrer les positions avant imputation. L'un des rares facteurs sur lesquels nous avons un certain pouvoir décisionnel est l'homogénéité de la couverture du panel de faible densité. Selon nos résultats en matière de qualité d'imputation, la méthodologie mise en œuvre dans ce chapitre permet d'acquérir un grand nombre de SNPs sur un grand nombre d'individus pour un coût inférieur à celui d'un séquençage complet. L'une des rares exigences concerne l'installation et les compétences pour l'utilisations des outils de bioinformatiques. De grandes capacité de calculs ne sont pas nécessaires pour effectuer l'imputation pour un jeu de données équivalent au nôtre.

3.2 Summary presentation of the chapter

This chapter is part of the second question of this thesis: Are we able to improve the conditions for a quality genomic prediction in poplar by integrating extra information? More precisely we investigated the advantages of genotyping densification through imputation. Preliminary tests were made on a part of the factorial mating design (table 2.2) to select the best imputation software, before performing substantial genotype imputation from sequences in our global dataset. According to other studies (Chud et al., 2015; Johnston et al., 2011; Toghiani et al., 2016; Ye et al., 2018), we selected two software tools to perform this preliminary test: BEAGLE 4.0 (Browning and Browning, 2007) and FImpute v2 (Sargolzaei et al., 2014). As a result, FImpute showed best results by masking either 10% or 50% of the genotypes of the factorial mating progenies with the chip dataset.

Our final data set comprised 1033 individuals divided among 35 families and 6 unrelated individuals from French wild populations. We sequenced 43 nodal individuals (grandparent, parents, progenies) and detected more than 1.4 millions of SNPs to impute a thousand individuals genotyped with Illumina custom 12K Bead chip (7540 SNPs). The imputation quality was assessed by a leave-one-out cross-validation of the sequenced individuals. The imputation error rate or accuracy was calculated at the individual and position levels. Different pre-imputation factors were studied and analyzed with a principal component analysis and a classification algorithm called Boruta, in order to

understand the factors affecting the imputation quality.

The preliminary results are shown in the first section of this chapter, before the first Article. This article has been submitted to BMC genomics in August 2018, and deposited to BioRxiv (Pegard et al., 2018). In this article, we presented the imputation quality results, together with an assessment of the explaining factors and the impacts of genotype imputation on linkage disequilibrium and annotation profile. Finally, to close the chapter, we discussed the impact and the advantages of genotype densification in relation to: (1) the detection of recombination events within chromosomes, (2) the estimation the recombination rate within families to improve subsequent predictions in silico of segregation; (3) the enrichment of genetic maps, and (4) the improvement of the accuracy of GS.

3.3 Preliminary tests

3.3.1 Imputation software selection

Material and methods quick overview

The tests were performed on the reduced factorial mating design (table 3.1), composed of 294 individuals genotyped with the 12K custom Illumina Bead Chip (Faivre-Rampant et al., 2016). We masked 10% and 50% of genotype information for 75% of the progenies. All the parents had their complete genotype information, as well as 25% of the progenies. The individuals and positions to be masked were randomly selected. The imputation was performed by the FImpute software (v 2.2) (Sargolzaei et al., 2014) and by BEAGLE 4.0 (Browning and Browning, 2007) software with the pedigree information provided. The imputation error rate was estimated as the proportion of imputed alleles differing from the true genotypes.

Results & discussion

The results reported in table 3.2 showed that both at 10% and 50% of masking FImpute performed better than BEAGLE 4.0. Fimpute showed similar results between 10% and 50% of masking rate with an error rate of around 1%. The error imputation rate of

Table 3.1: Factorial mating design used for the preliminary test of imputation

Male	SAN-GIORGIO		
Female	SRZ	BDG	71077-308
VGN-CZB25	55	57	54
71041-3-402		28	11
71072-501	25	28	29

BEAGLE 4.0 reached 13.24% when we masked 50% of the genotypes. The comparatively poorer results of BEAGLE 4.0 may be due to the small number of individuals provided. Indeed, Sargolzaei et al. (2014) indicate that BEAGLE works better with more than 300 individuals, slightly more than those used in this preliminary work. According to these first results, we decided to use FImpute for the next steps of imputation. FImpute has been already reviewed as being highly performant compared to other alternatives (Chud et al., 2015; Johnston et al., 2011; Toghiani et al., 2016; Ye et al., 2018).

Table 3.2: Factorial mating design used for the preliminary test of imputation

Imputation Software	Masking rate	Imputation error rate
FImpute	10%	1.06%
	50%	1.07%
BEAGLE 4.0	10%	2.16%
	50%	13.24%

3.3.2 Sequence: the first full-scale test

Material and methods quick overview

We used the complete Factorial mating design containing 392 individuals. Genotype imputation was performed with FImpute software (Sargolzaei et al., 2014), in order to impute genotypic data from 8K SNP (SNPchip) to 2.4 millions of SNP (SNPSeq). The assessment of the imputation accuracy, as the percentage of SNP correctly imputed, was calculated by cross-validation in a leave-one-out scheme. For each sequenced individual, the SNPSeq were masked and imputed from SNPchip data with all the information from the rest of the population and relatives. Some of the sequenced individuals belonged to unrelated populations and were used as challenging test for the imputation.

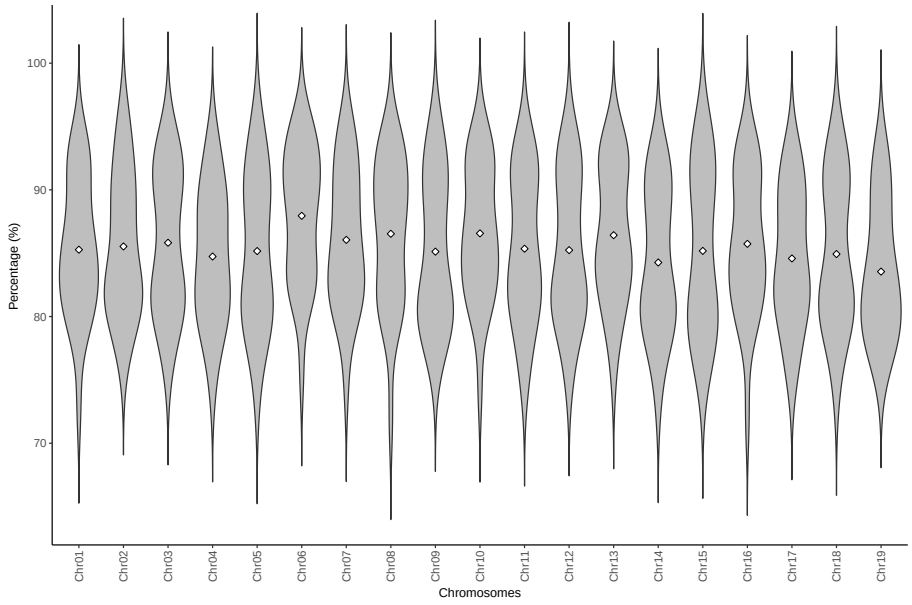


Figure 3.1: Percentage of positions correctly imputed by Chromosomes

Results & discussion

The imputation accuracy by chromosomes (Figure 3.1) varied from 60% to 100%, and with an average of 85%. All the chromosomes seemed to have similar ranges of variation for the accuracy. Chromosomes 6, 8 and 10 were slightly better imputed than the rest. The figure 3.2 represents the imputation accuracy by individuals, with large variations in distribution across individuals. In the worst case scenario, accuracies were slightly higher than 70%, while in the best cases values were higher than 90%. The figure 3.3 showed the results of accuracy depending on the individual's class, with three main classes (Parents, progeny or unrelated). The unrelated individuals had the lowest imputation accuracy, with the narrowest dispersion around the mean, whereas offspring was the group with the highest accuracies. Parents showed multimodal distribution for accuracies, clearly indicating heterogeneity in accuracies between some groups of parents. After trimming the imputed position and keeping only those positions with an imputation accuracy higher than 95%, there were 1 million of SNPs still available. When applying a filter on the minor allele frequency at 5%, there were 350K SNP with higher MAF still available for the first genomic selection test.

Several conclusions can be drawn from this full-scale imputation test. First, a good

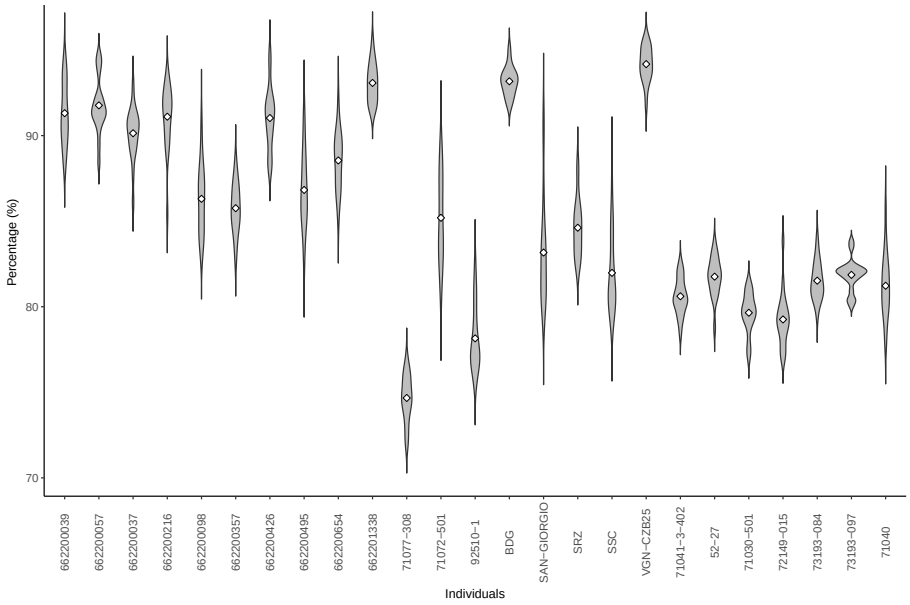


Figure 3.2: Percentage of positions correctly imputed by individuals

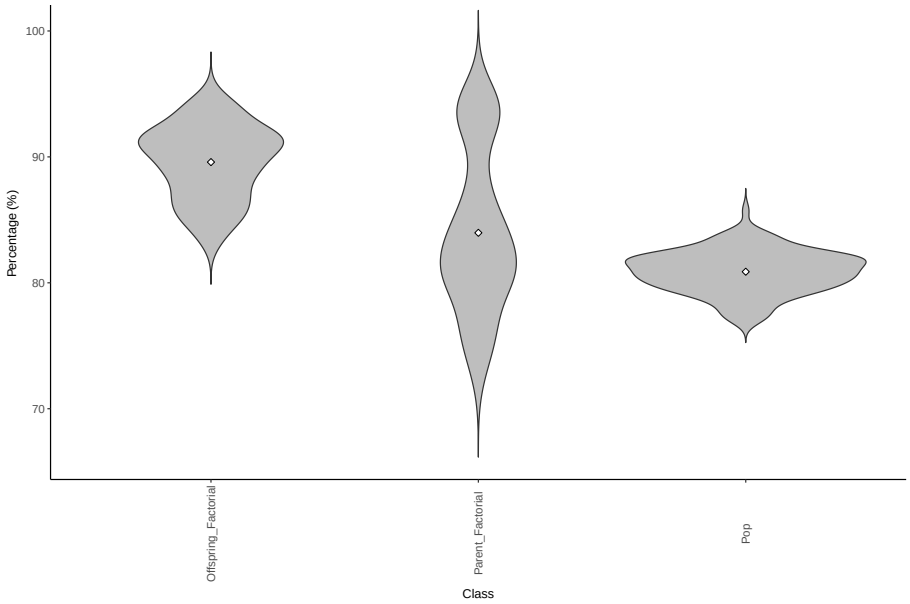


Figure 3.3: Percentage of positions correctly imputed by individual's class

imputation accuracy (more than 95% of similarity) was available when imputing from low density (8K SNPs) to high density (1 million of SNPs), even with unrelated individuals. There was a small variation between chromosomes in average and between individuals depending on their relatedness class. The most important part of correctly imputed positions involve rare alleles with a minor allele frequency lower than 5%. These results led us to ask which factors have a strong influence on the imputation quality, and how imputation works in a wider population. Some of these aspects were treated in the following article.

3.4 Article I: Sequence Imputation from Low-Density Single Nucleotide Polymorphism Panel in a Black Poplar Breeding population

This paper was Submitted to BMC genomics on August 2018 and accepted on March 2019.

RESEARCH ARTICLE

Sequence Imputation from Low Density Single Nucleotide Polymorphism Panel in a Black Poplar Breeding population

Marie Pégard[1], Odile Rogier[1], Aurélie Bérard[2], Patricia Faivre-Rampant[2], Marie-Christine Le Paslier[2], Catherine Bastien[1], Véronique Jorge[1] and Leopoldo Sánchez[1*]

Abstract

Background: Genomic selection accuracy increases with the use of high SNP (single nucleotide polymorphism) coverage. However, such gains in coverage come at high costs, preventing their prompt operational implementation by breeders. Low density panels imputed to higher densities offer a cheaper alternative during the first stages of genomic resources development. Our study is the first to explore the imputation in a tree species: black poplar. About 1000 pure-breed *Populus nigra* trees from a breeding population were selected and genotyped with a 12K custom Infinium Bead-Chip. Forty-three of those individuals corresponding to nodal trees in the pedigree were fully sequenced (reference), while the remaining majority (target) was imputed from 8K to 1.4 million SNPs using FImpute. Each SNP and individual was evaluated for imputation errors by leave-one-out cross validation in the training sample of 43 sequenced trees. Some summary statistics such as Hardy-Weinberg Equilibrium exact test p-value, quality of sequencing, depth of sequencing per site and per individual, minor allele frequency, marker density ratio or SNP information redundancy were calculated. Principal component and Boruta analyses were used on all these parameters to rank the factors affecting the quality of imputation. Additionally, we characterize the impact of the relatedness between reference population and target population.

Results: During the imputation process, we used 7,540 SNPs from the chip to impute 1,438,827 SNPs from sequences. At the individual level, imputation accuracy was high with a proportion of SNPs correctly imputed between 0.84 and 0.99. The variation in accuracies was mostly due to differences in relatedness between individuals. At a SNP level, the imputation quality depended on genotyped SNP density and on the original minor allele frequency. The imputation did not appear to result in an increase of linkage disequilibrium. The genotype densification not only brought a better distribution of markers all along the genome, but also we did not detect any substantial bias in annotation categories.

Conclusions: This study shows that it is possible to impute low-density marker panels to whole genome sequence with good accuracy under certain conditions that could be common to many breeding populations.

Keywords: Genotype Imputation; Low density arrays; Whole-Genome Resequencing; *Populus nigra*

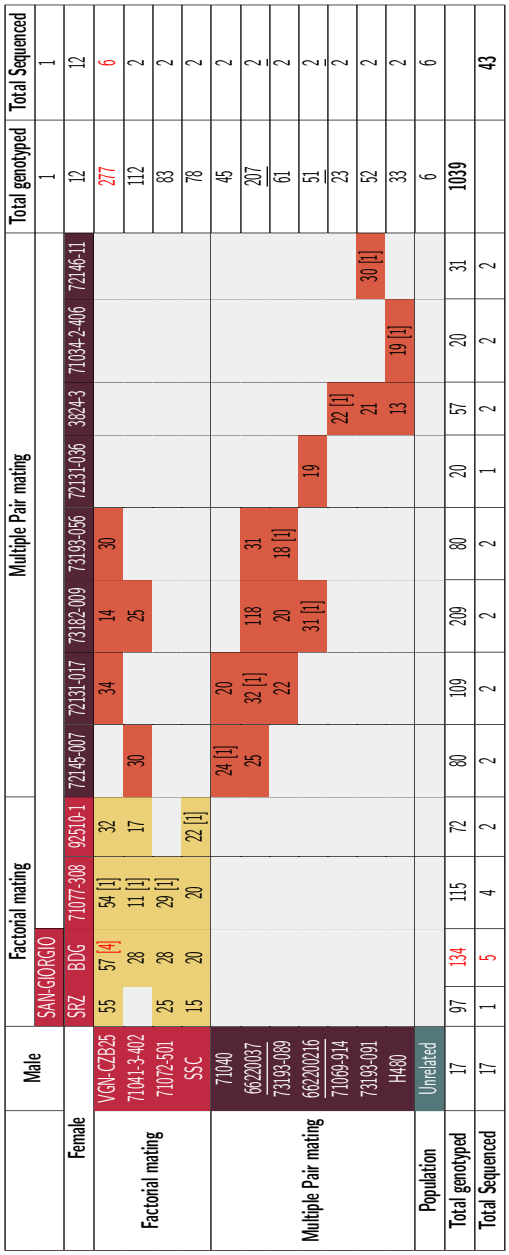
Background

In genome-wide analyses, the accuracy of genomic associations and predictions tends to increase with the density of marker coverage [1, 2]. Although the cost of genotyping has decreased steadily over the past decade, it still represents a significant investment for an improvement program. High-density genotyping of a large number of individuals remains unaffordable for non-domesticated and highly heterozygous species. Low-density panels imputed to higher densities offer an alternative to systematic genotyping or sequencing of the entire population, at least at the initial stages of compiling the minimum amount of genomic resources. The idea of genotype imputation as supplemental genotyping data was described by Burdick et al. [3], using the term "*in silico*" genotyping. In this context, imputation refers to the process of predicting genotyping data not directly available for an individual. Imputation uses a reference panel composed of genotyped individuals

*Correspondence: leopoldo.sanchez-rodriguez@inra.fr
[1]BioForA, INRA, ONF, 45075, Orléans, France, 2163 Avenue de la Pomme de Pin CS 40001 ARDON, 45075 Orléans Cedex 2, France
Full list of author information is available at the end of the article

with high marker density to predict all missing markers of another panel genotyped at lower density coverage [2]. Imputation can be used in at least three different scenarios: (i) to fill missing data that occurred due to technical problems, (ii) to correct for genotyping errors, and (iii) to infer data for non-genotyped SNPs on a set of individuals [4]. Another more extreme scenario involving imputation is to create all the genotype information of individuals that are no longer available from their extant relatives [5]. Imputation software uses two main strategies: the first is based on pedigree and Mendelian segregation [6–8], and the second relies on linkage disequilibrium [9, 10]. Some authors use sequentially or in a given combination both approaches [11]. The first strategy is the one implemented in algorithms like Lander-Green [12], Elston-Steward [13] or Monte-Carlo sampling algorithms [14, 15]. The second strategy is commonly used for samples with low levels of kinship and unknown ancestors, relying instead on the linkage disequilibrium between markers within the reference population. It uses heuristic algorithms as Expectation Maximization (EM) algorithm, coalescence models and Markov's hidden strings (HMM) [16, 17]. Recently, a study has compared eight machine learning methods to impute a genotype dataset, but results are of lower quality than those from Beagle, a reference software in the domain of imputation [18, 19] which is based on the forecited second strategy [20]. The imputation accuracy depends on several factors. Among them, there are the genotyping quality, the levels of linkage disequilibrium (LD), the marker density which in turn influences perceived linkage disequilibrium, and the relatedness between reference and imputed populations. Factors affecting imputation accuracy have already been studied both with simulated and empirical data. For instance, Hickey et al. [21] showed that imputation accuracy increases with marker density. The reference population constitution is also a decisive factor for the imputation accuracy. The reference population should be large enough to capture all relevant haplotypes [6] and recombination events, as well as to estimate correctly LD. The relatedness between the reference and the target panel favours imputation quality, with higher accuracies as relatedness increases between the two groups [22]. The effects of panel size, LD and relatedness become more important with decreasing marker density [6, 23]. Imputation of genotyping data has several advantages, the first being the reduction of genotyping costs [24], which can be very important depending on the species. In addition, imputation of genotyping data also improves the detection of QTLs and the model's prediction accuracy developed in association studies or genomic selection [2]. The imputation of genotyping data could be used in genetic mapping to enrich genetic maps for a higher coverage. Finally, imputation could correct to a certain degree the eventual heterogeneity in marker density related to constraints in chip design. Such heterogeneity in marker density across the genome happened to be the case of the chip used in our study here [25]. Often, imputation involves a difference in densities between reference and targeted panels of less than 10-fold (i.e. 5K to 50K [26–28] or around 10-fold 50K to 500K [29, 30]). With the increasing access to affordable genomic sequence data, the possibility to use full sequences in the reference panel for imputation becomes a reality, at least for a limited number of individuals. Two studies simulated sequences to find the better strategy between imputation accuracy, number of sequenced individuals and genome coverage [31, 32]. Both studies suggest that a good compromise is sequencing as many individuals as possible but at medium coverage (x8). To our knowledge, only three studies in animals have tried to impute successfully from low and medium densities (13 K and 50-60K) to real sequence data (350K and 13 millions) [33–35]. These studies show that inferring whole sequences from low-density marker panels with good accuracy is possible under certain conditions, notably with high levels of relatedness and persistence of LD between the markers across populations. Our study is one of the first to explore the benefits of imputation to densify SNP genotyping in a forest tree species, usually less favored than livestock in genomic resources. This paper is based on black poplar, specifically on one of the breeding populations that is used to produce hybrid poplars. In the context of this breeding effort, imputation is expected to enrich our knowledge, for the subsequent step of predicting and selecting candidates, in three different aspects: (1) to capture recombination events within families to improve subsequent *in silico* predictions of segregation; (2) to enriching the genetic map and (3) to improve genomic evaluation accuracy. The main objective of this study was to demonstrate to what extent high quality imputation was feasible from low density arrays. A complementary objective was to identify the factors that contributed to the quality of the imputation and its impact on the linkage disequilibrium and the annotation profile of covered positions.

Table 1: Number of individuals and pedigree information



Description of black poplar breeding resources used in the study, with mating designs involved and number of individuals per family in the inner cells. Parental and family cells are coloured by class in the mating regimes: yellow, factorial mating progenies; orange, multiple pair mating progenies; red, factorial mating parents; purple, multiple pair mating parents; and dark cyan, unrelated individuals. In brackets, some selected cells show the number of sequenced progenies, with the figure in red involving 2 progenies that were subsequently used as parental females (underlined codes) for the multiple pair mating.

Methods

Plant material

For this study, 1,039 *Populus nigra* were made available from the French breeding population. This sample was structured into 35 families resulting from 23 parents. Available families resulted from two mating sets. As shown in the **Table 1**, the first mating set corresponds to an almost complete factorial mating design involving 4 female and 4 male parents, and resulting in 413 F1 individuals structured into 14 full sib families. The second set involved multiple pair mating schemes involving 8 female and 7 male parents, with a number of crosses per parent ranging from 1 to 5, and resulting in 598 F1 individuals structured into 21 full sib families. Six individuals originated from a collection of French wild populations were also added to the population. All 1,039 individuals in this population were genotyped and 43 of them were also sequenced. Among the sequenced individuals, there were 1 grand-parent, 21 parents, 13 progenies and 2 female individuals that were both progenies in the factorial mating design and subsequently parents in the multiple pair mating set (**Table 1**). The progenies to sequence were chosen in such a way that all parents had at least one sequenced offspring. The six sequenced individuals originated from wild populations were added to assess the imputation ability with unrelated individuals. Detail of genotype list and origins are given in **table S1**[see Additional file 1].

Genotyping and sequencing

We used the sequences of 6 parents previously sequenced by Genome Analyzer IIx from Illumina [25]. For the others parents (17), 1 grandparent, 14 progenies and 6 unrelated the DNA extraction was made from leaf samples in the UMR0588-BioForA collection, by using the Macherey-Nagel Nucleospin®96 Plant II commercial kit. Illumina paired-end shotgun indexed libraries were prepared from one μg of DNA per accession, using Illumina TruSeq®DNA PCR-Free Sample Preparation kit. Briefly, indexed library preparation was performed with DNA fragmentation by AFA (Adaptive Focused Acoustics^{TM}) technology on Covaris focused-ultrasonicator, all enzymatic steps and clean up were realized according to manufacturer's instructions. Single or dual indexes were used. Final libraries were quantified by using qPCR using KAPA Library Quantification Kit and Life Technologies QuantStudio^{TM} Real-Time PCR system. Fragment size distribution of libraries was assessed by High Sensitivity DNA assay either on Agilent 2100 Bioanalyzer or on Caliper LabChip®GX nucleic acid

analyser. Equimolar pools of multiplexed samples, up to 11, were engaged in sequencing using 4 lanes. After clusters generation on CBot, paired-end sequencing 2 × 150 sequencing by synthesis (SBS) cycles was performed either on a Illumina HiSeq®2000/2500 running in high output mode (one lane) or on Illumina HiSeq®4000 (three lanes). Reads were trimmed with Trimmomatic (v. 0.32) [36], and mapped to the *P.trichocarpa* version 3.1 genome [37] using BWA-MEM 0.7.12- with default parameters [17]. Picard Tools (v. 2.0.1) [38] were used to remove duplicated reads. Local and Indel realignments were performed using Genome Analysis Toolkit (GATK v. 3.5) [39, 40]. The variant detection was performed on all individuals by two variant callers: (1) all individuals at the same time with Freebayes (V1.0.0) [41], and (2) by each individual separately with GATK HaplotypeCaller, to be subsequently assembled using GenotypeGVCFs (called later gVCF-GATK). We have used the VCFtools 0.1.15 [42] to filter variants with no missing data, with a minimum quality score of 30 and a minimum mean depth of 2. We allowed among selected SNPs those harboring three alleles, because mapping was done on another *Populus* species reference genome, so it was possible to have two alternative alleles and no reference allele in the aligned sequences. We finally kept only SNPs and Indels that were detected by both callers and consistent with Mendelian segregation. To simplify, SNPs and Indels were both called SNPs hereafter. All individuals were genotyped using the *Populus nigra* 12K custom Infinium Bead-Chip (Illumina, San Diego, CA) [25]. We applied the same quality filters as in Faivre-Rampant et al (2016): markers with more than 90% of missing data were removed and only Mendelian segregation consistent markers were selected.

Genotype imputation

We used the FImpute software (v 2.2) [11], as many studies have already pinpointed its good performance for imputation when compared to many other alternatives [16, 35, 43, 44]. FImpute can use different sizes of rolling windows with a given overlap to scan the genomes of target and reference datasets. The pedigree information is used to increase imputation accuracy. Therefore, FImpute combines both formerly stated strategies for imputation: that based on pedigree and that on LD. A first round of genotype imputation was performed to predict 1% of missing data still existing on the SNP chip panel. The second and most substantial imputation scheme was between the genotypic data from the chip SNP (SNPchip) and the sequence data (SNPseq). To assess imputation accuracy, a leave-one-out cross validation scheme was

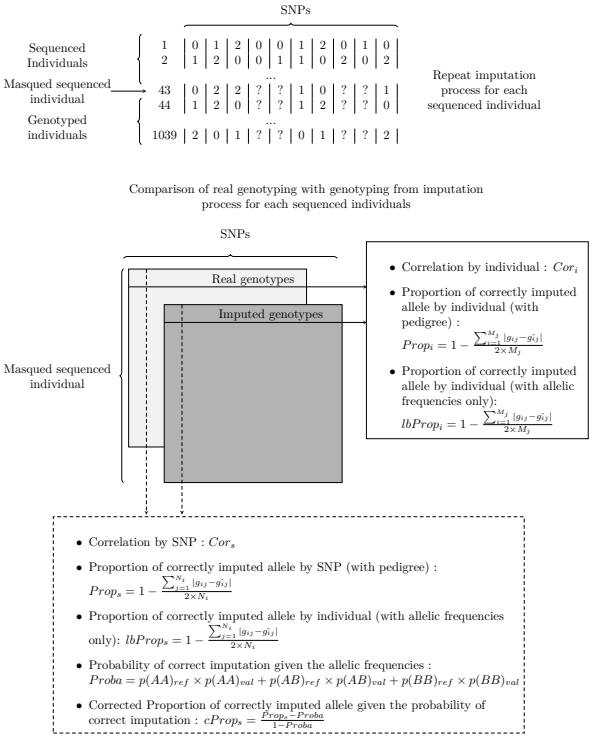


Figure 1: **Metrics for the assessment of imputation quality and accuracy by individuals and by SNPs.** The first upper panel depicts an example of a toy genotyping matrix containing the allelic doses, with markers in columns and individuals in rows. First two individuals correspond to complete genotypes from sequences; next two to sequences with masked positions to be imputed for quality assessment; and last individual to one genotype from the SNP array. The lower panel represents the two simplified genotyping matrices respectively with real and imputed genotypes. Associated boxes contain the different metrics that were used in the study: to the right and across markers (columns), the metrics by individual; at the bottom and across individuals (rows), it can be found the metrics by marker. The expressions for Prop-like metrics contain the following variables: gij the observed allelic dosage (0,1,2) of the SNP i in individual j; $\hat{g}ij$ the imputed allelic dosage (0,1,2) from FImpute; M the total number of SNP; Ni the number of individuals with called genotypes for SNP i; $p(AA)_{refi}$, $p(AB)_{refi}$, and $p(BB)_{refi}$ are the observed frequencies for genotypes AA, AB, and BB for SNP i in the reference and $p(AA)_{vali}$, $p(AB)_{vali}$, and $p(BB)_{vali}$ are the predicted genotypic frequencies in the testing population for SNP i.

performed among the 43 sequenced individuals. The SNPseq were masked for one individual at a time, and this individual with only SNPchip data was subsequently imputed with the rest of individuals. To challenge the imputation scheme, an additional set of 6 unrelated individuals with sequences were added to the target panel. We estimated imputation quality (or accuracy) using various statistics. One was the proportion of alleles correctly imputed by each leave-one-out individual (across SNPs, one proportion per individual and per chromosome: *Propi*), and by positions (across individuals, one proportion per position: *Props*) (further explanations in Figure 1). The proportion of alleles correctly imputed by SNP might be subjected to frequency-dependent bias, in the sense that imputation could be correct more often than not when the imputed allele is already highly frequent. To overcome this, Calus et al. [45] have proposed the use of an alternative statistic, the Pearson's correlation coefficient between true and imputed individuals (across SNPs, one correlation value per individual and per chromosome: *Cori*) and between true and imputed positions (across individuals, one value per SNP position: *Cors*). In our case, this latter correlation (*Cors*) was not always available for computation. The reason was that some SNPs had such a low allelic frequency that monomorphic outcomes happened after imputation, leading to zero variances. In order to account for this frequency-dependent outcome, alternatively, we used the option proposed by Badke et al. (2014) [46] to correct the error rate by the probability of correct imputation by chance (*cProps*: corrected SNP proportion). FImpute offers an imputation mode based on allelic frequency (option "random_fill"), which gives us a lower bound for imputation accuracy by individual (*lbPropi*: lower bound individual proportion) and by SNP (*lbProps*: lower bound SNP proportion).

Factors affecting SNP imputation

We considered different factors describing the heterogeneity between individuals and between markers imputations, and we checked to what extent these factors affected imputation. The first factors were at the individual level: the sequence depth (**MEAN_DEPTH**); and the level of relatedness defined according to the following categories : parent of factorial (Factorial_parents), parent of multiple pair mating design (MultiplePair_parents), progeny of factorial (Factorial_progenies), progeny of multiple pair mating design (MultiplePair_progenies) and French wild population (Unrelated). At SNP level, the following factors were considered: sequencing depth (**DEPTH**) across individuals; per-site SNP quality from the SNP calling step (column **QUAL** in the vcf file, extracted with vcftools v0.1.13 from the gVCF-GATK results files); minor allele frequency (**FreqOri**); the ratio between SNPchip density and SNPseq density in non-overlapping 500kb windows (**RatioDensity**); the p-value of an exact Hardy-Weinberg Equilibrium test (**hweOri**) for each site as defined by Wigginton et al. (2005) [47] and the level of unique information contributed by each SNP given the level of LD with neighbouring SNPs, and calculated as the weight (**Weight**) obtained by the LDAK5 software [48]. The variation of the imputation quality variables (*Props*, *lbProps* and *cProps*) were analysed according to the different factors by a principal component analysis. The factor's relevance to describe the imputation quality variables were quantified with a Boruta algorithm which is a wrapper built around the random forest classification algorithm implemented in the R (R Core Team 2015) package Borut [49]. This algorithm created "shadowMean", "shadowMax" and "ShadowMin" attribute values obtained by the shuffling of the original attributes across objects. This set of created attributes is used as a framework of reference. The value of the importance of the factors tested, must be different from the values of the attributes created, to be considered as having importance in explaining the observed variability.

Linkage Disequilibrium

Plink software [50, 51] was used to estimate the linkage disequilibrium parameter D' [52] in the SNPchip dataset and after imputation in the SNPseq dataset. Both sets were previously phased. The SNPseq dataset was further filtered based on Props (> 0.9) and cProps (> 0.8) variables, in order to provide for the LD analysis positions with few or no errors after imputation.

Annotation analysis

We were interested in assessing to what extent imputation could change the annotation profile of covered SNPs, notably given the fact that the process involved a substantial change in density. Changes in annotation profiles from enriched to non-enriched but denser genotypes could be of relevance when using the resulting genotypes to fit prediction models for a large spectrum of traits. To get an annotation profile, a gene annotation analysis was performed. The tool Annovar (v. 2017Jul16) [53] was used with the command "–geneanno -buildver" in the *Populus trichocarpa* v3.1 gene set.

Table 2: SNP Filter step

Filtering step	Freebayes	gVCF-GATK
No filter	26,489,941	27,475,756
vcftools (max allele=3, min allele=2, minQ=30)	5,011,303	10,474,367
Monomorphic within P. nigra individuals	1,246,546	2,504,973
Common positions between the two callers	2,488,736 (375,566)	
Homology between two callers more than 95%	1,612,432	
Consistent Mendelian Segregation	1,466,586 (208,217)	

Number of variants detected in the 43 sequenced individuals using two callers with no filter and after filtering with different parameters to obtain the input dataset used for imputation. In brackets, the number of Indels out of the total number of variants.

Results

Mapping and genotype calling results

Sequence datasets for every individual were mapped on the *P. trichocarpa* reference genome v.3.1. In average, 91.7% of reads were mapped, 76.5% were paired and only 2.2% were singletons. The genome coverage was calculated by individual, and it varied between 4X and 52X, with a mean coverage of 13X (**Table S1**[see Additional file 1]). A total of 27,475,756 SNPs and Indels were detected by gVCF-GATK, whereas 26,489,941 SNPs were detected by Freebayes (**Table 2**). After scoring the SNPs on a quality criterion (Phred score > 30), the number of trimmed positions were twice as many with gVCF-GATK than with Freebayes (**Table 2**). Among the remaining positions, some were monomorphic within *P. nigra* individuals but different from the reference sequence: about 1 million for gVCF-GATK and twice as much for Freebayes. A total of 2,488,736 positions were common between the two callers at that point of the filtering. Among these positions, 17% were Indels and 83% SNPs. To simplify, and given the relatively low frequency of Indels (17% of variants), SNPs and Indels in the study were both denoted under the same acronym of "SNPs" hereafter. To have the best quality in genotype calling, we kept the positions where the genotype calling was at least 95% similar between the two callers for all individuals. Mendelian segregation was checked on available trios, and 142,974 positions were removed for which the progeny were inconsistent with parents. For the chip, after applying quality filters, 7,540 SNPs were recovered for the population under study and were used to impute 1,466,586 SNPs from sequences along the 19 Chromosomes. In other words, we imputed 99% of the data.

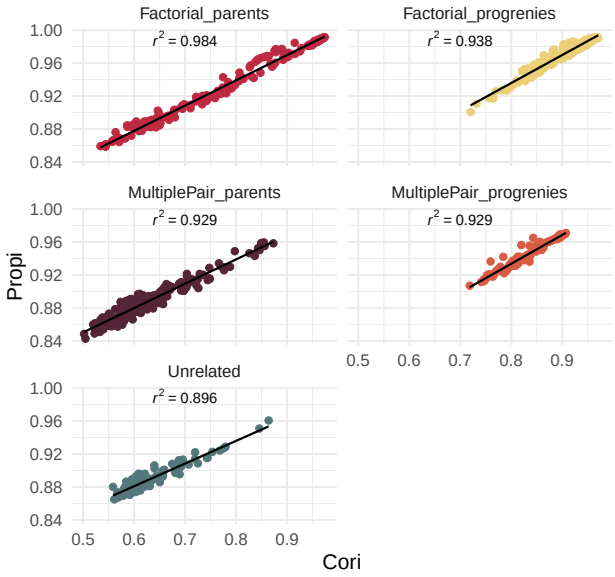


Figure 2: **Comparaison of two imputation accuracy variables.** Relationship between the proportion of alleles correctly imputed by each leave-one-out individual (*Propi*) and the Pearson's correlation coefficient between true and imputed individual genotypes (*Cori*). The different panels correspond to the different individual classes in the mating regimes, and each point represents the values for one chromosome and one individual. The correlation value is given in each panel and derives from the fitted regression line.

Imputation quality at the individual level

The Pearson's correlation between true and imputed individuals for each chromosome (*Cori*) was strongly correlated with the individual proportion of SNPs correctly imputed (*Propi*) per chromosome ($R^2 = 0.991$, **Figure 2**), with the former varying between 0.5 and 0.96, and the latter between 0.84 and 0.99. The coefficient of correlation between Cori and Propi was consistently high across individual classes (MultiplePair_parents: 0.929, Factorial_progenies: 0.938, MultiplePair_progenies: 0.929 and Factorial_parents: 0.984), even for unrelated individuals where it was slightly lower with 0.896 (**Figure 2**). Propi versus *Cori* relatedness clouds were differently clustered depending on the class of individuals (**Figure 2**). In general, factorial mating design progenies had higher *Propi* and *Cori* values (respectively from 0.94 to 0.98 and from 0.81 to 0.95) than those in the Multiple pair mating design progenies (from 0.93 to 0.96 and from 0.80 to 0.88).

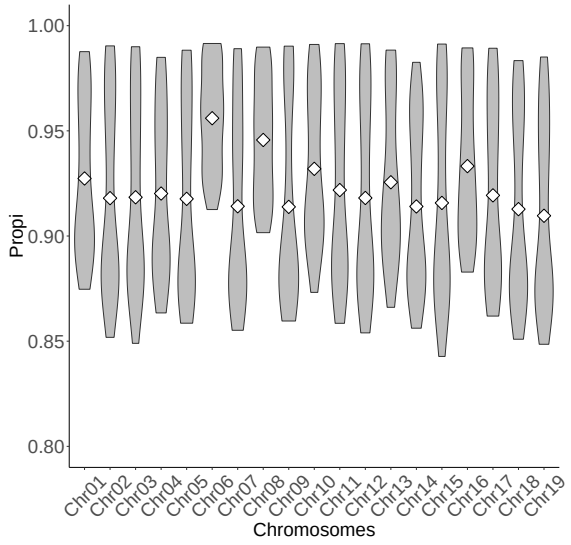


Figure 3: **Proportion of individual correctly imputed by chromosomes** Distribution of the proportion of SNPs correctly imputed by chromosomes (*Propi*). White diamond symbol stands for the mean.

Progenies from either of the two schemes had higher *Propi* and *Cori* values than those in the parental groups (from 0.87 to 0.90 and from 0.57 to 0.65). The parents of the factorial mating design resulted in the most variable ranges for *Propi* and *Cori* with respectively from 0.88 to 0.99 and 0.6 to 0.96, respectively, although that class had on average higher values than those found in parents in the multiple pair mating scheme. Finally, the unrelated individuals are in the lowest part of *Propi* and *Cori* variation (with respectively from 0.89 to 0.90 and from 0.62 to 0.63). There was no separate group within individual's categories (**Figure 2**) meaning that the individual class ranking was consistent along the chromosomes. The individual lower bound for imputation accuracy (*lbPropi*) was moderately correlated to *Propi* (**Figure S1**[see Additional file 2]). The ranking of individual classes was equivalent between *lbPropi* and *Propi*. However, there appears to be a higher gain in *Propi* with respect to *lbPropi* (i.e., using pedigree and LD versus frequencies) for the multiple pair-mating progenies, factorial progenies and factorial parents than for the multiple pair mating parents and unrelated individuals. In **Figure 3**, Propi distribution is shown per chromosome. This averaged imputation accuracy was roughly similar for all chromosomes, except for chromosomes 6 and 8 where means were substantially higher (respectively 0.96, and 0.95). No relationship between the sequencing depth (**MEAN_DEPTH**) and *Propi* was found at individual level whereas a poorly significant correlation seems to be present between depth (**MEAN_DEPTH**) and *lbPropi* and *Cori* (**Figure S2** [see Additional file 3]). In summary, at the individual level, imputation accuracy was high with a proportion of SNP correctly imputed ranging between 0.84 and 0.99. The variation was mostly due to the relatedness between individuals and to a lesser extent to sequencing quality or sequencing depth.

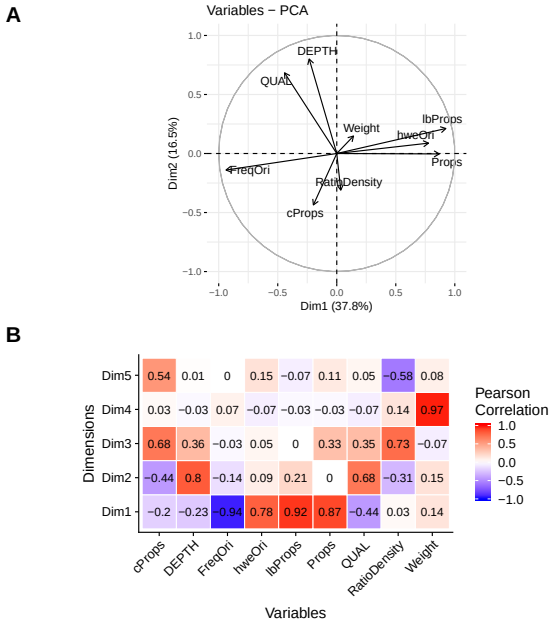


Figure 4: **Principal Component Analysis of Factors affecting SNP imputation** (A) Principal Component Analysis factor map of factors calculated at SNP level: *Props*: proportion of SNPs correctly imputed; *cProps*: proportion of SNPs correctly imputed and corrected by the minor allele frequency; *lbProps*: lower bound proportion of SNPs correctly imputed based only on allelic frequency; **hweOri**: p-value of a Hardy-Weinberg Equilibrium test for each site [47]; Weight: LD weight estimate obtained with the LDAK5 software; FreqOri: original allelic frequency in the sequenced individuals; **QUAL**: per-site SNP quality from the calling step; **DEPTH**: sequencing depth per site summed across all individuals ; **RatioDensity**: ratio between SNPchip density and SNPseq density in a 500kb window. (B) Correlations between parameters calculated at SNP level and dimension of the ACP from figure 3A.

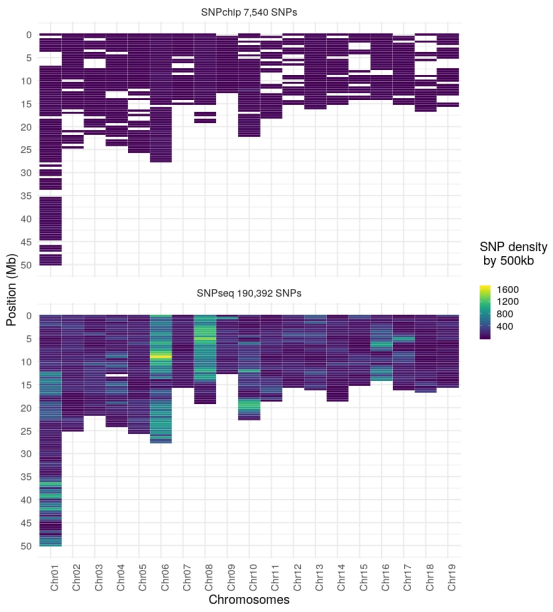


Figure 5: **Comparaison of density marker before and after imputation** SNP density map before imputation (top panel), corresponding to the SNP chip genotyping, and after imputation from sequence (bottom) in 500 kb windows. SNPs were selected on two different criteria based on the percentage of alleles correctly imputed: *Props* (> 0.90) and *cProps* (> 0.80). The scale colour represents the density of markers, with dark blue for low density and yellow for high density.

Imputation quality at the SNP level

A strong correlation between *Cors* and *cProps* (0.94) suggests that similar information was relayed by these two variables despite the frequency-based correction. The **Figure S3** [see Additional file 4] shows the variation of the three different estimates of imputation quality at the SNP level (*Props*, *lbProps*, *cProps*), as a function of different classes of minor allele frequency (**FreqOri**). While for low **FreqOri**, *Props* and *lbProps* distributions remained similar, with increasing frequencies their respective distributions tended to separate from each other. The frequency dependent correction applied to cProps was strongest at low frequencies, making *cProps* much lower on average than the other two counterparts. With increasing frequency, that correction was weaker with *cProps* getting closer to both *Props* and *lbProps*. This suggests that, while the problem of sensibility to frequencies can be easily overcome, *cProps* shows imputation qualities that can be far lower than what is actually observed.

The first 5 axes of the principal component analysis (PCA) considering the three estimates of imputation quality and six factors that potentially affect this quality, explained 90% of the variance (PC1 and PC2, explained respectively 37.8 and 16.5% of the variation; **Figure 4A**). *Props* showed the highest independence with respect to the sequence depth (**DEPTH**), the SNP quality (**QUAL**), *cProps*, the ratio between SNPchip density and SNPseq density (**RatioDensity**) and, to lesser extent, to the level of unique information contributed by each SNP (**Weight**). Props was negatively correlated to the **FreqOri** and positively correlated to the p-value of an exact Hardy-Weinberg Equilibrium test (**hweOri**) and to *lbProps*. In **Figure 4B**, correlation of each variable to the PCA dimensions are shown. The first dimension was negatively correlated to **FreqOri** (-0.94), and positively correlated to **hweOri** (0.78), *lbProps* (0.92) and *Props* (0.87). Sequencing quality parameter **QUAL** and **DEPTH** are highly correlated to the second dimension (respectively 0.68 and 0.8). **RatioDensity** and *cProps* were correlated to the third and fifth dimensions whereas the **Weight** variable was only strongly correlated to the fourth dimension. The Boruta analysis ranked the importance of the different factors considered to explain the variation in *Props*, *cProps* and *lbProps* variables (**Table 3**). All factors were quantified as being of higher importance than those of lower bond references in shadow attributes. **RatioDensity** resulted in the highest importance among all factors for *Props* and *cProps* with effects respectively being 1,351 and 1,182, largely ahead of the rest of factors, with effects ranging between 40 and 115 for *Props*, 33 and 132 for *cProps*. *lbProps* showed a different ranking of factors, dominated by **FreqOri** with the maximum effect among factors, which is expected given the fact that it is based on allele frequency. In summary, the quality of imputation at a SNPs level strongly depended on **RatioDensity** and to a lesser extent on **FreqOri**. By selecting SNP sets on *Props* and *cProps* simultaneously, we obtained 190,392 SNP with good imputation quality (*Props* > 0.90), while their level of polymorphism was not forced towards low allele frequencies (*cProps* > 0.80). The SNPs distribution along the genome after imputation was more homogeneous than what was initially available with the SNPchip (**Figure 5**).

Linkage Disequilibrium

The linkage disequilibrium (D') calculated in SNPchip and SNPseq sets is represented in **Figure 6A**, with density distributions showing that LD was lower in SNPseq than in SNPchip. This difference

Table 3: Estimation of importance of different explanatory factors by Boruta analysis

Factor	cProps (Mean ± SD)	lbProps (Mean ± SD)	Props (Mean ± SD)
shadowMax	1.44 ± 0.93	1.48 ± 0.70	1.80 ± 1.30
shadowMean	-0.05 ± 0.79	-0.22 ± 0.52	-0.01 ± 0.82
shadowMin	-2 ± 0.53	-2.22 ± 1.25	-1.57 ± 0.91
hweOri	32.96 ± 1.04	39.84 ± 0.73	40.33 ± 1.99
QUAL	98.95 ± 5.12	67.83 ± 1.70	67.90 ± 1.76
Weight	131.86 ± 3.56	92.78 ± 4.20	101.57 ± 4.23
FreqOri	64.28 ± 2.19	**110.92 ± 2.79**	115.02 ± 3.28
DEPTH	114.21 ± 5.08	75.51 ± 1.67	114.81 ± 4.81
RatioDensity	**1,182.87 ± 39.82**	36.68 ± 1.50	**1,351.57 ± 43.94**

Boruta analyses for the different explanatory factors assumed for imputation quality variables *Props*, *cProps* and *lbProps*. Values correspond to averaged effects and their corresponding standard deviations allowing for a ranking of importance of the factors. The maximum value is bolded. *Props*: proportion of SNPs correctly imputed; *cProps*: proportion of SNPs correctly imputed corrected by the minor allele frequency; *lbProps*: lower bound proportion of SNPs correctly imputed based only on the allelic frequency; **hweOri**: p-value of a Hardy-Weinberg Equilibrium test for each site [47] ; **Weight**: LD weight estimate with the LDAK5 software; **FreqOri**: original allelic frequency in the sequenced individuals; **QUAL**: per-site SNP quality from the calling step; **DEPTH**: sequencing depth per site summed across all individuals ; RatioDensity: ratio between SNPchip density and SNPseq density in a 500kb window. "ShadowMean", "shadowMax" and "ShadowMin" correspond to effects obtained by shuffling the original attributes across objects and used as a reference for deciding which factors are truly important.

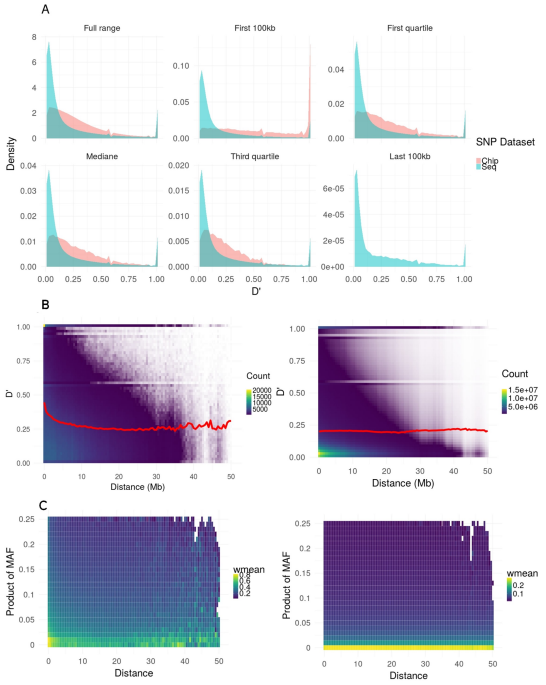


Figure 6: **Comparaison of linkage disequilibrium before and after imputation** Distribution of D' values of linkage disequilibrium for the two SNP sets in the study: SNPchip (pink) and SNPseq (blue) and over different ranges of physical distances (panel A). Panel B represents the distribution of D' values versus distances in a heat-plot with low densities in blue and high densities in yellow, respectively for SNPchip (left) and SNPseq (right). The red line is the average value of D' weighted by frequencies for a distance window of 500kb. Panel C represents the distribution of D' values as a function of distances between any two positions and the product of the corresponding minor allele frequencies in the pair of loci, with colour indicating the average value of D' weighted by frequencies for a distance window of 500kb from low range (blue) to high range (yellow), respectively for SNPchip (left) and SNPseq (right).

between sequence and chip sets was consistent over classes of distances across the genome. **Figure 6B** represents heat-maps for D' values according to physical distances. In general, D' decreased with increasing distances, as expected, although this trend was noticeably clearer for SNPchip than for SNPseq. For SNPchip, that D' decay was noticeable at the very shortest distance lags, with a bottom value for the mean sitting at 0.25. Some increases were observed at the highest distances, but this corresponded to very few number of points. For SNPseq, on the contrary, the weighted mean was almost invariable over distances with a mean value of 0.2. The very large numbers of short distance pairs with low D' had a high impact on the pattern of the weighted mean. **Figure 6C** presents the results under an alternative view in order to explain the differences in patterns between SNPchip and SNPseq. D' values are plotted as a function of distance and product of MAF of involved alleles, with the idea of checking to what extent the levels of D' was the result of low allelic frequencies in SNPseq. For the SNPchip set, the highest values of D' were found distributed over different distances and levels of MAF products, with a concentration of maximum values at very short distances and relatively low levels of MAF. The picture is substantially different with the SNPseq, where the highest values of D' were found exclusively at a very narrow band of low frequencies, suggesting that at least part of the levels in D' could be explained by the low polymorphisms brought by the sequence. As a consequence, the imputation did not appear to result in an increase of LD, but rather the opposite due to the differences in spectra of frequencies between SNPchip and SNPseq.

Annotation

Table 4: Proportion of Annotated SNP in genomic regions and mutation types

	Value % (number)	
	SNPchip	**SNPseq**
Region variant hit		
downstream	2,43 (183)	6,23 (12338)
exonic	36,18 (2728)	14,96 (29607)
intergenic	3,78 (285)	32,28 (63901)
intronic	37,02 (2791)	30 (59377)
splicing	0,04 (3)	0,1 (192)
exonic; splicing	0 (0)	0,001 (3)
upstream	1,8 (136)	5,71 (11307)
UTR3	8 (603)	6,3 (12470)
UTR5	3,55 (268)	3,02 (5968)
UTR5; UTR3	0 (0)	0,01 (20)
upstream; downstream	0,6 (45)	1,18 (2340)
Annotated Positions	93,4 (7042)	99,79 (197523)
Total number of Positions	100 (7540)	100 (197932)
Mutation type		
frameshift deletion	0 (0)	0,12 (246)
frameshift insertion	0 (0)	0,06 (118)
Non-frameshift deletion	0 (0)	0,08 (159)
Non-frameshift insertion	0,01 (1)	0,04 (84)
synonymous SNV	19,09 (1439)	6,64 (13133)
Non-synonymous SNV	16,92 (1276)	7,85 (15534)
Stop gain	0,15 (11)	0,15 (299)
Stop loss	0,01 (1)	0,02 (35)
Total number of exonic positions	**36,18 (2728)**	**14,96 (29608)**

Annotation results for SNPchip and SNPseq in percentage of counts per annotation category, and number of corresponding positions in brackets. For region variant hit: exonic;splicing corresponds to a variant within exon region but close to exon/intron boundary; UTR5;UTR3 corresponds to a variant positioned where two coding regions overlapped, one in forward and one in reverse; upstream;downstream corresponds to a variant positioned in an intergenic region between two neighbouring genes.

A total of 93.4% of SNPchip and 99.79% of SNPseq were annotated (**Table 4**). Most categories in the annotation catalog were enriched in the SNPseq compared to the corresponding levels of enrichment in the SNPchip. In the exonic region, SNPs were categorized depending on different mutation types. With SNPseq new locations, three new mutation types were represented: frameshift deletion, frameshift insertion and non-frameshift deletion. In summary, the genotype densification not only brought a better distribution of markers all along the genome, but also no loss in annotation categories.

Discussion

In this study, we have shown that substantial (26-fold) densification in marker coverage is possible in up to 1000 individuals through imputation from a few sequenced nodal individuals (43). Simultaneously, we have achieved imputation qualities higher than 0.84, which is sufficient for a heterozygous species like poplar but may be insufficient when working with species involving inbred lines. This imputation quality is similar to the one obtained on horses [34] with Impute2 software or in cattle [33], and higher than the one obtained on chickens [35]. The study is based on a subset of a breeding population in black poplar, with a relatively low effective number of contributing parents, which could explain partly the success of the imputation. However, this situation is far from exceptional and could be easily found in many other species going through breeding activities, where an elite of a few dozens of parents can contribute substantially to next generation [54]. Although relatedness between the group bringing marker density and the group to be imputed is key in the success of imputation [21, 24, 55], our study demonstrated also that imputation works with relatively small losses in quality when inferring unrelated individuals taken from a diversity collection of the natural range of the species in France. Moreover, such a substantial 26-fold imputation did not appear to increase artefactually the levels of LD. The annotation of imputed positions showed no loss in annotation categories compared to original low density coverage. These two results suggest that imputed data can be of enough quality to be the base of subsequent studies in genome-wide predictions.

The use of a "leave-one-out" cross validation scheme allowed us to ascertain the actual quality of the imputation, both by individuals and by SNP positions. The proportion of alleles correctly imputed by SNP gave the actual value of the imputation quality, although with the drawback of an allele frequency bias. Indeed, a selection based on that proportion by SNP alone could potentially favor positions with low MAF over the rest, as imputation is easier when one of the alternative alleles is rare. The correction we used based on the work of Badke et al. [46] compensated this bias. This measure is interesting whenever we wish to compare results between different imputation methods or between different software. However, it offers a less intuitive criterion, not easily connected to the actual values of imputation error. Therefore, we proposed to combine the actual value of the imputation quality and the frequency-based corrected measure to select SNPs that fulfil both criteria with high level values. Both criteria were given equal importance. The result in our study led to positions with the highest imputation quality while not necessarily resulting in an excess of rare alleles in the imputed population.

Many factors can affect imputation quality like LD, density ratio, minor allele frequency or relatedness between target and reference populations [56, 57]. Our results showed that all these factors considered in our study impacted to various degrees the quality

of imputation. It seems difficult to provide general predictor for the imputation quality based on these or other factors. For instance, [4] suggest that there is no obvious pre-imputation filter ensuring a good imputation quality. However, one of the factors with the highest impact on imputation quality in our study was the marker density in the neighborhood of the considered position for imputation. This is a somehow logical outcome, in the sense that numerous markers in dense regions would mutually facilitate their imputation through the extent of LD. These results were consistent with the fact that the imputation accuracy decrease with increasing distance between markers [58]. When designing a low-density chip, it is therefore important to choose SNPs regularly spaced. These results are consistent with the results of He et al. 2018 [59], which showed that an evenly-spaced SNPs combined with an increased minor allele frequencies SNP panel showed the best results.

Imputation requires some degree of LD in existing genomes to reconstruct missing positions [21]. Whenever the reconstruction comprises large chunks of genomes, like in our case here, one could hypothesize that there could be a risk of artefactually increasing the frequency of certain extant haplotypes and, therefore, exacerbate LD among imputed positions. A similar hypothesis has been already proposed by Pimentel et al. [27]. However, what we found appears to be the opposite, with a reduction in D' from 0.25 in the chip to less than 0.2 in the sequence, on average. The imputed sequence led to D' values in the low range (close to zero), which could be related to the fact that sequences harbor high number of rare alleles for many positions. Some studies [60, 61] showed that the upper limit of LD between two SNPs is mathematically determined by their difference in MAF. In case of extreme differences, alleles cannot match, even at small distances between SNPs, resulting in low LD. A decrease of LD between SNPs could be problematic for subsequent studies based on imputed data, especially at short distances. Indeed, LD is used to capture the effect of nearby quantitative traits loci (QTL), whenever SNPs are not directly placed on the QTL. This potential loss in capacity to capture QTL effects in the imputed sequences might be compensated for by the genotyping densification, which could extend the reach of markers to unexplored regions involving new QTLs. In summary, genotype densification allowed to have a better repartition of the markers along the genome and in different genomic regions. In our case, the proportion of SNPs in intergenic regions increased with the imputation, this compensated the bias of our low-density SNP chip which was enriched in coding regions [25]. Better marker repartition all along the genome could be useful to detect causal variants, as suggested by Jansen et al. [62]. They showed that with the imputation of missing data, the value of Phred-score genotype quality was improved. This lead to a better genotyping quality, a better causal variant identification in association studies and a better variant annotation. Sequences in our study have brought new spectra of allele frequencies, involving a much higher proportion of rare alleles compared to the chip data, which resulted from a carefully selected set of highly polymorphic markers [25]. While low frequencies could have some interest in diversity studies or kinship assignment [63], their use in the context of genomic evaluation or GWAS would be challenging because of power issues unless the involved rare alleles produce very large effects and are captured with large sample sizes.

From an operational point of view, our results showed that imputation can represent a good strategy to reduce genotyping costs. By using a few well-chosen sequenced individuals in the population, very good imputation results could be obtained and considerably increase the number of SNPs available. It is therefore possible to create a low-density chip to impute at high density via sequenced individuals. This could minimize differences in imputation quality along the genome and avoid any over-representation of certain chromosome regions. This type of strategy can be used in a breeding improvement program on several generations. Yet, it would be required to add high density genotyping or sequences every generation [64] in order to keep a high imputation accuracy. Not doing so could reduce the quality of imputation and result in accumulating errors over subsequent generations. Our study is a first step before using gathered genotypes for genome-wide predictions. The impact of imputation accuracy on genomic selection accuracy was studied by several authors. The genotype densification allowed to increase the genomic evaluation accuracy depending on the architecture of evaluated traits [65, 66]. Moreover, genomic selection accuracy increased with better imputation accuracies [26, 28]. The marker effect estimation could be biased and inbreeding levels could be under-estimated [27], if the imputation accuracy is too low.

Conclusion

In conclusion, we have demonstrated in this study that high imputation quality is possible even from low density marker sets. The relatedness had an important impact on the imputation quality at the individual level, but it is possible to impute unrelated individuals

with a good performance. All factors studied here had an impact on the imputation quality at the SNP level, but there is no obvious way to use their effects as criteria for a pre-imputation filter. The genotype densification towards sequences induced a decrease of linkage disequilibrium, due to the spectra of low allelic frequencies. The densification allowed to correct bias in variant annotation profile of the SNPchip marker set, with a better distribution in all genomic region categories.

Abbreviations

- AFA : Adaptive Focused Acoustics
- DNA : DeoxyriboNucleic Acid
- EM : Expectation Maximization
- F1 : first filial generation
- GWAS : Genome-wide association study
- HMM : Hidden Markov Model
- Indels : an Insertion or Deletion of bases in the genome of an organism
- LD : Linkage disequilibrium
- MAF : Minor Allele Frequency
- PCR: Polymerase Chain Reaction
- qPCR : quantitative Polymerase Chain Reaction
- QTL : Quantitative Trait Loci
- SBS : Sequencing By Synthesis
- SNP : Single Nucleotide Polymorphism
- SNV : Single Nucleotide Variation
- UTR3 : 3' Untranslated region
- UTR5 : 5' Untranslated region

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data
This Whole Genome resequencing project has been submitted to the international repository Sequence Read Archive (SRA) from NCBI RA as BioProject BreedToLast PRJNA483561.The datasets analysed during the current study are available in the INRA Data Portal repository, https://data.inra.fr/privateurl.xhtml?token=0f26535e-4c12-4907-8d2b-ce69e39c1ee0.

Competing interests
The authors declare that they have no competing interests.

Funding
This study was funded by the following sources : sequencing and genotyping data by the INRA AIP Bioressource, EU NovelTree (FP7 - 211868), EU Evoltree (FP6-16322), and INRA SELGEN funding program (project BreedToLast) ; the PhD grant of MP by INRA SELGEN funding program (BreedToLast) and by Region Centre - Val de Loire funding council.

Author's contributions
MP performed the analyses and drafted the manuscript. OR developped the variant calling pipeline scripts and helped for sequence data preparation and bioinformatics. AB, PFR and MCLP provided the sequence and genotyping datasets. CB provided access to plant material as scientist responsible for the *Populus nigra* breeding program. VJ and LS designed the study, assisted in drafting the manuscript, and obtained funding. All co-authors significantly contributed to the present study. All authors read and approved the final manuscript.

Acknowledgements
The authors acknowledge The authors acknowledge the "GIS peuplier", the UE GBFOR and PNRGF (ONF) for access, maintenance, and sampling of plant material. The authors want to thank Vincent Segura and Aurélien Chateigner for their valuable discussions; Vanina Guérin and Corrine Buret for their work on DNA extraction for the genotyping and the sequencing; Aurélie Chauveau and Isabelle Le Clainche for libraries construction, sequencing and Infinium genotyping; Elodie Marquand and Aurélie Canaguier for data processing and management. EPGV group acknowledges also CEA-IG/CNG by conducting the DNA QC and by providing access for their Illumina Sequencing and Genotyping platforms.

Author details
[1]BioForA, INRA, ONF, 45075, Orléans, France, 2163 Avenue de la Pomme de Pin CS 40001 ARDON, 45075 Orléans Cedex 2, France. [2]Etude du Polymorphisme des Génomes Végétaux (EPGV), INRA, Université Paris-Saclay, 91000, 2 rue Gaston Crémieux, 9100 Evry, France.

References

1. Marchini, J., Howie, B.N., Myers, S., McVean, G., Donnelly, P.: A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. **39**(7), 906–913 (2007). doi:10.1038/ng2088. 1110.6019
2. Marchini, J., Howie, B.: Genotype imputation for genome-wide association studies. Nat. Rev. Genet. **11**(7), 499–511 (2010). doi:10.1038/nrg2796. arXiv:1507.02142v2
3. Burdick, J.T., Chen, W.-M., Abecasis, G.R., Cheung, V.G.: In silico method for inferring genotypes in pedigrees. Nat. Genet. **38**(9), 1002–1004 (2006). doi:10.1038/ng1863
4. Roshyara, N.R., Kirsten, H., Horn, K., Ahnert, P., Scholz, M.: Impact of pre-imputation SNP-filtering on genotype imputation results. BMC Genet. **15**(1), 88 (2014). doi:10.1186/s12863-014-0088-5
5. Berry, D.P., McHugh, N., Randles, S., Wall, E., McDermott, K., Sargolzaei, M., O'Brien, A.C.: Imputation of non-genotyped sheep from the genotypes of their mates and resulting progeny. animal **12**(02), 191–198 (2018). doi:10.1017/S1751731117001653
6. Browning, S.S.R., Browning, B.B.L.: Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. **12**(10), 703–714 (2011). doi:10.1038/nrg3054
7. Howie, B.N., Donnelly, P., Marchini, J.: A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet. **5**(6), 1000529 (2009). doi:10.1371/journal.pgen.1000529
8. Scheet, P., Stephens, M.: A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am. J. Hum. Genet. **78**(4), 629–644 (2006). doi:10.1086/502802
9. Daetwyler, H.D., Wiggans, G.R., Hayes, B.J., Woolliams, J.A., Goddard, M.E.: Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing. Genetics **189**(1), 317–327 (2011). doi:10.1534/genetics.111.128082
10. Meuwissen, T., Goddard, M.: The Use of Family Relationships and Linkage Disequilibrium to Impute Phase and Missing Genotypes in Up to Whole-Genome Sequence Density Genotypic Data. Genetics **185**(4), 1441–1449 (2010). doi:10.1534/genetics.110.113936
11. Sargolzaei, M., Chesnais, J.P., Schenkel, F.S.: A new approach for efficient genotype imputation using information from relatives. BMC Genomics **15**(1), 478 (2014). doi:10.1186/1471-2164-15-478
12. Lander, E.S., Green, P.: Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. **84**(8), 2363–2367 (1987). doi:10.1073/pnas.84.8.2363
13. Elston, R.C., Stewart, J.: A General Model for the Genetic Analysis of Pedigree Data. Hum. Hered. **21**(6), 523–542 (1971). doi:10.1159/000152448
14. Heath, S.C.: Markov Chain Monte Carlo Segregation and Linkage Analysis for Oligogenic Models. Am. J. Hum. Genet. **61**(3), 748–760 (1997). doi:10.1086/515506
15. Huber, M., Chen, Y., Dinwoodie, I., Dobra, A., Nicholas, M.: Monte Carlo Algorithms for Hardy-Weinberg Proportions. Biometrics **62**(1), 49–53 (2006). doi:10.1111/j.1541-0420.2005.00418.x
16. Johnston, J., Kistemaker, G., Sullivan, P.G.: Comparison of Different Imputation Methods. Interbull Bull. (44), 25–33 (2011)

17. Li, Y., Willer, C., Sanna, S., Abecasis, G.: Genotype Imputation. Annu. Rev. Genomics Hum. Genet. **10**(1), 387–406 (2009). doi:10.1146/annurev.genom.9.081307.164242
18. Browning, S.R., Browning, B.L.: Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. **81**(5), 1084–1097 (2007). doi:10.1086/521987
19. Browning, B.L., Browning, S.R.: Genotype Imputation with Millions of Reference Samples. Am. J. Hum. Genet. **98**(1), 116–126 (2016). doi:10.1016/j.ajhg.2015.11.020
20. Mikhchi, A., Honarvar, M., Kashan, N.E.J., Aminafshar, M.: Assessing and comparison of different machine learning methods in parent-offspring trios for genotype imputation. J. Theor. Biol. **399**, 148–158 (2016). doi:10.1016/j.jtbi.2016.03.035
21. Hickey, J.M., Crossa, J., Babu, R., de los Campos, G.: Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. **52**(2), 654 (2012). doi:10.2135/cropsci2011.07.0358
22. Hickey, J.M., Gorjanc, G.: Simulated Data for Genomic Selection and Genome-Wide Association Studies Using a Combination of Coalescent and Gene Drop Methods. G3 Genes, Genomes, Genet. **2**(4), 425–427 (2012). doi:10.1534/g3.111.001297
23. Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E.: Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet. **6**(9), 1001139 (2010). doi:10.1371/journal.pgen.1001139
24. Huang, Y., Hickey, J.M., Cleveland, M.A., Maltecca, C.: Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. **44**(1), 25 (2012). doi:10.1186/1297-9686-44-25
25. Faivre-Rampant, P., Zaina, G., Jorge, V., Giacomello, S., Segura, V., Scalabrin, S., Guérin, V., De Paoli, E., Aluome, C., Viger, M., Cattonaro, F., Payne, A., PaulStephenRaj, P., Le Paslier, M.C., Berard, A., Allwright, M.R., Villar, M., Taylor, G., Bastien, C., Morgante, M.: New resources for genetic studies in Populus nigra : genome-wide SNP discovery and development of a 12k Infinium array. Mol. Ecol. Resour. **16**(4), 1023–1036 (2016). doi:10.1111/1755-0998.12513
26. Cleveland, M.A., Hickey, J.M.: Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation1. J. Anim. Sci. **91**(8), 3583–3592 (2013). doi:10.2527/jas.2013-6270
27. Pimentel, E.C.G., Edel, C., Emmerling, R., Götz, K.-U.: How imputation errors bias genomic predictions. J. Dairy Sci. **98**(6), 4131–4138 (2015). doi:10.3168/jds.2014-9170
28. Tsai, H.-Y., Matika, O., Edwards, S.M., Antolín–Sánchez, R., Hamilton, A., Guy, D.R., Tinch, A.E., Gharbi, K., Stear, M.J., Taggart, J.B., Bron, J.E., Hickey, J.M., Houston, R.D.: Genotype Imputation To Improve the Cost-Efficiency of Genomic Selection in Farmed Atlantic Salmon. G3 Genes, Genomes, Genet. **7**(4), 1377–1383 (2017). doi:10.1534/g3.117.040717
29. Hozé, C., Fouilloux, M.-N., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., Ducrocq, V., Phocas, F., Boichard, D., Croiseau, P.: High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. **45**(1), 33 (2013). doi:10.1186/1297-9686-45-33
30. Berry, D.P., McClure, M.C., Mullen, M.P.: Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. J. Anim. Breed. Genet. **131**(3), 165–172 (2014). doi:10.1111/jbg.12067
31. Druet, T., Macleod, I.M., Hayes, B.J.: Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb). **112**(1), 39–47 (2014). doi:10.1038/hdy.2013.13
32. VanRaden, P.M., Sun, C., O'Connell, J.R.: Fast imputation using medium or low-coverage sequence data. BMC Genet. **16**(1), 82 (2015). doi:10.1186/s12863-015-0243-7
33. Brøndum, R., Guldbrandtsen, B., Sahana, G., Lund, M., Su, G.: Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics **15**(1), 728 (2014). doi:10.1186/1471-2164-15-728
34. Frischknecht, M., Neuditschko, M., Jagannathan, V., Drögemüller, C., Tetens, J., Thaller, G., Leeb, T., Rieder, S.: Imputation of sequence level genotypes in the Franches-Montagnes horse breed. Genet. Sel. Evol. **46**(1), 63 (2014). doi:10.1186/s12711-014-0063-7
35. Ye, S., Yuan, X., Lin, X., Gao, N., Luo, Y., Chen, Z., Li, J., Zhang, X., Zhang, Z.: Imputation from SNP chip to sequence: a case study in a Chinese indigenous chicken population. J. Anim. Sci. Biotechnol. **9**(1), 30 (2018). doi:10.1186/s40104-018-0241-5
36. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014). doi:10.1093/bioinformatics/btu170
37. Tuskan, G.A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., DePamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., Rokhsar, D.: The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science (80-.). **313**(5793), 1596–1604 (2006). doi:10.1126/science.1128691
38. Picard tools (2015). https://broadinstitute.github.io/picard/
39. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J.: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. **43**(5), 491–498 (2011). doi:10.1038/ng.806. NIHMS150003
40. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**(9), 1297–1303 (2010). doi:10.1101/gr.107524.110
41. Garrison, E., Marth, G.: Haplotype-based variant detection from short-read sequencing (2012). 1207.3907
42. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group: The variant call format and VCFtools. Bioinformatics **27**(15), 2156–8 (2011). doi:10.1093/bioinformatics/btr330
43. Chud, T.C.S., Ventura, R.V., Schenkel, F.S., Carvalheiro, R., Buzanskas, M.E., Rosa, J.O., Mudadu, M.d.A., da Silva, M.V.G.B., Mokry, F.B., Marcondes, C.R., Regitano, L.C.A., Munari, D.P.: Strategies for genotype imputation in composite beef cattle. BMC Genet. **16**(1), 99 (2015). doi:10.1186/s12863-015-0251-7
44. Toghiani, S., Aggrey, S.E., Rekaya, R.: Multi-generational imputation of single nucleotide polymorphism marker genotypes and accuracy of genomic selection. animal **10**(07), 1077–1085 (2016). doi:10.1017/S1751731115002906
45. Calus, M.P.L., Bouwman, A.C., Hickey, J.M., Veerkamp, R.F., Mulder, H.A.: Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. animal **8**(11), 1743–1753 (2014). doi:10.1017/S1751731114001803
46. Badke, Y.M., Bates, R.O., Ernst, C.W., Fix, J., Steibel, J.P.: Accuracy of Estimation of Genomic Breeding Values in Pigs Using Low-Density Genotypes and Imputation. G3 Genes, Genomes, Genet. **4**(4), 623–631

(2014). doi:10.1534/g3.114.010504
47. Wigginton, J.E., Cutler, D.J., Abecasis, G.R.: A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am. J. Hum. Genet. **76**(5), 887–893 (2005). doi:10.1086/429864
48. Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., Balding, D.J.: Reevaluation of SNP heritability in complex human traits. Nat. Genet. **49**(7), 986–992 (2017). doi:10.1038/ng.3865
49. Kursa, M.B., Rudnicki, W.R.: Feature Selection with the Boruta Package. J. Stat. Softw. **36**(11), 1–13 (2010). doi:Vol. 36, Issue 11, Sep 2010
50. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., Lee, J.J.: Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **4**(1), 7 (2015). doi:10.1186/s13742-015-0047-8. 1410.4803
51. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C.: PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. **81**(3), 559–575 (2007). doi:10.1086/519795. arXiv:1011.1669v3
52. Gaunt, T.R., Rodríguez, S., Day, I.N.M.: Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool 'CubeX'. BMC Bioinformatics **8**(1), 428 (2007). doi:10.1186/1471-2105-8-428
53. Wang, K., Li, M., Hakonarson, H.: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. **38**(16), 164–164 (2010). doi:10.1093/nar/gkq603
54. Pâques, L.E. (ed.): Forest Tree Breeding in Europe. Managing Forest Ecosystems, vol. 25. Springer, Dordrecht (2013). doi:10.1007/978-94-007-6146-9
55. Roshyara, N.R., Scholz, M.: Impact of genetic similarity on imputation accuracy. BMC Genet. **16**(1), 90 (2015). doi:10.1186/s12863-015-0248-2
56. Hickey, J.M., Kinghorn, B.P., Tier, B., Van Der Werf, J.H., Cleveland, M.A.: A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. **44**(1), 1–11 (2012). doi:10.1186/1297-9686-44-9
57. Pei, Y.-F., Li, J., Zhang, L., Papasian, C.J., Deng, H.-W.: Analyses and Comparison of Accuracy of Different Genotype Imputation Methods. PLoS One **3**(10), 3551 (2008). doi:10.1371/journal.pone.0003551
58. van Binsbergen, R., Bink, M.C.A.M., Calus, M.P.L., van Eeuwijk, F.A., Hayes, B.J., Hulsegge, I., Veerkamp, R.F.: Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. **46**(1), 41 (2014). doi:10.1186/1297-9686-46-41
59. He, J., Xu, J., Wu, X.-L., Bauck, S., Lee, J., Morota, G., Kachman, S.D., Spangler, M.L.: Comparing strategies for selection of low-density SNPs for imputation-mediated genomic prediction in U. S. Holsteins. Genetica **146**(2), 137–149 (2018). doi:10.1007/s10709-017-0004-9
60. Lewontin, R.C.: The detection of linkage disequilibrium in molecular sequence data. Genetics **140**(1), 377–388 (1995)
61. Mueller, J.C.: Linkage disequilibrium for different scales and applications. Brief. Bioinform. **5**(4), 355–364 (2004). doi:10.1093/bib/5.4.355
62. Jansen, S., Aigner, B., Pausch, H., Wysocki, M., Eck, S., Benet-Pagès, A., Graf, E., Wieland, T., Strom, T.M., Meitinger, T., Fries, R.: Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics **14**(1), 446 (2013). doi:10.1186/1471-2164-14-446
63. Eynard, S.E., Windig, J.J., Leroy, G., Binsbergen, R.V., Calus, M.P.L.: The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genet. **16**(1), 1–12 (2015). doi:10.1186/s12863-015-0185-0
64. Judge, M.M., Purfield, D.C., Sleator, R.D., Berry, D.P.: The impact of multi-generational genotype imputation strategies on imputation accuracy and subsequent genomic predictions. J. Anim. Sci. **95**(4), 1489 (2017). doi:10.2527/jas2016.1212
65. Frischknecht, M., Meuwissen, T.H.E., Bapst, B., Seefried, F.R., Flury, C., Garrick, D., Signer-Hasler, H., Stricker, C., Bieber, A., Fries, R., Russ, I., Sölkner, J., Bagnato, A., Gredler-Grandl, B.: Short communication: Genomic prediction using imputed whole-genome sequence variants in Brown Swiss Cattle. J. Dairy Sci. **101**(2), 1–5 (2017). doi:10.3168/jds.2017-12890
66. Zhang, C., Kemp, R.A.R.A., Stothard, P., Wang, Z., Boddicker, N., Krivushin, K., Dekkers, J., Plastow, G.: Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. Genet. Sel. Evol. **50**(1), 14 (2018). doi:10.1186/s12711-018-0387-9

Additional Files

Additional file 1 : TableS1.pdf — Sequencing, pedigree and reference information's of each reference individuals.

Additional file 2 : FigureS1.pdf — Relationship between the proportion of alleles correctly imputed by each leave-one-out individual (Propi) and the lower bound individual proportion of SNP correctly imputed IbPropi).The different colors correspond to the different individual classes in the mating regimes, and each point represents the values for one chromosome and one individual.

Additional file 3 : FigureS2.pdf — Relationship between the sequencing depth and imputation quality variables at individual level. On the top of the diagonal: Pearson's correlations. The distribution of each variable is shown on the diagonal. On the bottom of the diagonal: the bivariate scatter plots.

Additional file 4 : FigureS3.pdf — Variation of the three different estimates of imputation quality at the SNP level (*Props* (Green), *IbProps* (Purple), *cProps* (Orange)), as a function of different classes of minor allele frequency (**FreqOri**).

3.5 Chapter Global discussion

All the results presented in this chapter converged to the fact that good quality genotype imputation is possible in the context of the training populations used in the study. A relatively restricted genome sequence strategy involving a few dozens of nodal individuals combined with an imputation process have allowed to multiply by 185 the number of available markers for more than thousand individuals, for which only low densities were available. However, the quality of imputation was to some extent heterogeneous across markers and individuals. The relatively important number of sequenced individuals compared to the population to impute, together with the levels of relatedness in that population involving parents, offspring and siblings crearly had a beneficial impact on the quality of the imputation results. Between the first test and the study in the article, the number of individuals had increased substantially. Maybe because of that factor, the range of variation in imputation quality was substantially reduced. In general, the quality of imputation was improved by 7%. Another difference between first and final tests was the fact that differences across chromosomes became more evident in the latter. In this chapter, we also showed that all pre-imputation factors can explain part of the differences in imputation quality but that none of them can be operationally used to filter positions before imputation. One of the few factors, on which we have some decisional power, is the coverage homogeneity of the low density panel. According to our results in quality imputation, the methodology implemented in this chapter can be used to acquire a large number of positions over a large number of individuals at a lower cost than going through sequencing all the way. One of the few requirements is computing facilities and skills. As for computing, there is no need of large clusters for performing imputation in a dataset of our size. Some restrictions may however apply and must be kept in mind. The main limitation of our methodology is the size of the data sets and their storage. Sequence data can take up to a considerable amount of space of several terabytes, especially assuming that some bioinformatics steps often multiply the versions of files. In addition, genome assembly evolves and improves over time, leading to the need to update sequence mapping regularly. In our case, this could happen if a reference sequence in *P. nigra* is published. A sequence in the same species would definitely improve the quality of all the steps from

mapping to detecting SNPs, their location and organization on the physical map and therefore of the resulting imputation. Each time new sequenced individuals are added to the panel, it will be necessary to make a new detection with all individuals to ensure the quality of the calling genotype. The previous data is not lost, and quality imputed positions can be used to help the different steps of the imputation. This is especially true since FImpute corrects the original genotyping when it turns out to disagree with the pedigree. One last risk is the dependency on the FImpute software and its evolutions. The set of imputed data represents a pool of information. Depending on the objectives and the selection criteria for the selected SNPs, different data sets may be defined with varying densities. The results showed already by several authors on the genotype densification are propitious for implementing GS (Cleveland and Hickey, 2013; Tsai et al., 2017).

Chapter 4

Genomic evaluation

4.1 Résumé du Chapitre

Dans ce chapitre 4 de la thèse, nous avons abordé la question de la faisabilité de l'évaluation génomique par rapport à la méthode d'évaluation basée sur le pedigree actuellement utilisée. Plus précisément, sommes-nous capables de prédire les phénotypes de nouveaux individus avec suffisamment de précision à partir de la population d'amélioration, et quels sont les principaux facteurs qui influencent sur la qualité de cette prédiction ?

Dans cette partie, nous présentons les résultats de notre étude de validation de principe pour l'évaluation génomique avec un millier d'individus et quatre ensembles de marqueurs de densités différentes (7K, 50K, 100K et 250K). Ce chapitre contient les travaux les plus avancés en termes d'évaluation de la prédiction génomique chez le peuplier noir, dans le but principal d'identifier les conditions dans lesquelles cette évaluation pourrait être compétitive.

La stratégie d'échantillonnage s'est avérée avoir un impact conséquent sur la précision de la prédiction, contrairement à la proportion d'individus dans la population d'entraînement. Un échantillonnage individuel avec une représentation de chaque famille dans la population d'entraînement a donné les meilleurs résultats, ce qui pourrait correspondre sur le plan opérationnel à l'augmentation de l'intensité de la sélection au sein des familles par l'évaluation génomique. L'optimisation de l'échantillonnage via le CDmean obtenue avec un rééchantillonnage classique ou via un un algorithm de

"simulated annealing" optimisant les rééchantillonnage n'a pas conduit à de meilleures populations d'entraînement.

Ce chapitre met en lumière plusieurs faits :

1. Premièrement, un avantage systématique en termes de prédiction n'est pas nécessairement corrélé à la proportion de variance additive captée par le modèle.

2. Les modèles basés sur le génome, et en particulier le GBLUP, ont donné les meilleures précisions de prédiction, suivis de près par les modèles basés sur le pedigree après correction des erreurs via l'information des marqueurs.

3. Les avantages apportés par un génotypage plus dense dépendent des caractères et sont plus évidents avec l'association de modèles de sélection de variable (wGBLUP).

4. Les modèles basés sur le génome ont d'une certaine manière mieux réussi que les modèles basés sur le pedigree dans les situations les plus difficiles, avec une population de validation totalement indépendante ou avec des phénotypes moins précisément évalués.

5. L'utilisation de différents critères de qualité a révélé d'autres avantages des modèles basés sur le génome par rapport aux modèles généalogiques.

D'après nos résultats sur l'exactitude de la prédiction, il est nécessaire d'utiliser plusieurs méthodes, plusieurs critères de qualité et plusieurs densités de marqueurs afin de trouver les conditions obtimales pour la prédiction génomique. Parmi tous ces modèles et ensembles de données, la combinaison du modèle wGBLUP à la première itération combinée avec l'ensemble des SNPs 50K a donné généralement les meilleurs résultats.

4.2 Summary presentation of the chapter

In this Chapter 4 of the thesis, we addressed the question of the feasibility of GS in relation to the pedigree-based evaluation method used currently. More specifically, are we able to predict phenotypes of new individuals with enough precision from the breeding population, and what are the main drivers affecting the quality of this prediction? These are surely not new questions and were already raised by all precedent experiences on GS on different

species. We needed to raise it again, as often the performance of GS is context-dependent. We identified the conditions for which the new evaluation methodology could eventually be feasible and competitive, after considering the benefits of new modelling approaches like multiple-trait models, models with non-additive effects, different genetic architectures, marker densification, or the use of less repeatable phenotypes. We considered different ways of designing the training and assessed the subsequent impacts on evaluation performances.

Preliminary tests were performed on a restricted part of the factorial mating design composed of 294 individuals, divided into eight families, from three males and three females. Different methods and cross-validation stratégies were tested. The results were confronted with the results obtained with a larger dataset in terms of individuals but especially with more SNP markers. The process is briefly presented and discussed with the addition of some complementary results in this section, and before showing the second article of this thesis. The submission of this second article is planned before the end of the year to the G3 journal. In this article, we presented the results of our proof of concept study for GS with a thousand individuals, and four marker sets of different densities (7K, 50K,100K, and 250K). This paper contains the most advanced work in terms of assessment of GS in the thesis, with the main aim of identifying the conditions in which this evaluation could be competitive.

4.3 Preliminary tests

4.3.1 First test

Methods

A subsample of the factorial mating design was used for this first test of GS (table 3.1) with the genotyping information from the chip (Faivre-Rampant et al., 2016). Seven traits were evaluated, as explained in chapter 2, in a complete randomized block design for height, circumference and rust resistance at one and two years of age, and for proleptic branch angle. The phenotypes were adjusted for micro-environmental effects with a pedigree-based model involving a bi-spline approach for capturing microenvironmental effects at

individual level (package R breedR; Muñoz and Sanchez 2017; R3.3.1 platform).

Two genomic methods were used: Ridge-Regression BLUP and BayesCπ, both implemented in the GS3 software. Gibbs sampling chains were obtained to derive the posterior distribution for the different parameters in the model (chain parameters: iterations = 100,000, burning = 20,000 and thinning = 1/1,000). Convergence of the Markov Chains was evaluated graphically for every trait and method as presented in Figure 4.1 by using the R package coda (Plummer et al., 2006).

Seven cross-validation schemes were used to estimate prediction abilities: T75V25Rand, 75% of individuals was used for model calibration and the remaining 25% of individuals were used for prediction, with a 4-fold cross validation and 10 repetitions; T50V50Rand, 50% for calibration and 50% for prediction, again with 4-fold cross validation and 10 repetitions; T75V25CDmean and T50V50CDmean, both with a CDmean optimization for the choice of individuals within each family (Rincent et al., 2012) and only one repetition; T75V25CDmeanG and T50V50CDmeanG, both with a CDmean optimization for the choice of individuals in the global population and with one repetition; and TFamily : two rows (or columns) in the factorial were used for calibration, and the remaining row (or column) for prediction, with six repetitions. This strategy allowed to assess the ability of the GS to predict new parental crosses with at least two other parents in the training population.

For each of the 7 cross-validation schemes, predicting ability was calculated as the correlation between the GEBV and the phenotype (spatially-adjusted clonal mean). To assess the benefits of GS, the same cross-validation scenario and data were used with a pedigree-based BLUP. GS3 software allowed the fitting of models with additive genetic effects (ADD) and with additive and dominance effects (ADDetDOM). At this step, PBLUP with additive and dominance effects was not obtained.

Results & Discussion

The results of average predicting ability were summarized in the figure 4.2. This first test showed that BayesCπ with a π value fixed at 1% gave the best results most of the time across traits, sampling strategies, and model complexity (additive versus dominance). The optimization of the population training sampling with the CDmean method seemed

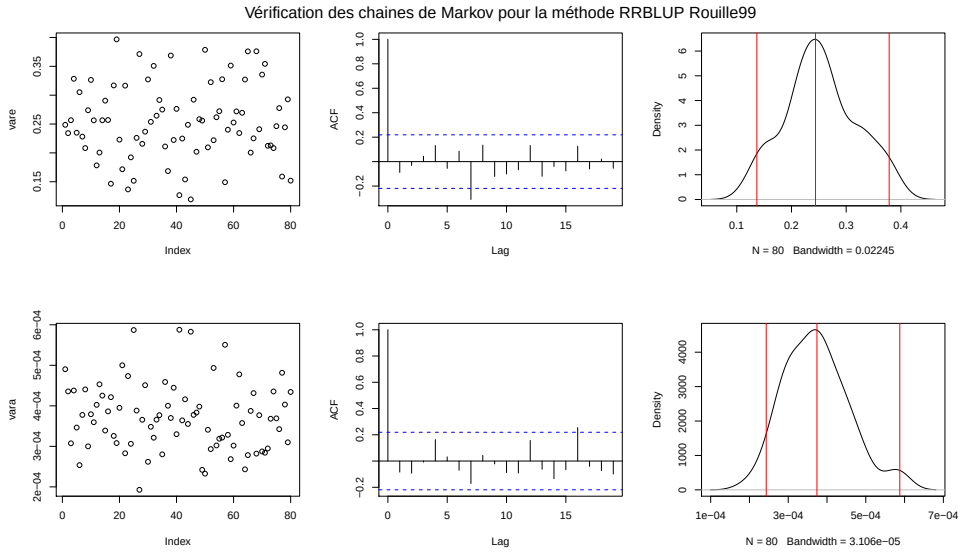


Figure 4.1: Markov chains verification obtained with Bayesian RRBLUP and the 7K SNP set for the trait Rust1. Upper panels represent different posterior parameters for the environmental variance, and bottom panels for the additive variance. Leftmost panels represent estimated variance distribution. Central panels are the autocorrelation analysis between estimated values evaluated with R Package Coda. Rightmost panels are the densities of the posterior distribution of the effect.

to be advantageous compared to the random method for growth and resistance traits with the additive model and for a training composed with 75% of the data. In all other cases, there was not a clear advantage of any of the training strategies.

In most of the cases, predicting unobserved descendants from a single parent in the factorial was less performing than a prediction of a random validation. Models including dominance were less performing and with larger variances among cross validations than their additive counterparts. Predicting abilities for the best of the models overall, BayesCπ1, fitted with dominance were 5% lower than those in the purely additive counterpart, although this arrived with large variations across scenarios, with up to 55% losses in predicting abilities and attaining at the best 11% advantages, always with respect to the additive model.

Results of this first tests are generally encouraging for the genomic selection implementation. They revealed the necessity to compare several evaluation methodologies and sampling strategies. It also showed that there was an interest in the CDmean optimization test. The next step was to do CDmean optimization with a simulating

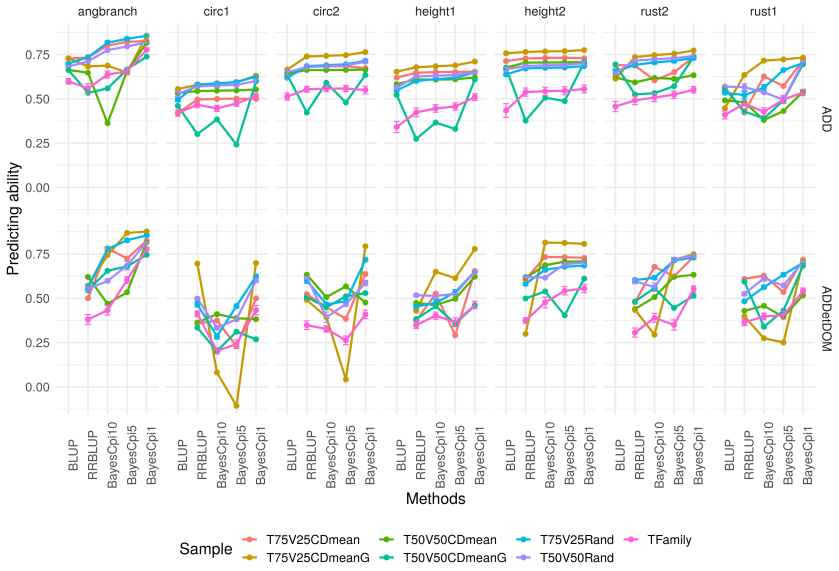


Figure 4.2: Predicting abilities for the factorial mating design with 7K SNP. The different evaluation methodologies used to estimated EBV are in columns, and organized by traits. Upper panels correspond to additive models (ADD), while botton panels are for dominance models (ADDetDOM). The sampling strategies for training are represented by different color trends.

annealing approach instead of a "greedy" simpler approach, and to use different random starts to reach the best combination of individuals between the training and the validation population.

4.3.2 First genomic selection test with sequences

Material and methods

For this test, the complete factorial mating design was used, it was composed of 392 individuals from four mothers and four fathers. Parents were initially assumed to be unrelated but genetic analysis showed that two of the fathers were half-brothers. The father of these two male parents was added to the analysis, and the pedigree was corrected in consequence. The factorial mating design contained fourteen families of 383 offspring.

Evaluated traits were growth (height and circumference), rust resistance at one and two years of age and proleptic branch angle. We tried two phenotype datasets. First, the original phenotype evaluated and adjusted with 6 blocks. Second, phenotypes were somehow "downgraded" by obtaining estimations from only 3 of the blocks. That latter

option allowed to check the impact of less precise phenotyping on both pedigree-based and genomic-based evaluations. We used the information from the 12K newly-developed custom Infinium BeadChip (Illumina, San Diego, CA) (Faivre-Rampant et al., 2016), which after filtering (Chapter 2) produced 8K SNPs. The 350K SNPs resulting from the first imputation test from the sequence (results Chapter 3 part 3.3.2) were used for this genomic selection test.

GS models The family structure was estimated with the A-matrix based on pedigree information, calculated with the R package nadiv (Wolak, 2012). This matrix was compared with a normalized genomic relationship matrix (GN 4.1) (VanRaden, 2007; Habier et al., 2007; Forni et al., 2011) obtained with the SNPchip dataset (GNchip) and with the SNPSeq dataset (GNSeq) following:

$$G_N = \frac{(M-P)(M-P)'}{trace[(M-P)(M-P)']/n} \tag{4.1}$$

where M was the genotype matrix with m markers in columns and n individuals rows, P a matrix $(n \times m)$ containing the frequency of the second allele (p_i) at the marker i. According to Forni et al. (2011) the denominator should assure compatibility with A. The matrix was computed in R3.3.1 platform. To assess the dominance genetic effect, a dominance matrix based on the pedigree information was calculated with the R package nadiv (Wolak, 2012) and the genomic dominance equivalent was calculated according to Su et al. (2012) with:

$$D_N = \frac{(X-W)(X-W)'}{\sum 2p_iq_i(1-2p_iq_i)} \tag{4.2}$$

where X was the genotyping $n \times m$ matrix with code 0 for homozygous and 1 for the heterozygous, and W a $n \times m$ matrix containing the heterozygous frequency ($2p_iq_i$). Three models were used for genomic estimated breeding values (GEBV) for each trait: GBLUP (Whittaker, 2000; Meuwissen et al., 2001), ridge regression best linear unbiased prediction (RR-BLUP) (Meuwissen et al., 2001) and BayesCπ (Kizilkaya et al., 2010; Habier et al., 2011). They were compared with our reference model, the best linear unbiased prediction based on pedigree information (PBLUP) (Henderson, 1975).

The model GBLUP and PBLUP were fitted within the package breedR, and those of RR-BLUP and BayesCπ were implemented in GS3 software. In the case of BayesCπ, we estimated π value jointly with the complete dataset by traits, and the obtained values were the ones used for the cross-validation subsequently. The parameters for the Gibbs sampling chain were: iterations = 5,000,000, burning = 20,000 and thinning = 1/5,000 for the SNP chip panel, and 10,000,000, 20,000 and 1/10,000, respectively for the SNPSeq panel.

Prediction accuracy and Cross-Validation As in the previous tests, five cross-validation scenarios were used for the assessment of prediction abilities: T75V25Rand, T50V50Rand, T75V25CDmeanG, T50V50CDmeanG, and TFamily, similar to those already described for the first test. CDmean optimization was made with an adhoc simulated annealing algorithm instead of the greedy algorithm that came by default. For each scenario, predicting abilities were calculated as in previous tests. To assess the benefits of GS, the same method was used with a pedigree-based BLUP. To compare BLUP and genomic results at a repetition basis, a Student's t-test for paired samples was performed.

Results & Discussion Results presented in Figure 4.3 showed a slight improvement of the predicting ability with GS. The difference was less important than our previous test with the chip overall, across traits, training strategy and model complexity (additive versus additive + dominance). Differences between additive and additive plus dominance models in performance was very small for most of the traits and evaluation methodologies, excepting RRBLUP for Angle01. There was a trend giving some advantage for RRBLUP over other methodologies for growth traits, while for resistance and architectural traits BayesCπ seemed to have a slight advantage. The optimization of CDmean sampling for training with a simulated annealing algorithm did not improve the predicting abilities over the random sampling strategies, and the slight advantage of CDmean training in the reduced dataset of previous test disappeared in this enlarged dataset. Results with the sequences set were not shown here for two reasons.

First, the RRBLUP model with a G matrix build from sequences gives the exact same results as those from a chip dataset. Second, BayesCπ and RRBLUP models did not

converge after several weeks of computing. This results led us to reconsider the use of BayesCπ and Bayesian RR-BLUP for the rest of the study. We decided to use GBLUP instead, extract marker effects from the GEBV (Strandén and Garrick, 2009) and use the weighted GBLUP approach proposed by Legarra et al. (2009) to test a method able to selecting variables. This simpler and faster method showed similar results to those from BayesCπ (Zhang et al., 2016; Teissier et al., 2018).

There was a slight reduction in predicting ability by using only 3 of the 6 available blocks in the obtention of the mean genotype, although differences were negligible compared to the variability across repetitions. This reduction affected similarly all 3 evaluation methodologies, pedigree-based and genomic-based (Figure 4.4). One effect of the downgrading to 3 field replicates was, however, the larger variance in prediction abilities within each scenario.

4.4 Article II: Conditions under which genomic evaluation outperforms classical pedigree evaluation are highlighted by a proof-of-concept study in Poplar

This paper was submitted at frontiers in plant sciences on November, 2018.

1 abstract

Forest trees like Poplar are particular in many ways compared to other domesticated species. They have long juvenile phases, ongoing crop-wild gene flow, extensive outcrossing, and slow growth. All these particularities tend to make the conduction of breeding programs and evaluation stages costly both in time and resources. Perennials like trees are therefore good candidates for the implementation of genomic selection (GS) which is a good way to accelerate the breeding process, by unchaining selection from phenotypic evaluation without affecting precision. In this study, we tried to compare GS to pedigree-based traditional evaluation, and evaluate under which conditions genomic evaluation outperforms classical pedigree. Several conditions were evaluated as the constitution of the training population by cross-validation, the implementation of multi-traits, single traits, additive and non-additive models with differents estimation methods (G-BLUP, weighted G-BLUP or BayesCπ), finally the impact of the marker densification was tested through four marker sensity set (7K, 50K, 100K and 250K). The population under study corresponds to a pedigree of 24 parents and 1011 offspring, structured into 35 full-sib cohorts. Four evaluation batches were planted in the same location and 7 traits were evaluated on one and two years old trees. The quality of prediction was reported by the accuracy, the Spearman rank correlation and prediction bias and tested with a cross-validation and an independent individual test set. Our results show that genomic evaluation performance could be comparable to the already well optimized pedigree-based evaluation under certain standard conditions. Genomic evaluation appeared to be advantageous when using independent test set and a set of less precise phenotypes. Our study also showed that looking at ranking criteria as Spearman rank correlation can reveal benefits to genomic selection hidden by biased predictions.

2 Background

Forest tree species of interest for domestication like Poplar are particular in many ways compared to other domesticated species, notably when it comes to breeding. Among the various particularities, forest trees have long juvenile phases, ongoing crop-wild gene flow, and extensive outcrossing (Miller and Gross, 2011). All of these hamper the process of *controlled* recombination by the breeder. Slow growth and cumbersomeness typical of trees do not facilitate either the conduction of breeding programs, notably with evaluation stages being costly both in time and resources. One of the poplar's particularities is clonality or the possibility of asexual reproduction, which is a powerful tool in evaluation and operational breeding (Bisognin, 2011). However, benefits rarely go hand in hand with facility. Typically for developing a new poplar variety, a first year is used for mating and seedling growth in nurseries. A second year is used to propagate the cuttings and install the designs to do evaluations in different environments, and many subsequent years pass before we can assess genotype-by-environment (G $\times$ E) interactions (Grüneberg et al., 2009), or late maturation traits like wood quality. Selection in poplars proceeds typically via independent level stages, with early stages involving screening for fast-growing, disease-resistant

individuals from large numbers of candidates. Late stages focus on a reduced remainder to select on final growth, architecture, disease resistance, and wood properties. This has been so far operationally efficient considering the constraints imposed by the particularities of trees, but it remains time consuming and lacks precision at the early stages.

For previous and additional reasons, perennials like trees are good candidates for the implementation of genomic selection (GS) (Meuwissen et al., 2001). GS can potentially accelerate the breeding process, by unchaining selection from phenotypic evaluation without affecting precision. When applied early at the seedling stage, GS could potentially save evaluation resources and reduce the time required for evaluation of late maturation traits. GS involves ranking and selecting individuals by using a genome-wide marker set and prediction models calibrated previously in a training set. GS has been made possible thanks to easy access to cheap genotyping data, and to recent developments in evaluation methodology (de los Campos et al., 2009). Recent studies of GS in forest trees were conducted on several species: eucalypts (Resende et al., 2012b; Müller et al., 2017; Tan et al., 2017, 2018), spruce (Gamal El-Dien et al., 2015; Ratcliffe et al., 2015; Gamal El-Dien et al., 2016; Lenz et al., 2017) and pines (Resende et al., 2012a; de Almeida Filho et al., 2016; Ratcliffe et al., 2017). Given the differences among forest species in general, and between their breeding programs in particular, assessments of GS feasibility at a case-by-case basis are often desirable.

According to Hayes et al. (2009), several parameters are involved in genomic evaluation accuracy. First, the extent of linkage disequilibrium in the population. Linkage facilitates the use of markers as *proxies* of unknown QTLs in estimating genetic effects. Two additional parameters affect linkage disequilibrium: effective population size and marker density (Grattapaglia and Resende, 2011; Wientjes et al., 2013). The second parameter of importance for accuracy is the composition of the training set. Such a set must be representative of the candidates for which a prediction is required. Several studies developed methods to optimize the composition of the training set (Rincent et al., 2012; Isidro et al., 2015; Akdemir et al., 2015). The third parameter is trait genetic architecture, usually unknown or poorly understood, but that has an influence on the performances of the different evaluation methods (Wimmer et al., 2013). Some evaluation methods, such as those using some efficient strategy to focus only on relevant variables like the family of bayesian methods, appear to be more efficient with traits with fairly uneven distributions of gene effects. Other methods with less stringent *a priori* on the distribution of gene effects work generally well with highly polygenic traits, like G-BLUP. Other modelling approaches intent to capture the underlying complexity of genetic architectures, by including non-additive effects like dominance and epistatic interactions (Toro and Varona, 2010; Su et al., 2012; Vitezica et al., 2013; Muñoz et al., 2014; Vitezica et al., 2017; Martini et al., 2017), and by considering multiple correlated traits. The latter have not been often used, despite some promising simulation studies (Calus and Veerkamp, 2011; Guo et al., 2014), empirical studies (Jia and Jannink, 2012), and the known fact from classical evaluation that genetic correlations can back accuracies of poorly heritable traits or those harbouring many missing values in the dataset (Gilmour et al., 2008).

In the present study, we intended to benefit from the large corpus of knowledge already established around the concept of GS to carry out a proof-of-concept study on the feasibility of the methodology in the context of the black poplar breeding program in France. Black poplar is the leading European species of riparian forest, with a wide distribution area, and contributing as a parent together with *Populus deltoides* to one of the most widely used hybrid (*Populus* × *canadensis*) tree in the wood industry. This study is the first GS study for a *Populus* species. One of the main objectives of the study was to compare GS to pedigree-based traditional evaluation, by assessing different modelling options including non-additive genetic effects and multiple-trait evaluation. The study also considered the role of marker densification in the performance of GS, by benefiting from a recent imputation study (Pegard et al., 2018). Finally, the design of the calibration and validation sets was taken into account as an additional factor in the comparison. Globally, the study intended to identify the situations in which GS could be a feasible option for poplar, and also the assessments required to reveal any eventual advantage.

3 Material and Methods

3.1 Plant material

The population under study corresponds to a pedigree of 24 parents and 1011 offspring, structured into 35 full-sib cohorts, and involving a 4 by 4 factorial mating design together with a series of multiple pair-mating designs (Pegard et al., 2018). Most of the parents were sampled from natural populations or were high-performance trees already used at nurseries. The population used corresponds therefore to the offsprings of these individuals obtained by controlled crosses at the glass house. We can consider that the population was relatively close to the natural population in diversity terms. The effective size was estimated between 4 and 12 from coancestry matrices computed respectively with a set of independent set of markers and with pedigree. Family size ranged from 10 to 118, with an average of 26 individuals per family. Field evaluations also corresponded to four different campaigns. The first batch (2000 and 2001) involved the factorial mating design with a total of 14 families and 413 offspring phenotyped. In second and third batches, 126 individuals in 6 families and 105 in 5 families were phenotyped (2012/2013 and 2014/2016, respectively). Finally, in order to reinforce the connectivity between the different evaluation batches, 10 additional full-sib families with some parents already in use in previous batches were added in 2015 and phenotyped in 2017/2018. In total, 367 individuals were phenotyped in this last batch.

3.2 Phenotyping

At their respective time-frames, all 1011 offspring and the 24 parents were vegetatively propagated, and field evaluated in separate experiments according to the same six randomized complete block design. All four evaluation batches were planted in the same location (47°37'59" N, 1°49'59" W, Gumn-Penfao, France) with small variations in plot orientation and with common

genotypes as controls across batches.

Phenotyping involved 7 different measurements over different years, and for 5 different traits. Growth was assessed as stem circumference and tree height. Stem circumference was considered at 1m for the second year (circ2). Height was assessed with a graduated rod after one (heigth1) and two years of growth (heigth2). Mean branching angle was scored on proleptic branches at the age of two years with a 1 to 4 scoring scale (angbranch), where score 1 was given to the narrowest angle between the branch and the trunk and score 4 to the widest angle. The scale for angbranch was calibrated in such a way that resulting measures in the same population of reference results in phenotyping distributions being close to normality. Rust resistance was assessed with a 1 (no symptom) to 9 (generalized symptoms) scale (Legionnet et al., 1999) at year one (rust1) and year two (rust2). Budburst phenology of the stem terminal bud was evaluated by measuring its kinetics (every 3 or 5 days from March to April) with a 0 to 5 scale, where stage 0 corresponded to a completely closed bud while stage 5 corresponded to the initiation of stem internode elongation (Castellani et al., 1967). A local polynomial regression model was fitted between stages and dates for each individual and this model was further used to predict the date in Julian days at which the terminal bud was at stage 3 and in order to assess individual susceptibility to late frosts (Howe et al., 2000). As a result of such fitting for Budburst, distributions were continuous and close to normality.

All seven phenotypes were independently adjusted to field micro-environmental heterogeneity with the breedR package (Muñoz and Sanchez (2018), implemented in R3.3.1 platform (R Core Team, 2018)). We used an individual-tree mixed model over all 4 evaluation batches with random effects to fit bi-splines surfaces (Cappa and Cantet, 2007; Cappa et al., 2015), which were nested within each evaluation batch (field experiment). Bi-splines were anchored at a given number of knots for rows and columns, which were optimised by an automated grid search based on the Akaike information criterion (Akaike 1974) provided by breedR. Data from all blocks were used as input to the adjustment model. The same adjustment was also performed with data from three of the blocks (blocks 1, 3 and 5), to assess the prediction models behaviour with a less precise phenotype. The model comprised genotyped and non-genotyped individuals according to a single-step formulation (Legarra et al., 2009), and in order to use all available information in field trials with minimum gaps to predict the micro-environmental individual effect. The micro-environmental individual effect was subtracted from the observed phenotype to obtain a spatially adjusted phenotype. A clonal mean of spatially adjusted phenotypes was estimated and used as raw phenotype for the rest of the study. All measurements were tested for deviations from normality by a randomised Q-Q plot.

3.3 Genotyping

All 1,033 individuals in this population were genotyped using the Populus nigra 12K custom Infinium Bead-Chip (Illumina, San Diego, CA) (Faivre-Rampant et al., 2016), and the genome of 43 individuals was also sequenced. The individuals selected for genome sequencing comprised: one identified grandparent, 22 parents, 14 progenies and 6 unrelated individuals from natural

populations. Progenies were chosen in such a way that all parents had at least one offspring with its genome sequenced. The set of unrelated individuals were used to assess the imputation ability under challenging conditions. In a previous study (Pegard et al., 2018) genotype imputation from 7K (effective SNPs out of 12K in array) to 1,466,586 SNPs was performed attaining imputation qualities higher than 0.84 per individual, and evaluated by a leave-one-out cross-validation scheme (CV). Resulting imputation was used in the present study to constitute alternative sets of selected markers for genotyping. For quality assessment and selection of the marker sets, we used the proportion of alleles correctly imputed by genomic position across individuals (Props), and Props corrected by the probability of correct imputation by chance (Badke et al. (2013); cProps). Among the imputed SNPs, we selected those with Props higher than 0.90, with cProps higher than 0.60 and a minor allele frequency (MAF) higher than 0.05, to obtain a set of 249,805 SNPs (250K). That latter set comprised the totality of the 7K from the chip. We selected five alternative downgraded marker sets: 100K (with 118,747 SNPs), 50K (with 50,565 SNPs), 25K (with 26,183 SNPs), 12K (with 12,906 SNPs) and 7K (with 7,048 SNPs) where coverage and homogeneity of density was optimised over the original 7K array. These five sets were composed by selecting respectively 1 SNPs every 500 bp, 1000 bp, 10,000 bp, 25,000 bp, 50,000 bp out of the 250K set. Whenever more than one candidate SNPs were available for the same window, we selected the one that had highest Props and cProps. We used as a medium-density panel the genotypes from the chip.

3.4 Models

We estimated variance components and heritabilities with the complete data set and single trait models, and genetic correlations with a genomic multiple-trait model (GBLUP). The Akaike Information Criterion (AIC) was used to assess for each given trait the quality of each model. Three alternative methods were used to calculate genomic estimated breeding values for each trait: the best linear unbiased prediction based on genomic information (GBLUP) (Whittaker et al., 2000; Meuwissen et al., 2001), the weighted GBLUP (wGBLUP; Legarra et al. (2009); Zhang et al. (2016)) and BayesCπ (Kizilkaya et al., 2010; Habier et al., 2011). They were all compared to the best linear unbiased prediction based on pedigree information (PBLUP) (Henderson, 1975). The models for GBLUP (and PBLUP) using matrix notation for additive and non-additive effects was given by:

$$y = X\beta + Zu + \varepsilon \tag{1}$$

$$y = X\beta + Zu + Wd + \varepsilon \tag{2}$$

where y was the clonal mean, β is a vector of fixed effects, u the vector of random additive effects following N(0,$G\sigma_a^2$) with σ_a^2 the additive variance and G (or A in PBLUP) the relationship matrix, d was the vector of random dominance effects following N(0,$D\sigma_d^2$) with σ_d^2 the dominance variance and D the dominance relationship matrix, ε the vector of residual effects following

N(0,$I\sigma_e^2$) with σ_e^2 the residual variance, and X, Z, W, and I are identity matrix relating the clonal mean to the fixed effects and random effects. The model for BayesCπ was given as follows:

$$y_{ij} = \mu + \sum_{j=1}^{k} x_{ij} b_j \delta_j + \varepsilon \tag{3}$$

where μ was the overall mean, x_{ij} an indicator covariate linking i-th individual to j-th locus allelic content, b_j the random additive effect, and δ_j an indicator variable taking value of 1 if marker has non-zero effect or 0 otherwise, following a binomial distribution with probability π (proportion of markers with non-zero effect). In our case, π was estimated by the model. The shortest Gibbs sampling chain, giving the same results as the longer chains, was used to obtain a sequence of observations to approximate the joint distribution with the following parameters: iterations = 20,000, burning = 5,000 and thinning = 1/10. The PBLUP and GBLUP single-trait analyses were performed with the R packages breedR (Muñoz and Sanchez, 2018) The BayesCπ single-trait model was performed with the package BGLR (de los Campos and Perez Rodriguez, 2018) while breedR was used for the multiple-trait analysis. We performed Multi-trait analysis with the pedigree information and the chip SNP panel for genomic evaluation with an additive model. Whereas for BayesCπ we performed single trait analysis with an additive model with SNP sets of 7k, 50K and 100K as shown in table 1.

3.5 Relationship matrix estimation

The ARM (additive relationship matrix) was built from the known pedigree at the moment of the controlled crossings, and denoted hereafter as A. However, a preliminary marker assessment in this study showed that there were errors in the pedigree. Pedigree was corrected based on these results and a new reconstructed ARM was obtained, denoted hereafter as A_{cor}. Pedigree errors involved in most cases a wrong paternity attribution, and less frequently individuals supposed to be different genetically. The total number of parent did not change because a father was added and a mother was removed. The main change concerned the number of families that went from 39 to 35. Both A and A_{cor} were calculated with the R package nadiv (Wolak, 2012). Both matrices were kept for the comparison in order to show the potential loss due to pedigree errors and the maximum performance attainable by pedigree. We used normalized genomic relationship matrix (G equ.4) calculated following VanRadens formulation (VanRaden, 2007; Habier et al., 2007) and the scaling proposed by Forni et al. (2011) to assure compatibility with A, for each genotyping set (7K, 50K, 100K and 250K) :

$$G = \frac{(M - P_1) W_a (M - P_1)'}{trace[(M - P_1) W_a (M - P_1)']/n} \tag{4}$$

where M was a genotyping matrix with m markers in columns and n individuals rows, P_1 was a matrix $(n \times p)$ containing the frequency of the second allele (p_i), at the marker i, and W_a was a matrix of weights described below. *Ad hoc* scripts in R were used to make the computations for G (R3.3.1 platform). To assess dominance effects, a dominance matrix based on the pedigree

information was calculated with the R package nadiv (Wolak, 2012) with expected and observed pedigree information (D and D_{cor}). The genomic dominance matrix was calculated as:

$$D = \frac{(X - P_2)W_d(X - P_2)'}{trace[(X - P_2)W_d(X - P_2)']/n} \tag{5}$$

where X was the genotyping $(n \times p)$ matrix code "0" for the homozygous and "1" for the heterozygous, P_2 was an $(n \times p)$ matrix containing the heterozygous frequency $(2p_iq_i)$ according to Vitezica et al. (2013) and normalized in the same way than G in equation 4, and W_d was a matrix of weights described below.. We used one of the procedures of Wang et al. (2012) for calculating weights in wGBLUP. Unlike GBLUP, where all markers have the same variance and therefore the same weight, the derivative wGBLUP uses transformed G according to marker weights to select markers. The weights were calculated as $w_j = \hat{u}_j^2$ where w_j was the weight for the SNP j and $\hat{u}_j$ was the estimated marker effect was obtained as

$$\hat{u}_a = W_a X G^{-1} \hat{g} \tag{6}$$

$$\hat{u}_d = W_d X D^{-1} \hat{d}, \tag{7}$$

where $W_{a,d}$ was a diagonal of weights, either a identity matrix (GBLUP) or a diagonal of w weights (wGBLUP) for additive (W_a) or dominance (W_d) relationship matrix, $\hat{g}$ the genomic estimated breeding values (GEBV) and $\hat{d}$ the estimated dominance effects. Several iterations of recomputed $\hat{u}_a$, $\hat{u}_d$, $\hat{g}$, and $\hat{d}$ were performed to update G, following recommendation by Wang et al. (2012), and according to the following steps:

1. Define $i = 1$, $WW_{(a,d)i} = I$ and G_i as eq 3
2. Compute $\hat{g}_i$ using GBLUP approach
3. Compute additive SNP effects as equation 6 and dominance SNP effects as equation 7
4. Calculate SNP weights as $w_{aj+1} = \hat{u}_{ai}^2$ and $w_{dj+1} = \hat{u}_{di}^2$
5. Scale w_{aj+1} and w_{dj+1}
6. Calculate G_{i+1} as equation 4
7. Calculate D_{i+1} as equation 5
8. $i = i + 1$
9. Iterate from 2 until $i = 3$

3.6 Prediction accuracy and Cross-Validation

We assessed the impact of the composition of the training (TS) and validation sets (VS) on the performance of the genomic evaluation by trying two TS/VS sizes and two different TS/VS compositions in a 4-fold cross-validation scheme. Two sizes for the samples of training set were

tested, with 50% (T50) and 25% (T25) of the individuals evaluated in the 2000/2001, 2012/2013, 2014/2016 field batches. The last field evaluation batch of 2017/2018 was used as an extra independent validation set for each of the four TS. The two composition scenarios for TS and VS involved: a sampling of individuals independently of their family membership, and a sampling of different family sets to be part of TS and VS. These sets were randomly composed by sampling individuals or families in such a way that the desired percentage (50 or 25%) was fulfilled. The performance of the models was evaluated following different criteria. Firstly, predictive ability (Predab), which was defined as the Pearson correlation coefficient between the phenotypes and the GEBVs of the samples in the VS, or in the test set (Predabtest). The accuracy (Accuracy and Accuracytest) of the models were estimated by dividing each predictive ability by the square root of the narrow sense heritability of the corresponding A model for the given trait. Additionally, the Spearman rank correlation between the phenotypes and the GEBVs of the individuals in the VS was calculated (Spearman). We estimated the Spearman and Pearson correlation of the top 5% of the trait, for the section between 5 and 10%, and between 10 and 50% within the VS. Finally, we assessed for potential bias in the genomic prediction by estimating the intercept and the slope of the linear regression between the phenotypes and the GEBV of each model in the VS and the test set.

3.7 Testing factor importance

In order to assess the main factors accounting for genomic evaluation performance, we applied the Random Forest algorithm (Liaw and Wiener, 2002) that is included in the Boruta R package (Kursa and Rudnicki, 2010). These main factors (or features) were: Trait, Matrix (A, Acor, G, Gw1, Gw2, Gw3, D, Dcor, Dw1, Dw2, Dw3), GeneticEffect (Additive, Additive and Dominance), ST_MT (Single-Trait, Multiple-Trait), GenoSet (none, 7K, 50K, 100K, 250K), Type (Individual, Family) and Perc (T50, T25). Classification of features was done for each of the performance variables available: predicting ability, Accuracy, Spearman correlation, Intercept and slope. The number of standard deviations from a reference point derived by the algorithm of each feature was used to determine the importance of each factor with respect to the performance variable.

4 Results

4.1 Estimated variances and heritabilities

All heritabilities with their corresponding variance components are shown for all models and traits in figure 1, while Akaike Information Criterion (AIC) are presented in the supplemental table S1 . In general, most traits showed intermediate to high heritabilities (average of 0.73), with height2 and rust2 showing the highest average values, and budburst correspondingly the lowest. In terms of models and under additive action, G-based analyses produced generally the highest heritabilities, followed by the first step weighted G and BayesCπ. Comparatively, pedigree-based analyses produced lower heritability estimates. However, under models involving

also dominance, G-based analyses resulted in a somehow reduced level of heritability compared to the additive counterpart, while the rest of the approaches (pedigree and weighted G) was comparatively unaffected. In general, dominance variation was detected for all traits under genomic based evaluations, with amounts of variation being less important than those for the additive counterpart or the residuals. On pedigree-based evaluations, however, the patterns for dominance variation across traits was more heterogeneous, with traits for which it represented most of the variance while for others no dominance was detected. The increase in the density of markers used in the genomic evaluation was accompanied almost monotonically by an increase in heritability, with the highest averages in heritability being for the 250K sets. All these genomic scenarios resulted in general with higher heritabilities than their pedigree-based counterparts.

4.2 Accuracies estimated by cross-validation with different models

Cross-validation accuracies are shown in Figure 2 for five traits (the remaining traits are shown in figure S1) and type of relationship matrix, and considering four different training scenarios (size and composition). Results correspond to single-trait additive models with ARM based on the 7K SNP panel. Accuracies varied between -0.3 and 1.3 across all scenarios and traits. It is important to note that, because of the choice of a particular model of reference to provide a basis heritability (pedigree-based model with the A matrix), accuracies larger than one could be obtained.

Accuracies responded greatly to changes in the way the training set was constituted, by the percentage and the composition. The fact of using different families for training than for validation had a large impact on the accuracy when compared to the alternative scenario where the splitting between training and validation occurred mostly within families. Basically, as expected, predicting different families was less accurate than predicting different individuals within the same cohort, with losses in accuracy averaging 25%. This pattern was found for all traits, except for one training scenario in angbranch, where differences between the two compositions were also the weakest. Concerning the percentage, the effect of reducing the training set from T50 to T25 had also an impact on accuracy, although mostly when training and validation involved different families. On average, reduction in accuracy with decreasing training set size was around 2.6% for the training composition based on individuals, and around 26% for that based on families.

To a lesser extent accuracies responded also to the differences in modelling the additive relationship (pedigree versus marker-based models). The behavior of the different models in terms of accuracies depended greatly on traits and on the training scenario. With T50, more often than not G-based models had advantages in accuracy over the pedigree-based counterparts, both for the individual and the family training compositions. Although these differences were only really conspicuous for rust1 and budburst. With T25, however, no global advantage of a single model was evident, with pedigree-based models performing similarly to their G-based equivalents. With the most challenging training scenario, however, the model based on uncorrected A had generally poorer accuracies than those shown by the corrected A.

Also, G-models based on weighting matrices had generally lower accuracies than the model with unweighted G, with each step of weighting reducing further the accuracy. Overall, the models based on Acor and plain G had the best accuracies.

Although not shown in detail in Figure 2, the four repetitions of the cross-validation had a ranking of accuracies that was generally well preserved for any given trait across the different models, and whenever training was based on individual sampling. With family sampling, however, such ranking was no longer kept across models, and differences amongst repetitions were larger at any given scenario than those for the individual sampling.

We compared the results obtained from weighted GBLUP models with those of BayesCπ models for two traits: angbranch and rust2, as these were traits for which the weighted alternative worked better (Figure 2 and Figure S1). The Bayesian model gave no real advantage over weighted methods in terms of accuracy (Figure S2). By adding a dominance effect to the single trait model for each trait with the 7K SNP panel, we did not observe a substantial change in accuracy with respect to the purely additive model (figure S3). Overall, dominance did not led to losses in accuracy, with similar to slight increases in performance across traits, except budburst for which accuracies were lower than those under the additive model.

We evaluated a multi-trait additive model in terms of accuracies (figure S4). The advantages of a multiple-trait approach over the single-trait counterpart were clearly dependent on the trait and the training scenario, although more often than not the single-trait approach had a superior performance. For instance, rust1 and circ2 showed clearly no benefit in using a multiple-trait prediction, while for height1 the multiple-trait prediction had an advantage when training over families. For the other two traits, budburst and angbranch, the multiple-trait approach brought a benefit in the most challenging training scenario, that for families and T25. Concerning the comparison over models, the multiple-trait approach did not seem to benefit one matrice modelling over the others (pedigree-based versus G-based). Therefore, the multiple-trait prediction did not bring a clear-cut advantage across traits and training scenarios. Genetic correlations between the traits involved in the multiple-trait analysis are shown in (figure S5 in supplementary data). In summary, the accuracy of unweighted G-based models appeared to be slightly better than with pedigree-based models, although in most cases Acor model obtained comparable levels of performance to the best G-based method. The cross-validation sampling strategy (individual/family) impacted the accuracy in all cases and for all traits, with individual scenarios being less challenging than family scenarios. The percentage of individuals in the training population (T50/T25) showed a less important impact on accuracy than that of composition. More advanced models involving dominance effects and multiple-traits did not improve the performance of genomic predictions.

4.3 Effect of of marker density on accuracies

Three out of seven traits (budburst, height1, and rust1) were selected to show the effect of an increase in marker density on prediction accuracy over different modelling approaches in figure 3 (the remaining traits are in figure S6 to S12). Selected traits were representative of

the patterns found across the seven traits for the effect of marker density. We compared the accuracies obtained with four markers sets of increasing density with a single-trait additive model, and T50_individual sampling scheme for cross-validation.

For some traits like budburst (add others in S6, like angbranch) densification resulted in decreased accuracies, and this happened no matter the modelling approach used to produce the G matrix. For this trait, Gw1 showed a less sensible reduction in accuracy across densities than that observed for G-BLUP. For some other traits like height1 (and also height2 in S6), however, densification brought an increase in accuracy, notably from 7K to 50K. For height1, weighted G benefited the most from densified marker sets compared to plain 7K sets. Finally, for traits like rust1 (or circ2 and rust2 in S5), the benefits of densification were dependent on the kind of modelling for the matrice G that was in use. Under unweighted G, density was not beneficial for accuracy to the point of making highest densities non competitive compared to pedigree-based methods. Contrarily, densification with weighted G modelling, notably first and second steps, had little or no detrimental effect on accuracies. Therefore, the benefit of densification for accuracy was trait-dependent and often results were more evident under models imposing some variable selection, resulting in a competitive scenario against pedigree-based methods.

A global analysis of accuracies pooling results from all traits, with a single-trait additive model and comparing marker densities across the four cross-validation strategies is shown in figure S13. No differences in accuracy are to be found between the different modelling strategies for relatedness (pedigree-based versus G-based), and this pattern is repeated across the different densities.

4.4 Challenging prediction models with new individuals and degraded phenotypes

We used a completely independent set of individuals representing the next generation of selection candidates to evaluate the different prediction models with 7K SNP and across two different training scenarios (T25 and T50). Results of accuracies from this independent set are presented for three traits in Figure 4.

Accuracies were substantially lower under the new more challenging testing scenario than those already shown for the cross-validation scheme (see Figure 2). In general, marker-based models resulted in a less affected level of accuracy compared to the pedigree-based counterparts, notably in the most challenging training scenarios involving different families and T25. Some exceptions are to be noted for height1, however, where Acor model obtained the best performances in the less challenging training schemes. For the rest of the cases, G-based model and Gw1 were overall the best performers with an independent validation set.

Adjusted phenotypes for all traits resulted from averaging 6 replicates. To test whether number of replications had an effect on the differences in performance between pedigree and genomic-based evaluations, new evaluations were produced based only on 3 out of 6 replicates. Resulting accuracies under this new evaluation scheme are presented in Figure 5, comprising

different cross-validation schemes (T50, T25, families and individuals), and two marker density sets (7K and 50K). The prediction accuracy was particularly affected by the reduction in repetitions with the 7K panel and across all models and training scenarios, with an average drop in accuracy from 0.75 to 0.25. However, with a denser panel of 50K, accuracies were greatly recovered for the marker-based models, attaining levels that were close to those with the full set of 6 repetitions (0.75). Therefore, downgrading the phenotype with less repetitions affected greatly pedigree-based predictions, while marker-based models remained almost unaffected whenever the SNP panel had a high enough marker density. This latter interaction of downgrading repetitions with marker density showed the limits of the basic 7K panel compared to the 50K in fully recovering the genetic signal.

4.5 Evaluation of prediction models with complementary criteria

Figure 6 shows the slope of the regression between phenotypes and GEBV in the validation population for the different models, cross-validation scenarios and marker densities. Trends for slope across models showed that the pedigree-based approaches had the most robust behavior with values always around 1. Contrarily, G-based approaches showed often higher slopes denoting biased predictions. This deviation was always more pronounced for G-BLUP than for weighted G-BLUP, with a decreasing trend in slope with increasing steps of weighting. Marker densities had the effect of increasing slopes, notably for G-BLUP and G-BLUP schemes with fewer steps of weighting. With a less pronounced effect, the change in training scenarios from individuals to families and from T50 to T25 increased slopes. In general, G-BLUP schemes showed the largest deviation in slopes due to changes in training scenarios. The intercept revealed a systematic bias for genomic evaluation models, which was not detected with the pedigree based model.

Contrary to the accuracy, when evaluating prediction quality by the Spearman correlation some advantages were observed with dominance and multiple-trait modeling over their simpler additive, single-trait counterparts (figure 7, figS14, and figS15). Some models involving dominance showed improvements in Spearman correlation with respect to their additive equivalents (0.8 and 0.7, respectively, for circ2), although this was not a general trend over traits, with examples like rust1 with no advantage. Regarding multiple-trait models, Spearman correlation offered a different picture than that of Pearson, with some favorable advantages shown for some traits like height1 with respect to the single-trait equivalent (0.7 and 0.5, respectively). Again, this advantage was not general across traits, and some like anglebranch did not show significant differences.

Figure 8 represents the prediction quality in terms of Spearman correlations between phenotypes and GEBV for a choice of three representative traits (budburst, height1, rust1), and across the four marker densities and four G-based models. Most traits (figure S16 to S22), including the 3 in the figure 8, showed a large advantage in Spearman of some G-based models over pedigree-based equivalents, notably for rust and height1. These advantages were maximized in the case of the weighted G models, notably Gw1. Although weighted G models with extra weighting cycles holded well against pedigree-based counterparts, they obtained somehow lower

performances than Gw1 and comparable to G-BLUP. The use of high density panels generally favoured the advantage of G-based models over their pedigree equivalents, notably between 7K and 50K panels as shown for height1. For the other traits, the contribution of increasing densities was especially noticeable under G-BLUP and, to a less extent, for Gw1. Therefore, Spearman correlation tended to show a clearer advantage of the G-BLUP over pedigree based models in terms of prediction quality based on ranking than that shown by the Pearson criterion. Moreover, Spearman criterion also revealed more clearly the benefits of densifying beyond the 7K panel.

In order to check the behavior of the two correlations, Pearson and Spearman, over different sections of the distribution of predicted individuals, we compared them for the top 5%, for the section between 5 and 10%, between 10 and 50%, and for the whole distribution. Results are shown in figure 9 (for other traits figures S23 to S28). With the 7K panel, Pearson and Spearman correlation were similar within each of the 3 distribution sections. With 50K, 100K, and 250K SNP and for the top 5%, Pearson correlation attained higher values than Spearman . The difference between the two correlations tended to disappear from 5% and onwards the middle of the distribution. Finally, there was a complete inversion between the two correlations, with higher values for Spearman than for Pearson, when considering the total validation population. These observations were consistent for all traits. In other words, Pearson was relevant when discerning prediction qualities for the top individuals, while Spearman showed higher values overall, with a trend that was accentuated by increasing marker densities. We tested intermediate genotyping densities for a trait (height at 1 year) in the hope of observing an inflection point. The results show that a set of 7K SNPs homogeneously distributed may be sufficient to improve prediction accuracy. More denser set of markers allows to increase slowly the quality of prediction before to reach a plateau.

4.6 Ranking of factors impacting prediction accuracies

The Boruta algorithm evaluated the different features explaining variability for accuracy, Spearman correlation and the slope. Results in terms of Z-score for all features are shown in Figure 10, with different results depending on the dependent variable. Although not with the same ranking, trait, type of sampling and percentage of individuals in the cross-validation strategy were the most explicative features (Z-score > 50) for the Accuracy and the Spearman correlation. Regarding the slope, the most important features were the matrix used for the model and the marker density (GenoSet).

5 Discussion

5.1 Genomics does not improve substantially prediction accuracy over pedigree in standard conditions

This study was conceived as a proof-of-concept of the genomic evaluation in the black poplar breeding program in order to evaluate feasibility and performance in a situation close to operational conditions. Several main messages could be drawn from this study. Firstly, genome-based models captured higher heritabilities and higher additive variances than their pedigree equivalents, although this did not led to a systematic advantage in terms of prediction accuracy for the former over the latter. G-BLUP obtained in general the best prediction accuracies, but it was very closely followed by the evaluation based in a genomically corrected pedigree. Secondly, the benefit of densification of the marker panel for the prediction quality was mostly trait-dependent, and its advantages were more easily revealed by models operating variable selection. Finally, the most clear advantages of genome-based methods and of marker densification were particularly found in challenging situations: when using an independent validation and when degrading the phenotype. The use of alternative criteria to assess prediction quality also pinpointed the combination of genome-based methods and denser marker sets as being generally better than the pedigree-based counterparts.

The genomic evaluation captured generally more genetic variance than pedigree evaluation, regardless of the trait. The number of markers fitted in the model generally increased the proportion of genetic variance explained by the model, but this occurred mostly under G-BLUP. When using a weighted GBLUP, the proportion of genetic variance explained by the model decreased with the cycles of weighting. Without variable selection, increasing the number of markers favoured a better coverage of all genomic regions, including those close or inside relevant QTLs. Variable selection in weighted GBLUP supposedly eroded relevant variation, affecting the proportion of captured variation. This type of behavior could reflect an infinitesimal-like trait architecture rather than a few underlying QTLs with a substantial effect (Zhang et al., 2016). Among all the traits studied, resulting heritabilities and genetic variances for budburst were the ones with the least equivalences with the literature. Often this is a trait found to have very high heritabilities (0.61 - 0.70: Teissier du Cros (1977); Pichot and Tessier Du Cros (1988), contrarily to our findings under pedigree and genomic evaluations. Apart from a fairly small genetic variance in our training set for that trait, no other likely explanation could be found for this discrepancy.

The fact of capturing more genetic variance with marker-based models did not result automatically into a better prediction of the phenotype than using plain A models. Our prediction accuracy was already relatively high under pedigree evaluation, probably due to the fact of using a good evaluation design with enough repetitions and spatial adjustments at individual level. Markers did not help to improve this scenario or very little. Globally, if there was a difference between pedigree-based and genomic predictions, the genomic prediction was better with G or Gw1 matrices. Using several weighting cycles (Gw2 and Gw3) did not show

in any case better results. Comparable results with decreasing efficiency of several cycles of weighting were found in other recent studies (Teissier et al., 2018).

Together with the fact that pedigree evaluations obtained already high levels of prediction accuracy, there is also the point that correcting pedigrees had generally a large beneficial effect, making resulting model truly competitive in some situations and with some traits compared to genome-based models. This is not new in forest assessments, given the fact that controlled crosses are cumbersome and prone to errors. In loblolly pine (Munoz et al., 2014) and in maritime pine (Bartholomé et al., 2016), pedigree errors led to decreases in predicting ability, and by completing or correcting the pedigree the predicting ability could be increased. In the maritime pine study (Bartholomé et al., 2016), the predicting ability was improved by the completion of the pedigree information in such a way that the genomic evaluation had little extra room for improvement in predicting ability. In our pedigree, the error rate was of 15%, involving in most cases wrong paternity attribution of complete or partial families, or individuals supposed to be different genetically. In the same way, more complex prediction models, as multi-trait evaluation or adding non-additive genetic effects in the prediction model, did not further improve the prediction accuracy. Our results were in line with other studies, where the gain in accuracy or predicting ability by adding non-additive genetic effects was negligible (Gamal El-Dien et al., 2016; Tan et al., 2017).

Apart from the general trends between pedigree *versus* genomic models, results of prediction accuracy were fundamentally trait-dependent and mostly driven by the kind of training scenario being applied. This is clearly shown by the results of the Boruta algorithm, which found trait and training scenarios to be key features in explaining predicting accuracies. Similarly to other authors (Norman et al., 2018),we observed that prediction accuracy performed better when the training and validation populations were closely related, as when the split between the two occured at within family levels. On the contrary, prediction accuracy could be greatly affected when resulting from distant, independent validation sets. In our study, the cross-validation with individual sampling performed better than with family sampling, and this somehow limited the use of genomic evaluations to predict unobserved crosses in our population with current approaches. The size of the training set used to develop prediction calibration is often cited as an important factor (Nakaya and Isobe, 2012). Curiously, the differences between our T50 and T25 schemes (50% and 25% of individuals to construct the calibration model, respectively) was not as large as one could expect. This is presumably very dependent on the properties of the populations being used for training.

5.2 Genomic prediction advantages are mostly observed in challenging conditions

The choice of the training and validation sets is known to have a non negligible impact on the prediction accuracy (Rincent et al., 2012). In that sense, our results showed that there was a substantial variation around each cross-validation realization, although often the ranking in performance between realizations was preserved across scenarios, notably for the individual

sampling. In general, these cross-validation cases corresponded to operational situations where validation contributes with extra selection intensities, for instance, with new crosses from known parents or additional sibs across families to select from. One additional scenario of training that could be considered as specially challenging, corresponded to the validation set of newly obtained crosses from parents that were mostly underrepresented in the cross-validation sets. This could be seen as an operational demand to incorporate comparatively new material for selection. Our results showed that such challenge affected substantially the prediction accuracy across models, although G-BLUP and Gw1 were generally the most robust performers and pedigree-based evaluations the ones with the greatest loss overall. In the cross-validation scheme, the factorial design had a relatively large influence in demographic terms in the training set. Being a system that creates a well interconnected network of families (Sørensen et al., 2005), the factorial design seemingly favoured pedigree predictions to a level that made it competitive compared to genomic predictions in the cross-validation. However, the new testing set possed a challenging prediction problem to pedigree-based models, as its relatedness was less populated as to support quality predictions. Despite of that, the situation was not always a clear-cut difference between pedigree and genome-based evaluations, as shown by traits like height1.

If the extent of relatedness thanks partly to the factorial design could have facilitated the competitiveness of pedigree-based predictions, the fact of using a high quality adjusted phenotype involving 6 repetitions was another element that could have a role in diminishing the differences between pedigree and genome-based performances in prediction terms. Indeed, our results showed that downgrading the quality of clonal means used as phenotypes clearly had a differential effect between pedigree and genome-based predictions, with the latter retaining prediction quality at a level without replicate reduction. It is important to note that this challenge revealed also the limitations of the 7K genotyping set, which did not allow a full recovery of prediction quality, as the one shown by the 50K set. This evaluation simplification has also important operational implications for field evaluation, which need to be balanced with the genomic investments.

5.3 Genomic prediction enables the ranking of candidates to selection

One of the main objectives of genetic evaluation is to rank individuals according to their breeding values, in order to use subsequently final selections as reproductors for the next generation. In that sense, pinpointing breeding values with accurate predictions is therefore a key element in genetic progress, and the use of predicting abilities based on a parametric correlation between predictions and true breeding values one of the most common means of quality assessment (Daetwyler et al., 2013). This latter correlation shows a linear relationship with the genetic response (Falconer, 1981). For the poplar breeding program, however, the stress is given to the selection of genotypes for clonal dissemination at the production stage directly, rather than for gametic dispersion in seed orchards. This essential difference leads to the importance of ranking in selection decisions for poplars, as for any other domesticated species with clonal

selection. When assessing the potential of genomic evaluations, it is essential to take into account the way predictions will be used for. In that sense, we used alternative measures of prediction quality, like the slope of the regression of true breeding values on estimated breeding values. This slope represents a way to assess departures due to bias in predictions, generally caused by unequal representations of lineages in the training (Patry and Ducrocq, 2011). Our results suggest that G-BLUP was particularly affected by biases problems, with large departures towards greater slopes, i.e. best phenotypes gave proportionally higher predictions than worst phenotypes. To a lesser extent, the best weighted G-BLUP (Gw1) also presented departures in slope. Comparatively, pedigree-based predictions were perfectly unbiased with slopes of one.

This result casted some doubts on the relevance of rankings derived from G-BLUP genomic predictions. We added an alternative measure of prediction quality, the Spearman correlation between predictions and true breeding values, which is a non-parametric estimate measuring the variation of the ranking. Moreover, this focus on ranking appeared as an appealing feature in the context of poplar breeding. Although less frequent in the literature than Pearson-based predicting abilities, a few authors used Spearman correlation to evaluate the prediction quality and to serve as criterion to select evaluation approaches (González-Recio et al., 2009; Mota et al., 2018). Some other authors suggest that individual ranking strategies could be more efficient (Blondel et al., 2015).

Our results using Spearman showed a quite different picture of the comparison between pedigree and genomic-based predictions to that obtained from the Pearson counterpart. Firstly, the comparison of Spearman correlations revealed some benefits of non-additive and multiple-trait G-BLUP models that were otherwise invisible under Pearson. The most clear picture brought by the Spearman correlation, however, came when comparing the benefits of densification of marker panels. The switch from 7K to 50K panels boosted substantially the Spearman correlation of G-BLUP to place it well above the levels of pedigree-based evaluations. Nevertheless, a plateau in Spearman coefficients was reached from 50K onwards in our study. These plateaus over densities were already reported in other species, although not for Spearman: in cocoa (Romero Navarro et al., 2017), wheat (Norman et al., 2018) and eucalyptus (Kainer et al., 2018). For eucalyptus, the plateau in correlation was still not reached at 500K, while for cocoa and wheat it was reached after thousands or tens of thousands of markers. In our case, it remains to be explored whether there is an inflexion point in correlation between 7K and 50K. Similarly, it seems from our homogenous 7K that the limits of the 7K chip (Faivre-Rampant et al., 2016) came from an irregular coverage of the genome. Our results suggested that with an homogeneous set of marker, the number of markers could be reduced and gives good results. These may be due the fact our effective size (between 4 and 12), is smaller than we expected from individuals of few selection generation.

Another factor of importance in the comparison of performances via Spearman versus Pearson is the fact that their values differed across the distribution of predictions in the validation set. This implies different outcomes for the two criteria according to the levels of selection intensity, or weights given to each trait in an index. Usually, the interesting part of the distribution is the

top percentiles. However, in some cases the interest lays in intermediate values, like for budburst. The goal here is to have trees that do not budburst too early to avoid late frosts, nor too late to avoid shortening the growing season. Basically, Pearson showed the highest correlations for the top percentile selections, while Spearman was generally better for the middle and whole distribution. In conclusion, a criterion based on ranking might be a reasonable option, notably when ranking is more important than prediction itself, or when multivariate compromises force selection differentials towards the middle of the distribution, there where Pearson appeared to be less advantageous.

6 Conclusions and perspectives

Our proof-of-concept study shows that genomic evaluation advantages are context-dependent. Its performance could be comparable to the already well optimized pedigree-based evaluation under certain standard conditions and with access to low to medium SNP density panels. Genomic evaluation appeared to be truly advantageous under less standard scenarios with a certain degree of challenge which have been clearly been pinpointed in present work. Our study focused on a fairly advanced stage of the evaluation in the breeding program, where a substantial part of the variation has already been let aside by using less efficient early selections at the nursery. We believe that genomic selection could be an interesting option at that early stage, where selection precision is typically poor and genetic variability abundant. Our study also showed that it is important to assess performances by looking at other alternative criteria, like those related to ranking, notably when these criteria respond to the operational context of the breeding program under scrutiny.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

MP performed the analyses and drafted the manuscript. FM developed the scripts for spatial adjustment of phenotypes. VS contributed to the discussion on analytical models and data preparation, providing as well valuable scripts. CB provided access to plant material and contributed to the view of the breeding program and ways of optimization as the scientist responsible for the *Populus nigra* breeding program. VJ and LS designed the study, discussed the analyses, assisted in drafting the manuscript and obtained funding. All co-authors significantly contributed to the present study, and read and approved the final manuscript.

Funding

This study was funded by the following sources : the INRA AIP Bioressource, EU NovelTree (FP7 - 211868), EU Evoltree (FP6-16322), and INRA SELGEN funding program (project BreedToLast) have funded sequencing and genotyping data. MP PhD grant was jointly funded by INRA SELGEN funding program (BreedToLast) and by Region Centre - Val de Loire funding council.

Acknowledgments

The authors acknowledge the GIS peuplier, the UE GBFOR and PNRGF (ONF) for access, maintenance, and sampling of plant material. The authors want to thank Vanina Gurin and Corinne Buret for their work on DNA extraction. The authors acknowledge Patricia Faivre-Rampant, Marie-Christine Le Paslier and Aurlie Brard from US EPGV for the genotyping and the sequencing.

Data Availability Statement

The datasets [GENERATED/ANALYZED] for this study can be found in the [NAME OF REPOSITORY] [LINK].

References

Akdemir, D., Sanchez, J. I., and Jannink, J. L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel. Evol.* 47, 1–10. doi:10.1186/s12711-015-0116-6

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., Fix, J., Van Tassell, C. P., et al. (2013). Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genet.* 14, 1–14. doi:10.1186/1471-2156-14-8

Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., et al. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics* 17, 1–14. doi:10.1186/s12864-016-2879-8

Bisognin, D. A. (2011). Breeding vegetatively propagated horticultural crops. *Crop Breed. Appl. Biotechnol.* 11, 35–43. doi:10.1590/S1984-70332011000500006

Blondel, M., Onogi, A., Iwata, H., and Ueda, N. (2015). A ranking approach to genomic selection. *PLoS One* 10, 1–23. doi:10.1371/journal.pone.0128570

Calus, M. and Veerkamp, R. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* 43, 26. doi:10.1186/1297-9686-43-26

Cappa, E. P. and Cantet, R. J. (2007). Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. *Can. J. For. Res.* 37, 2677–2688. doi:10.1139/X07-116

Cappa, E. P., Muñoz, F., Sanchez, L., and Cantet, R. J. C. (2015). A novel individual-tree mixed model to account for competition and environmental heterogeneity: a Bayesian approach. *Tree Genet. Genomes* 11, 120. doi:10.1007/s11295-015-0917-3

Castellani, E., Freccero, V., Lapietra, G., and Castellani, E and Freccero, V and Lapietra, G. (1967). Proposta di una scala di differenziazione delle gemme fogliari del pioppo utile per gli interventi antiparas sitari. *Plant Biosyst.* 101, 355–360. doi:10.1080/11263506709426301

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi:10.1534/genetics.112.147983

de Almeida Filho, J. E., Guimarães, J. F. R., e Silva, F. F., de Resende, M. D. V., Muñoz, P., Kirst, M., et al. (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity (Edinb).* 117, 33

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385. doi:10.1534/genetics.109.101501

[Dataset] de los Campos, G. and Perez Rodriguez, P. (2018). BGLR: Bayesian Generalized Linear Regression

Faivre-Rampant, P., Zaina, G., Jorge, V., Giacomello, S., Segura, V., Scalabrin, S., et al. (2016). New resources for genetic studies in *Populus nigra* : genome-wide SNP discovery and development of a 12k Infinium array. *Mol. Ecol. Resour.* 16, 1023–1036. doi:10.1111/1755-0998.12513

Falconer, D. S. (1981). *Introduction to quantitative genetics* (Longman.), 2nd edn.

Forni, S., Aguilar, I., and Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* 43, 1. doi:10.1186/1297-9686-43-1

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., and El-Kassaby, Y. A. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics* 16, 1–16. doi:10.1186/s12864-015-1597-y

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., and 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, 743–753. doi:10.1534/g3.115.025957

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Thompson, R., Butler, D., Cherry, M., et al. (2008). ASReml user guide release 3.0. *VSN Int Ltd*

González-Recio, O., Gianola, D., Rosa, G. J., Weigel, K. A., and Kranis, A. (2009). Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genet. Sel. Evol.* 41, 3. doi:10.1186/1297-9686-41-3

Grattapaglia, D. and Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7, 241–255. doi:10.1007/s11295-010-0328-4

Grüneberg, W., Diaz, F., Eyzaguirre, J., Espinoza, G., Burgos zum Felde, T., Andrade, M., et al. (2009). heritability estimates for an accelerated breeding scheme (abs) in clonally propagated crops-using sweetpotato as a model. In *Proceedings of the 15th symposium of the ISTRC (from 2-6 November 2009), Lima, Peru*

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet.* 15, 1–7. doi:10.1186/1471-2156-15-30

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi:10.1534/genetics.107.081190

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12. doi:10.1186/1471-2105-12-186

Hayes, B., Bowman, P., Chamberlain, A., and Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92, 433–443. doi:10.3168/jds.2008-1646

Henderson, C. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31, 423–447

Howe, G. T., Saruul, P., Davis, J., and Chen, T. H. (2000). Quantitative genetics of bud phenology, frost damage, and winter survival in an F2family of hybrid poplars. *Theor. Appl. Genet.* 101, 632–642. doi:10.1007/s001220051525

Isidro, J., Jannink, J. L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 128, 145–158. doi:10.1007/s00122-014-2418-4

Jia, Y. and Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192, 1513–1522. doi:10.1534/genetics.112.144246

Kainer, D., Stone, E. A., Padovan, A., Foley, W. J., and Külheim, C. (2018). Accuracy of Genomic Prediction for Foliar Terpene Traits in *Eucalyptus polybractea*. *G3: Genes, Genomes, Genetics* 8, 2573–2583. doi:10.1534/g3.118.200443

Kizilkaya, K., Fernando, R. L., and Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.* 88, 544–551. doi:10.2527/jas.2009-2064

Kursa, M. B. and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *J. Stat. Softw.* 36, 1–13. doi:Vol.36,Issue11,Sep2010

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92, 4656–4663. doi:10.3168/jds.2009-2061

Legionnet, A., Muranty, H., and Lefèvre, F. (1999). Genetic variation of the riparian pioneer tree species *Populus nigra*. II. Variation in susceptibility to the foliar rust Melampsora larici-populina. *Heredity (Edinb).* 82, 318–327. doi:10.1038/sj.hdy.6884880

Lenz, P. R., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M., and 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, 1–17. doi:10.1186/s12864-017-3715-5

Liaw, a. and Wiener, M. (2002). Classification and Regression by randomForest. *R news* 2, 18–22. doi:10.1177/154405910408300516

Martini, J. W., Gao, N., Cardoso, D. F., Wimmer, V., Erbe, M., Cantet, R. J., et al. (2017). Genomic prediction with epistasis models: On the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinformatics* 18, 1–16. doi:10.1186/s12859-016-1439-1

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi:11290733

Miller, A. J. and Gross, B. L. (2011). From forest to field: Perennial fruit crop domestication. *Am. J. Bot.* 98, 1389–1414. doi:10.3732/ajb.1000522

Mota, R. R., Silva, F. F. e., Guimarães, S. E. F., Hayes, B., Fortes, M. R. S., Kelly, M. J., et al. (2018). Benchmarking Bayesian genome enabled-prediction models for age at first calving in Nellore cows. *Livest. Sci.* 211, 75–79. doi:10.1016/j.livsci.2018.03.009

Müller, B. S., Neves, L. G., de Almeida Filho, J. E., Resende, M. F., Muñoz, P. R., dos Santos, P. E., et al. (2017). Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. *BMC Genomics* 18, 1–17. doi:10.1186/s12864-017-3920-2

[Dataset] Muñoz, F. and Sanchez, L. (2018). breedR: Statistical Methods for Forest Genetic Resources Analysts

Munoz, P. R., Resende, M. D. M. F., Huber, D. A., Quesada, T., Resende, M. D. M. F., Neale, D. B., et al. (2014). Genomic relationship matrix for correcting pedigree errors in breeding populations: Impact on genetic parameters and genomic selection accuracy. *Crop Sci.* 54, 1115–1123. doi:10.2135/cropsci2012.12.0673

Muñoz, P. R., Resende, M. F., Gezan, S. A., Resende, M. D. V., de los Campos, G., Kirst, M., et al. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* 198, 1759–1768. doi:10.1534/genetics.114.171322

Nakaya, A. and Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Ann. Bot.* 110, 1303–1316. doi:10.1093/aob/mcs109

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3: Genes, Genomes, Genetics* , g3.200311.2018doi:10.1534/g3.118.200311

Patry, C. and Ducrocq, V. (2011). Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci.* 94, 1011–1020

Pegard, M., Rogier, O., Bérard, A., Faivre-Rampant, P., Le Paslier, M.-C., Bastien, C., et al. (2018). Sequence Imputation from Low Density Single Nucleotide Polymorphism Panel in a Black Poplar Breeding population. *bioRxiv* Submitted

Pichot, C. and Tessier Du Cros, E. (1988). Estimation of genetic parameters in the European black poplar (*Populus nigra* L.). Consequence on the breeding strategy. *Ann. des Sci. For.* 45, 223–237

[Dataset] R Core Team (2018). R: A Language and Environment for Statistical Computing

Ratcliffe, B., El-Dien, O. G., Cappa, E. P., Porth, I., Klápště, J., Chen, C., et al. (2017). Single-Step BLUP with Varying Genotyping Effort in Open-Pollinated *Picea glauca*. *G3: Genes, Genomes, Genetics* 7, 935–942. doi:10.1534/g3.116.037895

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

Resende, M. D., Resende Jr, M. F., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., et al. (2012a). Genomic selection for growth and wood quality in Eucalyptus: Capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol.* 194, 116–128. doi:10.1111/j.1469-8137.2011.04038.x

Resende, M. F. R. D. V., Munoz, P., Resende, M. F. R. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., et al. (2012b). Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (*Pinus taeda* L.). *Genetics* 190, 1503–1510. doi:10.1534/genetics.111.137026

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (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, 715–728. doi:10.1534/genetics.112.141473

Romero Navarro, J. A., Phillips-Mora, W., Arciniegas-Leal, A., Mata-Quirós, A., Haiminen, N., Mustiga, G., et al. (2017). Application of Genome Wide Association and Genomic Prediction for Improvement of Cacao Productivity and Resistance to Black and Frosty Pod Diseases. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.01905

Sørensen, A. C., Berg, P., and Woolliams, J. A. (2005). The advantage of factorial mating under selection is uncovered by deterministically predicted rates of inbreeding. *Genetics Selection Evolution* 37, 57

Su, G., Madsen, P., Nielsen, U. S., Mäntysaari, E. A., Aamand, G. P., Christensen, O. F., et al. (2012). Genomic prediction for Nordic Red Cattle using one-step and selection index blending. *J. Dairy Sci.* 95, 909–917

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., and 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 Biol.* 17, 110. doi:10.1186/s12870-017-1059-6

Tan, B., Grattapaglia, D., Wu, H. X., and Ingvarsson, P. K. (2018). Genomic relationships reveal significant dominance effects for growth in hybrid Eucalyptus. *Plant Sci.* 267, 84–93. doi:10.1016/j.plantsci.2017.11.011

Teissier, M., Larroque, H., and Robert-Granié, C. (2018). Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: A quantitative trait influenced by a major gene. *Genet. Sel. Evol.* 50, 1–12. doi:10.1186/s12711-018-0400-3

Teissier du Cros, E. (1977). Aperçu de la transmission héréditaire de quelques caractères juvéniles chez *Populus nigra* L. *Ann. Sei. For.* 34, 311–322

Toro, M. A. and Varona, L. (2010). A note on mate allocation for dominance handling in genomic selection. *Genet. Sel. Evol.* 42, 1–9. doi:10.1186/1297-9686-42-33

VanRaden, P. M. (2007). Genomic Measures of Relationship and Inbreeding. *Interbull Bull.* 25, 111–114. doi:10.1007/s13398-014-0173-7.2

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics* 206, 1297–1307. doi:10.1534/genetics.116.199406

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195, 1223–1230. doi:10.1534/genetics.113.155176

Wang, H., Misztal, I., Aguilar, I., Legarra, A., and Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res. (Camb).* 94, 73–83. doi:10.1017/S0016672312000274

Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetics Research* 75, 249–252

Wientjes, Y. C., Veerkamp, R. F., and Calus, M. P. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193, 621–631. doi:10.1534/genetics.112.146290

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H. J., Wang, Y., and Schön, C. C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195, 573–587. doi:10.1534/genetics.113.150078

Wolak, M. E. (2012). nadiv : an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. *Methods Ecol. Evol.* 3, 792–796. doi: 10.1111/j.2041-210X.2012.00213.x

Zhang, X., Lourenco, D., Aguilar, I., Legarra, A., and Misztal, I. (2016). Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS. *Front. Genet.* 7, 1–14. doi:10.3389/fgene.2016.00151

Figure captions

6.1 Tables

Table 1: Combination of models and marker sets tested.

Methodes	ADD	ADD + DOM	MultiTrait	SNP set
P-BLUP	Yes	Yes	Yes	none
P-BLUPcor	Yes	Yes	Yes	none
GBLUP	Yes	Yes	Yes	7K
	Yes	Yes	No	50K
	Yes	Yes	No	100K
	Yes	Yes	No	250K
wGBLUP	Yes	Yes	No	7K
	Yes	Yes	No	50K
	Yes	Yes	No	100K
	Yes	Yes	No	250K
BayesCpi	Yes	No	No	7K
	Yes	No	No	50K
	Yes	No	No	100K
	No	No	No	250K

6.2 Figures

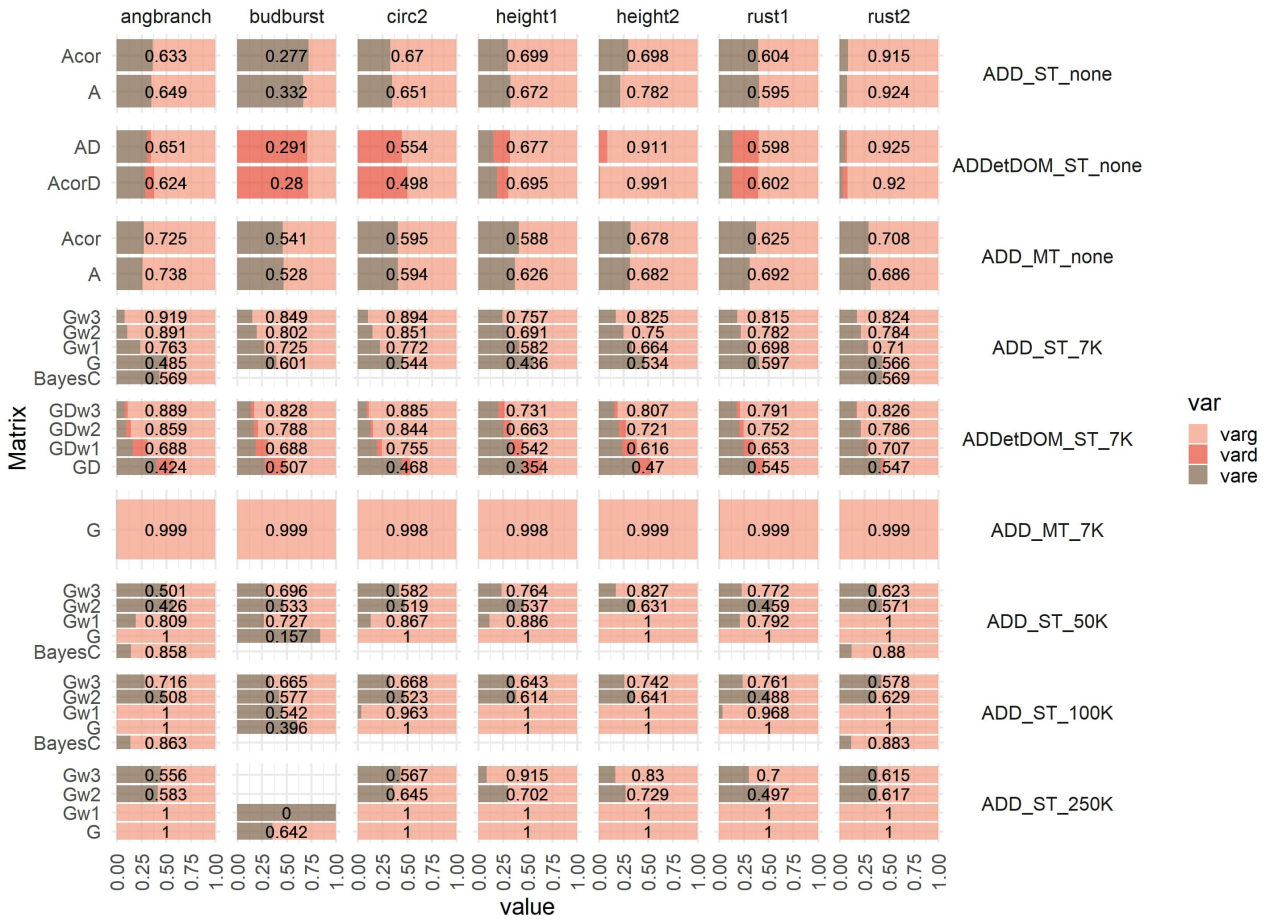


Figure 1: Heritability and variance components estimated with the complete dataset with different models and different traits. The results were organized with traits in columns and model matrices in rows. The matrices were classified by Model (Additive (ADD), Dominance (ADDetDOM), single-trait (ST), multi-trait (MT)) and by number of marker (none, 7K, 50K, 100K, 250K). The variance components were represented by the barplot: in light-grey the proportion of variance due to additive genetic effect, in medium-grey the proportion of variance due to the dominance genetic effect and in dark-grey the proportion of variance due to the residual effect.

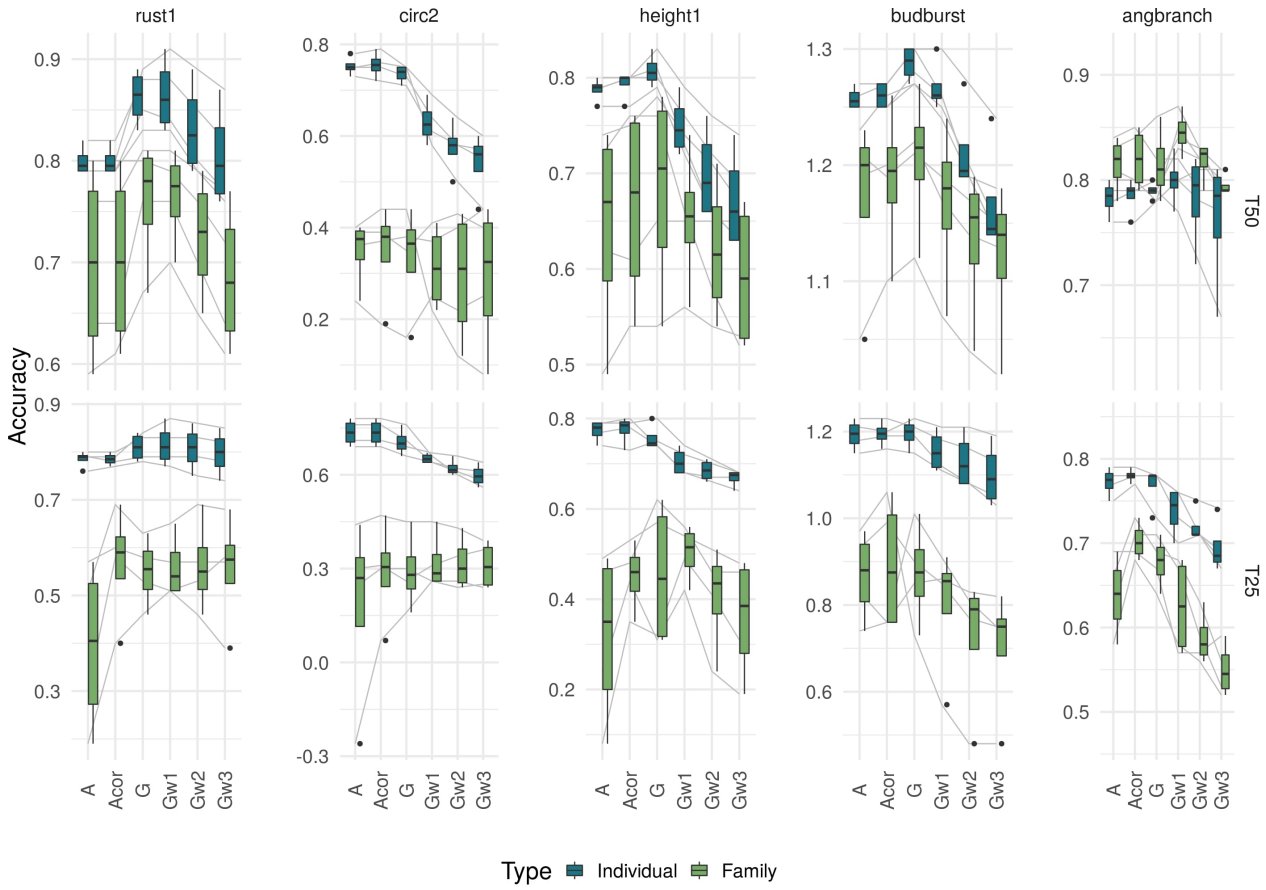


Figure 2: Cross-validation prediction accuracies using an additive model with 7K SNP for five traits grouped by the proportion of individuals in training sets 50% (T50) and 25% (T25). The color of boxplot showed the sampling strategy: in blue, the individual sampling strategy and in green the family sampling strategy. Each boxplot represented the accuracy of four repetitions for each relationship matrix. The gray lines represented the paired repetitions.

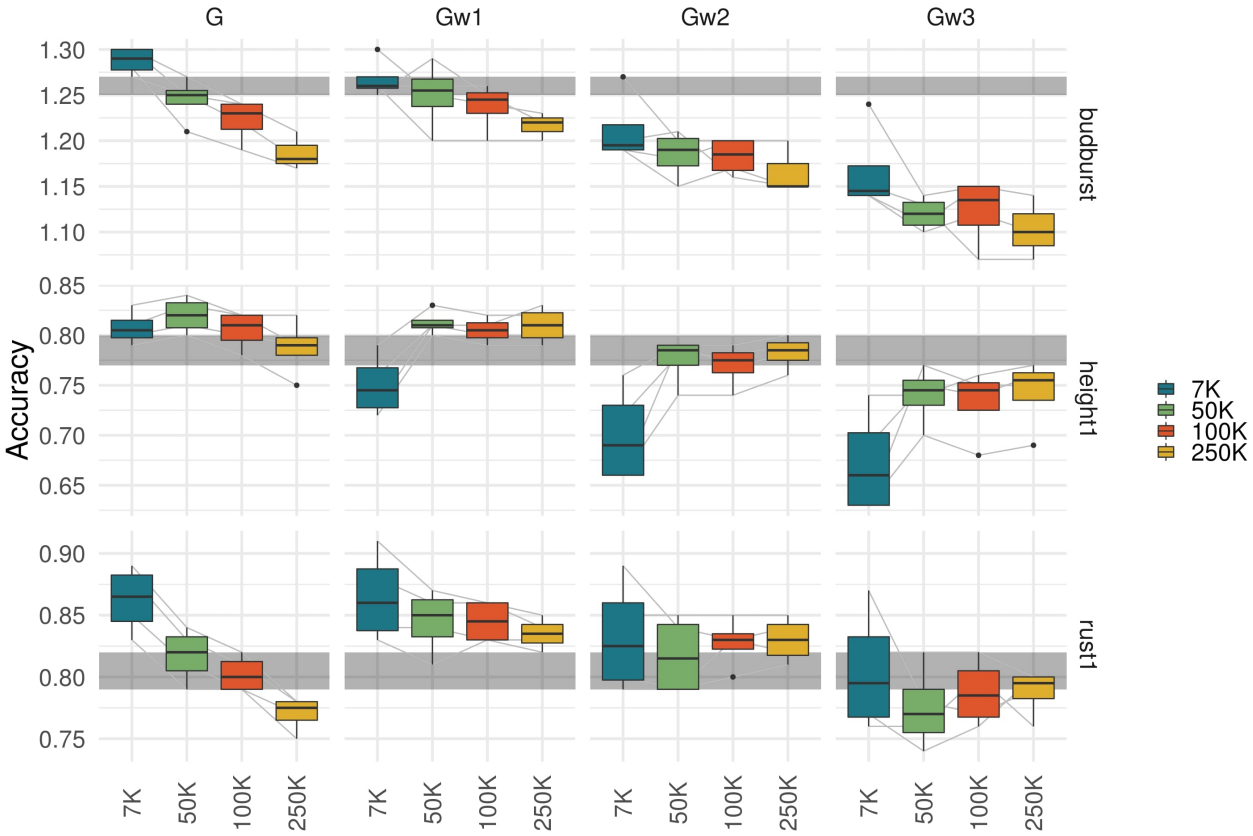


Figure 3: Marker densification impact on predictive accuracy of a single trait additive model with T50 individual cross-validation for four genomic relationships matrices (in columns) and three different traits (in rows) : height1, budburst and rust1. The range of accuracies obtained with the pedigree information was represented in grey and the boxplot colors represent the number of markers. The grey lines represent paired repetitions of cross-validations.

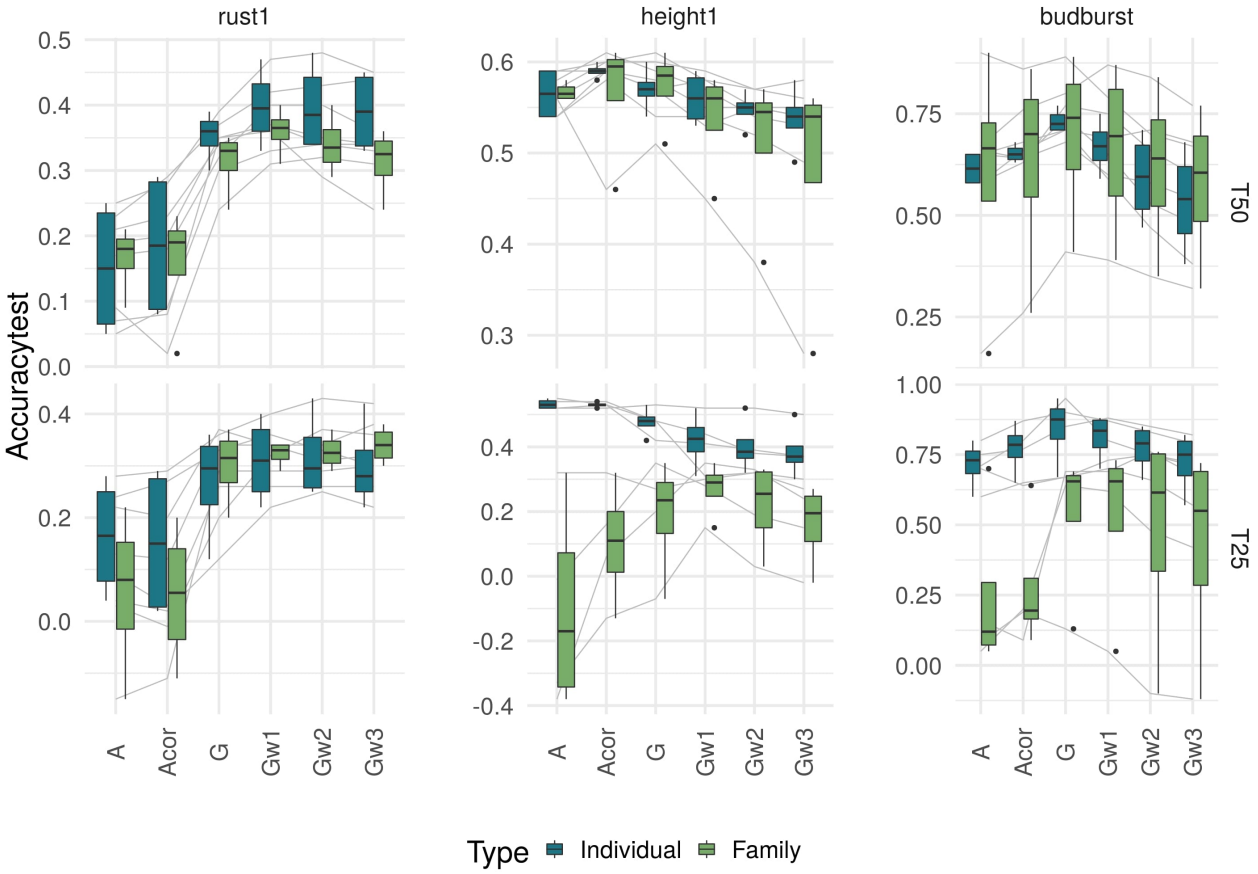


Figure 4: Prediction accuracy of the independent test sets for five traits grouped by the proportion of individuals in training sets 50% (T50) and 25% (T25) with 7K SNP. The color of boxplot showed the sampling strategy: in blue the individual sampling strategy and in green the family sampling strategy. Each boxplot represented the accuracy of four repetitions for a relationship matrix. The grey lines represented the paired repetitions.

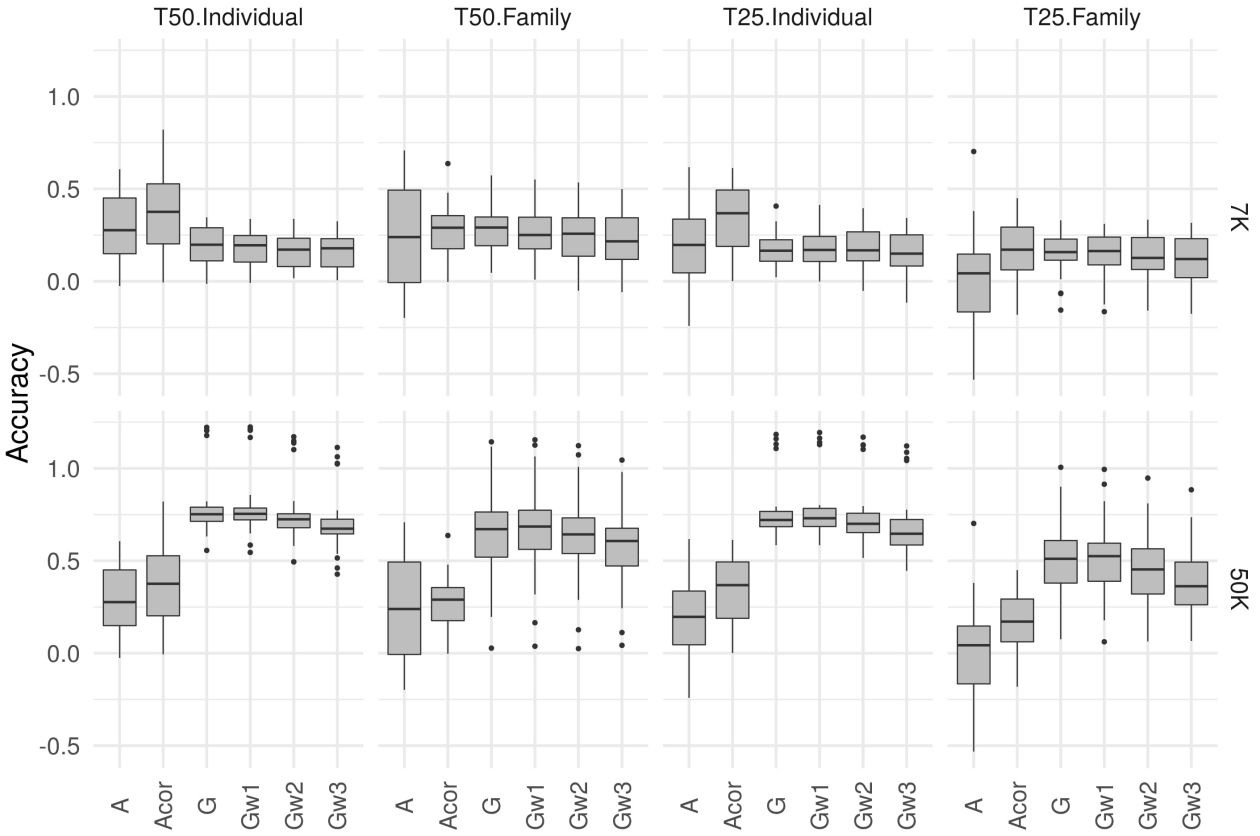


Figure 5: Prediction accuracy of cross-validation strategies with depreciated phenotypes (3 blocks) for 7K SNP and 50K SNP, and 6 matrices. Each boxplot represented accuracy values for pooled traits obtained with the different relationship matrices classified by cross-validation strategy in columns and by the number of markers in rows.

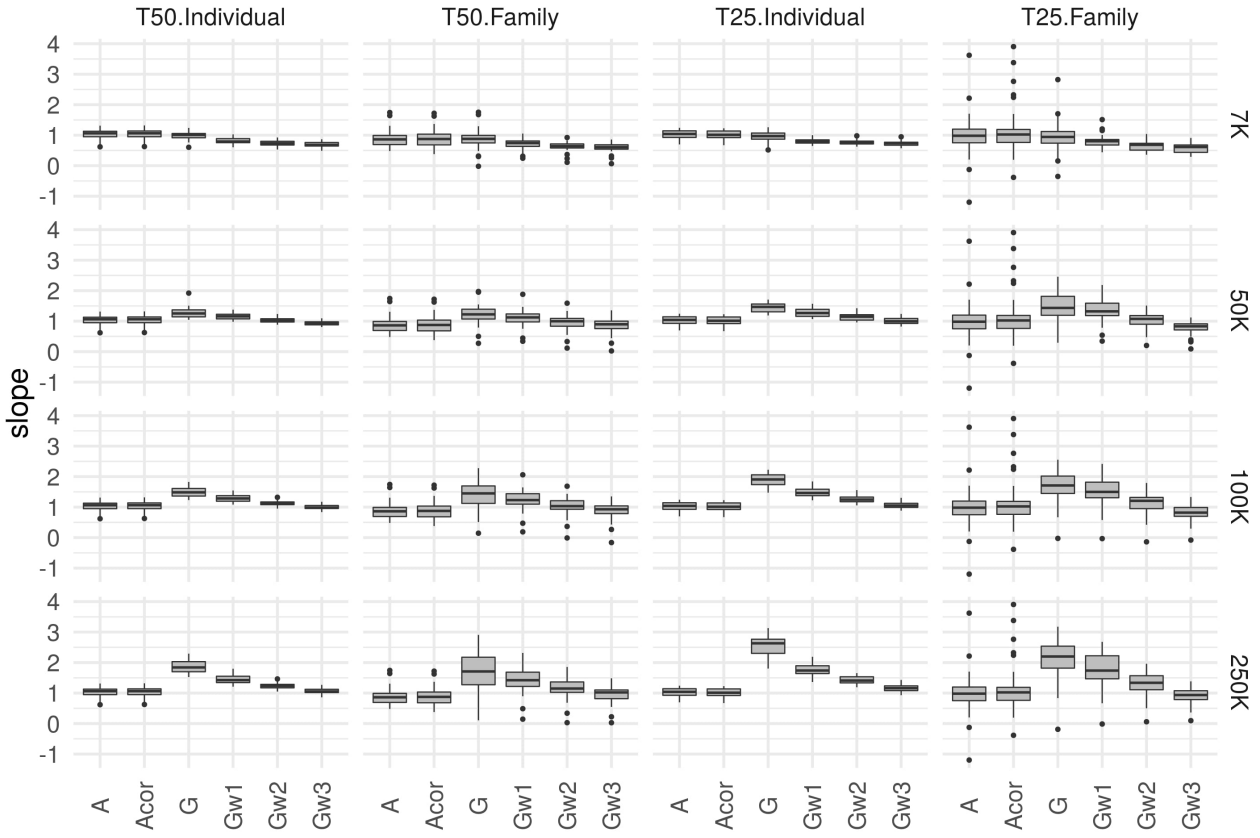


Figure 6: Regression slopes between phenotypes and estimated breeding value in a validation population. Each boxplot represented accuracy values for pooled traits according to relationship matrices used, and were classified by cross-validation strategy in columns and by the number of markers in rows.

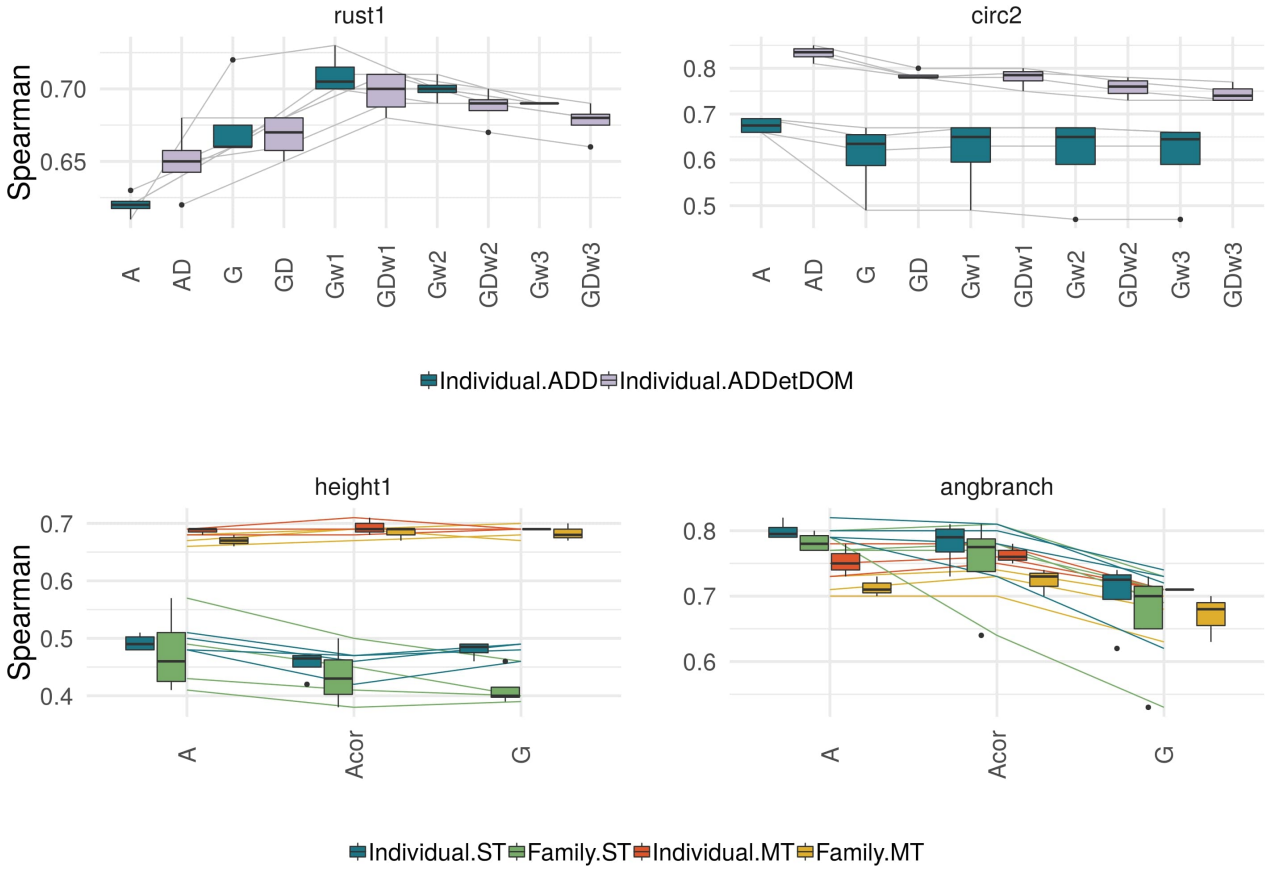


Figure 7: Spearman's correlation between phenotypes and estimated breeding value trained with phenotypes. The results were obtained by cross-validation type T50 with 7K SNP. The upper part of the figure compared single-trait models with additive genetic effect (blue) against single-trait with additive and dominance genetic effect (pink) for the rust1 and the circ2. Only the results of the individual sampling cross-validation were showed. The lower part, compared results of single-trait models with individual sampling (blue); single-trait models with family sampling (green); multiple-trait models with individual sampling (orange); multiple-trait models with family sampling (yellow).

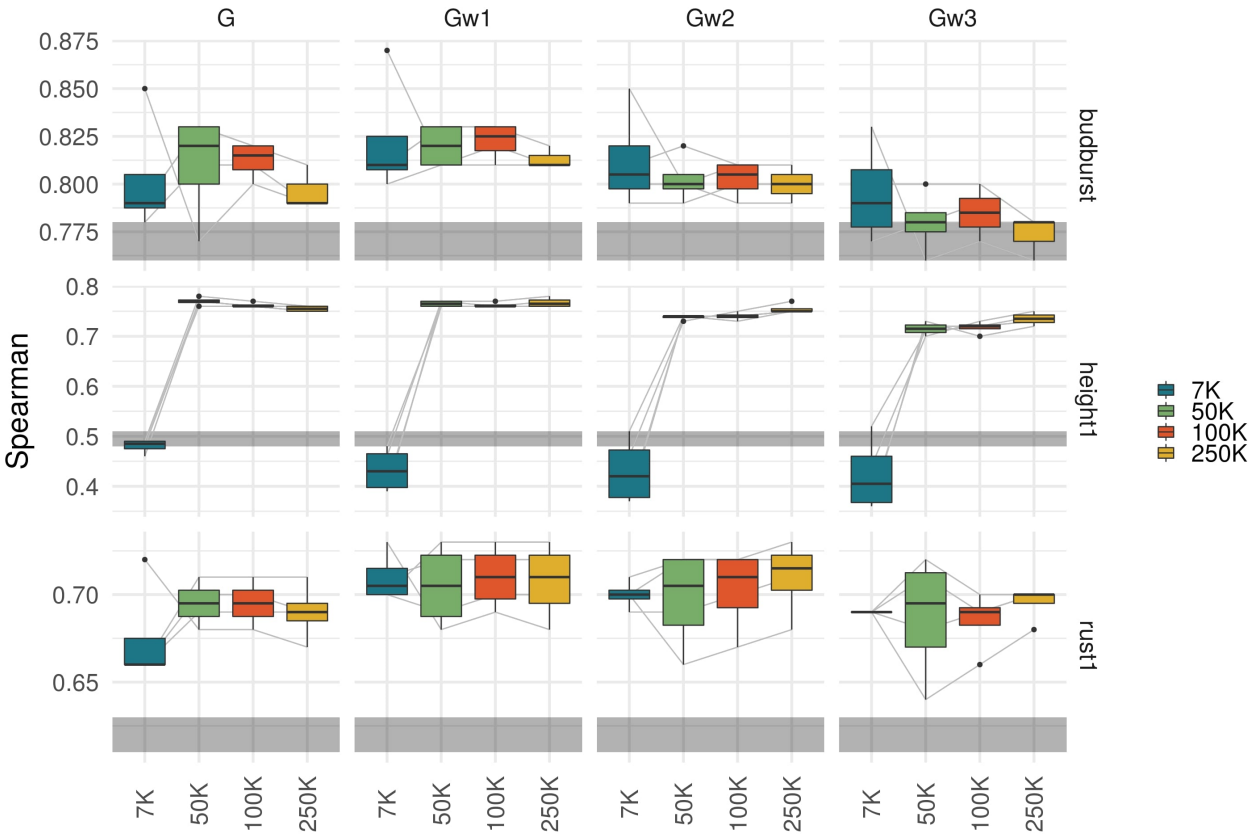


Figure 8: Marker densification impact on Spearman's correlation for four genomic relationships matrices (in columns) for three different traits (in rows) : height1, budburst and rust1. The Spearman's correlation obtained with the pedigree information was represented in grey and the boxplot colors represent the number of markers: in blue 7K, in green 50K, in orange 100K and in yellow 250K. The grey lines represent the repetitions. The results were obtained with T50 individual cross-validation scenario.

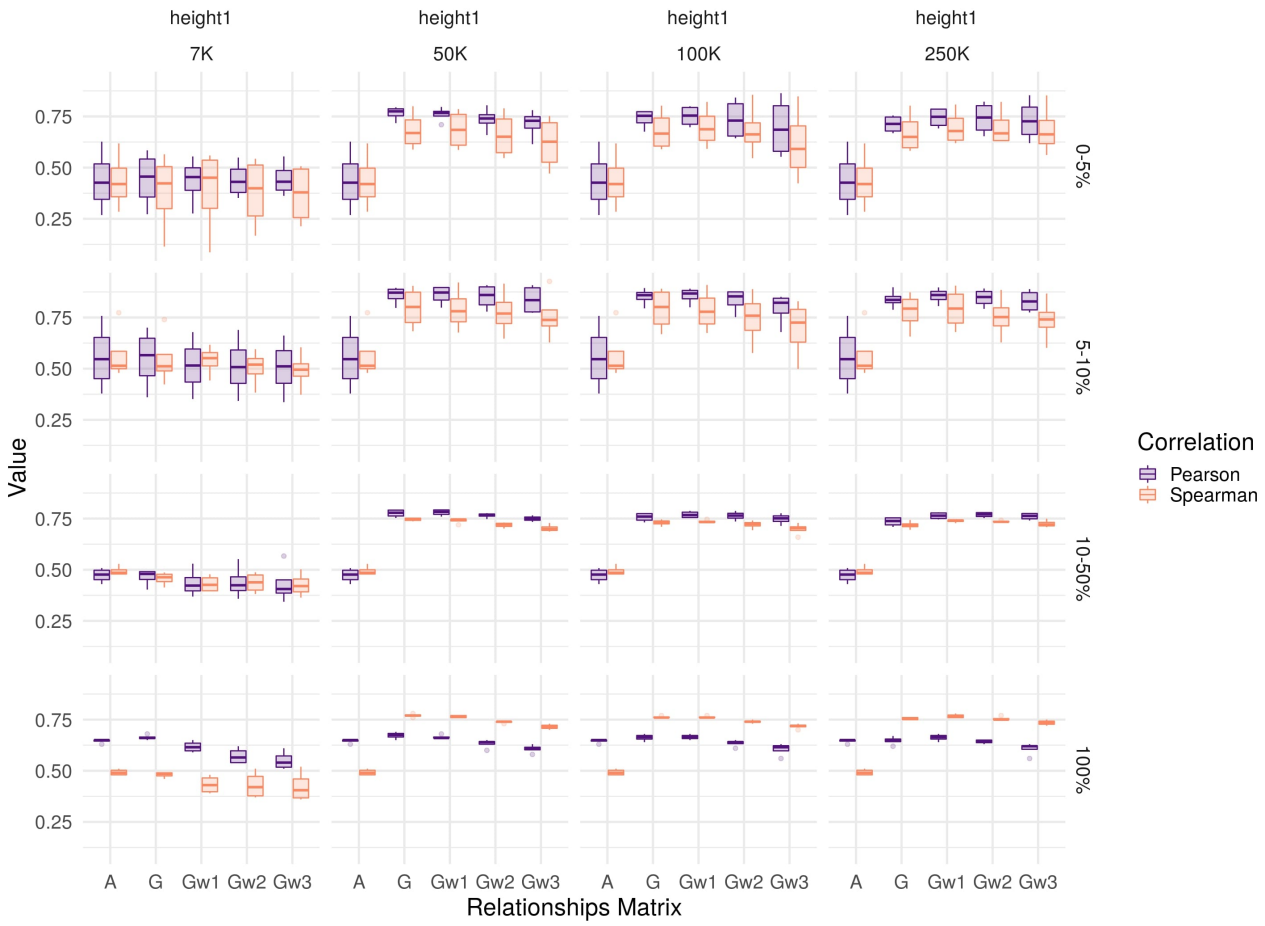


Figure 9: Comparison of Spearmans (orange) and Pearsons (purple) correlations between phenotypes and estimated breeding value. The matrices were in columns and grouped by the number of SNP. In rows the proportion of individuals to estimate the correlations. 0-5% were the 5% best individuals, 5-10% the 5% to 10% best individuals, 10-50% the 10% to 50% best individuals based on phenotypes inside the validation population (T50 individual sampling strategy). 100% represented the Spearman and Pearsons estimated using all the individuals in the validation population.

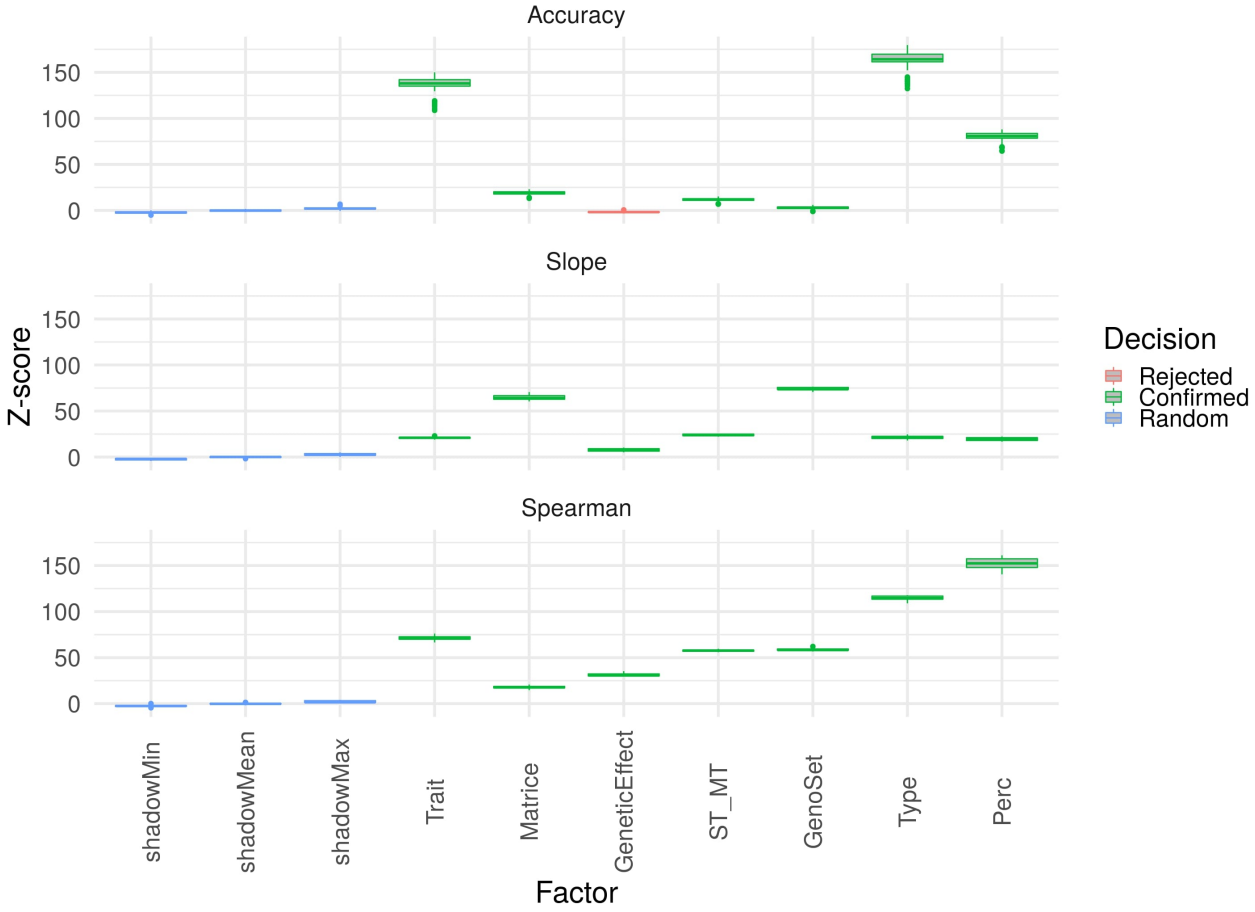


Figure 10: Importance (Z-score) for each features estimated with Boruta algorithm to explain Accuracy, slope and Spearman's correlation (Spearman) variability. Boruta shadow features were ShadowMin, ShadowMean and ShadowMax. The test factors were Trait (rust1, rust2, height1, height2, circ2, budburst, angbranch), Matrix (A, Acor, G, Gw1, Gw2, Gw3, D, Dcor, Dw1, Dw2, Dw3), GeneticEffect (Additive, Additive and Dominance), ST_MT (Single-Trait, Multiple-Trait), GenoSet (none, 7K, 50K, 100K, 250K), Type (Individual, Family) and Perc (T50, T25). Algorithm decision for each factor based on the significativity of the difference between factors and the shadow features are in color: Blue: Shadow features, Green: Confirmed and Pink Rejected.

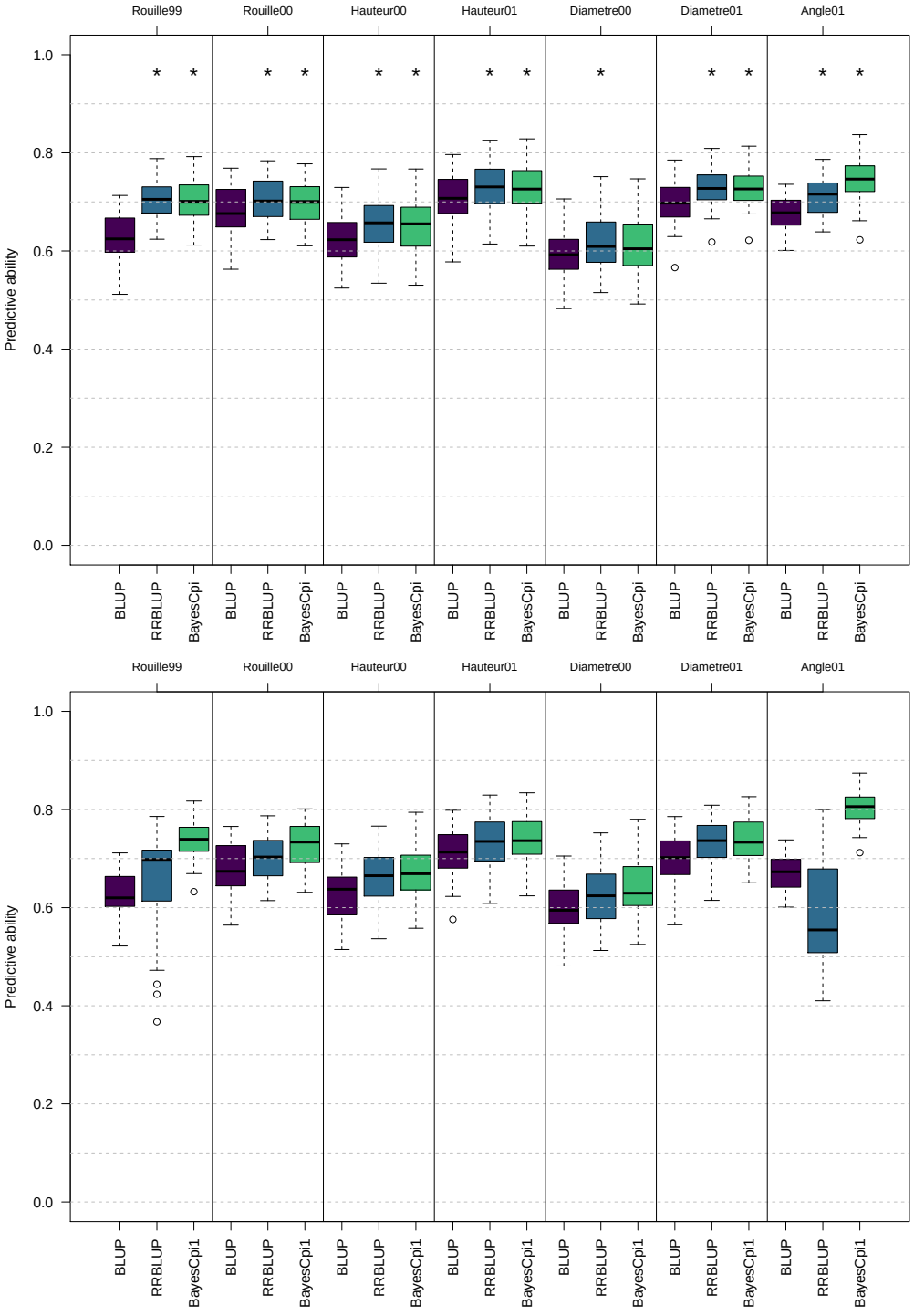


Figure 4.3: Predicting ability on the extended factorial mating design with 7K SNP, with 75% random sampling of population for training and 25% for validation. The methods used to estimate EBVs are in columns, organized by traits. Upper panel is for additive models and lower panel for additive + dominance models. The stars represented the Student's t-test p-value for paired samples p-value : " " > 0.05 "*" > 0.01 "**" > 0.001 > "***"

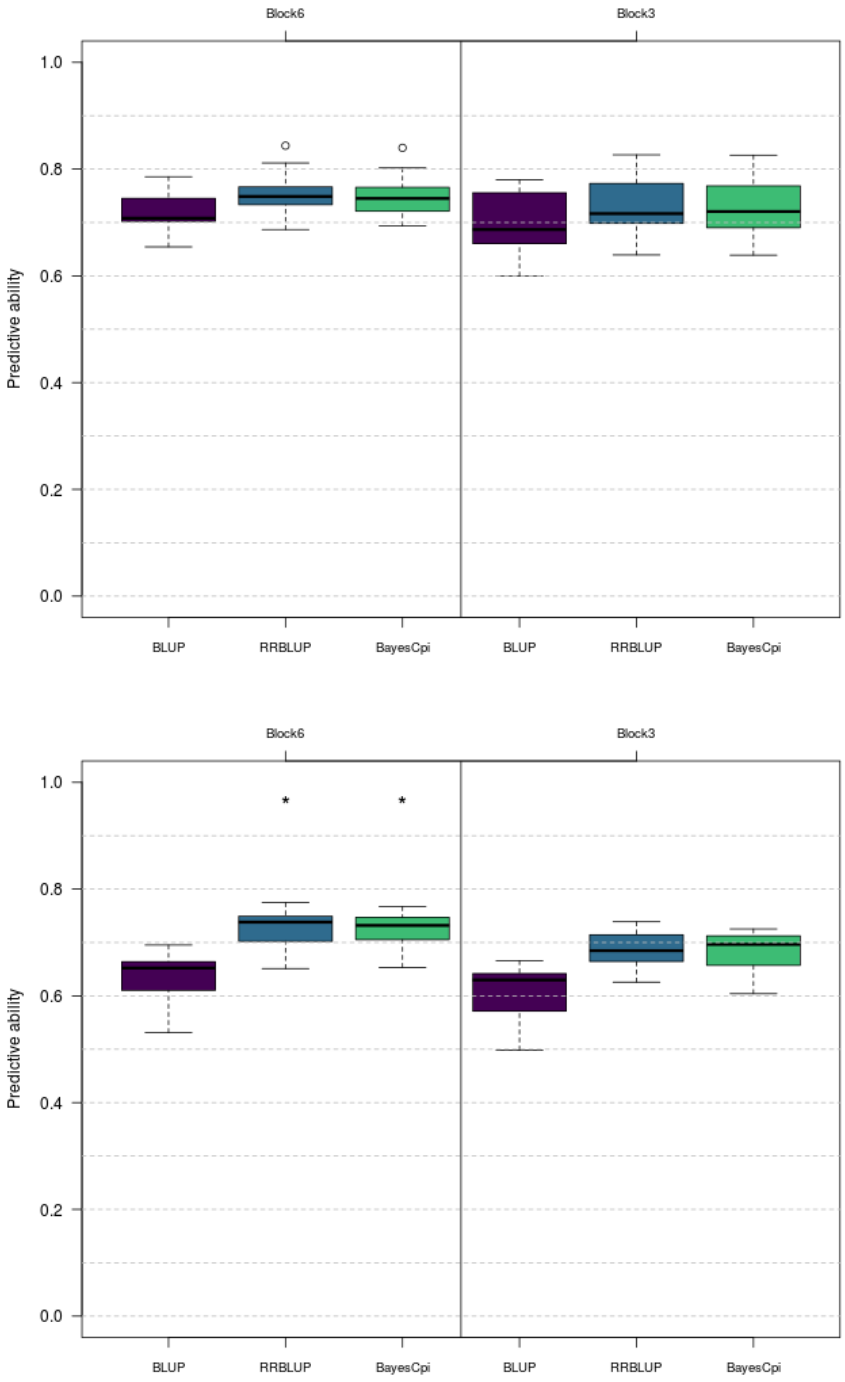


Figure 4.4: Predicting ability on the extended factorial mating design with 7K SNP, with 75% sampling with CDmean algorithm of population for training and 25% for validation, and an additive model. The methods used to estimate EBV are in columns, and organized by number of repetitions (Block6 : 6 repetitions and Block3 : 3 repetitions). Two traits: circumference at two years of age (left) and rust resistance in the first year of growth (right).The stars represented the Student's t-test p-value for paired samples p-value : " " $>$ 0.05 "*" $>$ 0.01 "**" $>$ 0.001 $>$ "***".

4.5 Additional results

4.5.1 Estimated markers effects within different shrinkage weights

The use of weighted GBLUP and the subsequent extraction of marker effects from GEBV allowed us to represent the distribution of effects across the genome. In (Figure 4.5), the distribution of marker effects for circumference at the age of 2 is shown for three steps of successive weightings in wGBLUP. Other traits presented similar patterns. At each new weighting, from the upper panel to the bottom panel in the figure the scale of the effects is augmented, given that a few markers accentuated their value with the weighting, while most of the remaining markers were pushed towards zero.

The trait, however, presented a rather flat landscape, with most effects being of the same magnitude. This is somehow a hint of the infinitesimal architecture of the trait, that had markers across all chromosomes. This infinitesimal aspect was confirmed with marker densification in figure 4.6, where wGBLUP at the first weighting iteration is shown over three different densities (7K, 50K and 250K). The fact of covering more efficiently all the genomic regions brought higher marker effects with densification. Marker effects were at least multiplied by a factor of six with the 50K SNP set and by a factor of ten for the 250K SNP set. The densest set with 250K SNPs brought more chromosomal regions with important effects.

4.5.2 Haplotype approach

The haplotypes of each individual were extracted from the 7K chip after imputation to remove the 1% of missing values. An ad-hoc script was developed to affiliate each of the two alleles in each marker and for each descendant to one of the four possible parental chromosomes (2 maternal homologous, 2 paternal).

The figure 4.7 represent the haplotype blocks along the chromosome 1 for a given progeny. The first relevant feature of the graph is the fact that there are relatively narrow regions with the same breaking point across many individuals. This could correspond to recombination hotspots, to mapping errors with wrong SNP location, to error in

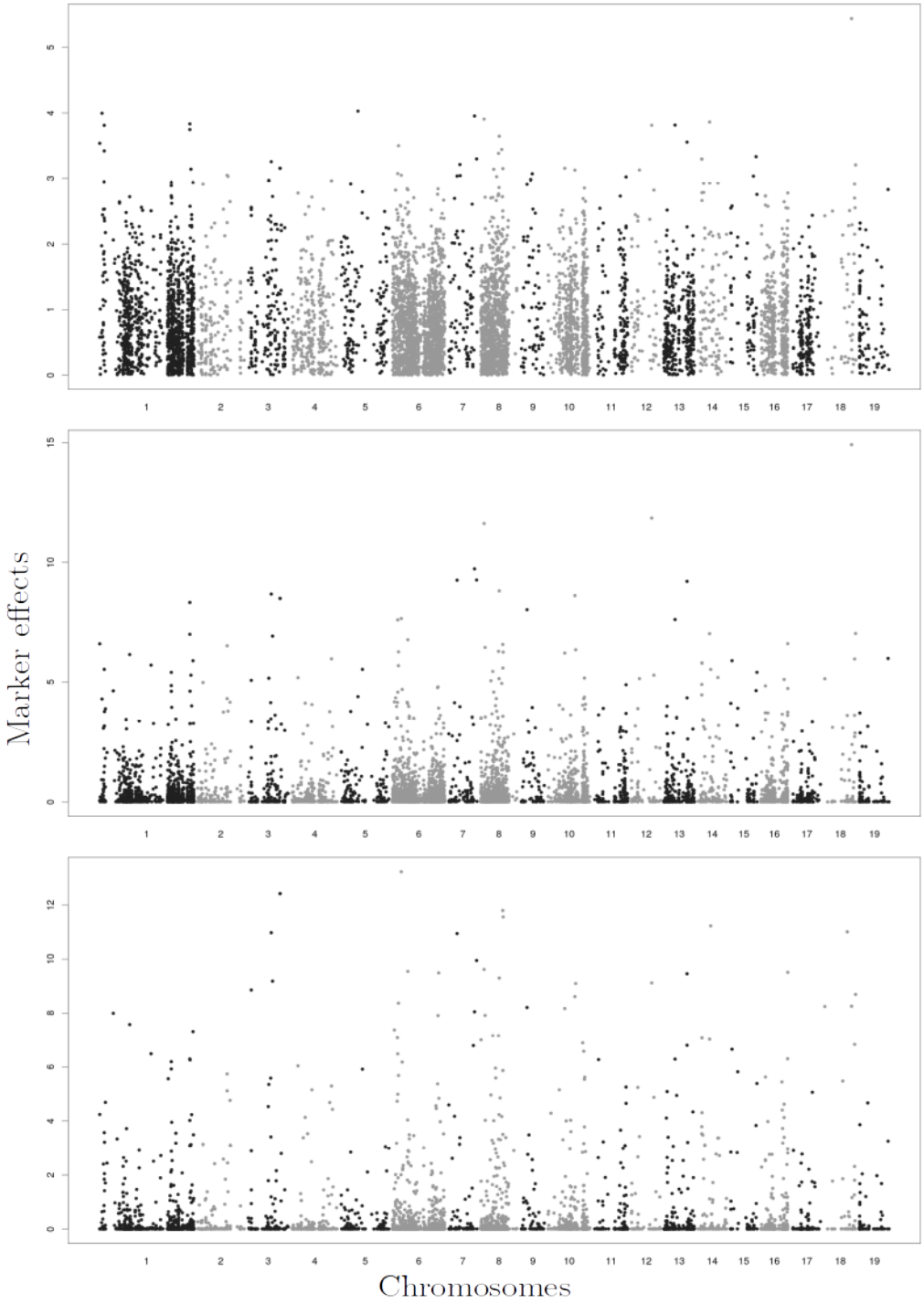


Figure 4.5: Marker effects estimated for the circumference at two years old with the 7K SNP with three weightings in the GBLUP: top panel, no weights in GBLUP; Middle panel, weights after a first iteration of wGBLUP (Gw1); bottom panel, third iteration of wGBLUP (Gw3).

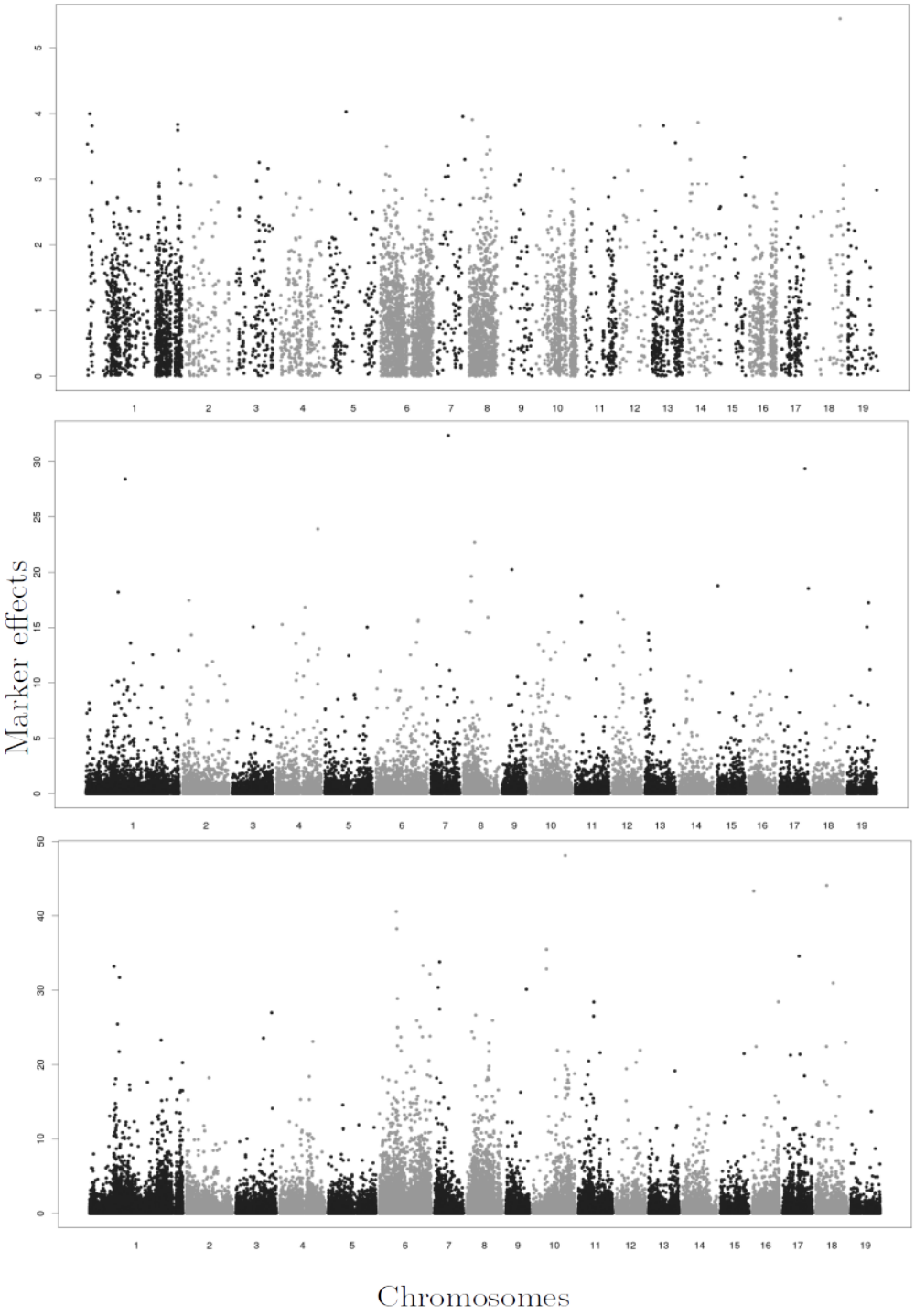


Figure 4.6: Markers effects estimated for the circumference at two years old with a first iteration of wGBLUP (Gw1) for three SNP sets : top panel, 7K SNPs; Middle panel, 50K SNPs; bottom panel, 250K SNPs.

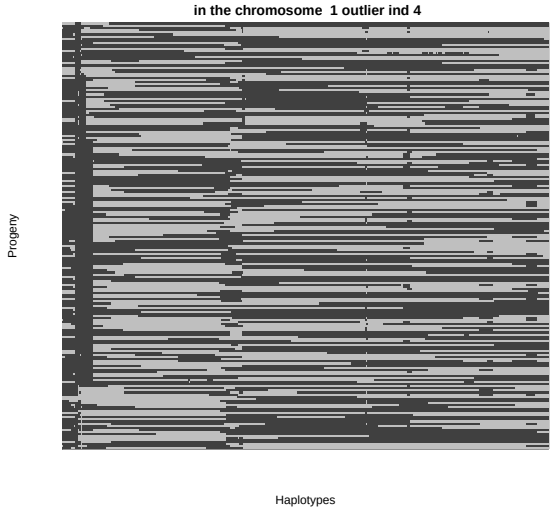


Figure 4.7: Haplotype visualization after phase reconstruction with the FImpute software of the Chromosome 1 (SNP in columns) for all the progenies (in rows) of the female parent 662200037. The dark gray and light gray represented the membership at either of the two homologous parental chromosomes. The SNPs were ordered following the physical position from the chip (Faivre-Rampant et al., 2016)

the reference sequence assembly or to lack of information in those particular regions to reconstruct the segregation information. Another feature is that there were in general a substantial amount of recombinations within the family.

The haplotype blocks reconstruction allowed to estimate the recombination rate in a half-sib family (Figure 4.8) and among all progenies (Figure 4.9) for the chromosome 1. Chromosomes 1, 6, 8 and 10 showed different hotspot recombination regions surrounded by regions with little or no recombination. However, the 7K SNP chip had too many gaps in terms of coverage to take these maps as representative of the actual distribution of recombination events. Some chromosomes like 9 or 18 had little or no density in some of their regions, for instance.

Despite this lack of coverage across all chromosomes, in the absence of the denser sets that were available latter on in the thesis, we used haplotype blocks to compute the A* matrix and compare resulting relationship with other relationship methods in terms of heritability and Akaike criterion (AIC). Figure 4 4.10 shows the comparison of different models with varying p, where p designed the similarity to A (p = 0) or to G ($p = 1$), and noting that G was set to be an hybrid matrix involving partly pedigree (25%) and

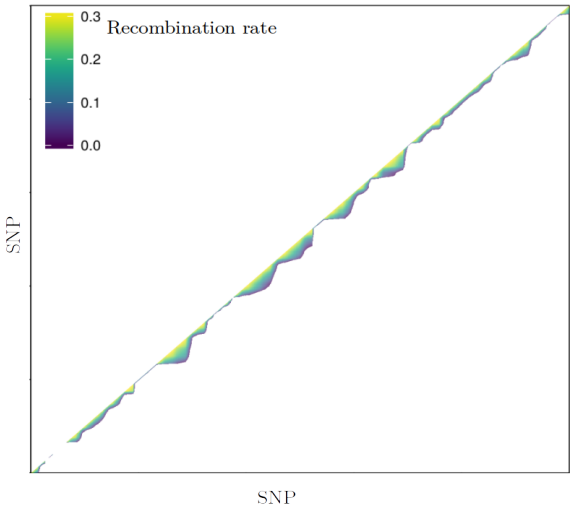


Figure 4.8: Recombination map between all the SNPs of Chromosome 1 ordered following the physical position from the chip (Faivre-Rampant et al., 2016), estimated from all the progenies of 662200037.

partly genomics (75%). Regardless the traits, heritabilities increased with of p, thus with the preponderance of G, except for the circumference at two years, where there was a very slight decline after p = 0.6. Regarding AIC, most traits presented the lowest values towards p = 1 (rust2, height1, height2), some had an inflexion point just before p = 1 (rust1 and angbranch), or at intermediate p (budburst), and one trait had best value contrarily at p = 0. However, for some traits the difference in AIC between models was very small being difficult to tell apart between them on that criterion alone. Results with p = 1 were not shown due to converge problems. We suspect inconsistencies between the genotyping and the haplotype information.

We compared the best A* based on the AIC to a H matrix (combining 25% of ancestral pedigree and 75% of marker-based relatedness), and to the G matrix from VanRaden (2007). The models are compared depending on heritabilities and AIC (Figure 4.11). Regardless the traits, G presented the lowest AIC, followed by A*, with H and A being the ones with the highest criterion. A* seemed to be more efficient than H when 25% of the individuals were not genotyped. Models were more performant when all individuals were genotyped.

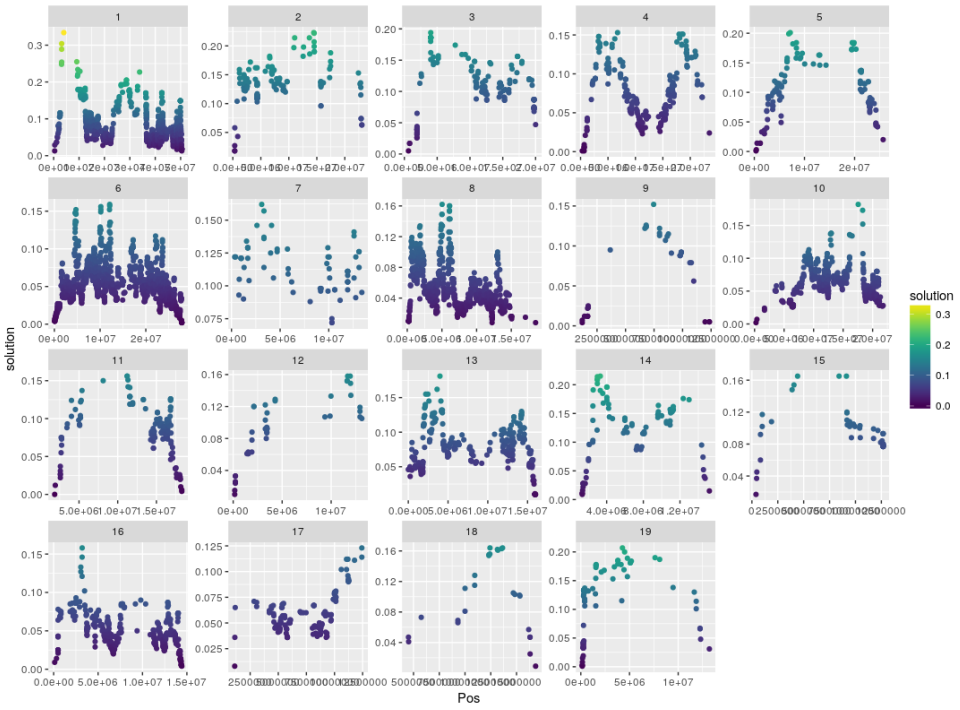


Figure 4.9: Recombination map by chromosomes, an average recombination was estimated via non-overlapping window of 100 base pair among all progenies.

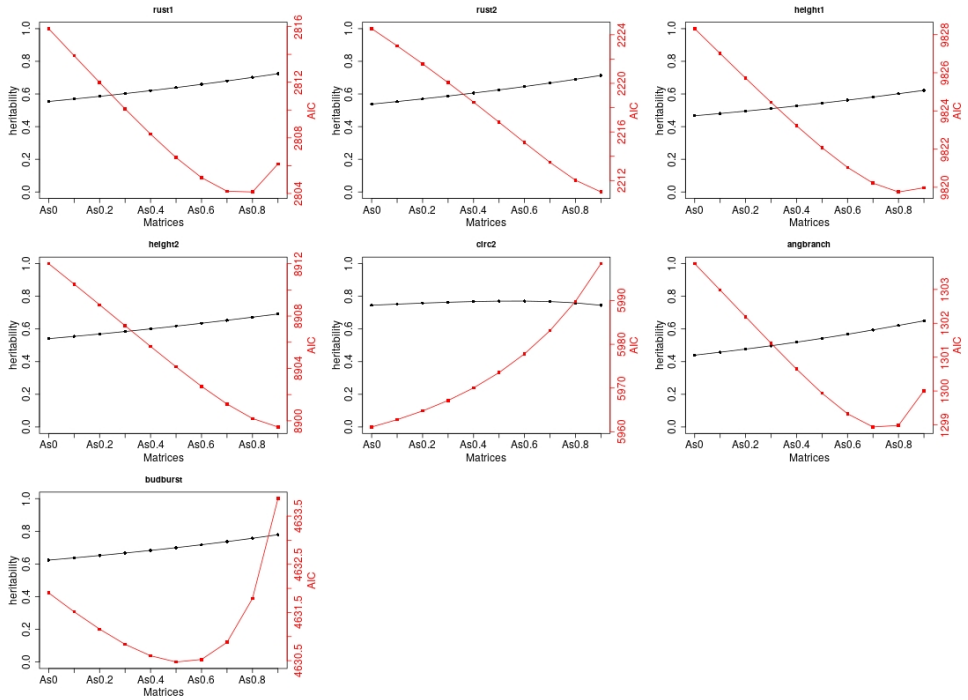


Figure 4.10: Heritabilities and Akaike Information Criterion (AIC) for seven traits obtained following BLUP models with an A* matrix for varying values of p.

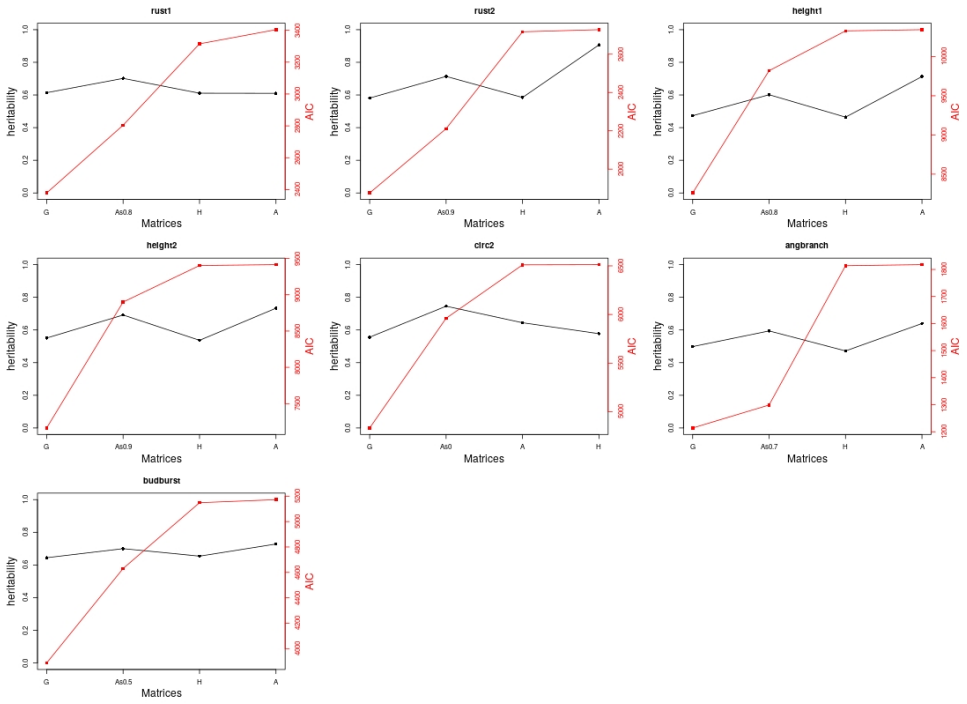


Figure 4.11: Heritabilities and Akaike Information Criterion (AIC) for seven traits following BLUP models with the G Matrix (VanRaden, 2007), best A*, the H matrix and the A matrix.

4.6 Chapter Global discussion

This chapter presented the results of the first genomic selection proof-of-concept in a black poplar breeding population. From the first test on a restricted number of individuals (292) to the latest test with more than thousand individuals, results indicated that the advantages of genomic selection were context dependent, and highlighted the need to try and compare different models and sampling strategies for cross-validation.

The sampling strategy proved to have a substantial impact on the prediction accuracy, in a more substantial way than the proportion of individuals in the training set. An individual sampling with a representation of each family in the training population gave the best results, which could correspond operationally to increase the selection intensity within families via GS. CDmean optimization obtained with a greedy optimizer or via a more efficient simulated annealing did not lead necessarily to better training sets.

Concerning the models, the BayesCπ was a promising option in the first test, but with the augmentation of the population the results were not different from the infinitesimal

methods. This suggests that QTLs with substantial effects may be present in the restricted context of the factorial mating design, and that could become less important in a larger context. The second test with the first implementation of a larger set of SNPs from imputation had firstly an impact on the computing time. It also revealed some convergence difficulties with BayesCπ. This motivated the use of faster and easier to implement alternatives like wGBLUP (Legarra et al., 2009; WANG et al., 2012). These two preliminary tests improved our strategy to setup the basis for the second article of this thesis presented in this chapter.

This article highlighted several facts:

- First, a systematic advantage in term of prediction is not necessarily correlated to the proportion of additive variance captured.
- Genome-based models, and in particular GBLUP, gave the best prediction accuracies, closely followed by pedigree-based models with marker-corrected pedigree.
- The benefits brought by denser SNPs set were trait-dependent and were more evident with the association of variable selection models (wGBLUP).
- Genome-based models succeeded somehow better than the pedigree counterparts in the most challenging situations, with a completely independent set of validation or with downgraded phenotypes.
- Using different quality criteria revealed some other advantages of genome-based models over the pedigree models.

According to our results in prediction accuracy, this study showed the need for multiple methods, prediction quality criteria and different density marker sets to find comfort zones in the use of genomic prediction. Among all these models and datasets, the combination of wGBLUP model at the first iteration combined with the 50K SNPs set gave generally the best results. However, predictions in genome-based models were affected by upward bias in the slope leading to overdispersion of the distribution of GEBV. Bias can lead eventually to wrong selection of individuals.

Model deficiencies can a priori explain this bias, notably due to wrong variance estimation (Sorensen and Kennedy, 1984), unbalanced data (Blair and Pollak, 1984) or pre-selected individuals (Patry and Ducrocq, 2011b). In our case, one of the possible causes of bias could be the use of biased variances in the model. Sometimes, models were hard to get converged, although all results corresponded to converged runs, and there was also some lack of correlation between the additive variance explained by the model and the prediction accuracy that was obtained.

When the prediction quality was assessed through non parametric estimates like Spearman rank correlations, the performance of the different models was substantially altered, and advantages in correlation of G-based methods, notably wGBLUP, over between pedigree-based were clearer. Spearman's criterion also revealed more clearly the benefits of densifying beyond the 7K panel, with optima in 50K. One could argue that for species with clonal dissemination as poplars, predicting accurately the ranking is as important as getting correctly the phenotype.

Further investigations are still necessary to improve the model prediction in terms of accuracy, but also to reduce systematic and overdispersion biases. The slope bias seemed to be positively correlated with the number of markers, while the use of variable selection models like wGBLUP was able to reduce somehow the slope bias. It appeared that the density and marker distribution of original 7K chip did not allow GS to get a clear advantage over the pedigree-based counterpart. Other marker densities between 7K and 50K could be tested to find an eventual inflexion point in performance, while keeping low bias. The marker selection can be optimized to select the best number of the marker but also their repartition. There are many possible ways to optimize a genotyping chip composition in markers, among which a homogeneous repartition seems one of the first logical options. We tested a constant repartition all along the genome, but some regions would need more SNPs than others. The marker repartition could follow the recombination rate in the genome with more SNPs around recombination hotspots. In that sense, the preliminary work with haplotypes could be of help. Apart from being useful for targeting recombinant regions, haplotype derived relationships could also be of help in prediction. Some adjustment and further test with a denser SNP set, however, are necessary to know whether A* can improve predictabilities or not. Additionally, a denser

chip could eventually help the performance of the CDmean method for designing training sets, although it seems difficult to believe that a 7K chip does not capture already in an efficient way the relatedness in the population.

As it stands, genome-based models appear to give some advantages in prediction accuracy, notably when considering ranking individuals. The implementation of genomic selection in the poplar breeding program could be a way to increase genetic gain, shorten the selection cycles and valorize the genetic variability at an early stage.

Chapter 5

General discussion

5.1 Résumé du Chapitre

Cette thèse a été conçue pour évaluer la faisabilité de l'évaluation génomique dans un sous-échantillon du programme de sélection du peuplier noir dans une situation proche des conditions opérationnelles. La méthode d'évaluation de référence, celle qui est actuellement utilisée et fondée sur l'information contenue dans le pedigree, possède déjà une précision de sélection relativement élevée, du moins dans ses dernières étapes. Cependant, cela n'implique pas un manque complet d'intérêt et d'avantages dans la mise en œuvre de la prédiction génomique dans le programme de sélection. En effet, les avantages de l'évaluation génomique dépendent du contexte, et certaines de ces situations sont intéressantes sur le plan opérationnel. Les modèles basés sur le génome capturent généralement des héritabilités et des variances additives plus élevées que leurs équivalents généalogiques, mais l'avantage en matière de précision des prédictions n'est pas systématique. La qualité de la prédiction peut tirer parti de la densification des marqueurs et de l'utilisation de modèles capables de sélectionner les marqueurs pertinents. Dans cette thèse, nous avons montré qu'une densification substantielle de la couverture des marqueurs est possible pour des milliers d'individus par imputation à partir de quelques individus nodaux séquencés. Les étapes de densification par un processus d'imputation ont permis d'avoir une meilleure distribution des marqueurs le long du génome, dans différentes régions génomiques, sans augmentation du déséquilibre de liaison résultant. Les avantages des méthodes basées sur le génome ont été démontrés dans des situations difficiles lors

de l'utilisation d'un ensemble de validation indépendant ou lors de la dégradation du phénotype, ces deux méthodes pouvant être considérées comme intéressantes et réalistes sur le plan opérationnel. Les modèles fondés sur le pedigree n'ont pas été en mesure de prédire correctement les phénotypes d'un jeu de données complètement indépendant, contrairement aux modèles fondés sur le génome. L'utilisation de moyennes clonales calculées sur un nombre plus restreint de copie pour l'entraînement du modèle a eu un effet différentiel entre les prédictions fondées sur le pedigree et les prédictions fondées sur le génome, ces dernières conservant une qualité prédictive similaire à celle d'un entraînement du modèle avec l'ensemble des répétitions.

La densification a également mis en évidence les limites du jeu de données issus du génotypage avec la puce de SNPs de 7K. Celle-ci ne semble pas adapté à la mise en œuvre de la prédiction génomique dans notre programme d'amélioration. Nos résultats ont montré que les SNPs du jeu de données 50K semblent être suffisants, même si le choix de ces SNP de 50K aurait sans doute besoin d'être amélioré, et d'autres que celui basé exclusivement sur la couverture. Il est également nécessaire de rechercher le point d'inflexion entre 7K et 50K SNPs, et de définir qu'elles sont les limitations de la puce, sont-elles due à une couverture hétérogène, à la composition particulière des SNPs ou à une combinaison des deux.

Un autre point intéressant, compte tenu de la façon dont la sélection se fait dans le programme sur le peuplier, est le fait que les modèles basés sur le génome ont été en mesure d'atteindre une bonne performance dans la prédiction précise du classement des candidats à la sélection. Pour rappel, la corrélation Spearman de wGBLUP était bien supérieure aux niveaux des évaluations fondées sur le pedigree pour tous les caractères sauf pour l'évaluation de la rouille. Dans le cas d'une espèce où le gain génétique est disséminé par des clones, la prédiction précise du classement des phénotypes candidats non observés est une caractéristique précieuse. Cette caractéristique était particulièrement évidente après la densification des marqueurs de SNP de 7K à 50K et au-delà. Ces résultats semblent encourageants, en ce sens que l'on pourrait obtenir un gain génétique important en apportant une telle précision dès les premiers stades. Reste cependant la question du biais dans la prédiction de la valeur des candidats, de ses causes et de la façon dont il pourrait nuire à la performance dans les modèles basés sur le génome.

L'une des principales limites de l'étude est probablement l'utilisation d'une population d'entraînement dont la conception ne correspondait pas nécessairement à ce qui se fait habituellement dans l'amélioration du peuplier. En effet, le modèle d'accouplement factoriel, bien que potentiellement intéressant du point de vue de la variabilité parentale, était plus orienté pour les études cartographiques. Ceci a été en partie surmonté par l'ajout de familles et de croisements supplémentaires, bien liés au programme d'amélioration. En ce sens, une population d'entraînement véritablement représentative de la population de base du programme d'amélioration aurait pu permettre de généraliser plus facilement les résultats de l'étude. Un autre aspect est lié à l'étape à laquelle l'évaluation génomique a été mise en œuvre : à un stade de la sélection où une grande partie de la variation a déjà été examinée.

De la même manière, nous avons utilisé une puce disponible qui a été principalement conçue pour des études de cartographie et d'association avec un certain enrichissement en caractères d'intérêt. Cette situation a également été partiellement surmontée par l'approche d'imputation.

5.2 In relation to the objectives of the thesis

This thesis was conceived as a proof-of-concept of the genomic evaluation in a subsample of the black poplar breeding program, and in order to evaluate the feasibility and performances in a situation close to operational conditions. The evaluation method of reference, the one in use currently and based on pedigree-based BLUPs, had already a relatively high selection accuracies, at least in its final steps. This high accuracy may appear counterintuitive in highly heterozygous species and with the assumption of families harbouring very large Mendelian segregation variances for an evaluation system based on pedigrees. Important Mendelian sampling variation are known to be more challenging for pedigree-based models (Hill and Weir, 2011). Part of the competitiveness of this pedigree-based system could be in the use of proper designs, with high number of clonal repetitions and spatial adjustments at the individual level. Another part of the explanation may lie in the well-interconnected network of families given by the factorial system. The factorial design seemingly favored pedigree predictions to a level that made it competitive compared

to genomic predictions in the cross-validation. With an increasing number of crosses, the risk of labelling errors could also increase. However, when correcting pedigrees with the use of markers, which could be a few for parental analyses, such a pedigree-based evaluation became truly competitive compared to genome-based models.

However, such a competitiveness of the pedigree-based methodology currently in use does not imply a complete lack of interest and absence of benefit in the implementation of genomic prediction in the breeding program. Indeed, the genomic evaluation advantages were context-dependent, and some situations where genomic predictions were advantageous are operationally interesting. Genome-based models captured generally higher heritabilities and additive variances than their pedigree equivalents, but the advantage regarding prediction accuracy was not systematic. The prediction quality can take advantage of marker densification, and using models that are able to do selection of relevant markers or efficient shrinkage. In this thesis, we showed that substantial densification in marker coverage is possible for thousand individuals through imputation from a few sequenced nodal individuals. The densification steps through an imputation process allowed to have a better distribution of the markers along the genome, in different genomic regions, with no increase in the resulting linkage disequilibrium. The advantages of genome-based methods were shown in challenging situations when using an independent validation set or when *degrading* the phenotype, which both could be considered as operationally interesting and realistic. The pedigree-based models were not able to correctly predict the phenotypes of a completely independent set of individuals, unlike to the genome-based models. The *downgrading* of the quality of clonal means used as phenotypes for calibration had a differential effect between pedigree and genome-based predictions, with the latter retaining predicting quality at a level similar to that without replicate reduction.

The densification also highlighted the limitations of the 7K SNPs genotyping set. It seemed to be not well adapted for the implementation of genomic prediction in our breeding programme. Our results showed that 50K SNPs could be sufficient, even if the choice of this 50K SNPs would undoubtedly need improvement, and other than one based exclusively on coverage. It is still necessary to explore where the inflexion point lays between 7K and 50K SNPs, and validate whether the limitations of the lower density chip

are due to heterogeneous coverage or to the particular composition in SNPs.

Another point of interest, considering the way selection is done in the poplar program, is the fact that genome-based models were able to attain a good performance in predicting accurately the ranking in the validation sets. As a reminder, the Spearman correlation of wGBLUP was well above the levels of pedigree-based evaluations for all traits except for rust evaluation. With a species where the genetic gain is disseminated by clones, predicting accurately the ranking of unobserved candidate phenotypes is a valuable feature. Such feature was particularly evident after marker densification from 7K to 50K SNPs and beyond. These appeared to be encouraging results, in that gain could be produced by nailing accurately the ranking and by bringing such accuracy at early stages. It remains, however, the question of the bias in predicting abilities, its causes and the way it could impair performance in genome-based models.

One of the main limitations of the study was probably the use of a training population with a design that did not correspond necessarily to what is routinely done in poplar breeding. Indeed, the factorial mating design, although potentially interesting in terms of the parental variability, was more oriented for mapping studies. This was partially overcome by the addition of extra families and crosses, well connected to the breeding program. In that sense, a training population truly representative of the base population for the breeding program could have made more easily generalizable the results of the study. Another aspect connected to the population and that will be treated later on is the step at which the genomic evaluation was implemented: at a stage of the selection where a big part of the variation has already been screened. In the same way, we used an available chip that was mainly designed for mapping and association studies with some enrichment in traits of interest. This was also partially overcome by the imputation approach.

5.3 How to increase the genomic selection accuracy

5.3.1 Increase and Optimization of the training set size

The size and composition of the training population is known to have a large impact on the accuracy (Grattapaglia and Resende, 2011). The constitution of an international

consortium could increase the size of a reference population for any given species. For instance, an international collaboration for the pooling of reference populations has been set up for Holstein dairy cattle (Eurogenomics Consortium). The constitution of this consortium allowed for a population being 3 to 4 times larger than each of the national populations. The accuracy of genomic evaluations has thus been increased by 10% (Lund et al., 2011). In the same way, several selection programs exist at European level for tree species and for poplar, all with very similar selection criteria. An eventual grouping of improvement populations would certainly bring a panel with greater diversity and representativeness. However, the case of cattle is certainly unique because of their very good pregenomic organization, including a pre-existing system to collect evaluation data in a very centralized way.

Resampling in the existing population is a good way to improve the training population and increase the prediction accuracy. For instance, to include 6 to 8 trees per family and evaluation site appears to be sufficient to guarantee an accurate estimation of genetic parameters for wood density and growth in an open pollinated test of black spruce (Perron et al., 2013). For some species (Cros et al., 2015; Tayeh et al., 2015), CDmeans has given good results in optimizing the training population (Rincent et al., 2012). Our test did not allow to find an advantage for such an optimal procedure, and one of the reasons could be the lack of differentiation within the population to derive truly different training sets. The optimal procedure could also be tried with a denser SNP set, like the 50K. Another strategy to optimize the training step would be to integrate existing information in the pedigree and from genetic association studies in the way proposed by Cericola et al. (2017).

5.3.2 New ways of modelling phenotypes

With the rapid expansion of genomic selection, new methods and models have emerged. Among them, the integration of major QTLs as fixed effects in the model showed good results in plants (Moore et al., 2017). The weighted GBLUP showed also a good performance when integrating directly SNPs effects of a major QTL (Teissier et al., 2018), or even with no a priori information other than what is implicitly derived from the iterative process. Other methods go beyond the traditional modelling by, for instance,

taking into account the results of inferred gene-based networks to weight relevant SNPs (Riedelsheimer and Melchinger, 2013; Westhues et al., 2017).

Some low-cost methods have recently been proposed with encouraging results, like the "*phenomic selection*" in wheat and poplar (Rincent et al., 2018). This approach uses NIRS (Near Infrared Spectroscopy) to predict different traits by using a full range of different spectra as markers instead of genotypes to compute relationship matrix. Results suggest that *phenomic selection* could be competitive compared to genomic evaluation, with the possibility also of using it at very early ages with good accuracy.

Multi-trait and multi-environment evaluations are essential in plant and tree breeding programs, although performing single-step analyses in these circumstances could be methodologically and computationally challenging. In that sense, Montesinos-Lopez et al. (2018) have proposed efficient heuristic methods based on multi-trait deep learning (MTDL), which appear to be well adapted when data is highly unbalanced, contain missing values data and there is a need for accommodating different design factors.

5.4 Proposition of a genomic breeding scheme

5.4.1 Actual breeding scheme

The present proof-of-concept study fits at a particular step in the poplar breeding program, as illustrated in (Figure 5.1), specifically when evaluating selected candidates on juvenile traits in the nursery. The current selection scheme was the result of optimizing for many constraints derived from the phenotypic evaluation and operational factors over the years. It comprises several steps of selection conducted at the greenhouse, at the nursery, in the laboratory via *in vitro* tests and later in field trials, with each step implying different selection intensities and notably different selection accuracies. It is important to note that each selection step is done sequentially and conditionally onto the precedent (i.e. independent culling levels), instead of jointly and simultaneously, leading to inefficiencies with the risk of losing in the first steps important variation for subsequent steps. First steps of selection at the greenhouse and nursery are the less accurate, but the ones that screen most of the variation. Conversely, later steps at the lab and in the fields are relatively accurate but screen through a subsample of original variation. Therefore

accuracy and genetic variation do not meet in a single same step for maximum efficiency in the current scheme.

Our test of genomic selection was performed with moderate to high heritability traits, well evaluated in field trials, and on a relatively reduced set of individuals that were the result of two previous steps of selection conducted typically with a low precision and at a relatively high selection intensity (see Figure 5.1, with the red circle indicating where genomic evaluation was tested). Therefore, there is room for improvement in the way genomic evaluation is integrated in the scheme, *there* where extra precision is specially required (see Figure 5.2 with new proposition at earlier stages). In that sense, our test was placed on a step that was not particularly favourable for the genomic evaluation. There would be, however, in the new schemes proposed in Figure 5.2 the challenge of training a relevant genomic evaluation model in the diversity rich context of the first steps of the breeding program.

5.4.2 What is the genomic selection added value?

By looking at each of the parameters of the classic breeder's equation, $G = ir\sigma_A/L$, we can review briefly where genomic selection brings potentially an added value to the genetic improvement:

- The first parameter (i) is the selection intensity, representing the proportion of individuals in the candidate population that are selected. GS can have a direct impact on i by selecting at a seedling stage a greater number of candidates than those phenotyped, making it particularly interesting for costly phenotyping and for late maturation traits.

- The parameter (r) is the accuracy of selection. Well designed genome-based models allow to increase r by capturing extra genetic variance relevant for the target traits, either for single traits taken independently or for traits evaluated simultaneously and genetically correlated. Multi-trait evaluation can help the prediction at compensating too many missing values in different traits and poor heritabilities (Calus and Veerkamp, 2011; Jia and Jannink, 2012) and can reduce prediction bias (Kadarmideen et al., 2003). The implementation of non-additive

effects, as dominance and epistasis, can also boost selection accuracy, notably when selecting mates is relevant for breeding (Toro and Varona, 2010), or for clonal dissemination of particular genotypic combinations.

- The third parameter is the genetic variation (σ_A), usually not modulable by the breeder in an advanced breeding program. However, the construction of the training population and the decision of the place where genomic evaluation takes place in the improvement scheme would have a real impact on the amount of genetic variation that are made available for selection.

- The last parameter is the generation interval (L), and one on which genomic evaluation could have a large impact by radically reducing the time between two generations (Heffner et al., 2009). Candidates can be selected at very early ages, providing that they have enough biomass for yielding DNA. In poplars, this could happen at the stage of seedlings with few weeks of growth in the greenhouse.

5.4.3 How implementing genomic evaluation?

GS can contribute to accelerate genetic gain by increasing the individual selection accuracy at early stages, thus shortening the generation interval, and by increasing the selection intensity. We propose to implement GS sooner in the cycle, at the seedling stage, than what was assessed in this thesis.

Short term horizon

In the short term, a genomic selection scheme at the seedling stage (Figure 5.2A), when there is a great number of individuals taking up the least space, would be of great benefit to the breeding program. Such an early scheme combined to a multi-trait approach with a selection index can increase the genetic gain in the short term for most traits simultaneously, even for those phenotyped at maturity like wood properties. For now, only the *P. nigra* parents could be selected with such early genome-based approach, and in order to identify the best black poplar parents at the same year as the controlled-crosses to produce both pure species descendants and hybrids with other species. Time-consuming and resource-intensive evaluations could then take place only

on those genomically preselected parents, with the possibility to enlarge the panel of preselections.

Long term horizon

In the longer term (Figure 5.2B), GS can be implemented in the other parental species, *P. deltoides*, and even at the hybrid progeny, depending on the breeding strategy for hybrids. In this case, in addition to the step at the nursery evaluation, new steps at the laboratory can focus on other targeted traits, like interaction genotype × rust strain and woolly aphid resistance for hybrids, increasing the accuracy of prediction for costly traits related to resistance. Such proposition could save eventually from 5 up to 9 years in the breeding program. One of the evaluations for which time gains are expected is that related to wood quality, with the interesting possibility of predicting potential uses at the individual level according to the wood properties.

However, there are limits to the rapid advancements of the cycle. One is a regulatory constraint, another is biology. Even if accurate genomic evaluation is available at very early stages, the release of varieties will require under current regulations evaluations under production conditions in several environments, which usually takes 10 years. Biological constraints are related to the sexual maturity. Indeed, if we want to use a selected individual from a parental species for hybridization, it is necessary to wait until sexual maturity at around seven years of age.

5.4.4 What is missing for this?

The first point in the implementation of genomic selection in *P. deltoides* or in the hybrids is the construction of adapted genotyping resources. At the moment only *P. trichocarpa* and *P. nigra* have some genotyping tools at their disposal. The construction of a chip for *P. nigra* is not necessarily the best cost-wise solution, even if high quality genotyping is available. This solution is still expensive requiring usually the engagement of several thousands of individuals to be genotyped. When starting from scratch, the genotyping-by-sequencing (GBS) approach can be a good option if it is combined with efficient imputation and an optimization of the restriction enzymes. The quality of genomic predictions from this genomic resource could be similar to those typically from chip-arrays (Elbasyoni et al.,

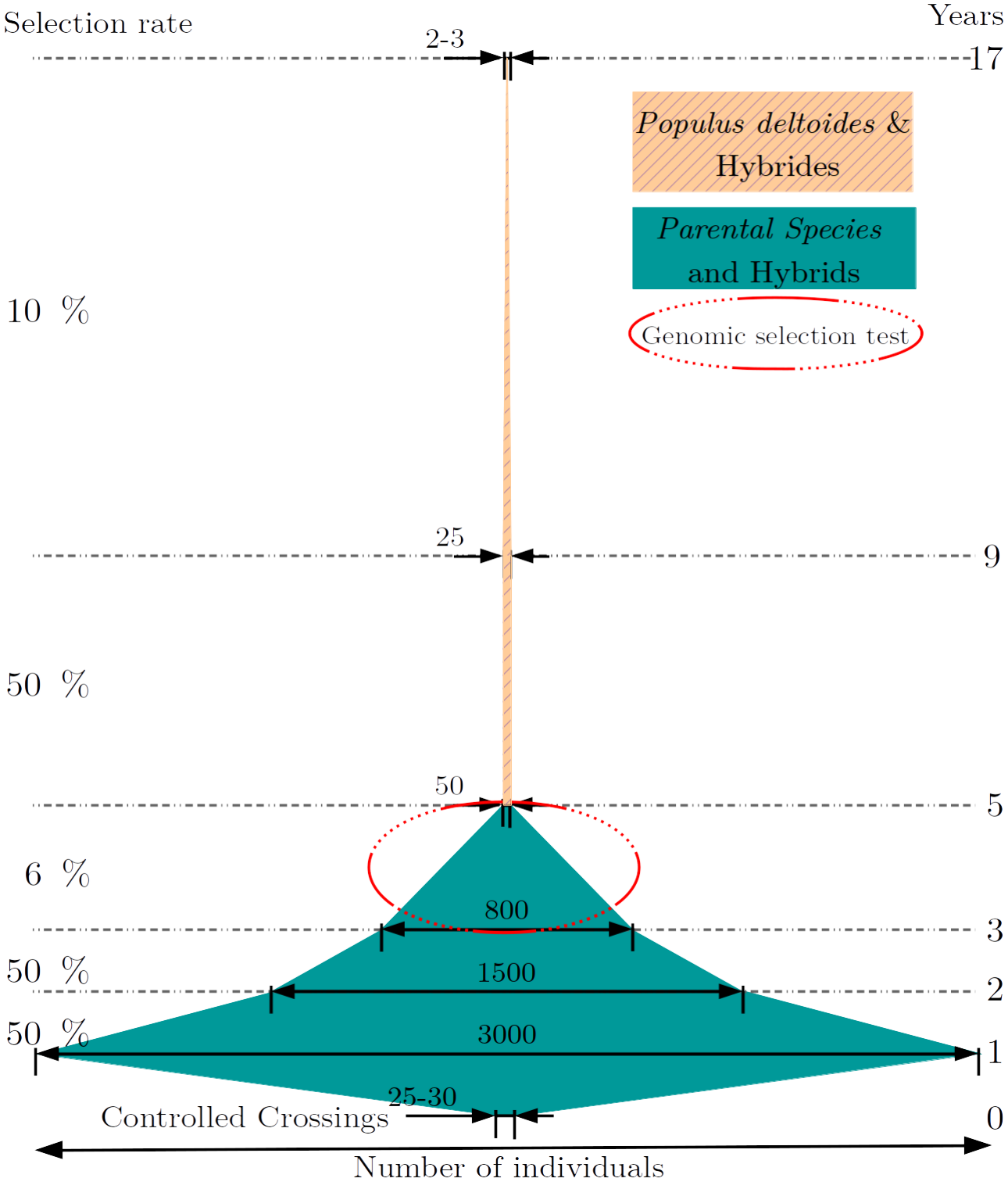


Figure 5.1: Evolution of the number of individuals and the selection rate during the different steps of selection after crossings (year 0). Numbers correspond to one cycle of selection. Selection rate values correspond to a rate relative to the previous step. The genome-based evaluation test was represented in the red circle.

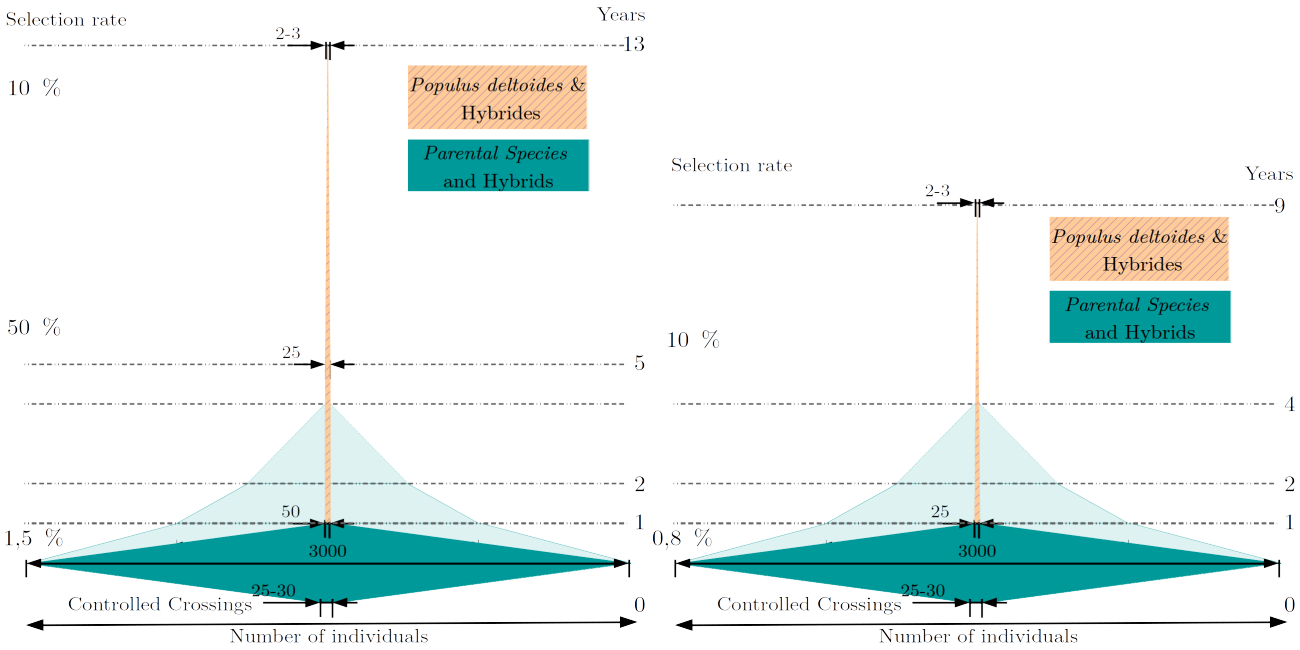


Figure 5.2: Representation of the number of individuals and the different levels of selection depending on the timeline of the implementation of genomic selection in Short term horizon (left) and Long term horizon (right).

2018). A cost-efficient alternative could be the use of a multi-species chip for all populus species or for a combination of economically important species including populus in an international consortium. In that sense, the parental species and their hybrids could be genotyped with the same tool, and the fact of including a large portfolio of species in the tool would increase the portfolio of clients, reducing genotyping cost. A complete implementation of genomic selection in the poplar breeding program would need also new investments in phenotyping. Selection at the seedling stage requires to guarantee ability to root and to resprout of most candidates, and the knowledge whether this traits have or not any antagonism with traits of interest. Similar requirements exist for the sexual determinism, which is not known at early stages. Therefore, predicting at an early age the sex of individuals would allow maintaining even sex proportions in the candidate population, to select parents in a cross and to maintain genetic diversity in the breeding population.

There is also a lack of fast and cost effective phenotypic evaluations. Laboratory tests are time consuming and expensive. For example, the woolly aphid resistance test on hybrids takes 85 days to evaluate ten genotypes. Similarly, wood quality can only be assessed after ten years of growth. For wood quality evaluation at an early stage, the use of NIRS (Near Infra-Red Spectroscopy) can be a feasible alternative. A recent study showed that NIRS prediction of wood quality and wood chemical contents is feasible with reasonable accuracy from 2 years old trees (Gebreselassie et al., 2017). If we want to combine this approach with a genomic selection at the seedling stage, it will be necessary to build a NIRS calibration model from leaves instead of wood and use it extensively in the training population. Another limitation is the ability to predict hybrid performance with a genome-based approach, which so far has not been done in poplar. Examples from eucalyptus show that genomic evaluation of hybrids have little advantage over classical evaluation schemes (Tan et al., 2017).

5.5 Genomic selection impacts

5.5.1 Impacts on costs

Genomic selection represents a significant investment. In the best case, for instance, very competitive genotyping could be obtained at 14€ per individual for a 12K SNP chip and with a minimum use of 8 000 individuals (H2020 project B4EST). Such conditions could well be unattainable for many breeding programs in forestry, requiring international consortia to attract providers. Gorjanc et al. (2017) showed that the effective cost of genomic selection can be optimized by using imputation methods and by increasing the intensity of selection leding to a fast return on investment. However, implementing a cost-effective genomic selection scheme could often require a whole rethinking of the breeding program, as proposed already in this discussion, requiring probably additional investments to those of genotyping. In any case, it is difficult to believe that cost-effective implementation is possible into an already operationally optimized program.

We demonstrated that genome-based evaluation could be efficient with a large enough marker set, even if phenotypes were obtained less accurately. From an operational point of view, an evaluation simplification reducing the number of replicates and simplifying the design could be used to save monetary resources for genotyping or to invest in increasing the number of evaluated individuals. Genotyping could be done cost-effectively by using extensively imputation. Our results showed that imputation could represent an excellent strategy to reduce genotyping costs. We used few well-chosen sequenced individuals in the population, resulting in excellent imputation quality and considerably increasing the number of SNPs available. This would require the availability of a well designed low-density chip and a permanent monitoring and optimization to choose new individuals for sequencing to support future imputation. Indeed, the fact of not updating the reference population of sequences could reduce the quality of imputation and thus result in accumulating errors over subsequent generations. After the first investment, the costs in subsequent years would represent the maintenance of this training population in terms of new accessions or the phenotyping of new traits. The information accumulation is expected to lead to increasing accuracies along the selection cycles, with a potential return of investment.

However, current forestry sector is not necessarily ready to absorb a rapid renewal of varieties and a large number of them. Poplar wood producers have their favorite cultivars and they are known to secure their investments by keeping their habits. Out of the 40 poplar varieties available in the commercial catalogue, roughly 10 are used to some extent although with great differences. Something similar existed for the dairy cattle before the arrival of genomic selection, with a system of "*star*" bulls that heavily affected the acceptance of new accessions. Changements could be initiated by the peeling industry, which could be interested in cultivars with particular wood characteristics. In the long term, genomic selection could represent the way to screen and to provide a choice of varieties with customized characteristics according to production and industrial objectives. If the forest sector is interested in customized cultivars, part of the research to produce them could be financed by a tax on producers. This kind of tax exists already in France in other sectors, as fruit and vegetable production, although it does not exist for the French Forest sector which relies entirely on publicly funded breeding programs.

In a longer-term perspective, with increasing protective environment policies and the objective of reducing plastics, poplar wood has a promising future as a biodegradable packaging material. Moreover, poplar production does not take the place of arable land for food, as it is often produced in seasonally flooded areas that are of no use for conventional agriculture. Also, poplar production does not require the addition of inputs to ensure proper production and only requires weeding in the early years. A negative impact of poplar production that is controversial depending on author sources is the genetic pollution of the wild compartment by the cultivated compartment. While some studies showed that pollution to be negligible, others estimated it to be around 13% (Pospíšková and Šálková, 2006; Rathmacher et al., 2010). In any case, this risk could be somehow minimized by increasing the portfolio of varieties thanks to genomic selection, partially avoided by cultivars with floral phenologies being extreme or different to those in the wild, or completely avoided by producing sterile hybrids.

5.5.2 Impacts on genetic diversity and how to turn it into benefit

Recent research in genomic selection has been mostly focused on factors affecting prediction accuracies. Comparatively, little was done on the impacts that this extra accuracy might have on genetic diversity at whole genome scale, and through the accelerated cycles of selection propitiated by the approach. Genomic selection could accelerate the development of inbreeding per year in the population, as a consequence of the shortening of the generation interval (Daetwyler et al., 2007). The study of the impacts of several genomic selection cycles on the genetic diversity is therefore essential. If the implementation of genomic evaluation is going to produce a rethinking of the breeding strategy, it is essential to incorporate the management of the genetic diversity to the improvement of genetic gain.

Indeed, our study population consisted on a series of multiple-pair matings, with advantages for evaluating parental performance and also non-additive components in the genetic variation. Such a system is also useful for mate selection, or the choice of the best performing crosses among all available between selected candidates. It would be interesting to evaluate the impact of these mating choices in terms of genetic diversity, by the use of algorithms to optimize the genetic contribution of parents in the next generation or to select mates. Solutions do exist, mostly from the pedigree-based era, that optimize breeding for maximum gain over the long-term, like optimum contribution selection (OCS, Meuwissen and Goddard (1997); Howard et al. (2018) or mate selection, Toro and Varona (2010)). However, these existing solutions rely on the setting of constraints on average information, like marker-based coancestry. Thus, the information across markers or genomic regions is neglected when devising the diversity constraints. Although some recent solutions go beyond average-based constraints and account for marker variation (Sun et al., 2013; Gómez-Romano et al., 2016), they do not correspond to an integrated selection and mating system that can fit conveniently the breeding scheme for all species. There is therefore a need to devise a genomic selection strategy that is able to maximize Mendelian segregation at the phase of mate allocation, by accounting for parental information across genomes. In poplar, where dissemination

is by clonal selection, the mate choice is essential together with a maximum exploitation of the Mendelian segregation.

Unlike most field crops, forest trees including poplars have breeding populations that are still very close and genetically linked to wild populations. The incorporation of new accessions if required by optimal algorithms would not require a complex system of backcrosses to recover the previous genetic gain. There is currently no protocol for storing poplar seeds, which puts a higher emphasis on the realization of controlled crosses. These are easily managed as already stated in Chapter 2. Currently, our results indicated that predictions of unobserved crosses are not accurate, limiting somehow the use of a mate selection system. Further research would be required with new models and denser genotyping to validate a mate selection strategy as feasible. A single controlled cross of poplar can easily provide several thousands of seeds, which allows for efficient exploitation of Mendelian segregation, a priori only available to genome-based evaluations. Such an use and management of the Mendelian sampling term is also the key to long-term gain with little impact on inbreeding (Avendaño et al., 2004).

The management of the genetic diversity is not simple, even by proposing unrelated and diversified cultivars, producers are still able to plant only one or two per plot. This has been shown already as risky, mostly due to the facilitation of pest dispersal in case of outbreak, with all the economic consequences of losses in productivity. The "*star*" system in cattle was changed by genomic selection, with the proposal of a wider portfolio of young elite bulls that changes often. There are different ways to compel producers to use a diversified portfolio of cultivars. At the political level, the legislation can force to plant different cultivars at the landscape level. *A priori*, GS allows for a widening of the set of selections without the need of extra phenotypic evaluation, with the possibility to select different profiles of resistances across cultivars without loss of performances for the rest of the traits. Such a coctel of cultivars could be proposed as mixed clones varieties (MCV). A highly performing MCV can more easily attract producers, while assuring high levels of diversity at the production and landscape levels.

Finally, the use of genomic information not only allows for genomic evaluation, as this thesis has shown, but also opens the door to a level of diversity management that was not at hand before, by accessing to the level of variation across the genomes. Such a double

use of genomics is probably the key to short and long-term breeding in poplar.

Final Conclusion

The results indicate that genomic selection (GS) implementation is feasible in the black poplar breeding program and should provide the opportunity to select multiple traits at a seedling stage. In doing so, a decrease in the average generation interval and an increase in the intensity of breeding, resulting in a genetic gain beyond that of the traditional breeding method. The use of imputation as a method of optimizing genotyping effort to reduce the cost of implementing GS has proven its effectiveness, significantly increasing the number of markers. However, further studies are required to complete the results obtained in this thesis before the practical application of GS.

The main objective is now to complete phenotyping in order to test the prediction of vegetative propagation ability and recovery from the seedling stage as well as the sex of the individual. This will allow for a larger study confirming the ability of GS to predict large numbers of individuals at an early age. In the perspective of the implementation of SG, the feasibility study of the complementary parent (*P. deltoides*) and the prediction of the hybrid value should be considered. From a methodological point of view, future research should use the new models that have been developed because they are potentially more effective, particularly through the consideration of other non-additive effects or the integration of transcriptomic information. Secondly, the construction of models for the implementation of genetic diversity management in poplar genomic selection is necessary to ensure that genetic gain is maintained in the long term. Finally, more ambitious approaches should be considered, such as the use of SG models combining the results of crossover tests performed in different environments, molecular information and environmental variables. This would make it possible to predict the interactions between Genotypes $\times$ Environment.

In conclusion, the study shows that, in the current context where agricultural

production must increase at an unprecedented rate to meet demand while ensuring the maintenance of the genetic diversity necessary to cope with health risks, SG undeniably has a role to play in genetic improvement in general and in poplar in particular.

Conclusion générale

Les résultats indiquent que l'implémentation de sélection génomique (SG) est faisable dans le programme d'amélioration du peuplier noir et devrait apporter la possibilité de sélectionner plusieurs caractères à un stade plantule. Ce faisant, une diminution de l'intervalle moyen de génération et un accroissement de l'intensité de sélection, aboutissant à un gain génétique dépassant celui de la méthode de sélection traditionnelle. L'utilisation de l'imputation comme méthode d'optimisation de l'effort de génotypage dans le but de réduire le coût de mise en place de la SG a montré son efficacité, permettant de multiplier considérablement le nombre de marqueurs. Cependant, des études supplémentaires sont requises pour compléter les résultats obtenus dans cette thèse avant l'application pratique de la SG. Le principal objectif étant désormais de compléter le phénotypage afin de tester la prédiction de l'habileté à la multiplication végétative et la reprise dès le stade plantule ainsi que le sex de l'individu. Cela permettra de faire une étude plus grande ampleur confirmant la capacité de la SG de prédire un grand nombre d'individus au plus jeune âge. Dans la perspective de la mise en œuvre de la SG, l'étude de la faisabilité chez le parent complémentaire (*P. deltoides*) et la prédiction de la valeur hybrides sont à considérer. Du point de vue méthodologique, les futures recherches devraient utiliser les nouveaux modèles qui ont été développés car ceux-ci sont potentiellement plus efficaces, notamment grâce à la prise en compte d'autres effets non additifs ou d'intégration informations de transcriptomiques. Dans un second temps, la construction de modèles permettant la mise en place d'un gestion de la diversité génétique dans le cadre de la sélection génomique du peuplier est nécessaire pour garantir le maintien du gain génétique à long terme. Dans un dernier temps, des approches plus ambitieuses devraient être envisagées, comme l'utilisation de modèles de SG combinant les résultats de tests en croisements réalisés dans différents environnements, des informations moléculaires et des variables

environnementales. Ceci permettrait de prédire les de prédire les interactions Génotypes x Environnement. Enfin, il ressort de cette étude que, dans le contexte actuel où la production agricole doit augmenter à un rythme jamais atteint pour faire face à la demande tout en garantissant le maintien d'une diversité génétique nécessaire pour faire face aux risques sanitaires, la SG a indéniablement un rôle à jouer pour l'amélioration génétique en générale et tout particulièrement pour le peuplier.

Appendix

Appendix A

Chapitre 2 : Estimated micro-environnemental effects

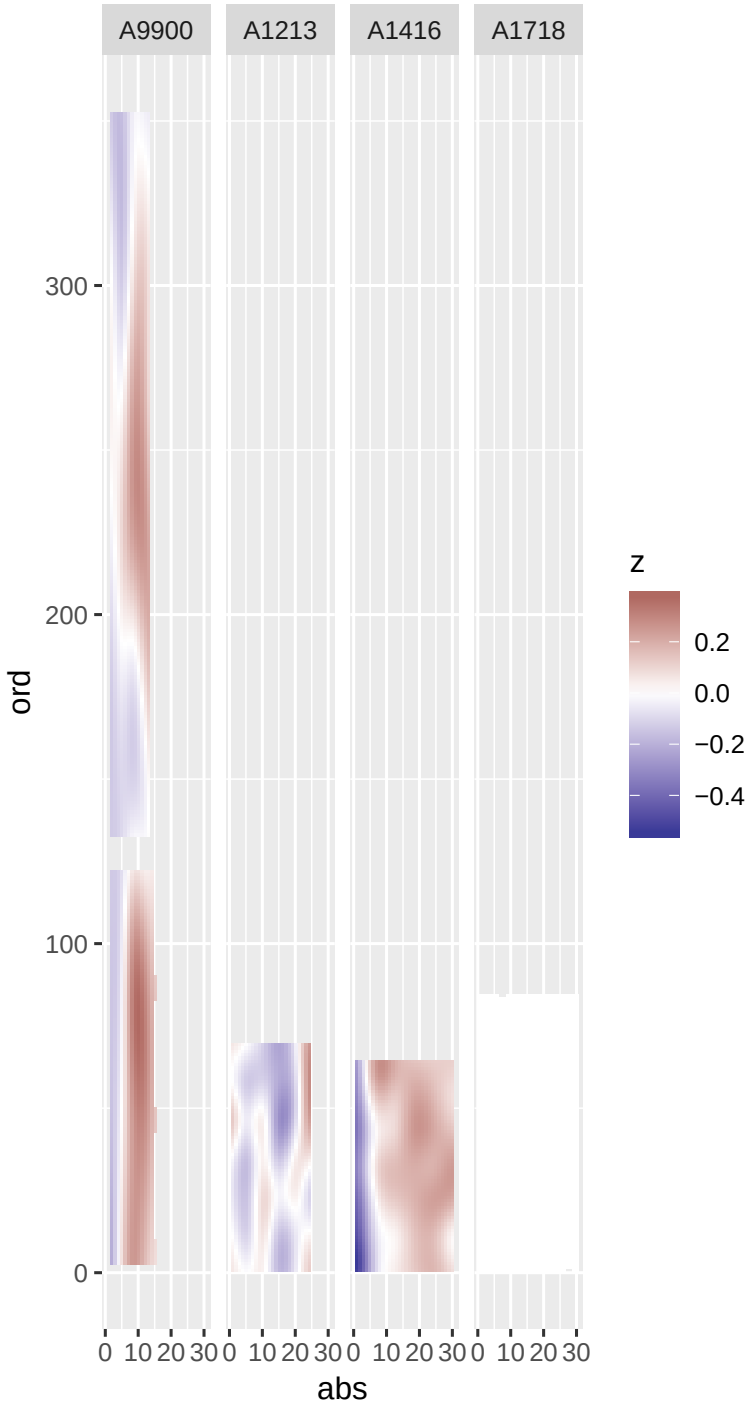


Figure A.1: Micro-environmental effects estimated with BreedR (Muñoz and Sanchez, 2018) for the Branch angle and by evaluation batch (columns), with a H matrix (Legarra et al., 2009). The effect's magnitude was represented in color: blue: the environment tend to gives lower values and red : environment tend to gives higher values.

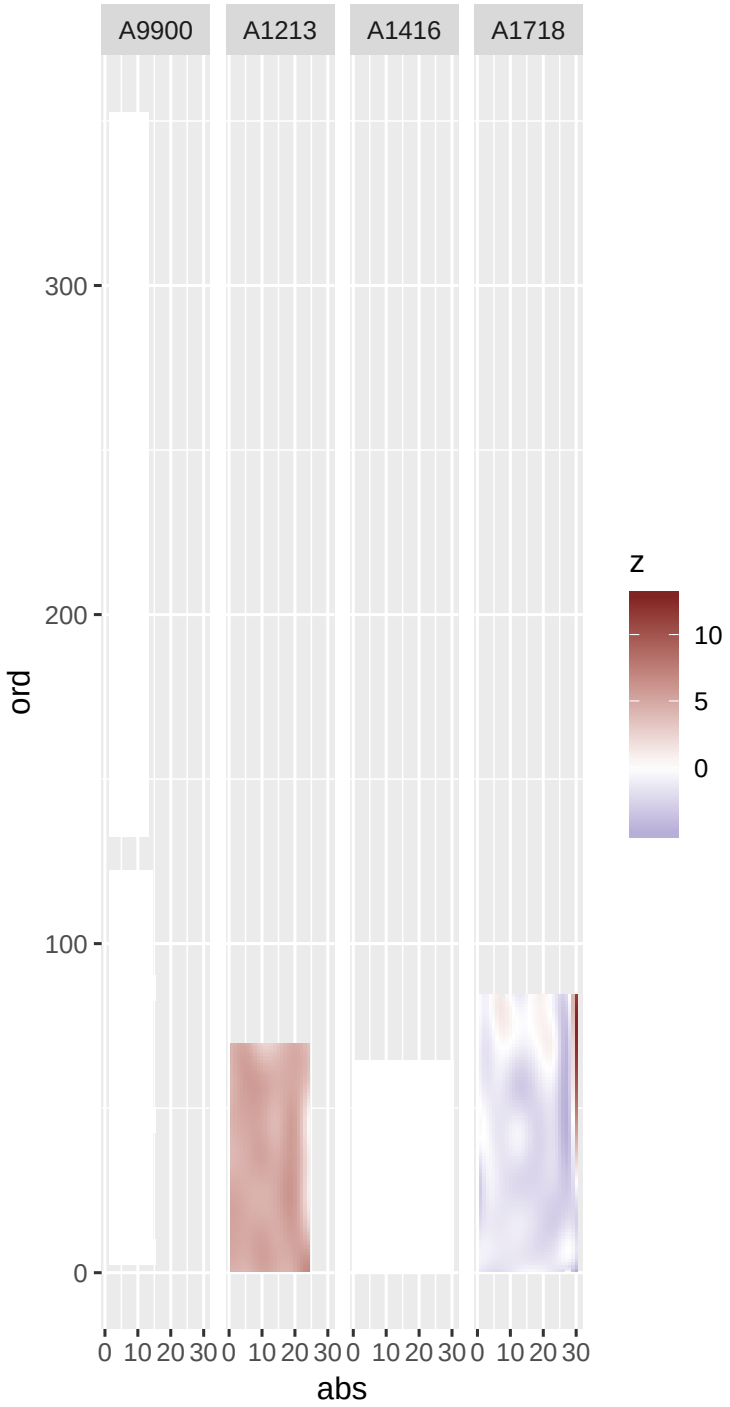


Figure A.2: Micro-environmental effects estimated with BreedR (Muñoz and Sanchez, 2018) for the budburst and by evaluation batch (columns), with a H matrix (Legarra et al., 2009). The effect's magnitude was represented in color: blue: the environment tend to gives lower values and red : environment tend to gives higher values.

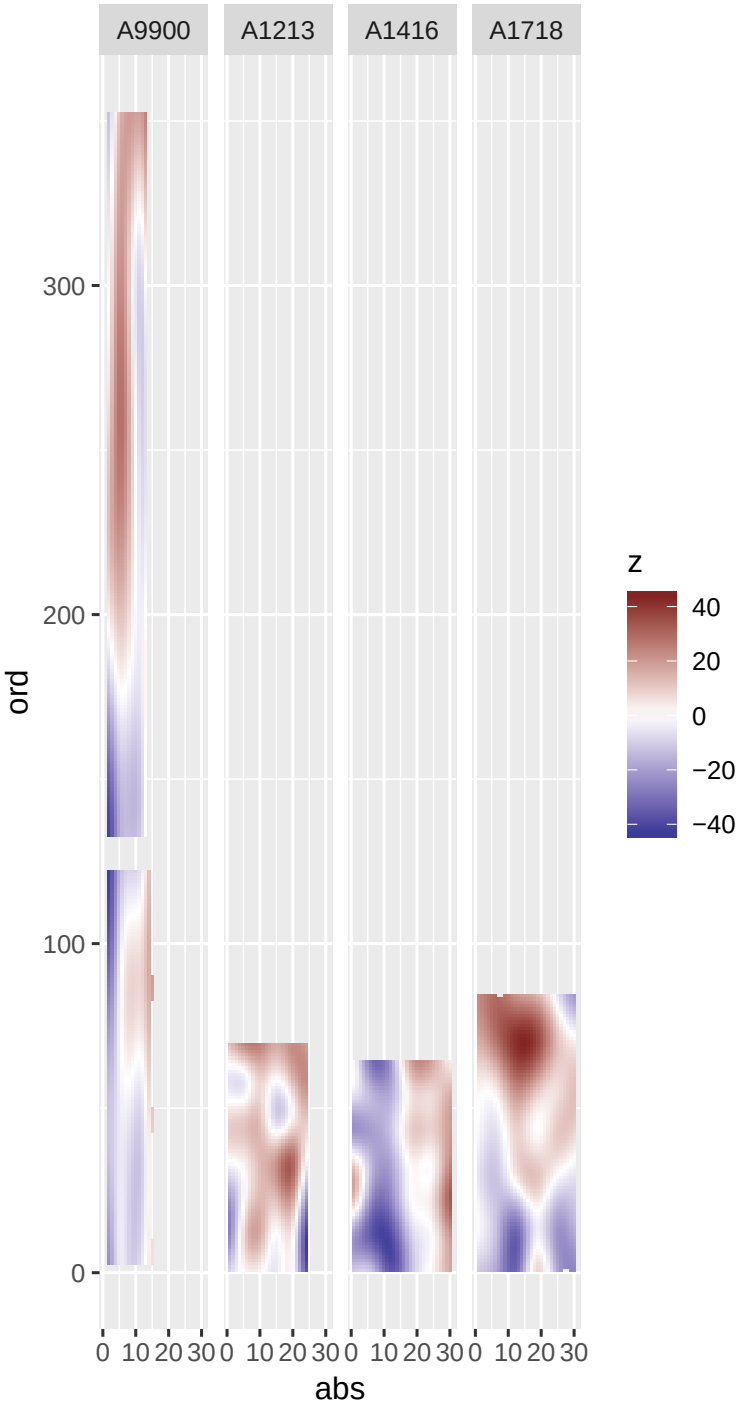


Figure A.3: Micro-environmental effects estimated with BreedR (Muñoz and Sanchez, 2018) for the height at one year old and by evaluation batch (columns), with a H matrix (Legarra et al., 2009). The effect's magnitude was represented in color: blue: the environment tend to gives lower values and red : environment tend to gives higher values.

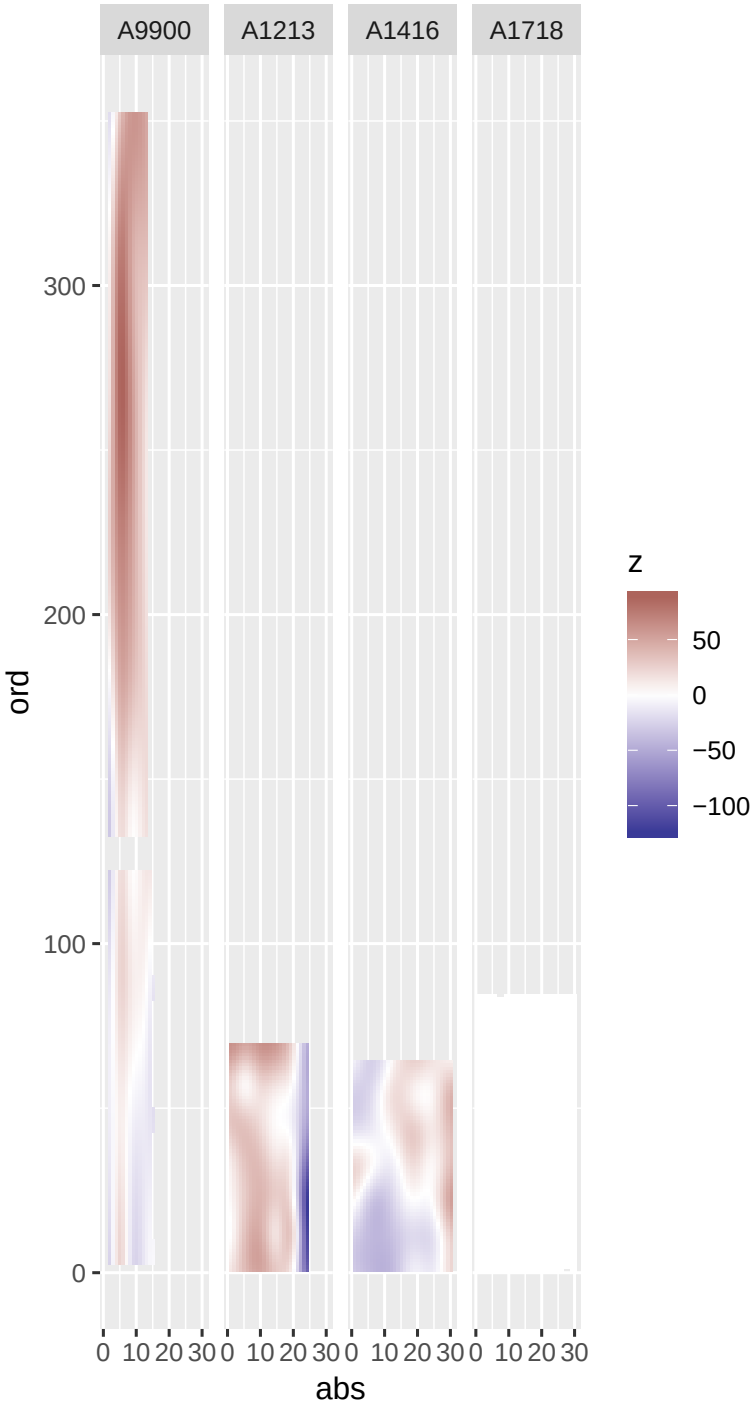


Figure A.4: Micro-environmental effects estimated with BreedR (Muñoz and Sanchez, 2018) for the height at two year old and by evaluation batch (columns), with a H matrix (Legarra et al., 2009). The effect's magnitude was represented in color: blue: the environment tend to gives lower values and red : environment tend to gives higher values.

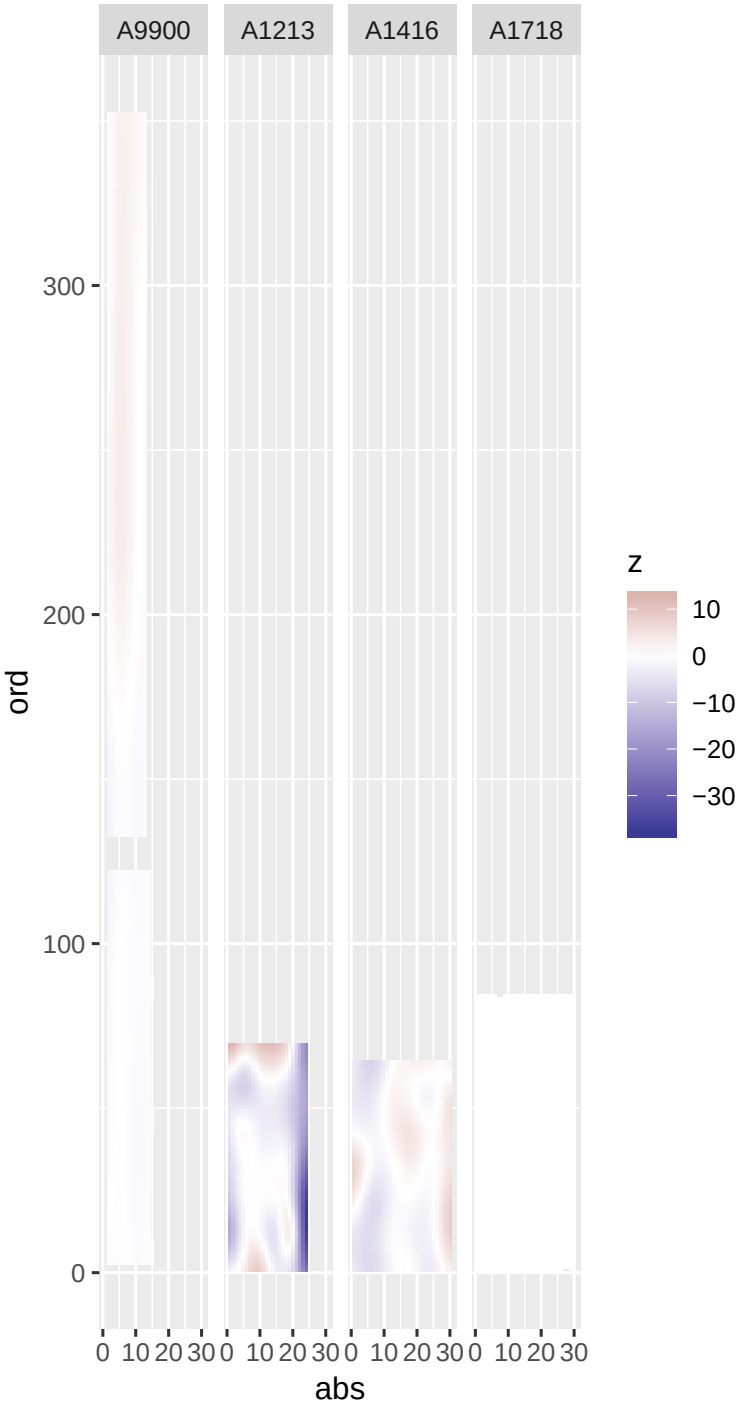


Figure A.5: Micro-environmental effects estimated with BreedR (Muñoz and Sanchez, 2018) for the circumference at two year old and by evaluation batch (columns), with a H matrix (Legarra et al., 2009). The effect's magnitude was represented in color: blue: the environment tend to gives lower values and red : environment tend to gives higher values.

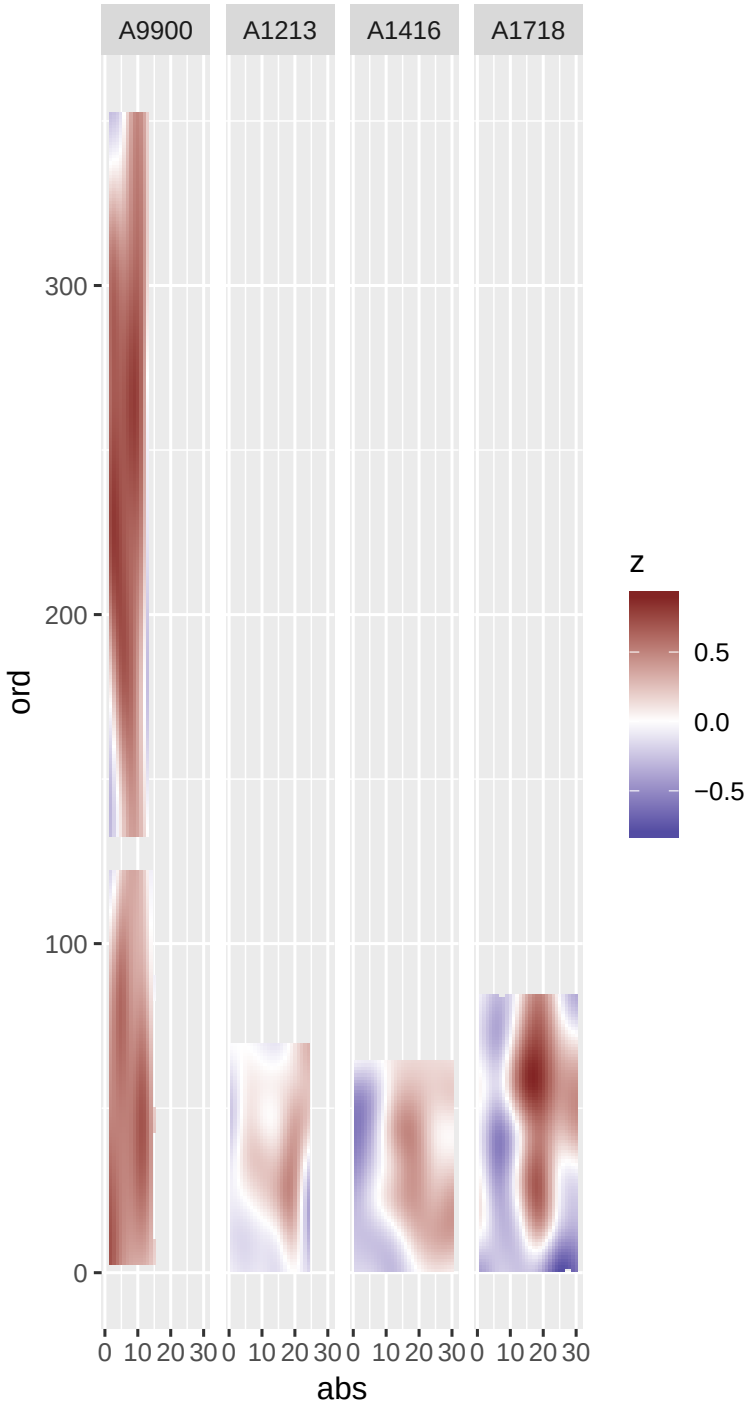


Figure A.6: Micro-environmental effects estimated with BreedR (Muñoz and Sanchez, 2018) for the rust resistance at one year old and by evaluation batch (columns), with a H matrix (Legarra et al., 2009). The effect's magnitude was represented in color: blue: the environment tend to gives lower values and red : environment tend to gives higher values.

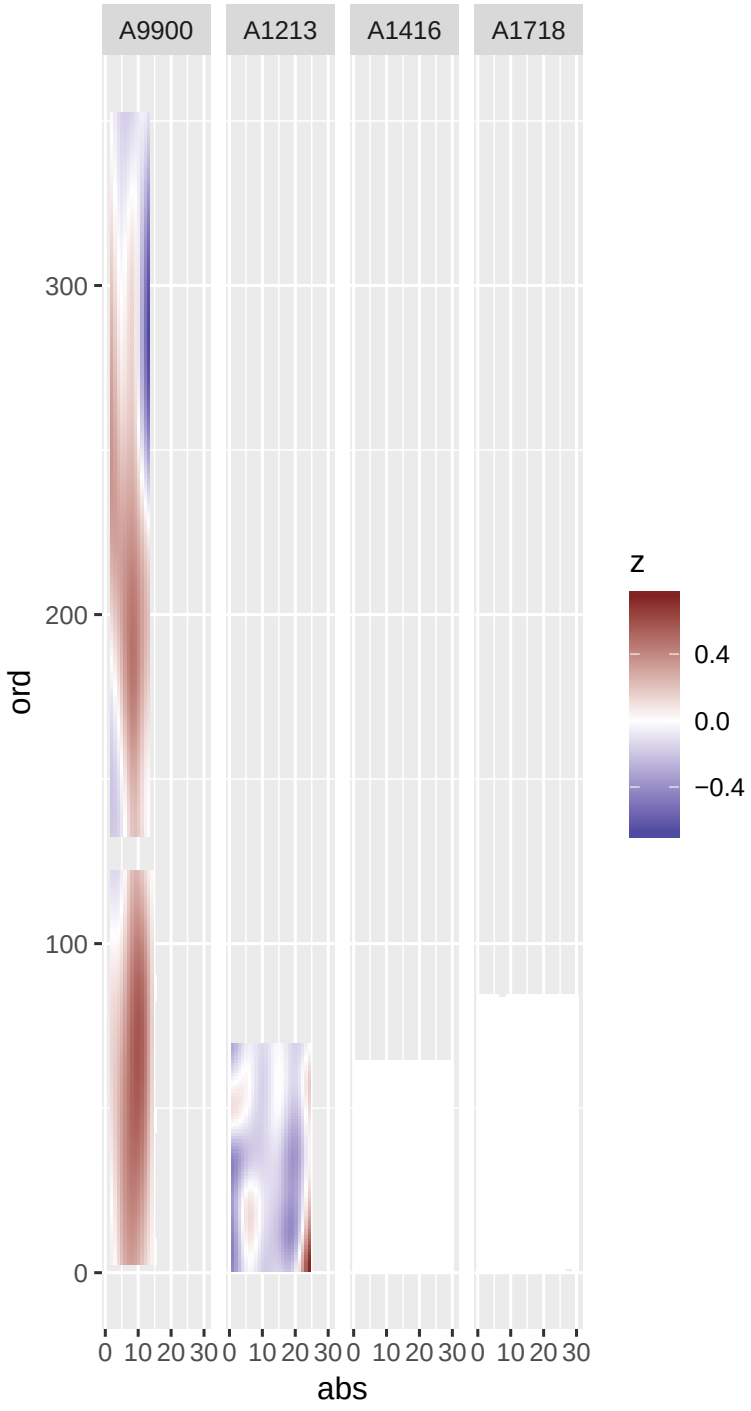


Figure A.7: Micro-environmental effects estimated with BreedR (Muñoz and Sanchez, 2018) for the rust resistance at two year old and by evaluation batch (columns), with a H matrix (Legarra et al., 2009). The effect's magnitude was represented in color: blue: the environment tend to gives lower values and red : environment tend to gives higher values.

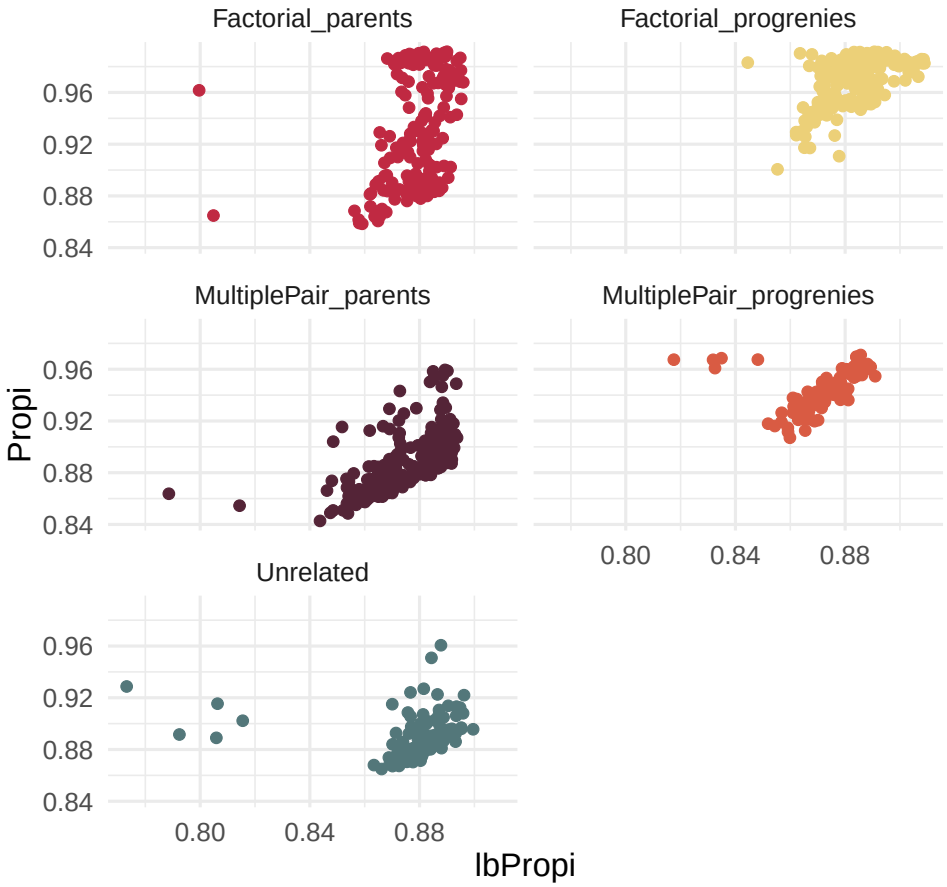
Appendix B

Chapitre 3 : Article Supplementary material

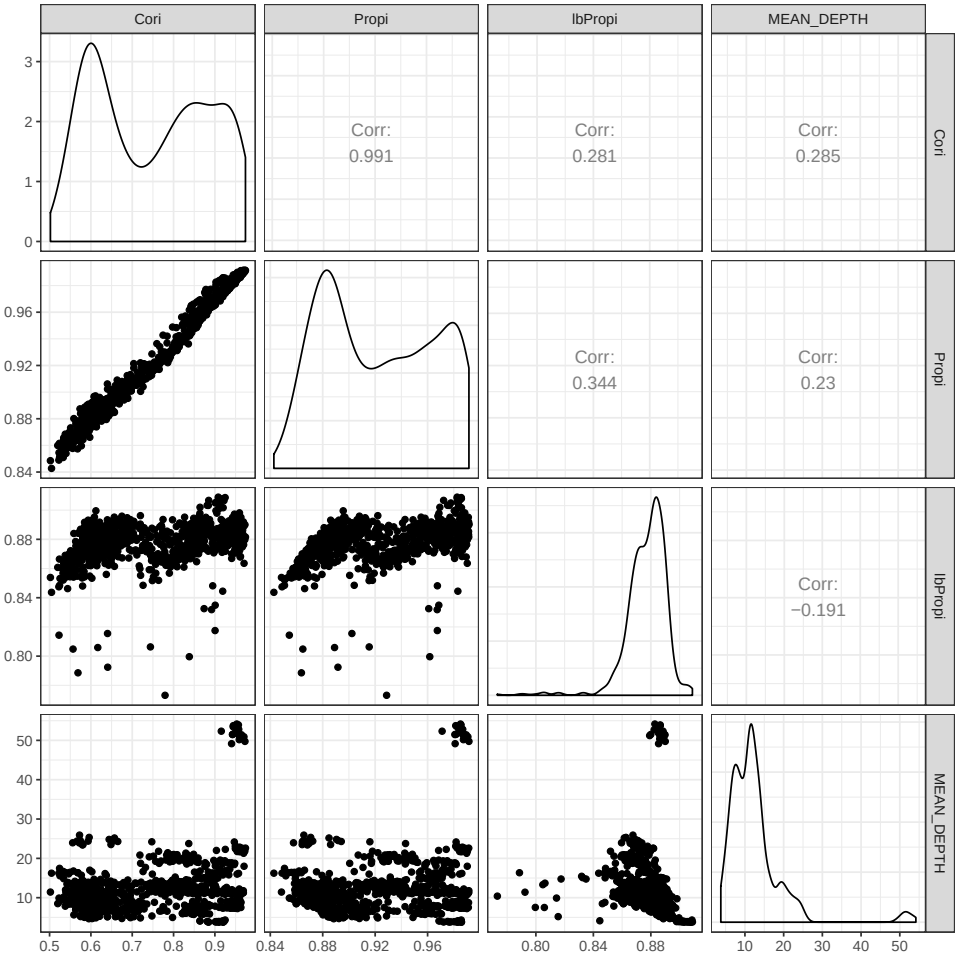
Additional file 1 : TableS1.pdf — Sequencing, pedigree and reference information's of each reference individuals.

Name	Grand-parent	Parent	Progeny	Unrelated	MEAN_DEPTH	Class	Reference Genbank SRA Run number or Bioproject number	Origin
73182-009		yes			5.21	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Pyrénées-Atlantiques, Licq-Arthérey
73193-091		yes			6.53	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Hautes Pyrénées, Soulon
71034-2-406		yes			6.84	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Savoie
73193-089		yes			7.16	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Hautes Pyrénées, Soulon
72131-017		yes			8.79	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Puy de Dôme, Coudes
73193-056		yes			8.83	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Hautes Pyrénées, Soulon
72146-11		yes			8.85	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Gard, Dions
H480		yes			9.66	MultiplePair_parents	BioProject BreedToLast PRJNA483561	NA
3824-3		yes			11.37	MultiplePair_parents	BioProject BreedToLast PRJNA483561	NA
72131-036		yes			12.08	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Puy de Dôme, Coudes
72145-007		yes			16.36	MultiplePair_parents	SRR3045889	Gard, Brignon
71069-914		yes			24.33	MultiplePair_parents	BioProject BreedToLast PRJNA483561	Haute-Savoie, Seyssel
GIS89-052			yes		11.40	MultiplePair_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
GIS68-098			yes		11.72	MultiplePair_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
GIS70-046			yes		12.78	MultiplePair_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
GIS60-076			yes		14.16	MultiplePair_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
GIS84-001			yes		14.95	MultiplePair_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
GIS88-018			yes		20.73	MultiplePair_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
71041-3-402		yes			6.75	Factorial_parents	BioProject BreedToLast PRJNA483561	Savoie, Saint-Jean-de-Maurienne
71077-308		yes			7.84	Factorial_parents	SRR3045875 + SRR3045876 + SRR3045878 + SRR3045880	Ain, Meximieux
SAN-GIORGIO	yes				10.63	Factorial_parents	BioProject BreedToLast PRJNA483561	Ornemental, origin unknown
SSC (Selle sur cher)		yes			11.39	Factorial_parents	BioProject BreedToLast PRJNA483561	Loir-et-Cher, Selles-sur-Cher
71072-501		yes			12.94	Factorial_parents	SRR3045961 + SRR3045962	Savoie, Saint-Genix-sur-Giers
SRZ (Sarazin)		yes			13.56	Factorial_parents	SRR3045895	Lot, Cahors
92510-1		yes			13.58	Factorial_parents	BioProject BreedToLast PRJNA483561	Nièvre, La-Celle-sur-Loire
VGN-CZB25 (Vert de Garonne – Cazebon 25)		yes			22.40	Factorial_parents	SRR3045896 + SRR3045902	NA
BDG (Blanc de Garonne)		yes			51.79	Factorial_parents	SRR3045881 + SRR3045882 + SRR3045883 + SRR3045885 + SRR3045886 + SRR3045887	NA
662200216		yes	yes		3.94	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662200357			yes		6.05	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662200037		yes	yes		7.54	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662200039			yes		9.04	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662200426			yes		11.37	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662201338			yes		11.37	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662200057			yes		11.54	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662200098			yes		16.57	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662200495			yes		18.87	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
662200654			yes		19.33	Factorial_progrenies	BioProject BreedToLast PRJNA483561	controlled crossbreeds
52-27				yes	5.06	Unrelated	BioProject BreedToLast PRJNA483561	NA
71040				yes	7.34	Unrelated	BioProject BreedToLast PRJNA483561	France, Savoie, Aiton
73193-084				yes	9.33	Unrelated	BioProject BreedToLast PRJNA483561	Hautes Pyrénées, Soulon
71030-501				yes	10.94	Unrelated	BioProject BreedToLast PRJNA483561	Savoie, Rognaix
73193-097				yes	13.68	Unrelated	BioProject BreedToLast PRJNA483561	Hautes Pyrénées, Soulon
72149-015				yes	13.78	Unrelated	BioProject BreedToLast PRJNA483561	Gard, Codolet

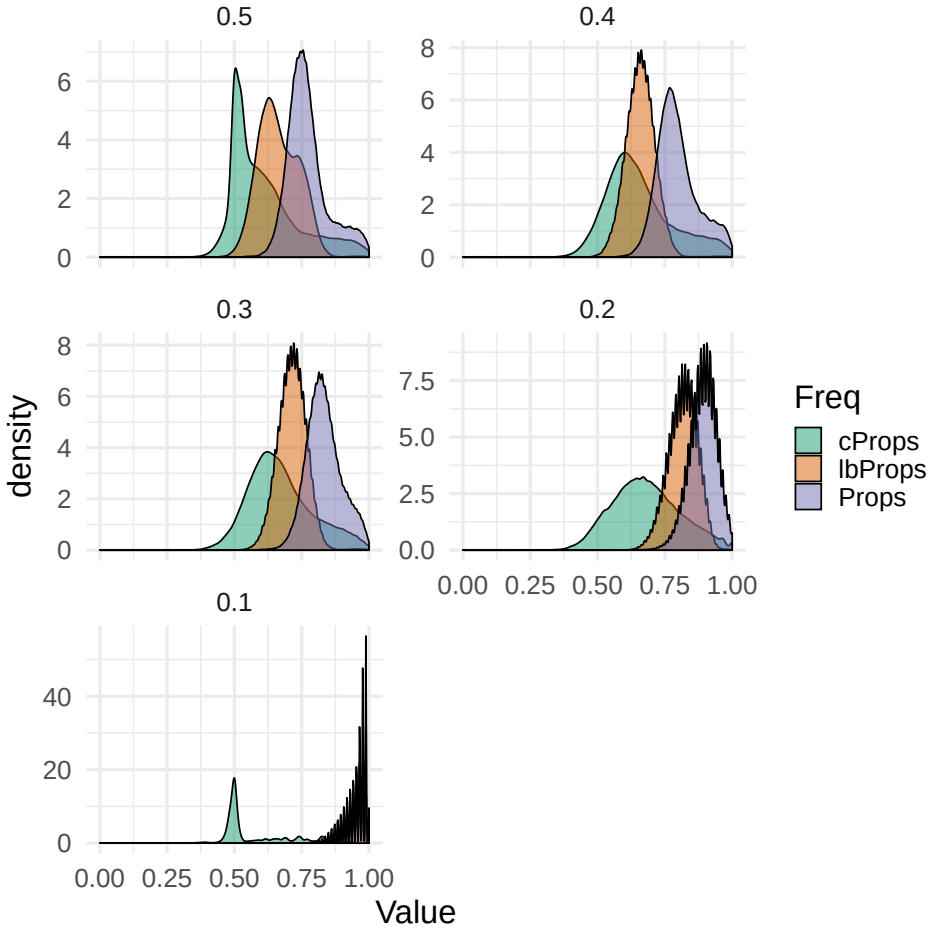
Additional file 2 : FigureS1.pdf — Relationship between the proportion of alleles correctly imputed by each leave-one-out individual (Propi) and the lower bound individual proportion of SNP correctly imputed lbPropi).The different colors correspond to the different individual classes in the mating regimes, and each point represents the values for one chromosome and one individual.



Additional file 3 : FigureS2.pdf — Relationship between the sequencing depth and imputation quality variables at individual level. On the top of the diagonal: Pearson's correlations. The distribution of each variable is shown on the diagonal. On the bottom of the diagonal: the bivariate scatter plots.



Additional file 4 : FigureS3.pdf — Variation of the three different estimates of imputation quality at the SNP level (*Props* (Green), *lbProps* (Purple), *cProps* (Orange)), as a function of different classes of minor allele frequency (FreqOri).



Appendix C

Chapitre 4 : Article Supplementary material

1 Supplementary Tables and Figures

1.1 Tables

Table 1: Akaike Information Criterion (AIC) and nerrow sense heritabilities by Traits, Matrices, Models, Single or Multiple trait (ST_MT), and GenoSet

Trait	Matrice	h^2	AIC	Model	ST_MT	GenoSet
angbranch	A	0.74	19313.62	ADD	MT	none
angbranch	A	0.65	347.87	ADD	ST	none
budburst	A	0.53	19313.62	ADD	MT	none
budburst	A	0.33	367.89	ADD	ST	none
circ2	A	0.59	19313.62	ADD	MT	none
circ2	A	0.65	390.46	ADD	ST	none
height1	A	0.63	19313.62	ADD	MT	none
height1	A	0.67	463.02	ADD	ST	none
height2	A	0.78	181.48	ADD	ST	none
height2	A	0.68	19313.62	ADD	MT	none
rust1	A	0.69	19313.62	ADD	MT	none
rust1	A	0.60	673.54	ADD	ST	none
rust2	A	0.69	19313.62	ADD	MT	none
rust2	A	0.92	214.17	ADD	ST	none
angbranch	Acor	0.72	19397.86	ADD	MT	none
angbranch	Acor	0.63	343.82	ADD	ST	none
budburst	Acor	0.54	19397.86	ADD	MT	none
budburst	Acor	0.28	351.58	ADD	ST	none
circ2	Acor	0.60	19397.86	ADD	MT	none
circ2	Acor	0.67	391.97	ADD	ST	none
height1	Acor	0.59	19397.86	ADD	MT	none
height1	Acor	0.70	453.53	ADD	ST	none
height2	Acor	0.70	178.94	ADD	ST	none
height2	Acor	0.68	19397.86	ADD	MT	none
rust1	Acor	0.62	19397.86	ADD	MT	none
rust1	Acor	0.60	657.22	ADD	ST	none
rust2	Acor	0.71	19397.86	ADD	MT	none
rust2	Acor	0.91	212.21	ADD	ST	none
angbranch	AcorD	0.62	341.89	ADDetDOM	ST	none
budburst	AcorD	0.28	288.12	ADDetDOM	ST	none
circ2	AcorD	0.50	381.79	ADDetDOM	ST	none
height1	AcorD	0.70	447.49	ADDetDOM	ST	none
height2	AcorD	0.99	256.03	ADDetDOM	ST	none

Continued on next page

Table 1 – *Continued from previous page*

Trait	**Matrice**	h^2	**AIC**	**Model**	**ST_MT**	**GenoSet**
rust1	AcorD	0.60	630.26	ADDetDOM	ST	none
rust2	AcorD	0.92	211.05	ADDetDOM	ST	none
angbranch	AD	0.65	347.35	ADDetDOM	ST	none
budburst	AD	0.29	301.82	ADDetDOM	ST	none
circ2	AD	0.55	384.70	ADDetDOM	ST	none
height1	AD	0.68	450.25	ADDetDOM	ST	none
height2	AD	0.91	174.46	ADDetDOM	ST	none
rust1	AD	0.60	649.73	ADDetDOM	ST	none
rust2	AD	0.92	214.02	ADDetDOM	ST	none
angbranch	BayesC	0.57	NA	ADD	ST	7K
angbranch	BayesC	0.86	NA	ADD	ST	50K
angbranch	BayesC	0.86	NA	ADD	ST	100K
rust2	BayesC	0.57	NA	ADD	ST	7K
rust2	BayesC	0.88	NA	ADD	ST	50K
rust2	BayesC	0.88	NA	ADD	ST	100K
angbranch	G	1.00	19142.88	ADD	MT	7K
angbranch	G	0.49	335.57	ADD	ST	7K
angbranch	G	1.00	356.74	ADD	ST	50K
angbranch	G	1.00	393.98	ADD	ST	100K
angbranch	G	1.00	436.33	ADD	ST	250K
budburst	G	1.00	19142.88	ADD	MT	7K
budburst	G	0.60	233.29	ADD	ST	7K
budburst	G	0.40	548.09	ADD	ST	100K
budburst	G	0.64	579.20	ADD	ST	250K
budburst	G	0.16	595.34	ADD	ST	50K
circ2	G	1.00	19142.88	ADD	MT	7K
circ2	G	0.54	395.44	ADD	ST	7K
circ2	G	1.00	459.44	ADD	ST	50K
circ2	G	1.00	515.81	ADD	ST	100K
circ2	G	1.00	583.57	ADD	ST	250K
height1	G	1.00	19142.88	ADD	MT	7K
height1	G	1.00	207.99	ADD	ST	50K
height1	G	1.00	276.60	ADD	ST	100K
height1	G	1.00	353.06	ADD	ST	250K
height1	G	0.44	427.42	ADD	ST	7K
height2	G	0.53	163.27	ADD	ST	7K
height2	G	1.00	19142.88	ADD	MT	7K
height2	G	1.00	433.59	ADD	ST	50K

Continued on next page

Table 1 – *Continued from previous page*

Trait	**Matrice**	h^2	**AIC**	**Model**	**ST_MT**	**GenoSet**
height2	G	1.00	471.10	ADD	ST	100K
height2	G	1.00	505.24	ADD	ST	250K
rust1	G	1.00	19142.88	ADD	MT	7K
rust1	G	0.60	565.66	ADD	ST	7K
rust1	G	1.00	654.30	ADD	ST	50K
rust1	G	1.00	693.77	ADD	ST	100K
rust1	G	1.00	751.20	ADD	ST	250K
rust2	G	1.00	19142.88	ADD	MT	7K
rust2	G	0.57	197.93	ADD	ST	7K
rust2	G	1.00	237.35	ADD	ST	50K
rust2	G	1.00	281.48	ADD	ST	100K
rust2	G	1.00	330.80	ADD	ST	250K
angbranch	GD	0.42	327.54	ADDetDOM	ST	7K
budburst	GD	0.51	209.57	ADDetDOM	ST	7K
circ2	GD	0.47	394.77	ADDetDOM	ST	7K
height1	GD	0.35	407.39	ADDetDOM	ST	7K
height2	GD	0.47	156.59	ADDetDOM	ST	7K
rust1	GD	0.54	552.75	ADDetDOM	ST	7K
rust2	GD	0.55	197.33	ADDetDOM	ST	7K
angbranch	GDw1	0.69	4.28	ADDetDOM	ST	7K
budburst	GDw1	0.69	-161.36	ADDetDOM	ST	7K
circ2	GDw1	0.76	131.99	ADDetDOM	ST	7K
height1	GDw1	0.54	148.35	ADDetDOM	ST	7K
height2	GDw1	0.62	-81.14	ADDetDOM	ST	7K
rust1	GDw1	0.65	154.49	ADDetDOM	ST	7K
rust2	GDw1	0.71	-61.11	ADDetDOM	ST	7K
angbranch	GDw2	0.86	-208.87	ADDetDOM	ST	7K
budburst	GDw2	0.79	-392.56	ADDetDOM	ST	7K
circ2	GDw2	0.84	-97.26	ADDetDOM	ST	7K
height1	GDw2	0.66	-42.96	ADDetDOM	ST	7K
height2	GDw2	0.72	-231.63	ADDetDOM	ST	7K
rust1	GDw2	0.75	-32.31	ADDetDOM	ST	7K
rust2	GDw2	0.79	-195.47	ADDetDOM	ST	7K
angbranch	GDw3	0.89	-319.15	ADDetDOM	ST	7K
budburst	GDw3	0.83	-487.40	ADDetDOM	ST	7K
circ2	GDw3	0.89	-219.60	ADDetDOM	ST	7K
height1	GDw3	0.73	-128.21	ADDetDOM	ST	7K
height2	GDw3	0.81	-299.85	ADDetDOM	ST	7K

Continued on next page

Table 1 – *Continued from previous page*

Trait	**Matrice**	h^2	**AIC**	**Model**	**ST_MT**	**GenoSet**
rust1	GDw3	0.79	-114.58	ADDetDOM	ST	7K
rust2	GDw3	0.83	-253.65	ADDetDOM	ST	7K
angbranch	Gw1	0.76	21.99	ADD	ST	7K
angbranch	Gw1	0.81	364.04	ADD	ST	50K
angbranch	Gw1	1.00	371.72	ADD	ST	100K
angbranch	Gw1	1.00	373.16	ADD	ST	250K
budburst	Gw1	0.72	-142.79	ADD	ST	7K
budburst	Gw1	0.73	285.06	ADD	ST	50K
budburst	Gw1	0.54	345.67	ADD	ST	100K
budburst	Gw1	0.00	947.00	ADD	ST	250K
circ2	Gw1	0.77	127.99	ADD	ST	7K
circ2	Gw1	0.87	485.08	ADD	ST	50K
circ2	Gw1	0.96	507.25	ADD	ST	100K
circ2	Gw1	1.00	513.21	ADD	ST	250K
height1	Gw1	0.58	158.88	ADD	ST	7K
height1	Gw1	0.89	193.29	ADD	ST	50K
height1	Gw1	1.00	205.53	ADD	ST	100K
height1	Gw1	1.00	207.17	ADD	ST	250K
height2	Gw1	0.66	-70.89	ADD	ST	7K
height2	Gw1	1.00	424.94	ADD	ST	50K
height2	Gw1	1.00	444.60	ADD	ST	100K
height2	Gw1	1.00	451.98	ADD	ST	250K
rust1	Gw1	0.70	168.49	ADD	ST	7K
rust1	Gw1	0.79	640.43	ADD	ST	50K
rust1	Gw1	0.97	644.99	ADD	ST	100K
rust1	Gw1	1.00	668.57	ADD	ST	250K
rust2	Gw1	0.71	-61.33	ADD	ST	7K
rust2	Gw1	1.00	218.99	ADD	ST	50K
rust2	Gw1	1.00	227.90	ADD	ST	100K
rust2	Gw1	1.00	238.85	ADD	ST	250K
angbranch	Gw2	0.89	-206.02	ADD	ST	7K
angbranch	Gw2	0.43	406.81	ADD	ST	50K
angbranch	Gw2	0.58	409.20	ADD	ST	250K
angbranch	Gw2	0.51	425.60	ADD	ST	100K
budburst	Gw2	0.80	-382.93	ADD	ST	7K
budburst	Gw2	0.53	351.18	ADD	ST	50K
budburst	Gw2	0.58	368.47	ADD	ST	100K
circ2	Gw2	0.85	-96.32	ADD	ST	7K

Continued on next page

Table 1 – *Continued from previous page*

Trait	**Matrice**	h^2	**AIC**	**Model**	**ST_MT**	**GenoSet**
circ2	Gw2	0.64	601.57	ADD	ST	250K
circ2	Gw2	0.52	615.29	ADD	ST	50K
circ2	Gw2	0.52	633.28	ADD	ST	100K
height1	Gw2	0.69	-30.28	ADD	ST	7K
height1	Gw2	0.70	244.76	ADD	ST	250K
height1	Gw2	0.61	273.16	ADD	ST	100K
height1	Gw2	0.54	275.37	ADD	ST	50K
height2	Gw2	0.75	-226.69	ADD	ST	7K
height2	Gw2	0.73	475.29	ADD	ST	250K
height2	Gw2	0.63	478.13	ADD	ST	50K
height2	Gw2	0.64	483.64	ADD	ST	100K
rust1	Gw2	0.78	-35.86	ADD	ST	7K
rust1	Gw2	0.46	713.36	ADD	ST	50K
rust1	Gw2	0.49	719.52	ADD	ST	100K
rust1	Gw2	0.50	723.59	ADD	ST	250K
rust2	Gw2	0.78	-192.35	ADD	ST	7K
rust2	Gw2	0.62	262.17	ADD	ST	250K
rust2	Gw2	0.63	264.76	ADD	ST	100K
rust2	Gw2	0.57	268.75	ADD	ST	50K
angbranch	Gw3	0.92	-322.43	ADD	ST	7K
angbranch	Gw3	0.50	463.80	ADD	ST	50K
angbranch	Gw3	0.56	491.10	ADD	ST	250K
angbranch	Gw3	0.72	508.82	ADD	ST	100K
budburst	Gw3	0.85	-479.80	ADD	ST	7K
budburst	Gw3	0.70	450.33	ADD	ST	50K
budburst	Gw3	0.67	475.67	ADD	ST	100K
circ2	Gw3	0.89	-215.48	ADD	ST	7K
circ2	Gw3	0.58	764.45	ADD	ST	50K
circ2	Gw3	0.57	774.30	ADD	ST	250K
circ2	Gw3	0.67	797.75	ADD	ST	100K
height1	Gw3	0.76	-116.36	ADD	ST	7K
height1	Gw3	0.92	360.97	ADD	ST	250K
height1	Gw3	0.64	382.85	ADD	ST	100K
height1	Gw3	0.76	383.75	ADD	ST	50K
height2	Gw3	0.82	-303.17	ADD	ST	7K
height2	Gw3	0.83	556.50	ADD	ST	250K
height2	Gw3	0.74	558.77	ADD	ST	100K
height2	Gw3	0.83	563.18	ADD	ST	50K

Continued on next page

Table 1 – *Continued from previous page*

Trait	Matrice	h^2	AIC	Model	ST_MT	GenoSet
rust1	Gw3	0.82	-127.37	ADD	ST	7K
rust1	Gw3	0.77	818.41	ADD	ST	50K
rust1	Gw3	0.76	828.20	ADD	ST	100K
rust1	Gw3	0.70	830.30	ADD	ST	250K
rust2	Gw3	0.82	-249.37	ADD	ST	7K
rust2	Gw3	0.62	304.75	ADD	ST	250K
rust2	Gw3	0.58	323.48	ADD	ST	100K
rust2	Gw3	0.62	327.09	ADD	ST	50K

1.2 Figures

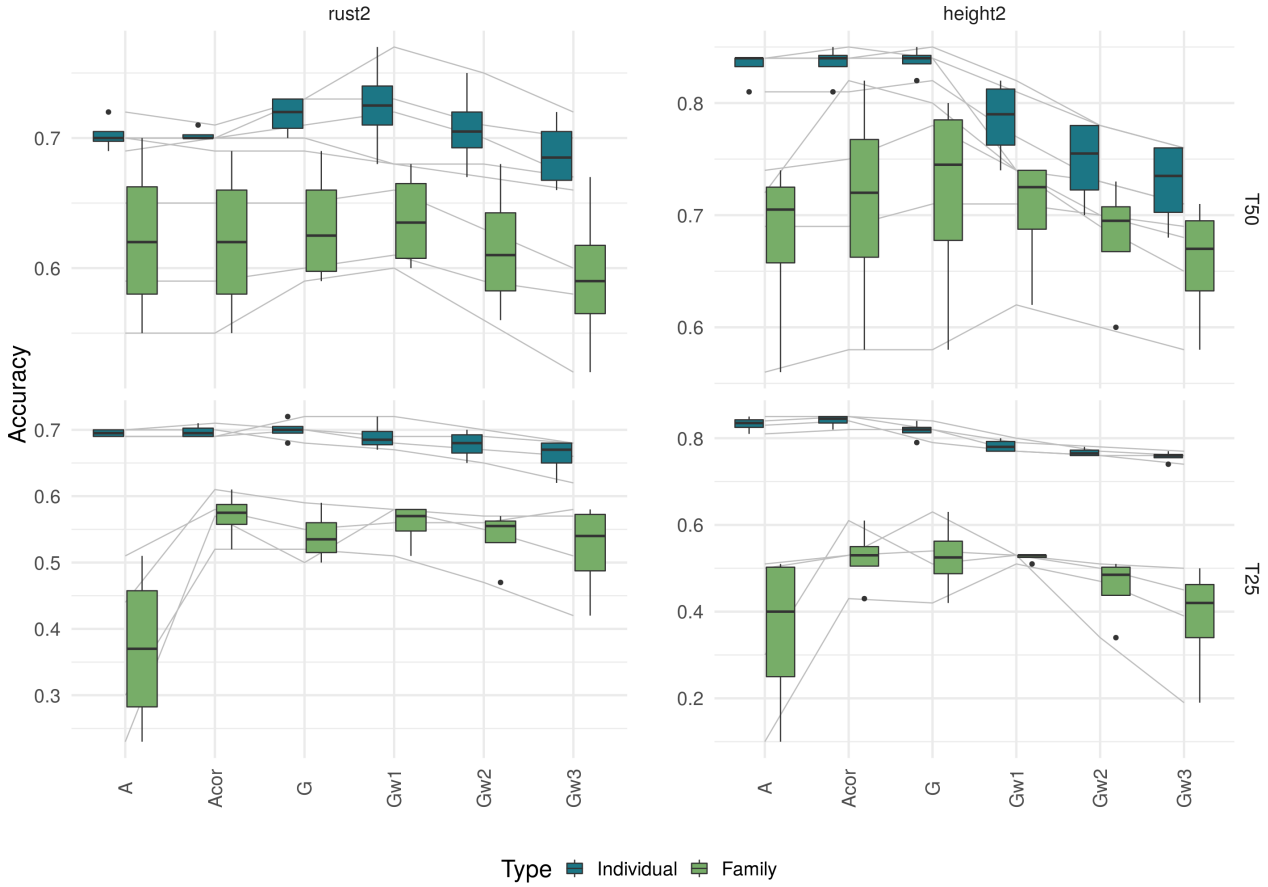


Figure 1: Cross-validation prediction accuracies using an additive model with 7K SNP for two traits grouped by the proportion of individuals in training sets 50% (T50) and 25% (T25). The color of boxplot showed the sampling strategy: in blue, the individual sampling strategy and in green the family sampling strategy. Each boxplot represented the accuracy of four repetitions for each relationship matrix. The gray lines represented the paired repetitions.

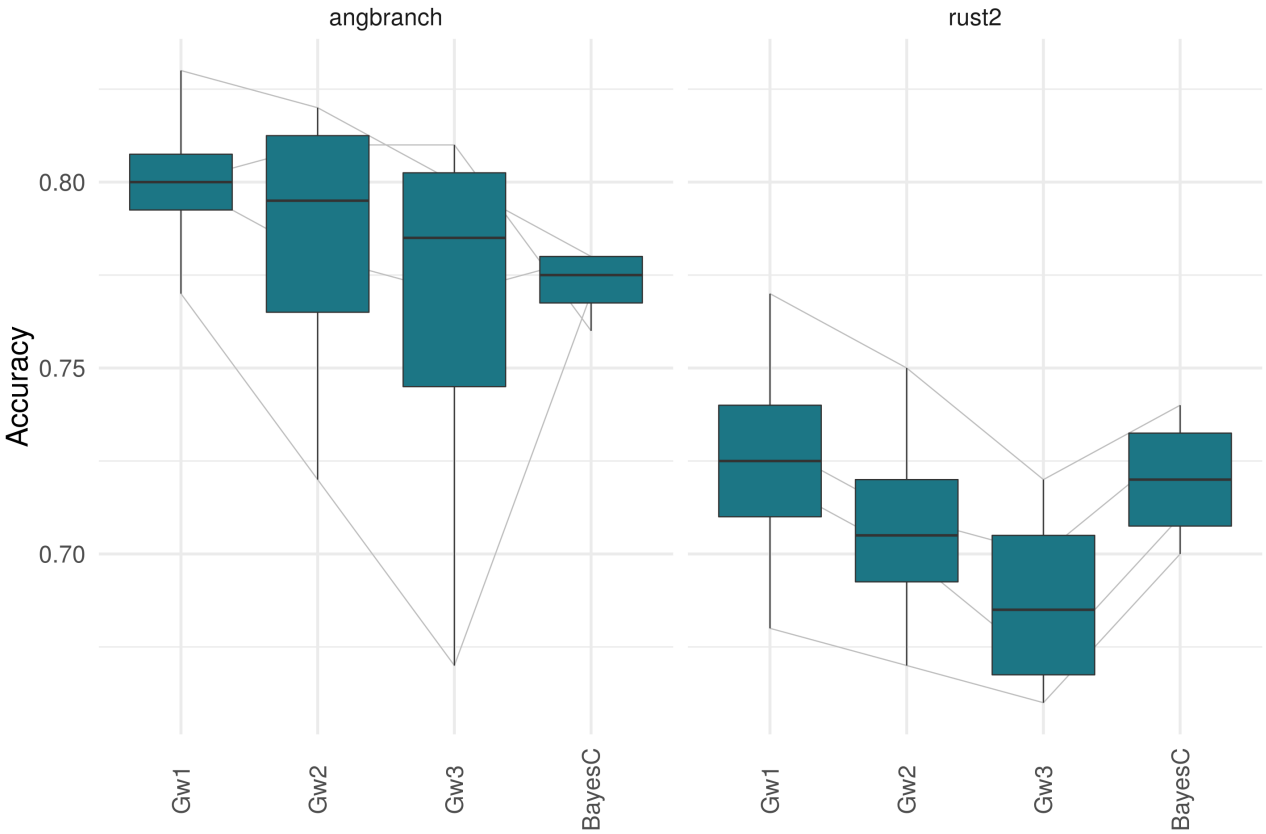


Figure 2: Cross-validation prediction accuracies using a single-trait additive model with 7K SNP for two traits with a training set of 50% sampled by individuals. Each boxplot represents the accuracy of four repetitions : wGBLUP from weighting iteration 1 to 3, and BayesCπ. The grey lines between boxplots link paired repetitions.

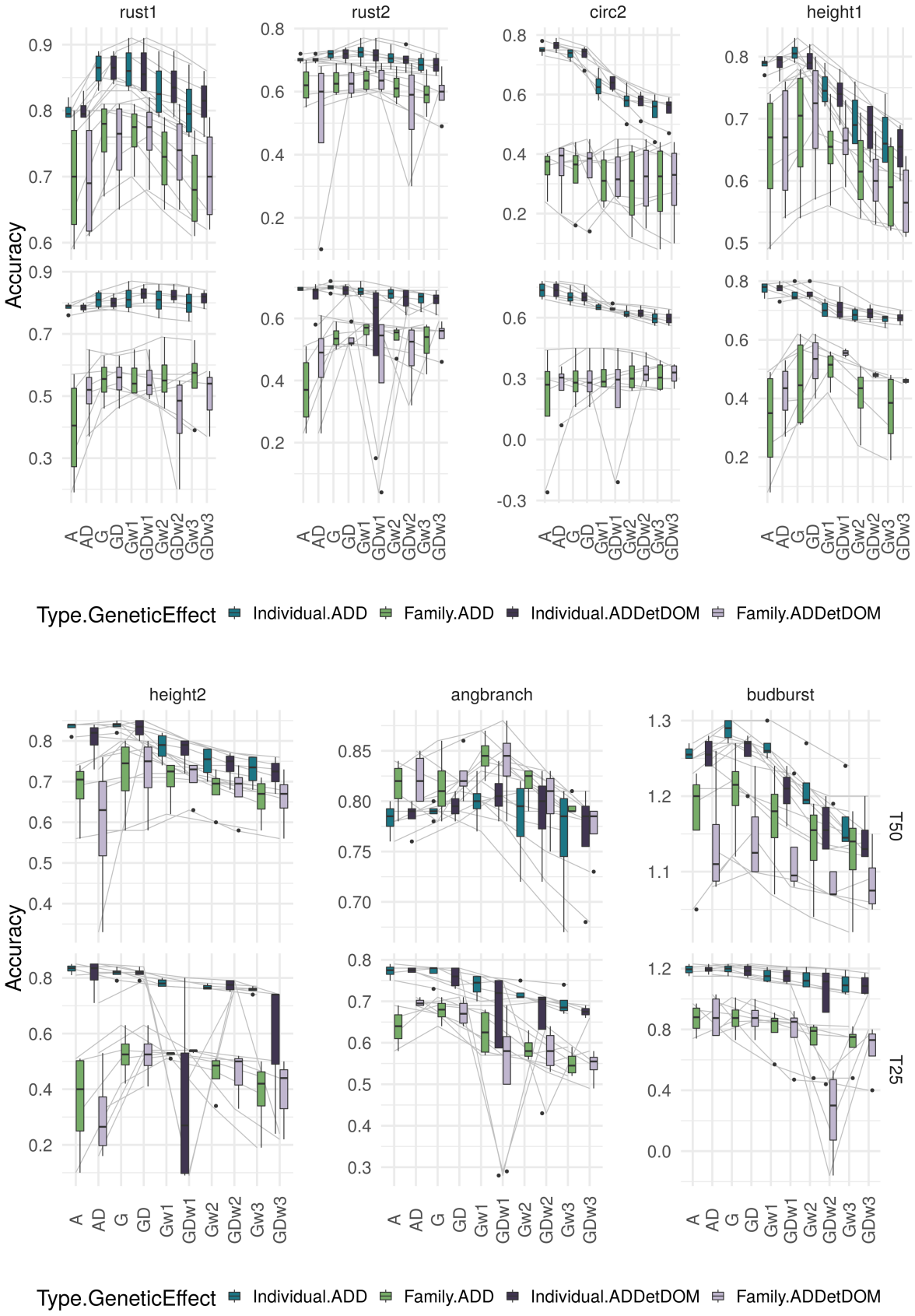


Figure 3: Cross-validation prediction accuracies using a single-trait additive model and additive plus dominance model with 7K SNP, for three traits and proportions of individuals in training sets of 50% (T50) and 25% (T25). The blue and light purple boxplots correspond to the sampling strategy based on individuals, and green and dark purple on families. The additive model is represented in blue and green boxplots, whereas the additive plus dominance model is in light and dark purple. Each boxplot represented the accuracy of four repetitions. The grey lines between boxplots link paired repetitions.

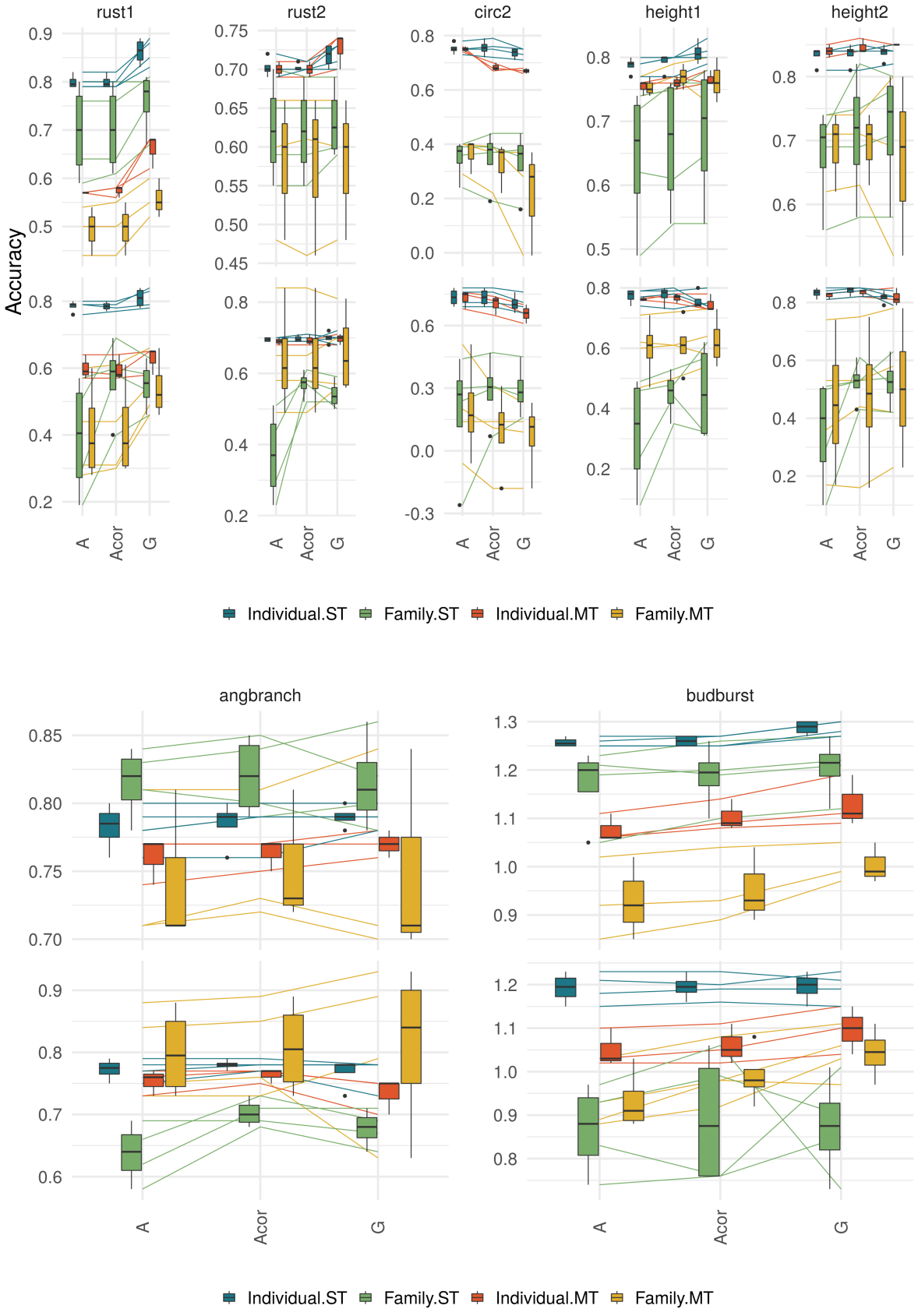


Figure 4: Cross-validation prediction accuracies using a single-trait and multiple-trait additive model with 7K SNP for two traits and proportions of individuals in training sets of 50% (T50) and 25% (T25). The blue and yellow boxplots correspond to the individual sampling strategy, green and orange boxplots to the family sampling strategy. Blue and green boxplots are for single trait model, and yellow and orange for the multiple-trait model. Each boxplot represented the accuracy of four repetitions. The grey lines between boxplots link paired repetitions.

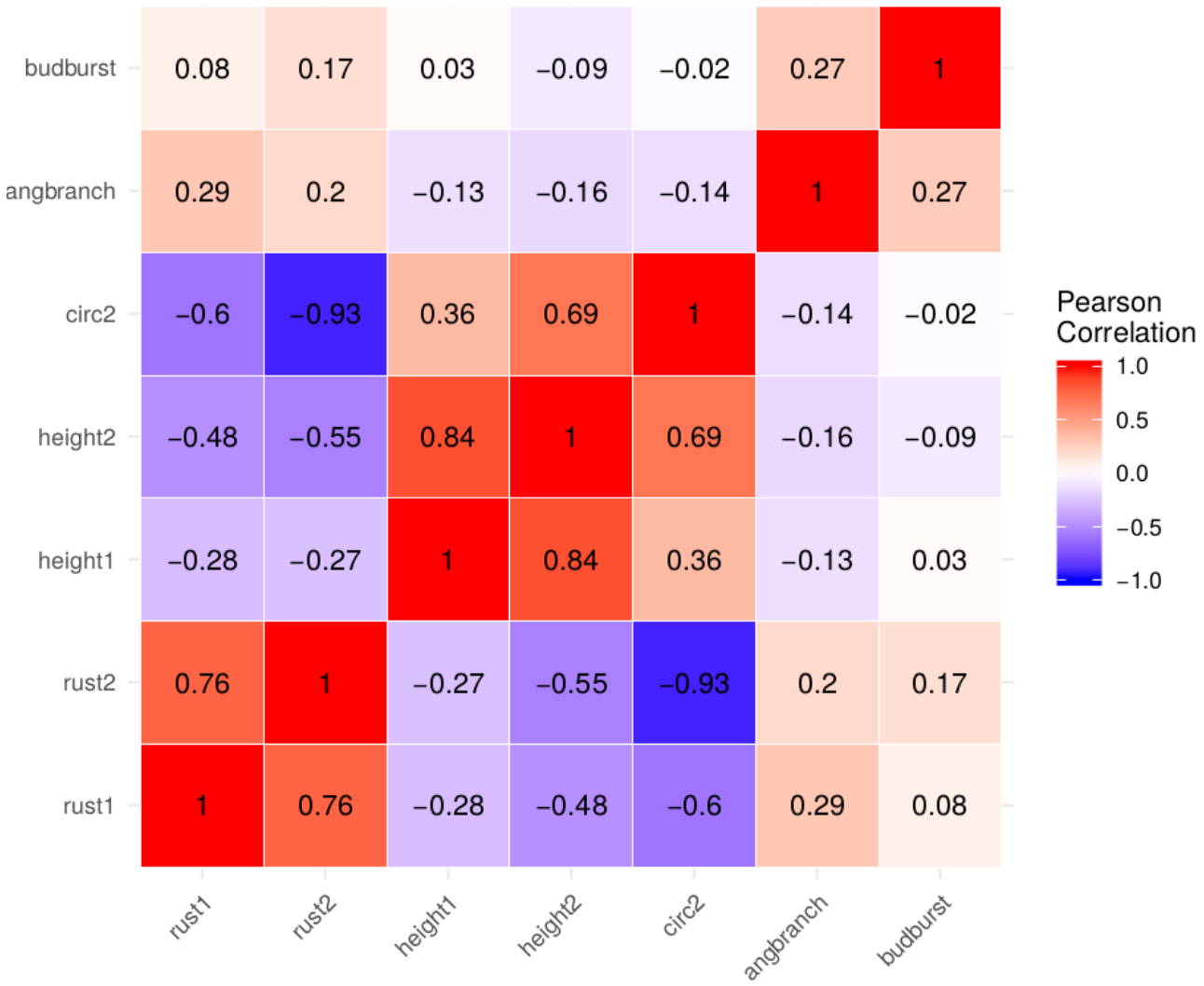


Figure 5: Genetic correlation between traits estimated with a GBLUP additive multiple trait model with a 7K SNPs genotyping.

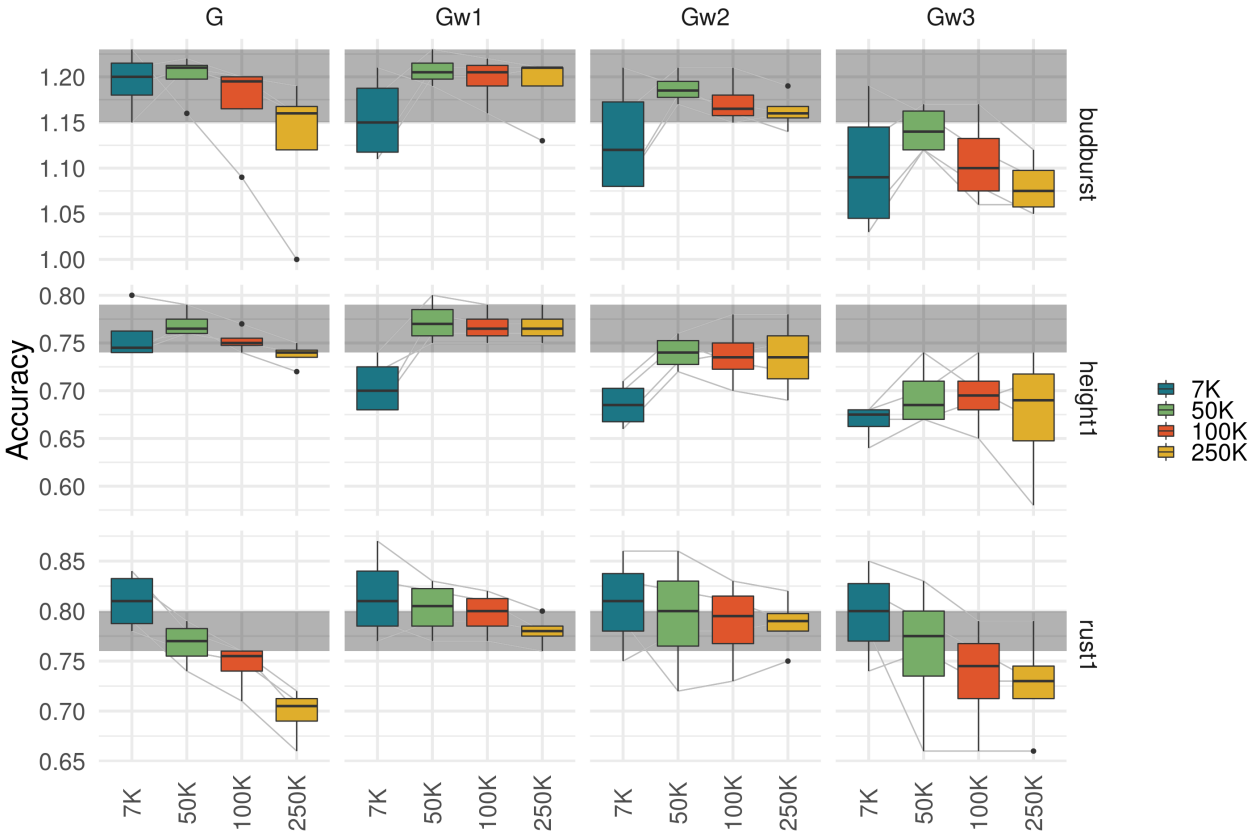


Figure 6: Marker densification impact on predictive accuracy of a single trait additive model with T50 family cross-validation, considering four genomic relationships matrices derived from different marker densities (in columns), and for three different traits (in rows) : height1, budburst and rust1. The range of accuracies obtained with the pedigree information was represented in a grey band. The grey lines between boxplots link paired repetitions.

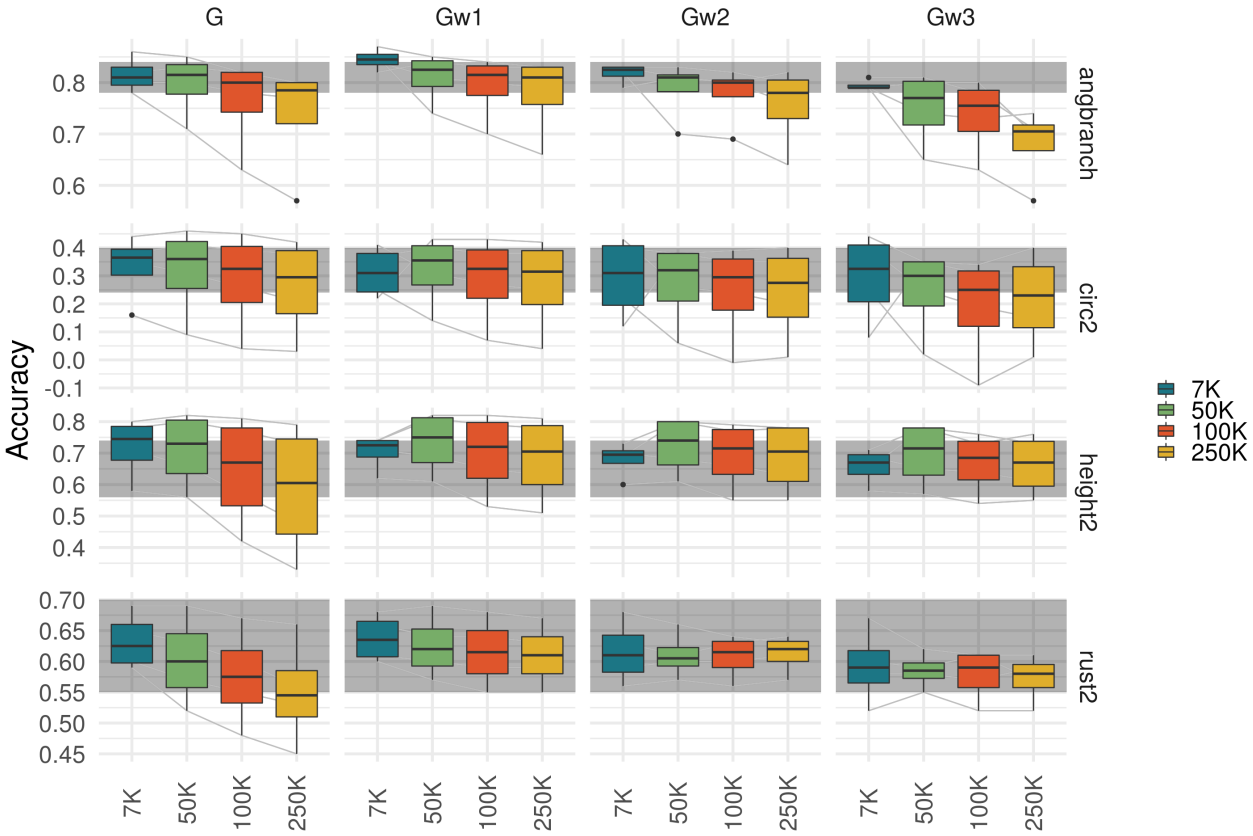


Figure 7: Marker densification impact on predictive accuracy of a single trait additive model with T50 individual cross-validation, considering four genomic relationships matrices derived from different marker densities (in columns), and for four different traits (in rows) : angbranch, circ2, height2, and rust2. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

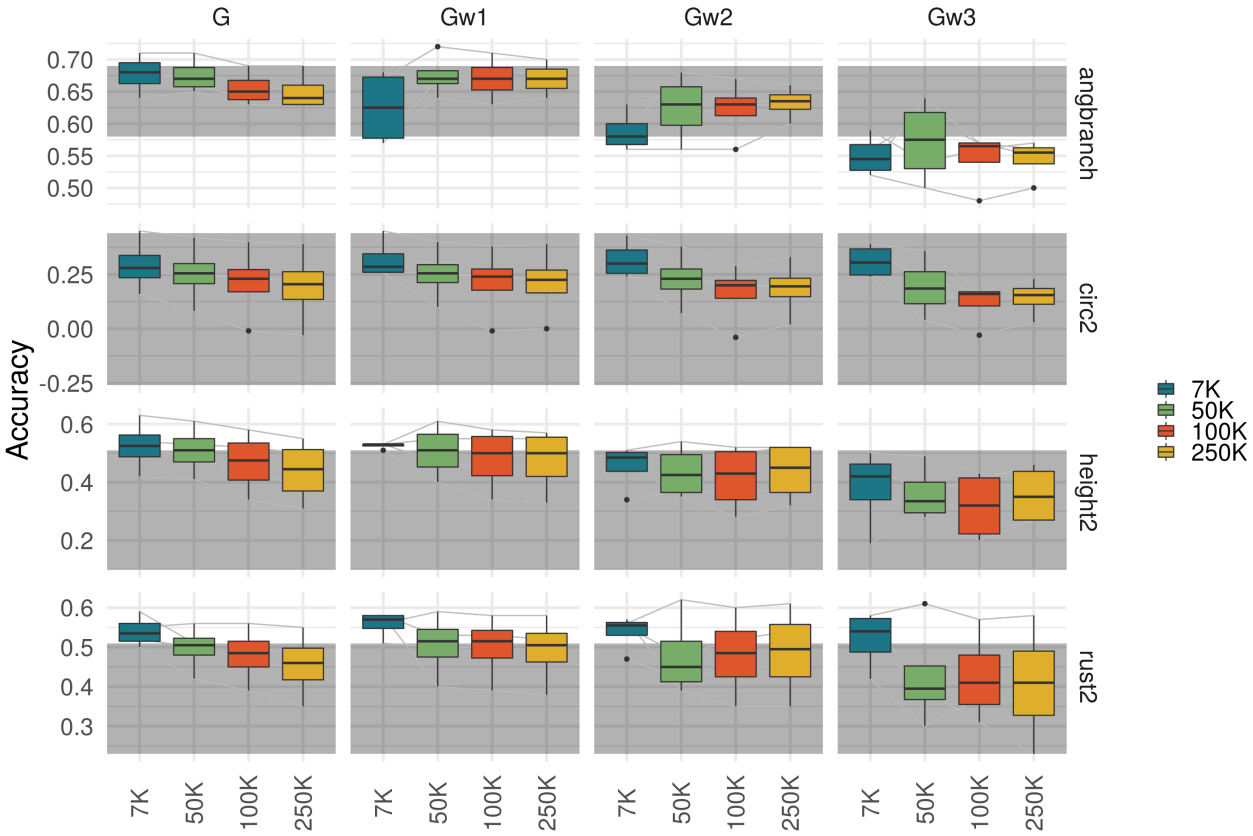


Figure 8: Marker densification impact on predictive accuracy of a single trait additive model with T50 family cross-validation, considering four genomic relationships matrices derived from different marker densities (in columns), and for four different traits (in rows) : angbranch, circ2, height2, and rust2. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

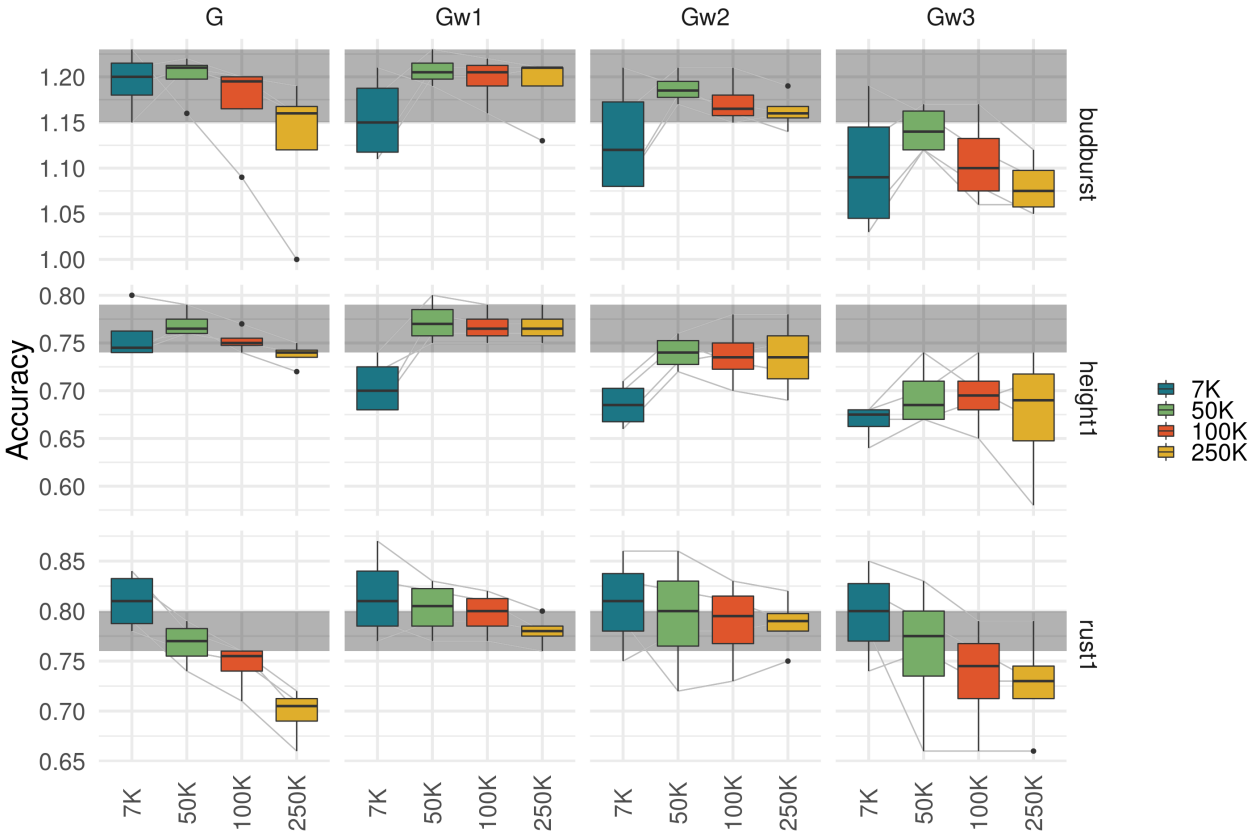


Figure 9: Marker densification impact on predictive accuracy of a single trait additive model with T25 individual cross-validation, considering four genomic relationships matrices derived from different marker densities (in columns), and for four different traits (in rows) : budburst, height1, and rust1. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

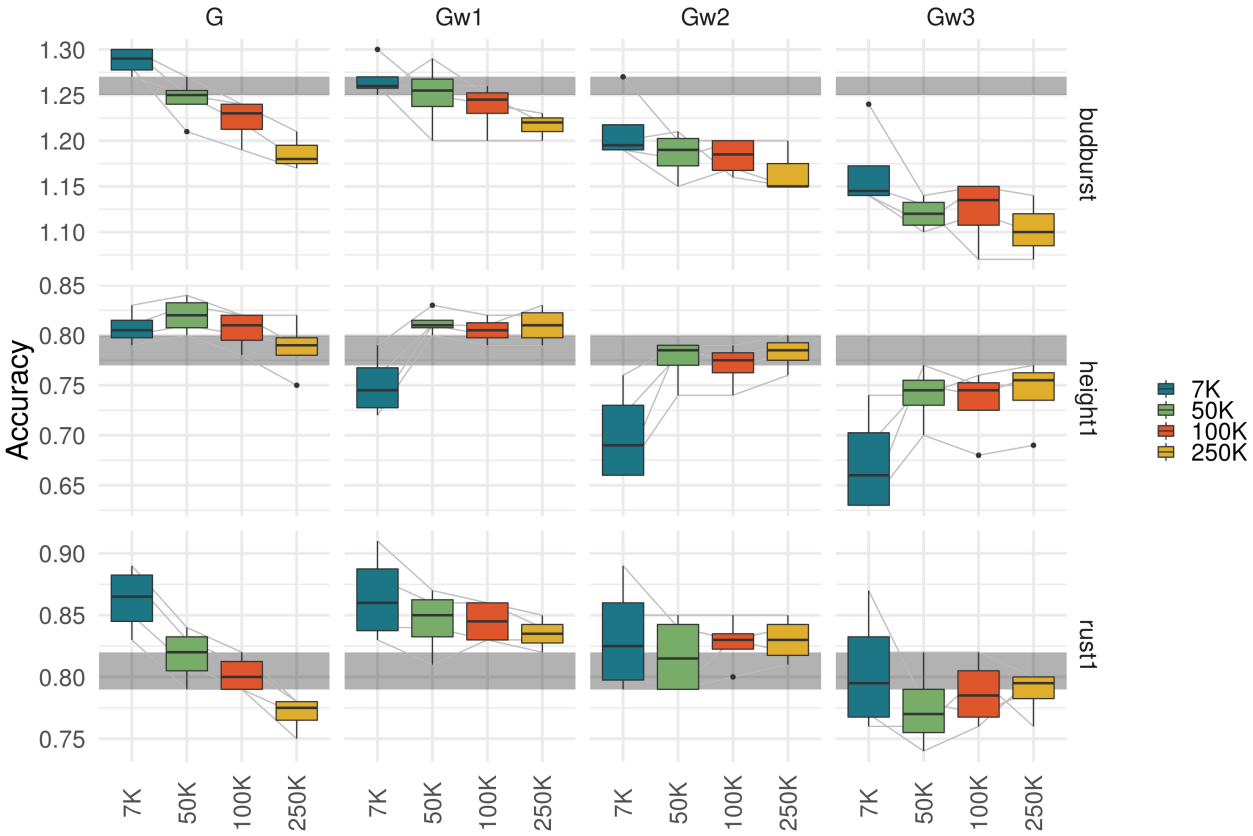


Figure 10: Marker densification impact on predictive accuracy of a single trait additive model with T25 family cross-validation, considering four genomic relationships matrices derived from different marker densities (in columns), and for four different traits (in rows) : budburst, height1, and rust1. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

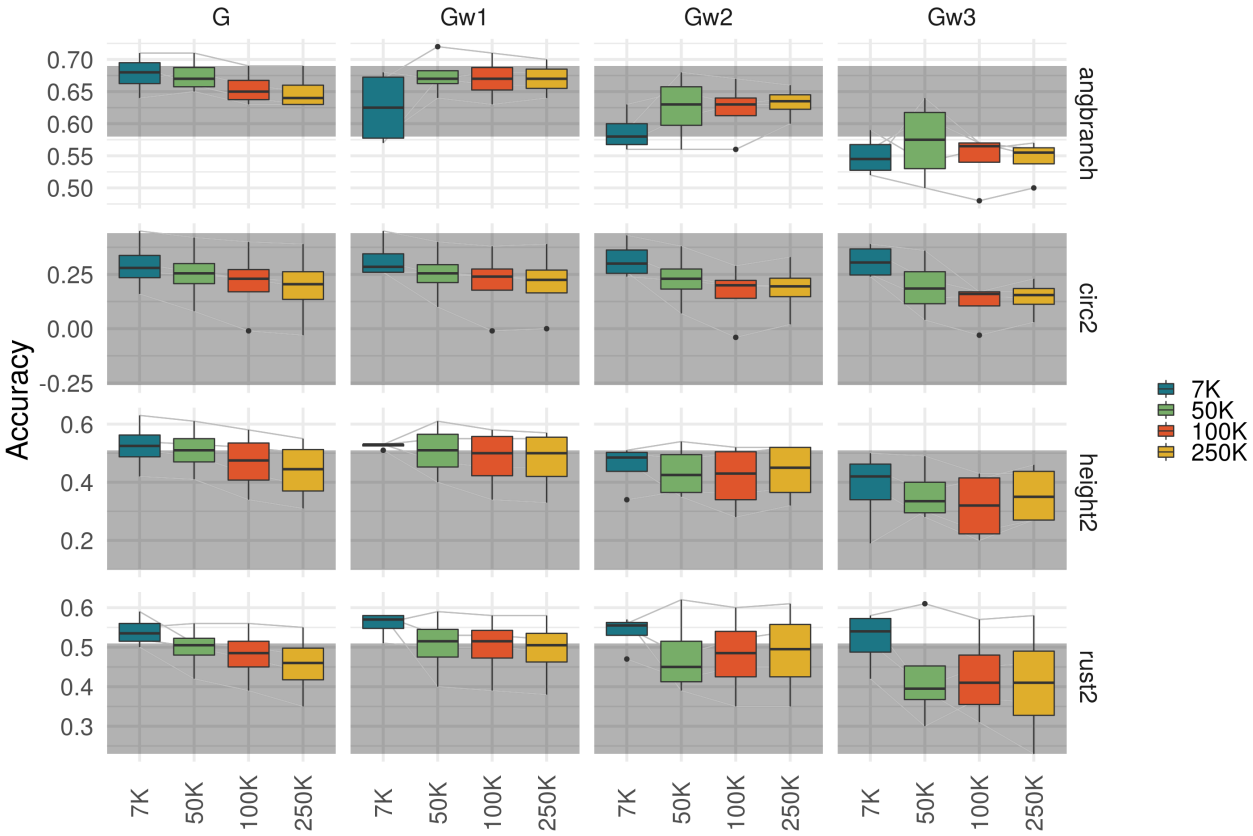


Figure 11: Marker densification impact on predictive accuracy of a single trait additive model with T25 individual cross-validation, considering four genomic relationships matrices derived from different marker densities (in columns), and for four different traits (in rows) : angbranch, circ2, height2, and rust2. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

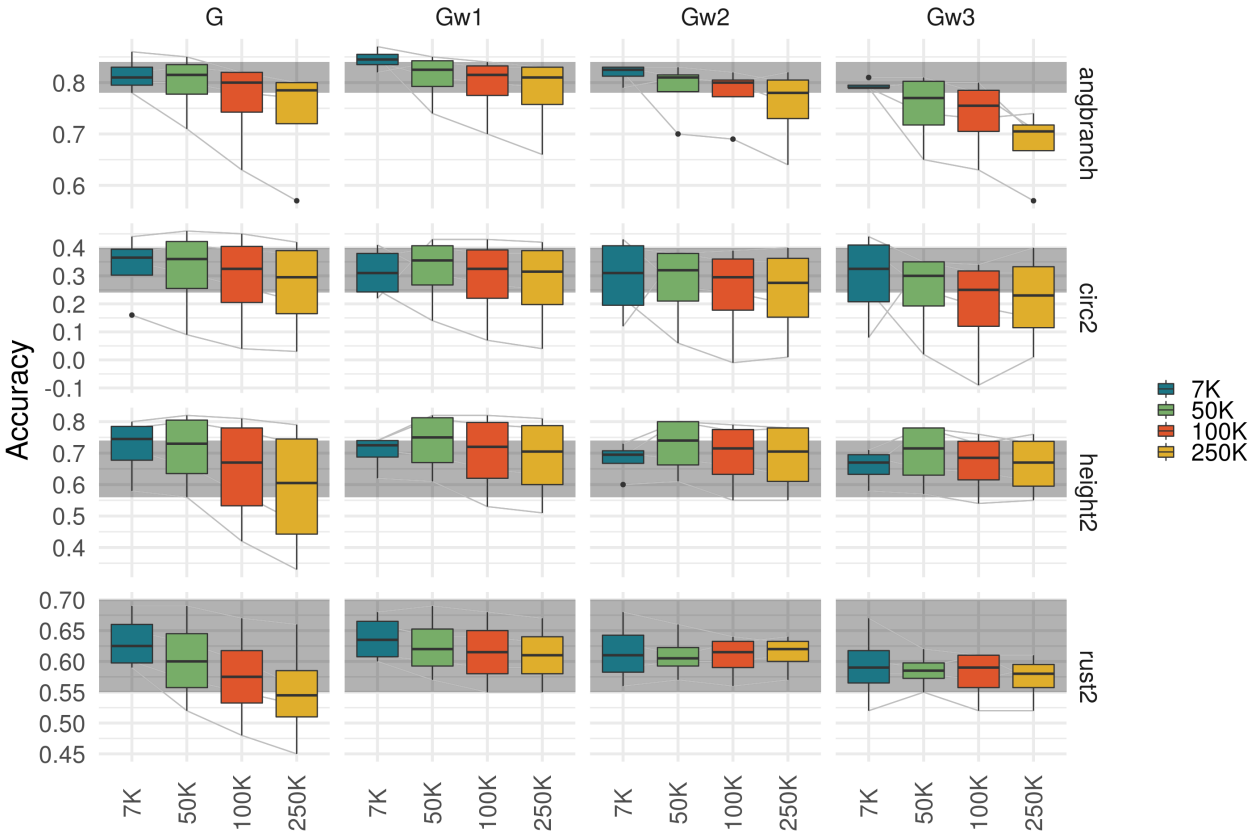


Figure 12: Marker densification impact on predictive accuracy of a single trait additive model with T25 family cross-validation, considering four genomic relationships matrices derived from different marker densities (in columns), and for four different traits (in rows) : angbranch, circ2, height2, and rust2. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

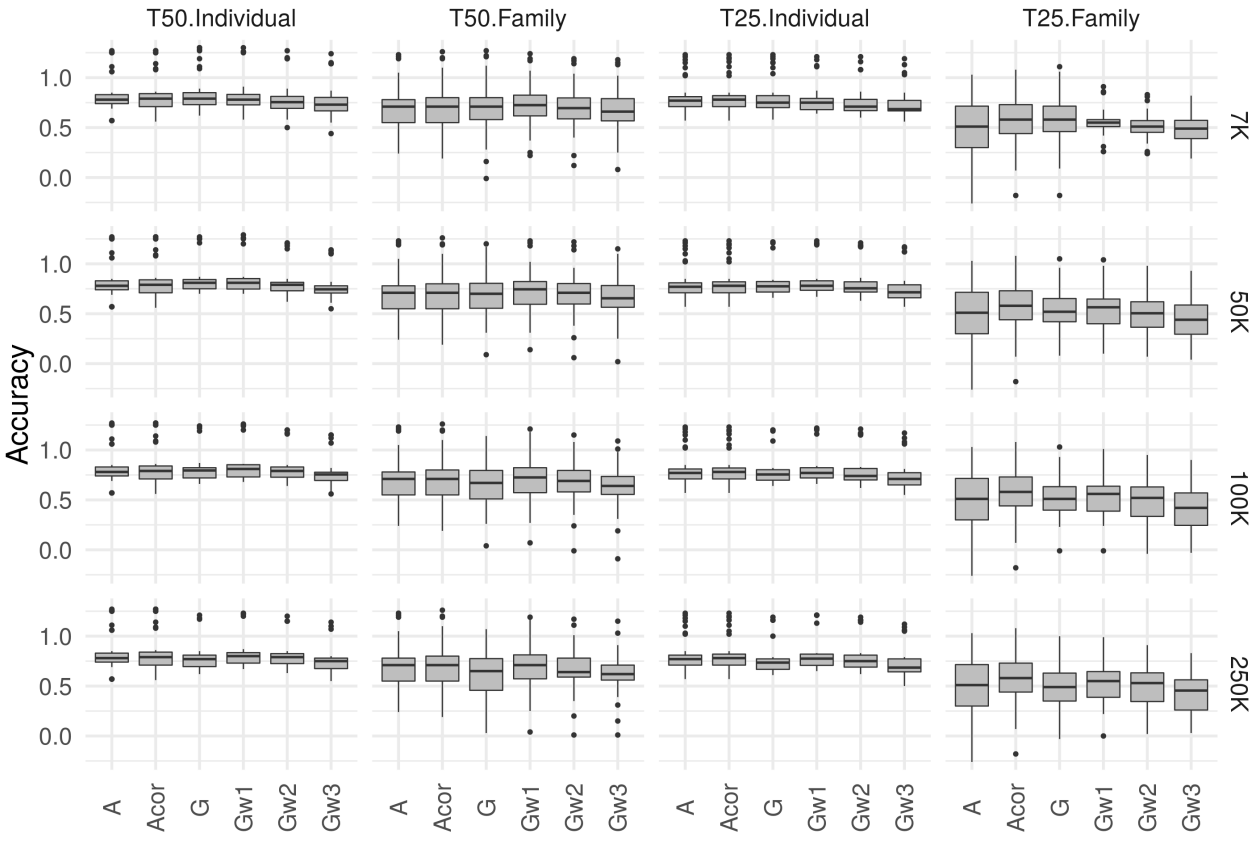


Figure 13: Prediction accuracy of cross-validation strategies by different relationship matrix strategies, by marker sets of different density, and by cross-validation strategies. Boxplot represent pooled results across traits.

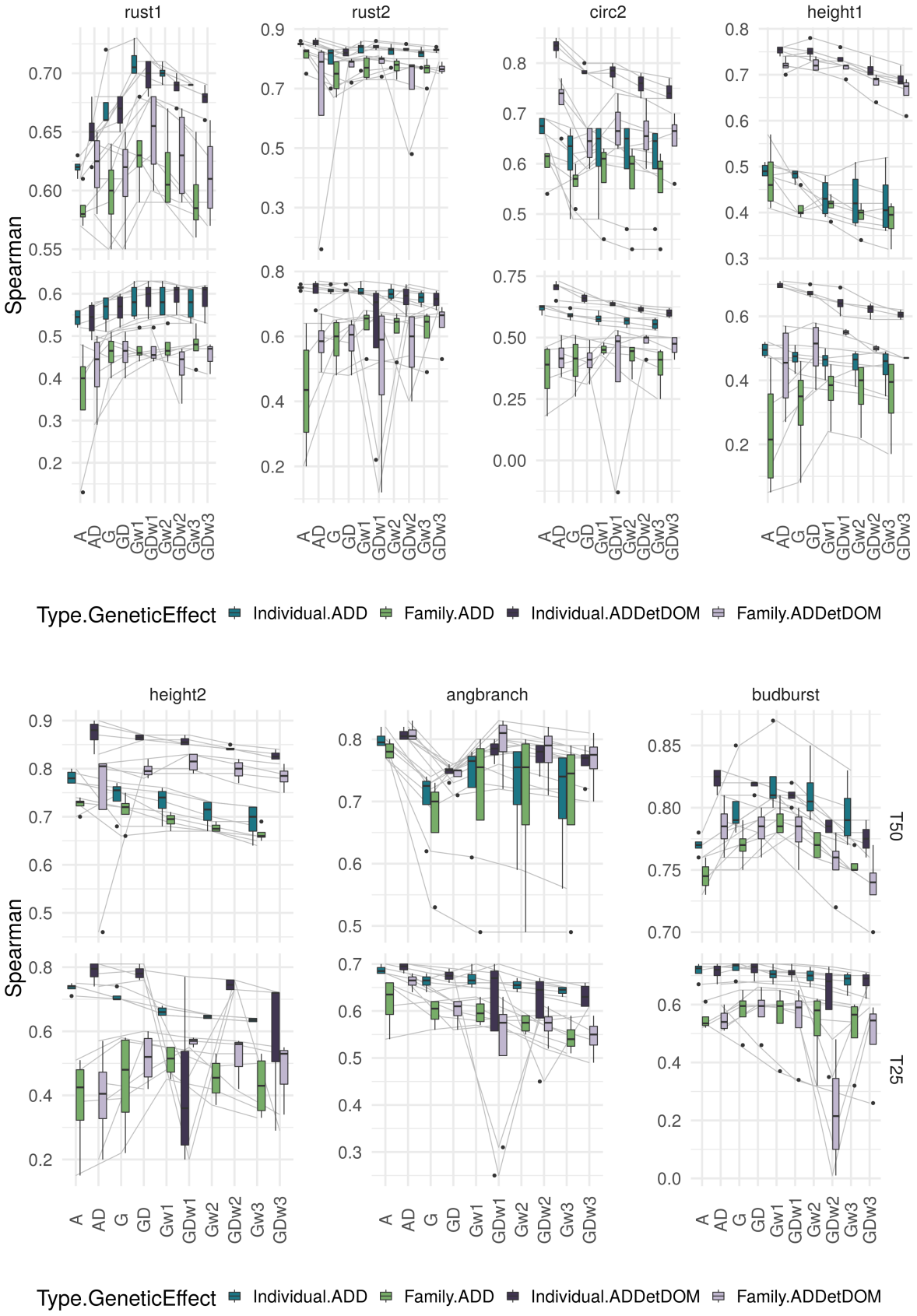


Figure 14: Independent test prediction accuracies using a single-trait additive model and additive plus dominance model with 7K SNP, for different traits grouped and proportions of individuals in training sets of 50% (T50) and 25% (T25). The blue and light purple boxplots correspond to the individual sampling strategy, green and dark purple to the family sampling strategy, blue and green to the additive model, and light and dark purple to additive plus dominance model. Each boxplot represented the accuracy of four repetitions. The grey lines between boxplots link paired repetitions.

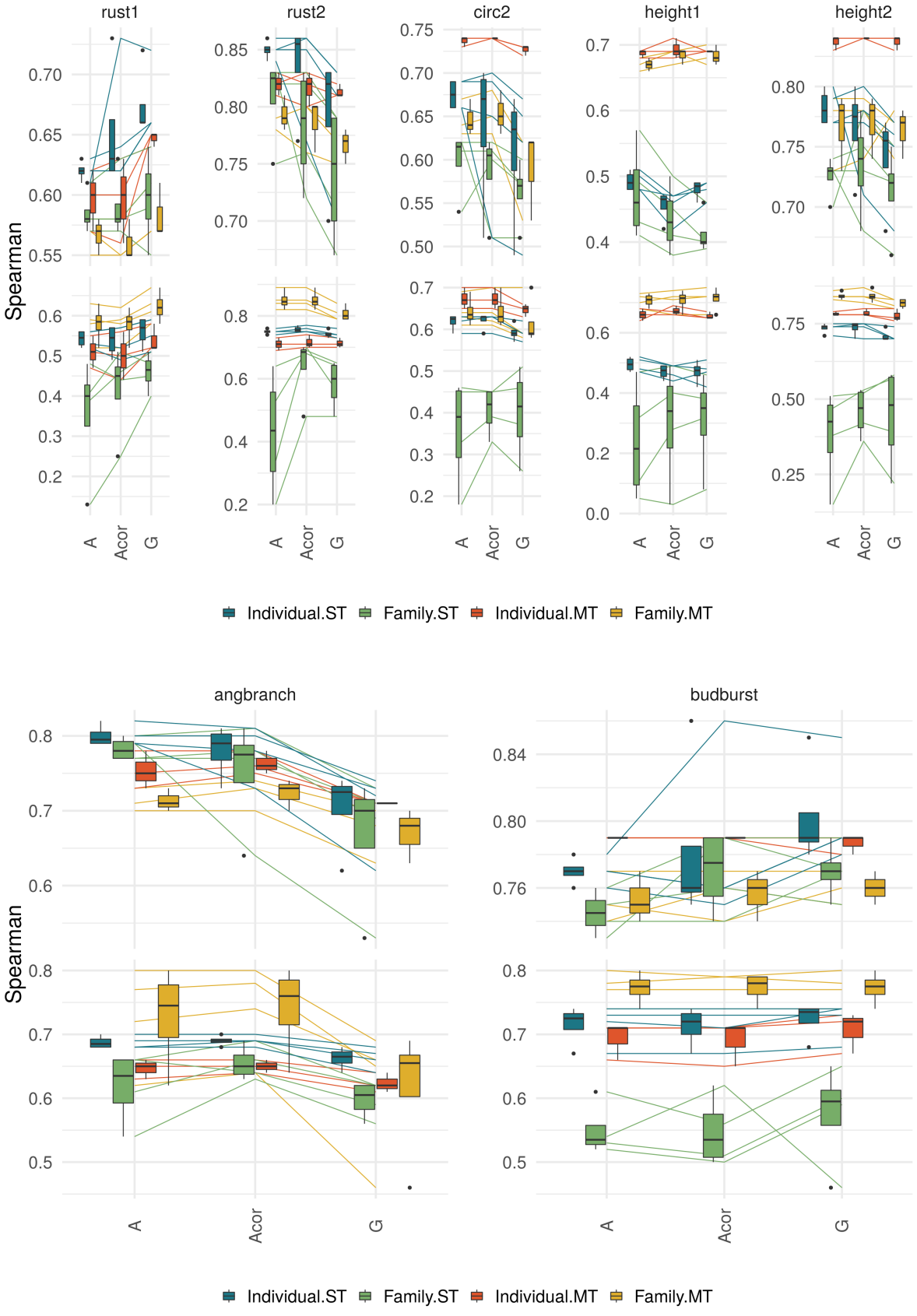


Figure 15: Independant test prediction accuracies using a single-trait and multiple-trait additive model with 7K SNP, for different traits and proportions of individuals in training sets of 50% (T50) and 25% (T25). The blue and yellow corresponds to the individual sampling strategy, green and orange to the family sampling strategy, blue and green for the single trait model, and yellow and orange for the multiple-trait model. Each boxplot represented the accuracy of four repetitions. The grey lines between boxplots link paired repetitions.

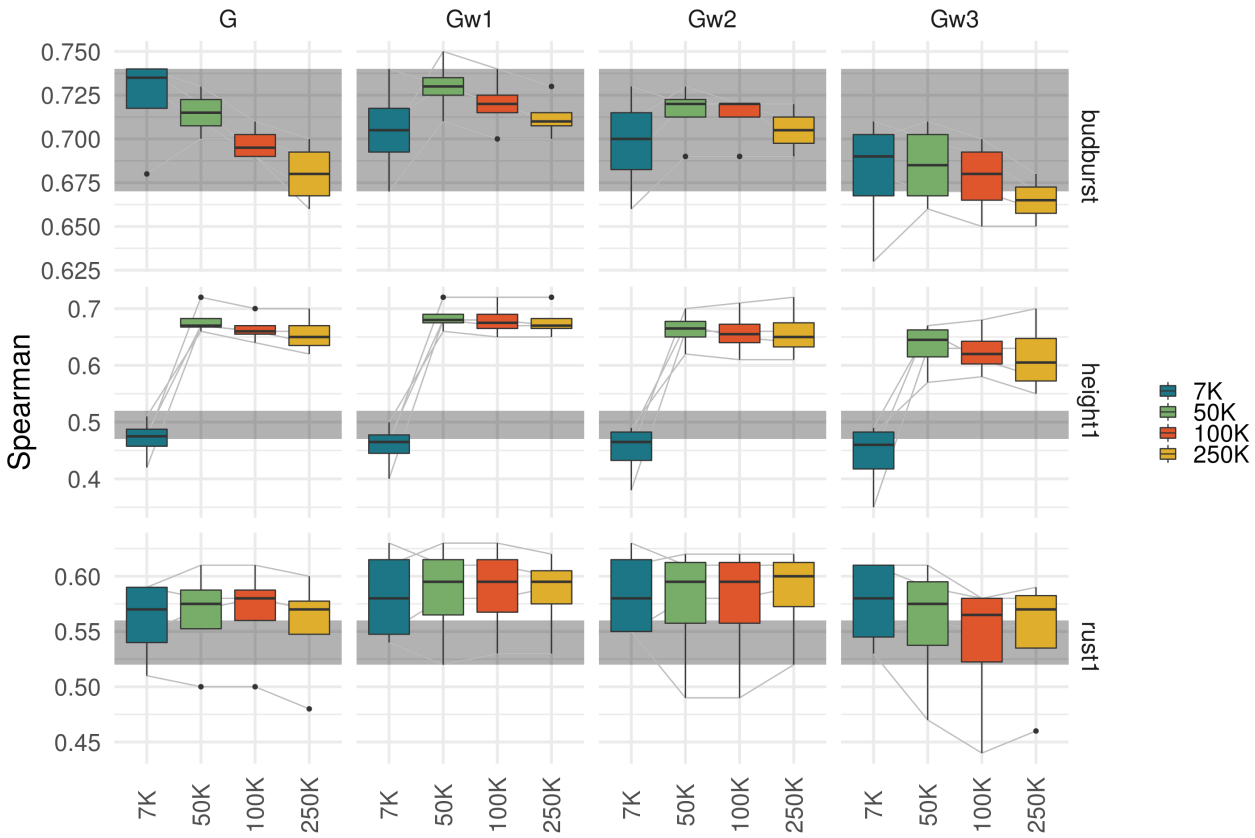


Figure 16: Marker densification impact on Spearman correlation of a single trait additive model with T50 family cross-validation, for four genomic relationships matrices (in columns) and three different traits (in rows) : height1, budburst and rust1. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

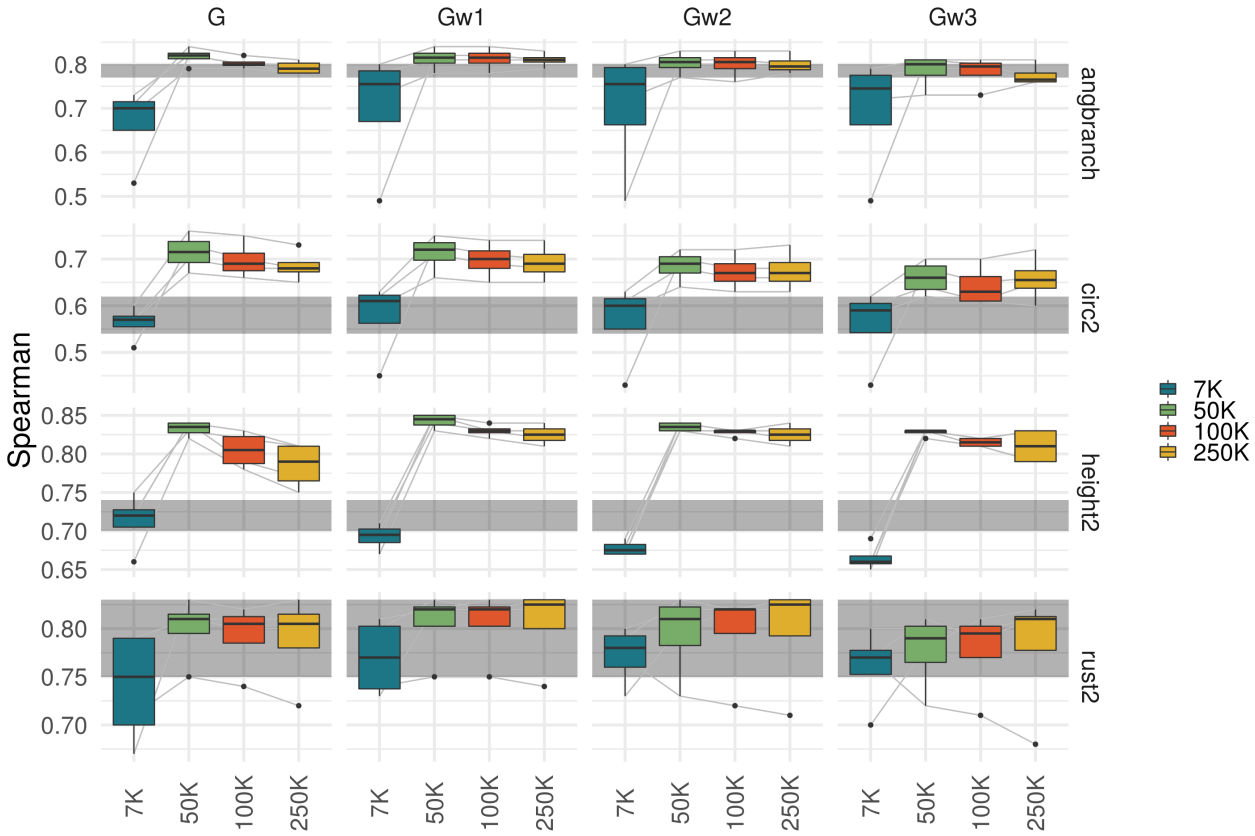


Figure 17: Marker densification impact on Spearman correlation of a single trait additive model with T50 individual cross-validation, for four genomic relationships matrices (in columns) and for four different traits (in rows) : angbranch, circ2, height2, and rust2. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

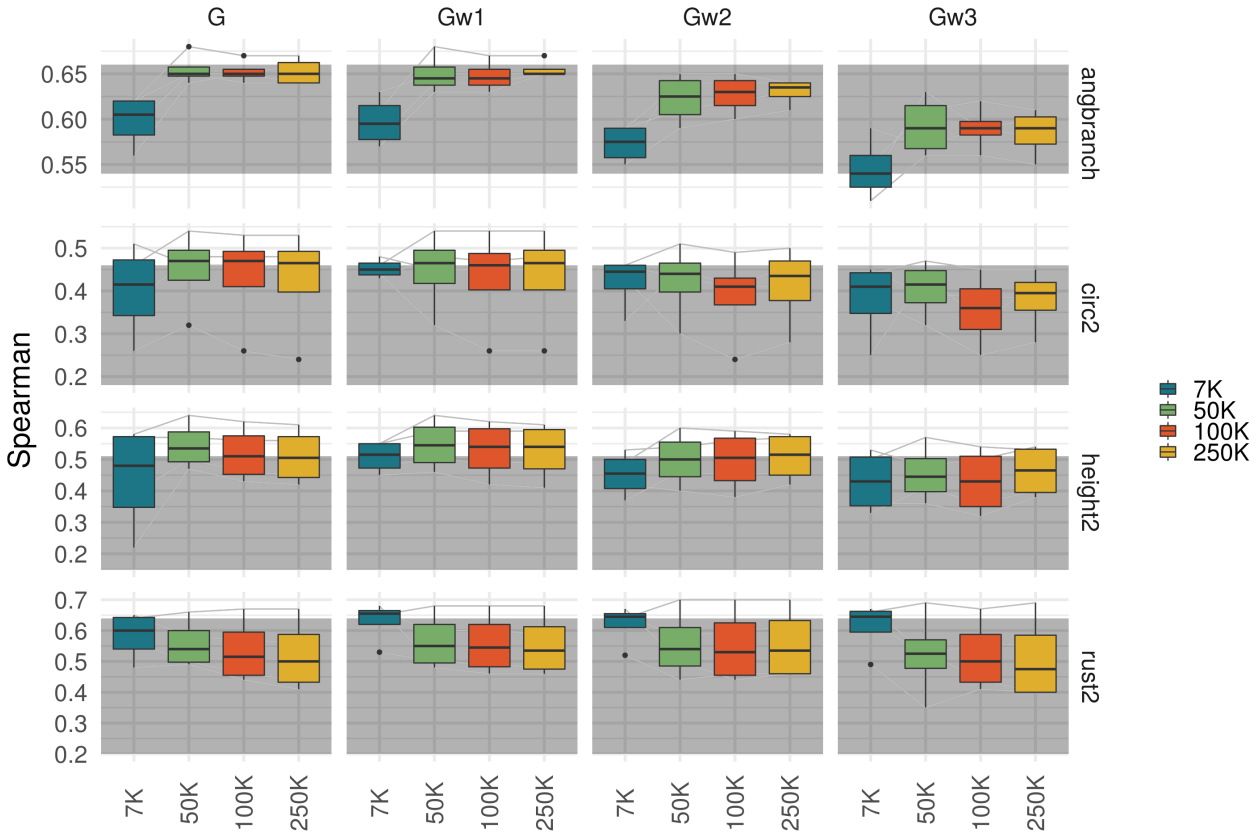


Figure 18: Marker densification impact on Spearman correlation of a single trait additive model with T50 family cross-validation, for four genomic relationships matrices (in columns) and for four different traits (in rows) : angbranch, circ2, height2, and rust2. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

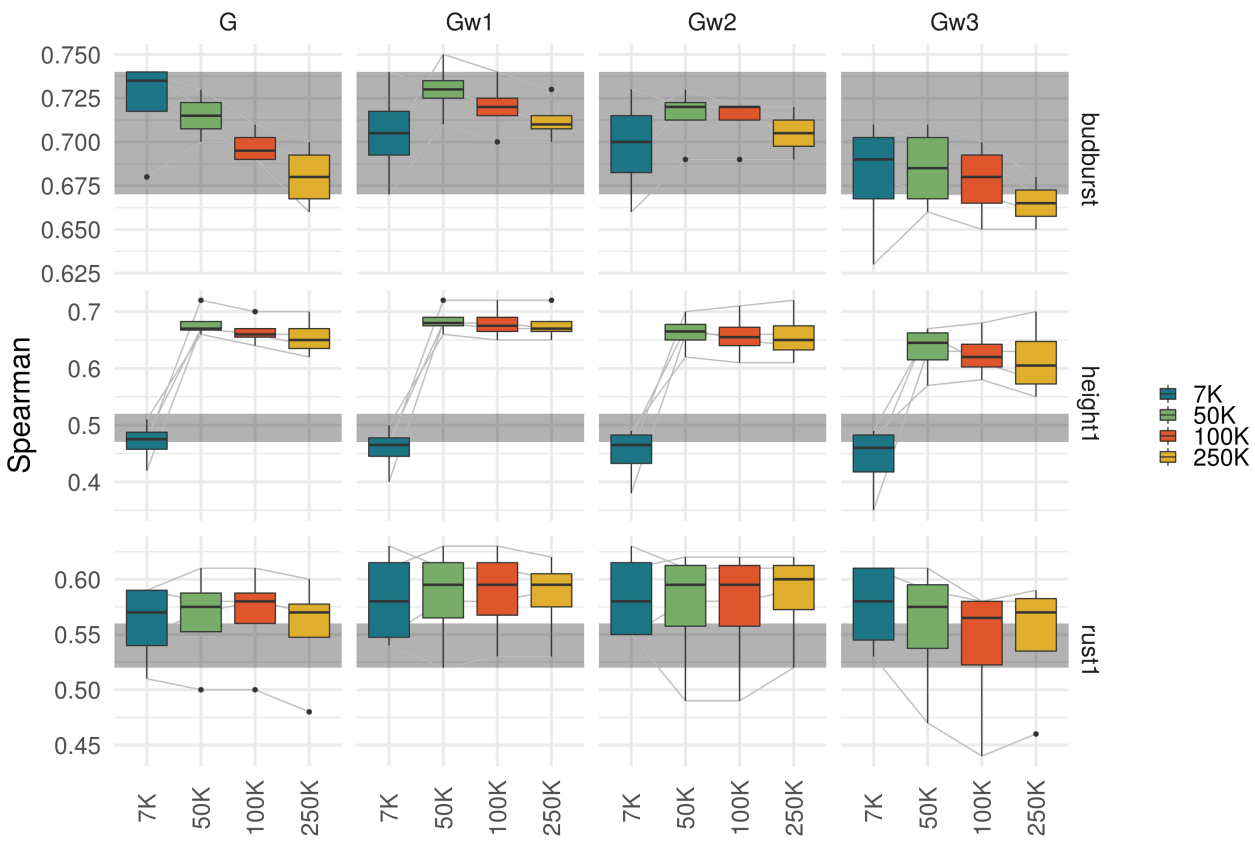


Figure 19: Marker densification impact on Spearman correlation of a single trait additive model with T25 individual cross-validation, for four genomic relationships matrices (in columns) and for four different traits (in rows) : budburst, height1, and rust1. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

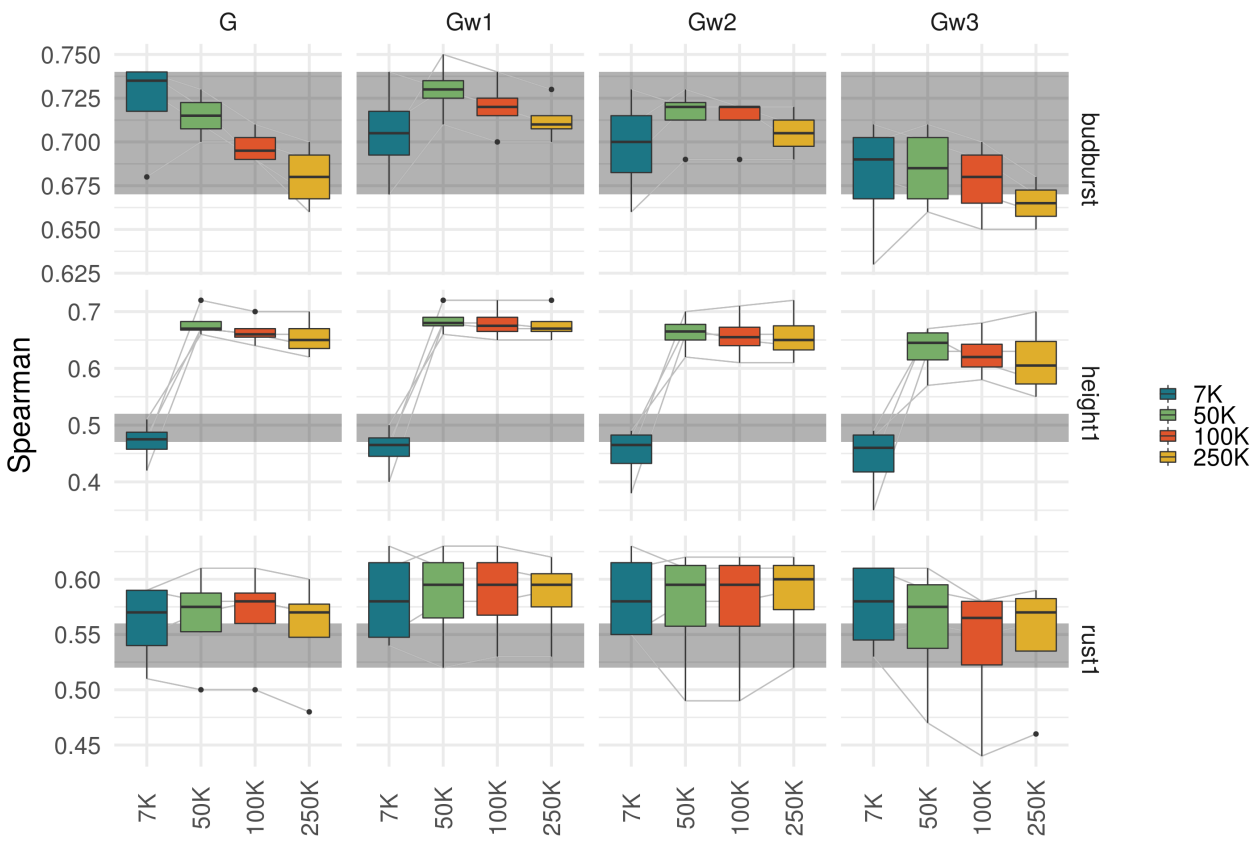


Figure 20: Marker densification impact on Spearman correlation of a single trait additive model with T25 family cross-validation, for four genomic relationships matrices (in columns) and for four different traits (in rows) : budburst, height1, and rust1. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

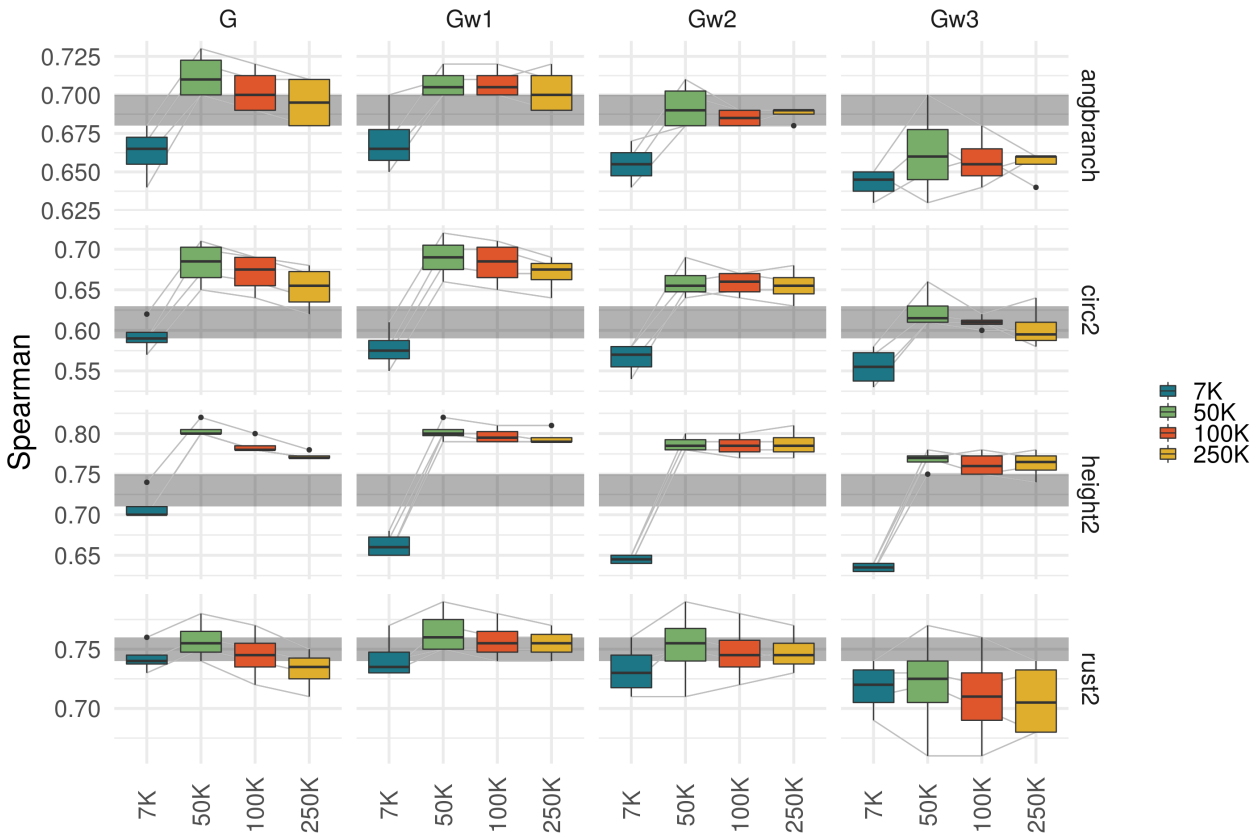


Figure 21: Marker densification impact on Spearman correlation of a single trait additive model with T25 individual cross-validation, for four genomic relationships matrices (in columns) and for four different traits (in rows) : angbranch, circ2, height2, and rust2. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

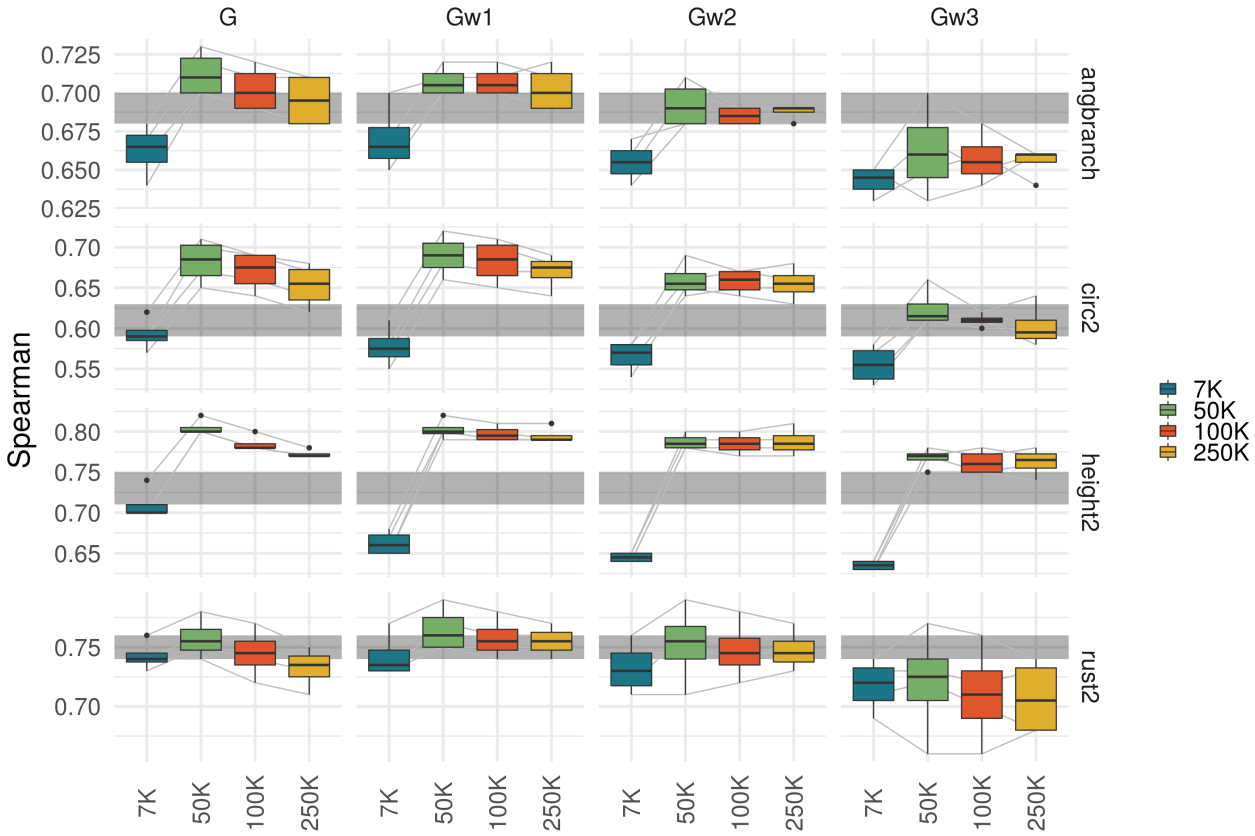


Figure 22: Marker densification impact on Spearman correlation of a single trait additive model with T25 family cross-validation, for four genomic relationships matrices (in columns) and for four different traits (in rows) : angbranch, circ2, height2, and rust2. The range of accuracies obtained with the pedigree information was represented by a grey band. The grey lines between boxplots link paired repetitions.

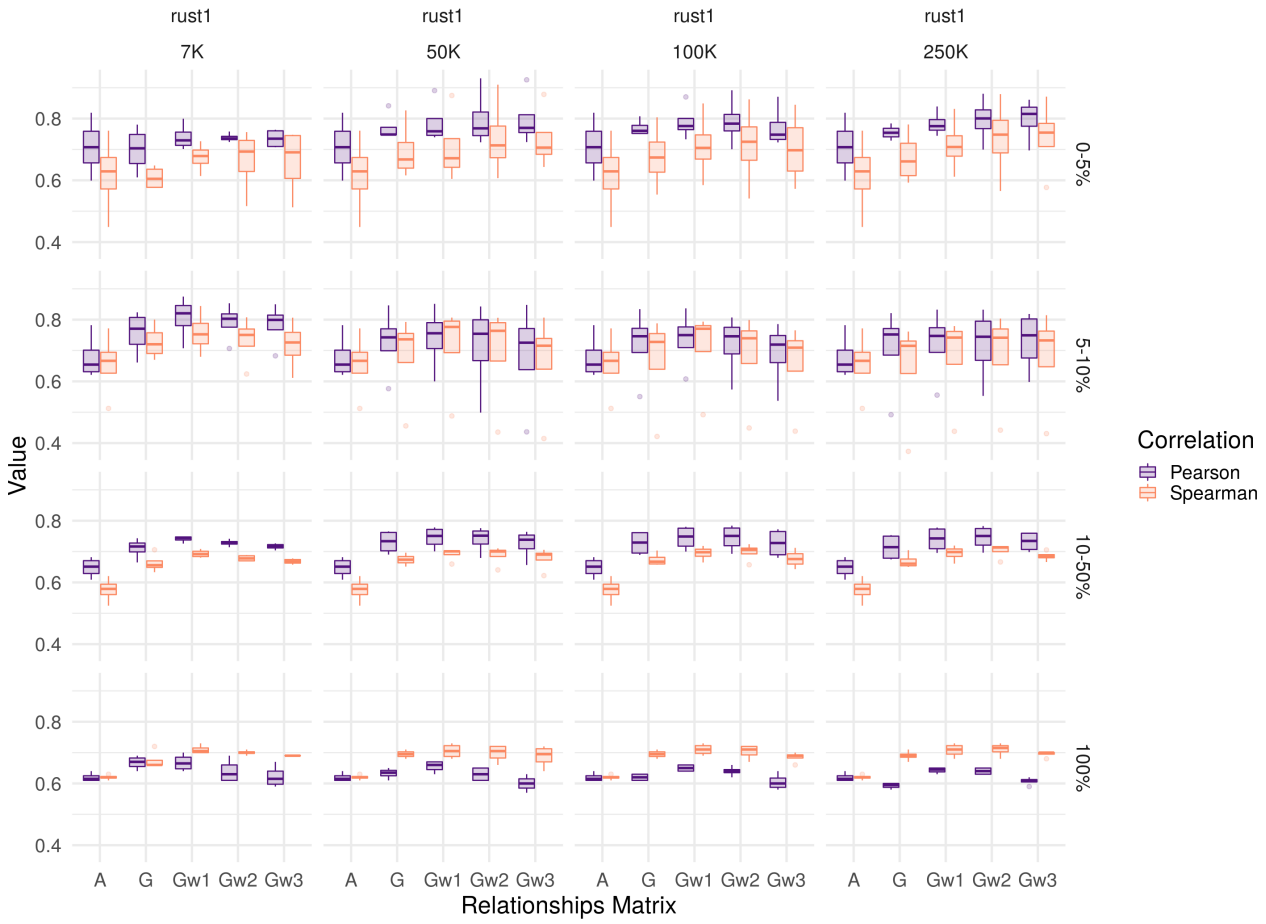


Figure 23: Comparison of Spearmans (orange) and Pearsons (Purple) correlations between phenotypes and estimated breeding values of the first year of rust evaluation. Relationship matrixes are in columns, grouped by the number of SNPs. In rows the proportion of individuals to estimate the correlations: 0-5% were the 5% best individuals, 5-10% the 5% to 10% best individuals, 10-50% the 10% to 50% best individuals based on phenotypes inside the validation population (T50 individual sampling strategy). 100% represented the Spearman and Predab values for all the individuals in the validation population.

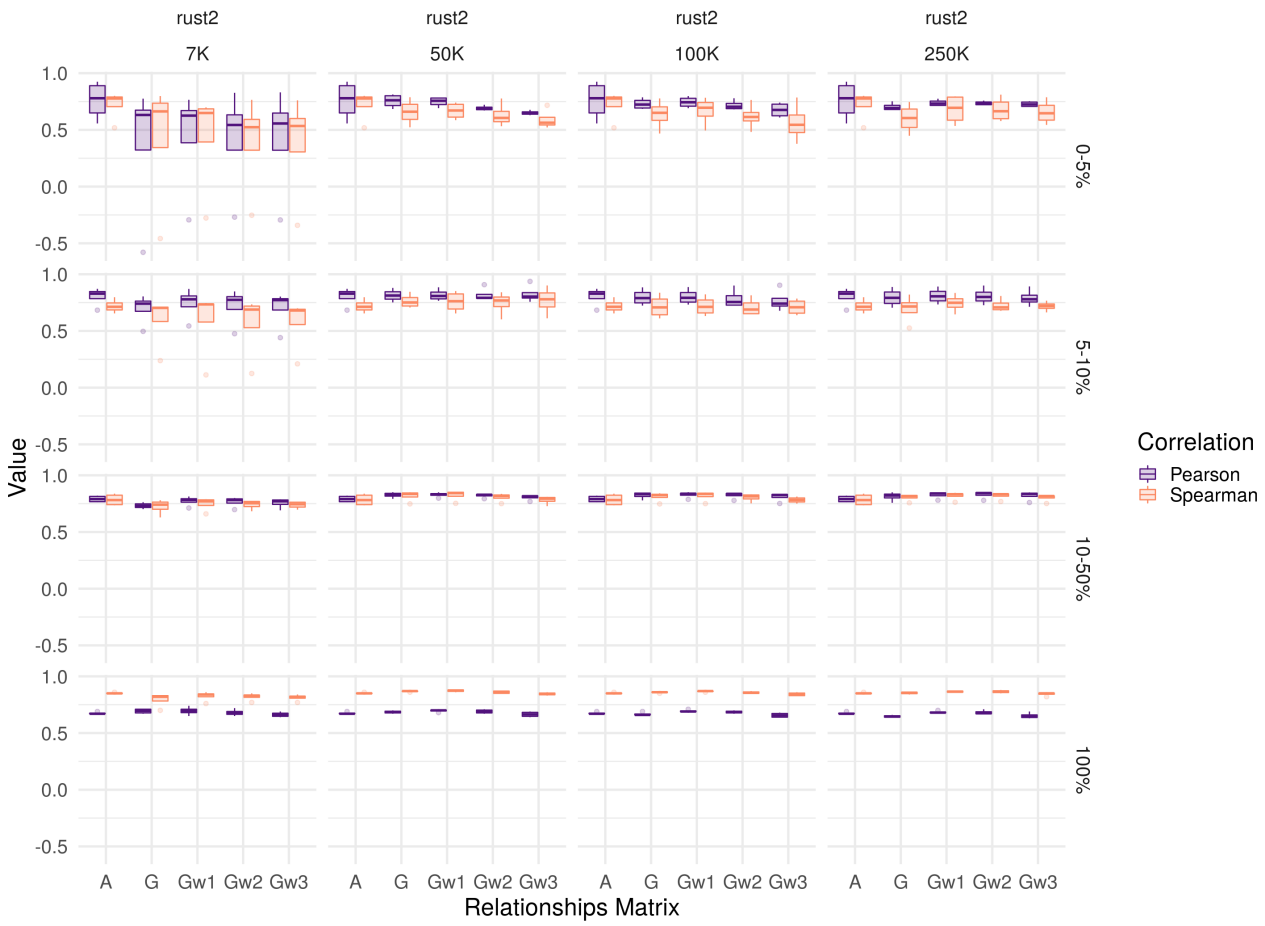


Figure 24: Comparison of Spearmans (orange) and Pearsons (Purple) correlations between phenotypes and estimated breeding values of the second year of rust evaluation. Relationship matrixes are in columns, grouped by the number of SNPs. In rows the proportion of individuals to estimate the correlations: 0-5% were the 5% best individuals, 5-10% the 5% to 10% best individuals, 10-50% the 10% to 50% best individuals based on phenotypes inside the validation population (T50 individual sampling strategy). 100% represented the Spearman and Predab values for all the individuals in the validation population.

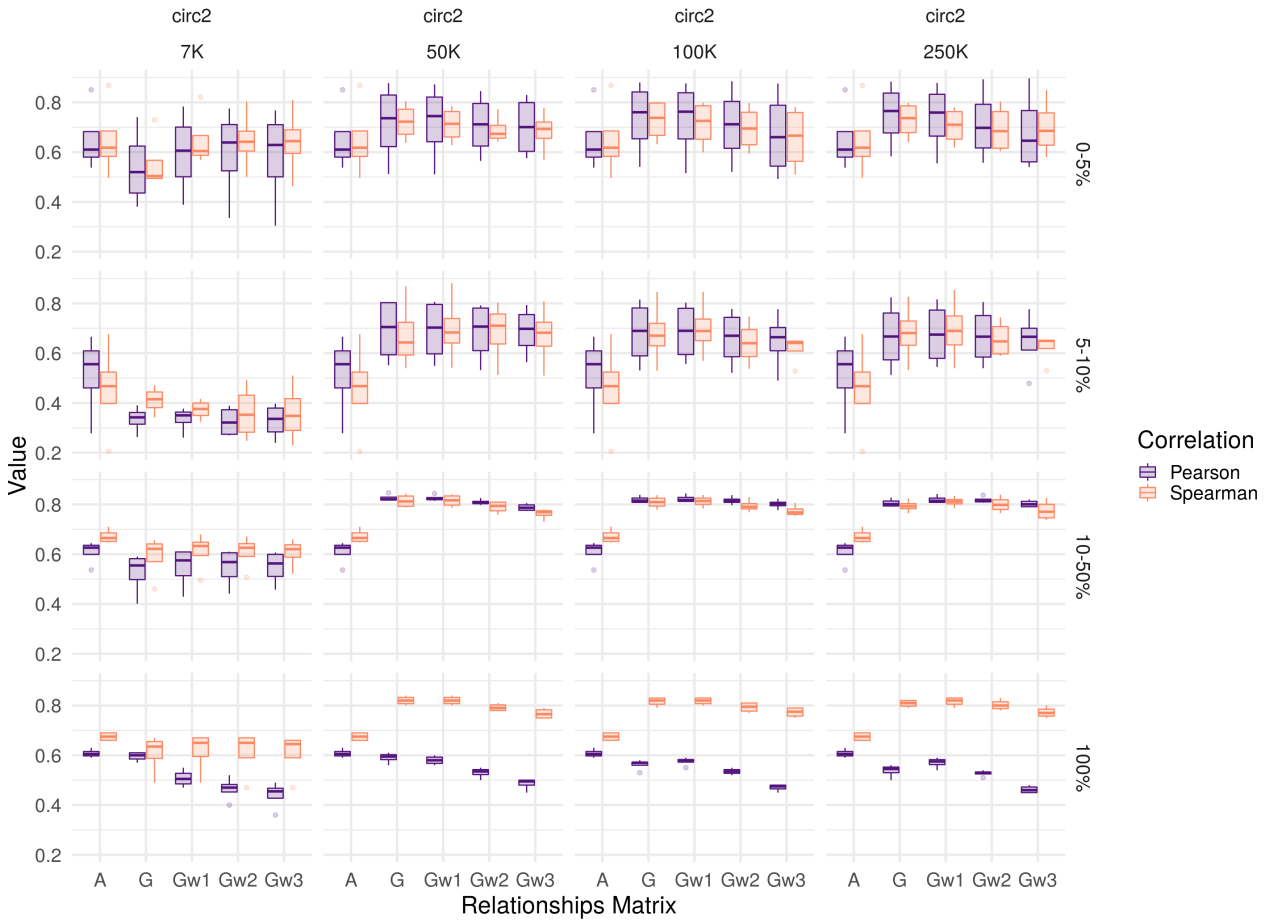


Figure 25: Comparison of Spearmans (orange) and Pearsons (Purple) correlations between phenotypes and estimated breeding values of the second year of circumference evaluation. Relationship matrixes are in columns, grouped by the number of SNPs. In rows the proportion of individuals to estimate the correlations: 0-5% were the 5% best individuals, 5-10% the 5% to 10% best individuals, 10-50% the 10% to 50% best individuals based on phenotypes inside the validation population (T50 individual sampling strategy). 100% represented the Spearman and Predab values for all the individuals in the validation population.

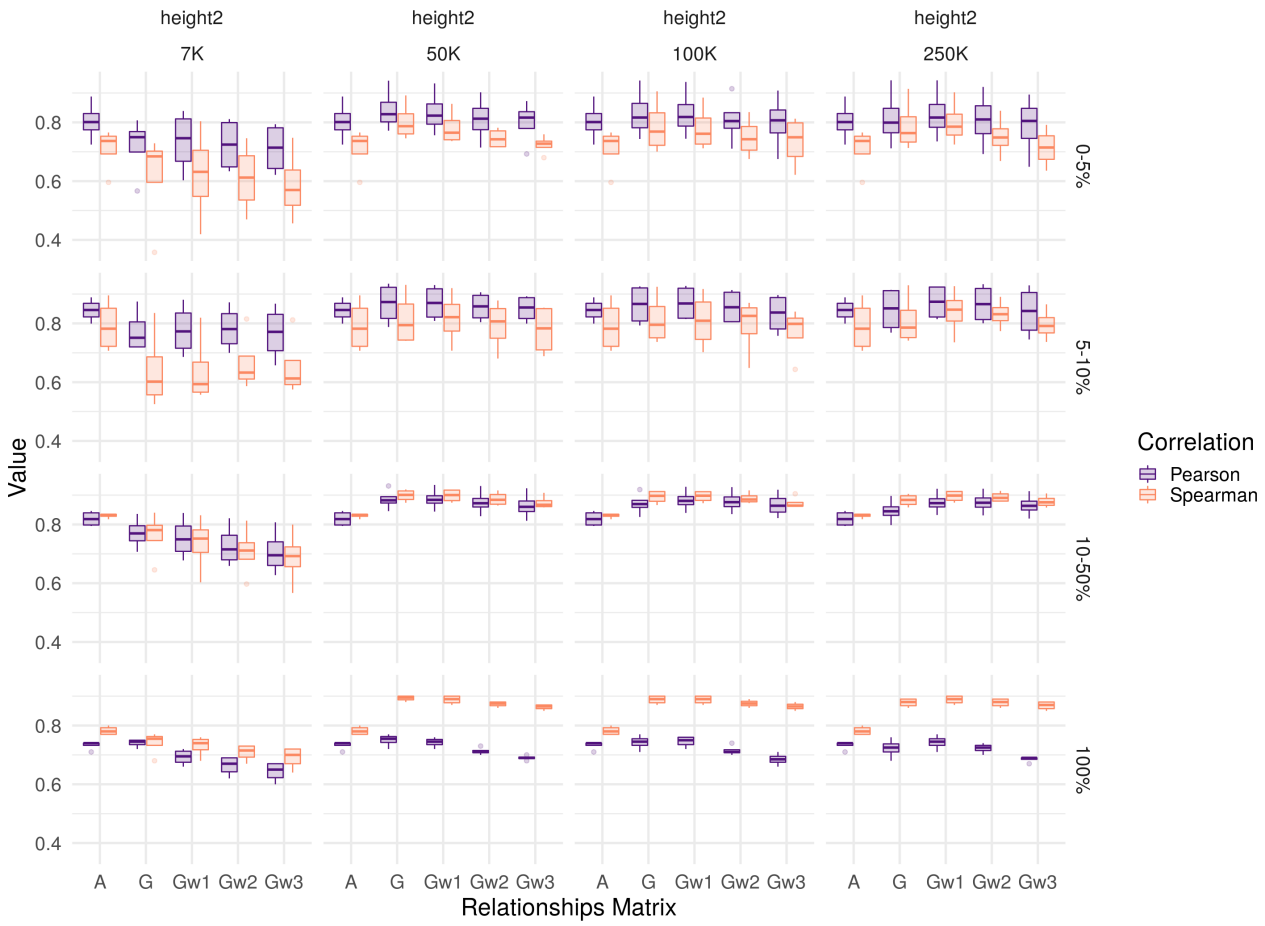


Figure 26: Comparison of Spearmans (orange) and Pearsons (Purple) correlations between phenotypes and estimated breeding values of the second year of height evaluation. Relationship matrixes are in columns, grouped by the number of SNPs. In rows the proportion of individuals to estimate the correlations: 0-5% were the 5% best individuals, 5-10% the 5% to 10% best individuals, 10-50% the 10% to 50% best individuals based on phenotypes inside the validation population (T50 individual sampling strategy). 100% represented the Spearman and Predab values for all the individuals in the validation population.

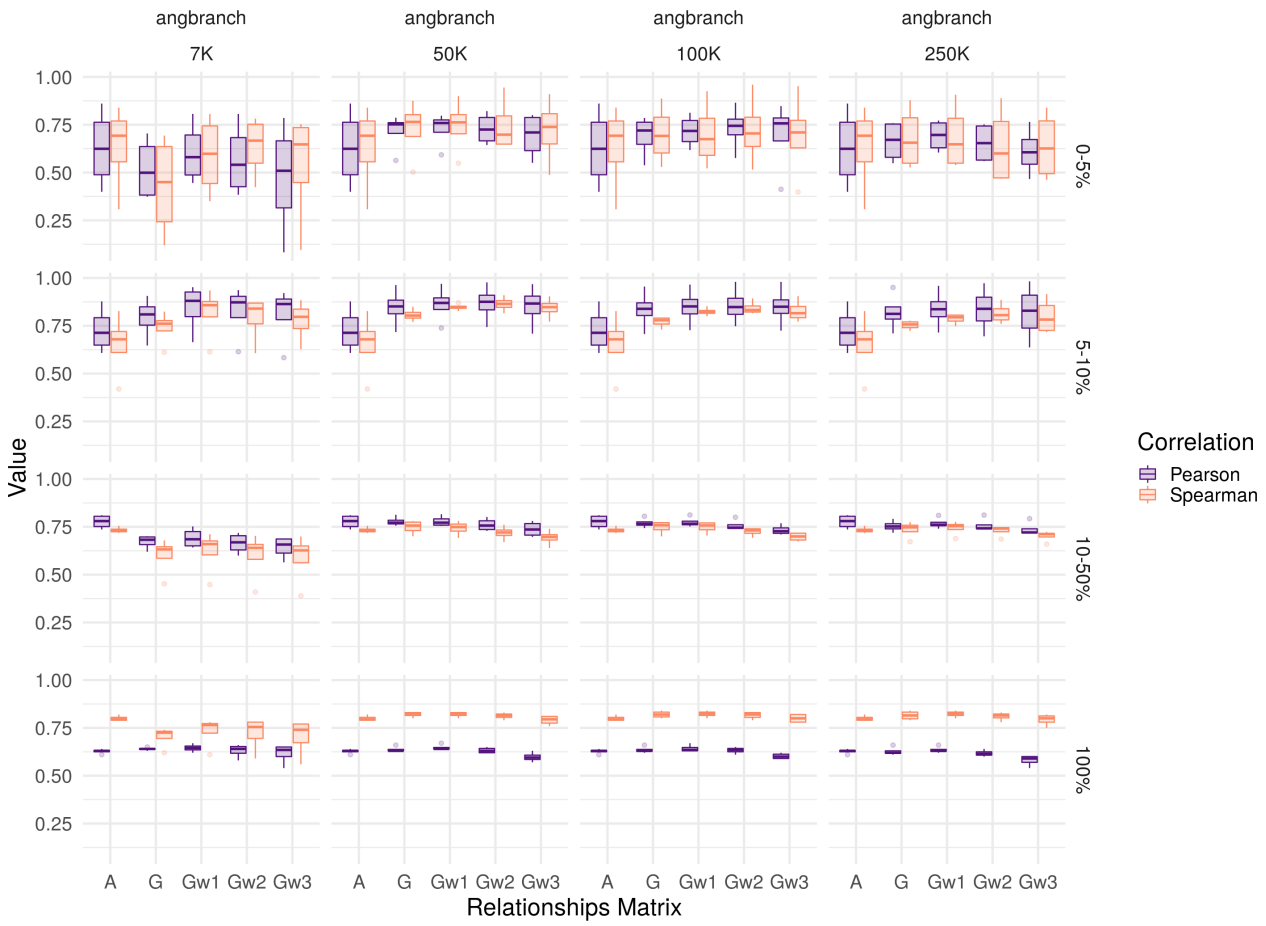


Figure 27: Comparison of Spearmans (orange) and Pearsons (Purple) correlations between phenotypes and estimated breeding values of angbranch evaluation. Relationship matrixes are in columns, grouped by the number of SNPs. In rows the proportion of individuals to estimate the correlations: 0-5% were the 5% best individuals, 5-10% the 5% to 10% best individuals, 10-50% the 10% to 50% best individuals based on phenotypes inside the validation population (T50 individual sampling strategy). 100% represented the Spearman and Predab values for all the individuals in the validation population.

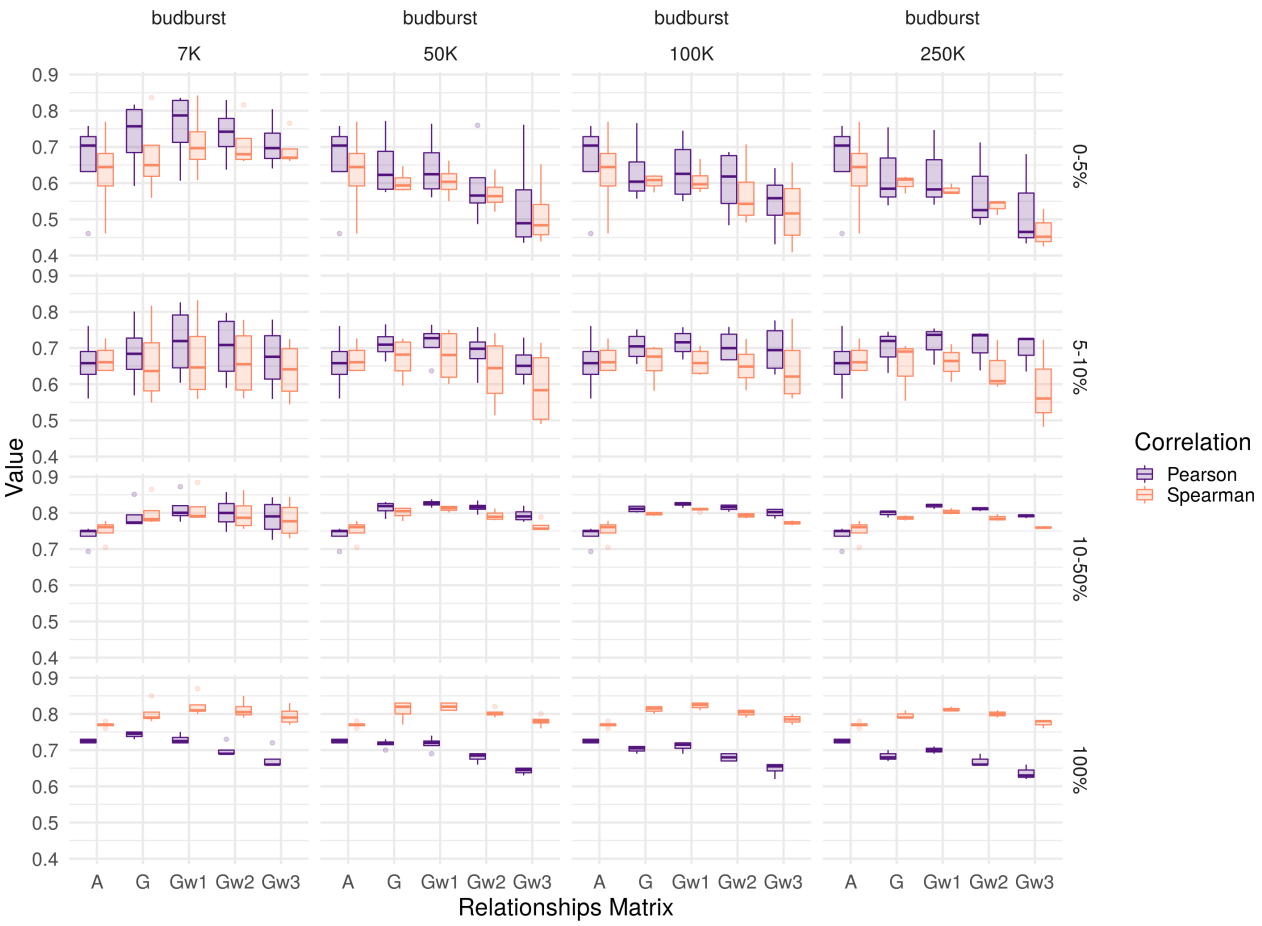


Figure 28: Comparison of Spearmans (orange) and Pearsons (Purple) correlations between phenotypes and estimated breeding values of the budburst evaluation. Relationship matrixes are in columns, grouped by the number of SNPs. In rows the proportion of individuals to estimate the correlations: 0-5% were the 5% best individuals, 5-10% the 5% to 10% best individuals, 10-50% the 10% to 50% best individuals based on phenotypes inside the validation population (T50 individual sampling strategy). 100% represented the Spearman and Predab values for all the individuals in the validation population.

Bibliography

[Dataset] (2016). Poplars and Other Fast-Growing Trees - Renewable Resources for Future Green Economies. Synthesis of Country Progress Reports.

[Dataset] (2016). Technoguide du peuplier : Le peuplier une richesse pour l'avenir

[Dataset] (2016). Working party on Taxonomy, Nomenclature and Registration

(2017). résultats de l'enquête statistique annuelle MAAF/IRSTEA sur les ventes en France de plants forestiers pour la campagne de plantation 2015-2016

Akdemir, D., Sanchez, J. I., and Jannink, J. L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel. Evol.* 47, 1–10. doi:10.1186/s12711-015-0116-6

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., and Hayes, B. J. (2016). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genet. Sel. Evol.* 48, 1–11. doi:10.1186/s12711-016-0186-0

Avendaño, S., Woolliams, J. A., and Villanueva, B. (2004). Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. *Genet. Res.* 83, 55–64. doi:10.1017/S0016672303006566

Badke, Y. M., Bates, R. O., Ernst, C. W., Fix, J., and Steibel, J. P. (2014). Accuracy of Estimation of Genomic Breeding Values in Pigs Using Low-Density Genotypes and Imputation. *G3 Genes, Genomes, Genet.* 4, 623–631. doi:10.1534/g3.114.010504

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., Fix, J., Van Tassell, C. P., et al. (2013). Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genet.* 14, 1–14. doi:10.1186/1471-2156-14-8

Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., et al. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics* 17, 1–14. doi:10.1186/s12864-016-2879-8

Bastien, C., Chenault, N., Dowkiw, A., Villar, M., Klein, E. K., and Frey, P. (2009). Interactions entre populations naturelles et cultivés: l'exemple du peuplier. *Biofutur 305 (28), 31-34.(2009)*

Beaulieu, J., Doerksen, T., Clément, S., Mackay, J., and Bousquet, J. (2014a). Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity (Edinb).* 113, 343–352. doi:10.1038/hdy.2014.36

Beaulieu, J., Doerksen, T. K., MacKay, J., Rainville, A., and Bousquet, J. (2014b). Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genomics* 15, 1–16. doi:10.1186/1471-2164-15-1048

Bernardo, R. and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47, 1082–1090. doi:10.2135/cropsci2006.11.0690

Berry, D. P., McHugh, N., Randles, S., Wall, E., McDermott, K., Sargolzaei, M., et al. (2018). Imputation of non-genotyped sheep from the genotypes of their mates and resulting progeny. *animal* 12, 191–198. doi:10.1017/S1751731117001653

[Dataset] BERTHELOT, A., BASTIEN, C., VILLAR, M., PINON, J., HEOIS, B., BOURLON, V., et al. (2005). Le Gis Peuplier, 4 ans après sa création

[Dataset] Berthelot, A., Villar, M., Pinon, J., and Breton, V. (2001). Création d'un Groupement d'Intérêt Scientifique (GIS) "Peuplier"

Bisognin, D. A. (2011). Breeding vegetatively propagated horticultural crops. *Crop Breed. Appl. Biotechnol.* 11, 35–43. doi:10.1590/S1984-70332011000500006

Blair, H. T. and Pollak, E. J. (1984). Estimation of Genetic Trend in selected Population with and Without the use of Control Population. *J. Anim. Sci.* 58, 878–886

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Bouquet, A. and Juga, J. (2013). Integrating genomic selection into dairy cattle breeding programmes: a review. *Animal* 7, 705–713

Browning, B. L. and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* 88, 173–182. doi:10.1016/j.ajhg.2011.01.010

Browning, S. R. and Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi:10.1086/521987

Burdick, J. T., Chen, W.-M., Abecasis, G. R., and Cheung, V. G. (2006). In silico method for inferring genotypes in pedigrees. *Nat. Genet.* 38, 1002–1004. doi:10.1038/ng1863

Cagelli, L., Lefèvre, F., and Others (1995). The conservation of Populus nigra L. and gene flow with cultivated poplars in Europe. *For. Genet.* 2, 135–144

Cain, N. P. and Ormrod, D. P. (1984). Hybrid vigor as indicated by early growth characteristics of Populus deltoides, P. myra, and P. x eurameucana. *Can. J. Bot.* 62, 1–8

Calus, M. and Veerkamp, R. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* 43, 26. doi:10.1186/1297-9686-43-26

Calus, M. P. L., Bouwman, A. C., Hickey, J. M., Veerkamp, R. F., and Mulder, H. A. (2014). Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. *animal* 8, 1743–1753. doi: 10.1017/S1751731114001803

Cappa, E. P. and Cantet, R. J. (2007). Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. *Can. J. For. Res.* 37, 2677–2688. doi:10.1139/X07-116

Cappa, E. P., Muñoz, F., Sanchez, L., and Cantet, R. J. C. (2015). A novel individual-tree mixed model to account for competition and environmental heterogeneity: a Bayesian approach. *Tree Genet. Genomes* 11, 120. doi:10.1007/s11295-015-0917-3

Castellani, E., Freccero, V., Lapietra, G., and Castellani, E and Freccero, V and Lapietra, G. (1967). Proposta di una scala di differenziazione delle gemme fogliari del pioppo utile per gli interventi antiparas sitari. *Plant Biosyst.* 101, 355–360. doi: 10.1080/11263506709426301

Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., and Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. A case of study in advanced wheat breeding lines. *PLoS One* 12, e0169606

Cervera, M. T., Gusmão, J., Steenackers, M., Peleman, J., Storme, V., Broeck, A. V., et al. (1996). Identification of AFLP molecular markers for resistance against Melampsora larici-populina in Populus. *Theor. Appl. Genet.* 93, 733–737

Chamaillard, S., Fichot, R., Vincent-Barbaroux, C., Bastien, C., Depierreux, C., Dreyer, E., et al. (2011). Variations in bulk leaf carbon isotope discrimination, growth and related leaf traits among three Populus nigra L. populations. *Tree Physiol.* 31, 1076–1087. doi:10.1093/treephys/tpr089

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi:10.1186/s13742-015-0047-8

Chu, Y., Su, X., Huang, Q., and Zhang, X. (2009). Patterns of DNA sequence variation at candidate gene loci in black poplar (Populus nigra L.) as revealed by single nucleotide polymorphisms. *Genetica* 137, 141–150. doi:10.1007/s10709-009-9371-1

Chud, T. C. S., Ventura, R. V., Schenkel, F. S., Carvalheiro, R., Buzanskas, M. E., Rosa, J. O., et al. (2015). Strategies for genotype imputation in composite beef cattle. *BMC Genet.* 16, 99. doi:10.1186/s12863-015-0251-7

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* 44, 4. doi:10.1186/1297-9686-44-4

Cleveland, M. A. and Hickey, J. M. (2013). Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation1. *J. Anim. Sci.* 91, 3583–3592. doi:10.2527/jas.2013-6270

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39, 859

Comstock, R. E., Robinson, H. F., and Harvey, P. H. (1949). Breeding procedure designed to make maximum use of both general and specific combining ability. *Agron. J.*

Cros, D., Denis, M., Sánchez, L., Cochard, B., Flori, A., Durand-Gasselin, T., et al. (2015). Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). *Theor. Appl. Genet.* 128, 397–410. doi:10.1007/s00122-014-2439-z

Crossa, J., Pérez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., and Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. *J. Crop Improv.* 25, 239–261. doi:10.1080/15427528.2011.558767

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi:10.1534/genetics.112.147983

Daetwyler, H. D., Hickey, J. M., Henshall, J. M., Dominik, S., Gredler, B., Van Der Werf, J. H., et al. (2010). Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Anim. Prod. Sci.* 50, 1004–1010. doi:10.1071/AN10096

Daetwyler, H. D., Villanueva, B., Bijma, P., and Woolliams, J. A. (2007). Inbreeding in genome-wide selection. *J. Anim. Breed. Genet.* 124, 369–376. doi:10.1111/j.1439-0388.2007.00693.x

Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A., and Goddard, M. E. (2011). Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing. *Genetics* 189, 317–327. doi:10.1534/genetics.111.128082

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–8. doi:10.1093/bioinformatics/btr330

de Almeida Filho, J. E., Guimarães, J. F. R., e Silva, F. F., de Resende, M. D. V., Muñoz, P., Kirst, M., et al. (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity (Edinb).* 117, 33

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi:10.1534/genetics.112.143313

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385. doi:10.1534/genetics.109.101501

Dekkers, J. C. M. and Hospital, F. (2002). Multifactorial Genetics The Use of Molecular Genetics in the Improvement of Agricultural Populations. *Nat. Rev. Genet.* 3, 22–32. doi:10.1038/nrg701

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi:10.1038/ng.806

Desta, Z. A. and Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* 19, 592–601. doi:10.1016/j.tplants.2014.05.006

Dickman, D. L. and Kuzovkina, J. (2014). Poplars and Willows of the World, With Emphasis on Silviculturally Important Species. In *Poplars willows Trees Soc. Environ.*, eds. J. Isebrands and J. Richardson (The Food and Agriculture Organization of the United Nations and CABI). 8–91

Dillen, S. Y., Storme, V., Marron, N., Bastien, C., Neyrinck, S., Steenackers, M., et al. (2009). Genomic regions involved in productivity of two interspecific poplar families in Europe. 1. Stem height, circumference and volume. *Tree Genet. Genomes* 5, 147–164. doi:10.1007/s11295-008-0175-8

Dos Santos, J. P. R., De Castro Vasconcellos, R. C., Pires, L. P. M., Balestre, M., and Von Pinho, R. G. (2016). Inclusion of dominance effects in the multivariate GBLUP model. *PLoS One* 11, 1–21. doi:10.1371/journal.pone.0152045

Douglas, C. J. (2015). Populus as a Model Tree. In *Plant Genet. Genomics Crop. Model.* 1–26. doi:10.1007/7397

Dowkiw, A. and Bastien, C. (2004). Characterization of two major genetic factors controlling quantitative resistance to Melampsora larici-populina leaf rust in hybrid poplars: strain specificity, field expression, combined effects, and relationship with a defeated qualitative resistance ge. *Phytopathology* 94, 1358–1367

Dowkiw, A. and Bastien, C. (2007). Presence of defeated qualitative resistance genes frequently has major impact on quantitative resistance to Melampsora larici-populina leaf rust in P. x interamericana hybrid poplars. *Tree Genet. Genomes* 3, 261–274. doi:10.1007/s11295-006-0062-0

Dowkiw, A., Chenault, N., Guérin, V., Borel, C., Bastien, C., and Villar, M. (2014). Post-pollination paternal reproductive success in Populus nigra: A male affair. *Tree Genet. Genomes* 10, 565–572. doi:10.1007/s11295-014-0704-6

Dowkiw, A., Voisin, E., and Bastien, C. (2010). Potential of Eurasian poplar rust to overcome a major quantitative resistance factor. *Plant Pathol.* 59, 523–534. doi:10.1111/j.1365-3059.2010.02277.x

Eding, H. and Meuwissen, T. H. E. (2001). Marker-based estimates of between and within population kinships for the conservation of genetic diversity. *J. Anim. Breed. Genet.* 118, 141–159. doi:10.1046/j.1439-0388.2001.00290.x

Eggen, A. (2012). The development and application of genomic selection as a new breeding paradigm. *Anim. Front.* 2, 10–15

Einspahr, D. W., Van Buijtenen, J. P., and Peckham, J. R. (1963). Natural variation and heritability in triploid aspen. *Silvae Genet* 12, 51–58

El Malki, R. (2013). *Architecture génétique des caractères cibles pour la culture du peuplier en taillis à courte rotation.* Ph.D. thesis, Université d'Orléans

Elbasyoni, I. S., Lorenz, A. J., Guttieri, M., Frels, K., Baenziger, P. S., Poland, J., et al. (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Sci.* doi:10.1016/j.plantsci.2018.02.019

Every, A. D. and Wiens, D. (1971). Triploidy in Utah aspen. *Madrono* 21, 138–147

Eynard, S. E., Croiseau, P., Laloë, D., Fritz, S., Calus, M. P. L., and Restoux, G. (2017). Which Individuals To Choose To Update the Reference Population? Minimizing the Loss of Genetic Diversity in Animal Genomic Selection Programs. *G3: Genes|Genomes|Genetics* 8, g3.1117.2017. doi:10.1534/g3.117.1117

Fabbrini, F., Gaudet, M., Bastien, C., Zaina, G., Harfouche, A., Beritognolo, I., et al. (2012). Phenotypic plasticity, QTL mapping and genomic characterization of bud set in black poplar. *BMC Plant Biol.* 12, 47. doi:10.1186/1471-2229-12-47

Faivre-Rampant, P., Zaina, G., Jorge, V., Giacomello, S., Segura, V., Scalabrin, S., et al. (2016). New resources for genetic studies in Populus nigra : genome-wide SNP discovery and development of a 12k Infinium array. *Mol. Ecol. Resour.* 16, 1023–1036. doi:10.1111/1755-0998.12513

Falconer, D. S. (1981). *Introduction to quantitative genetics* (Longman.), 2nd edn.

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinburgh* 52, 399–433. doi:DOI:10.1017/S0080456800012163

Forni, S., Aguilar, I., and Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* 43, 1. doi:10.1186/1297-9686-43-1

Foussadier, R. (2003). Les systèmes racinaires des arbres de la ripisylve: effets des contraintes physiques et exemples. *Les forêts riveraines des cours d'eau, écologie, Fonct. Gest.* , 124–133

Frischknecht, M., Neuditschko, M., Jagannathan, V., Drögemüller, C., Tetens, J., Thaller, G., et al. (2014). Imputation of sequence level genotypes in the Franches-Montagnes horse breed. *Genet. Sel. Evol.* 46, 63. doi:10.1186/s12711-014-0063-7

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., and El-Kassaby, Y. A. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics* 16, 1–16. doi:10.1186/s12864-015-1597-y

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., and 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, 743–753. doi:10.1534/g3.115.025957

Gastine, F., Berthelot, A., Servant, H., Roy, B., and Bouvet, A. (2003). Is the phytosanitary protection of the cultivar'Beaupre'effective. *Informations-Foret, Afocel (667)*

Gaunt, T. R., Rodríguez, S., and Day, I. N. (2007). Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool 'CubeX'. *BMC Bioinformatics* 8, 428. doi:10.1186/1471-2105-8-428

Gebreselassie, M. N., Ader, K., Boizot, N., Millier, F., Charpentier, J. P., Alves, A., et al. (2017). Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from Populus nigra (L.) natural populations. *Ind. Crops Prod.* 107, 159–171. doi:10.1016/j.indcrop.2017.05.013

KEY: Gebreselassie2017
ANNOTATION: Discussion et introduction. le cote nirs et prediction plus discuté de la qualité du bois et interaction de celle-ci.

Geraldes, A., Difazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., et al. (2013). A 34K SNP genotyping array for Populus trichocarpa: Design, application to the study of natural populations and transferability to other Populus species. *Mol. Ecol. Resour.* 13, 306–323. doi:10.1111/1755-0998.12056

Gianola, D. and van Kaam, J. B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Thompson, R., Butler, D., Cherry, M., et al. (2008). ASReml user guide release 3.0. *VSN Int Ltd*

Goddard, M. E. and Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10, 381–391. doi:10.1038/nrg2575

Gómez-Romano, F., Villanueva, B., Fernández, J., Woolliams, J. A., and Pong-Wong, R. (2016). The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. *Genetics Selection Evolution* 48, 2

González-Camacho, J. M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125, 759–771

Gorjanc, G., Battagin, M., Dumasy, J.-F., Antolin, R., Gaynor, R. C., and Hickey, J. M. (2017). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Sci.* 57, 216–228

Gowda, M., Zhao, Y., Würschum, T., Longin, C. F. H., Miedaner, T., Ebmeyer, E., et al. (2014). Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. *Heredity (Edinb).* 112, 552

Grattapaglia, D. (2014). Breeding Forest Trees by Genomic Selection: Current Progress and the Way Forward. In *Genomics Plant Genet. Resour.* (Dordrecht: Springer Netherlands), July. 651–682. doi:10.1007/978-94-007-7572-5_26

Grattapaglia, D. (2017). Status and Perspectives of Genomic Selection in Forest Tree Breeding. In *Genomic Sel. Crop Improv.* (Cham: Springer International Publishing). 199–249. doi:10.1007/978-3-319-63170-7_9

Grattapaglia, D. and Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7, 241–255. doi:10.1007/s11295-010-0328-4

Guet, J. and Bastien, C. (2011). Structuration géographique de la variabilité génétique de la phénologie de croissance et de l ' efficience d ' utilisation de l ' eau chez le peuplier noir (Populus nigra L .) Structuration géographique de la variabilité génétique de la phénologie de cro

Guet, J., Fichot, R., Lédée, C., Laurans, F., Cochard, H., Delzon, S., et al. (2015). Stem xylem resistance to cavitation is related to xylem structure but not to growth and water-use efficiency at the within-population level in Populus nigra L. *J. Exp. Bot.* 66, 4643–4652. doi:10.1093/jxb/erv232

Guillaume, F., Fritz, S., Boichard, D., and Druet, T. (2008). Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle (Open Access publication). *Genet. Sel. Evol.* 40, 91–102. doi:10.1186/1297-9686-40-1-91

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet.* 15, 1–7. doi:10.1186/1471-2156-15-30

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi:10.1534/genetics.107.081190

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12. doi:10.1186/1471-2105-12-186

Hayes, B., Bowman, P., Chamberlain, A., and Goddard, M. (2009a). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92, 433–443. doi:10.3168/jds.2008-1646

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., and Goddard, M. E. (2009b). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41, 1–9. doi:10.1186/1297-9686-41-51

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010). Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. *PLoS Genet.* 6, e1001139. doi:10.1371/journal.pgen.1001139

Heffner, E. L., Jannink, J. L., Iwata, H., Souza, E., and Sorrells, M. E. (2011). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51, 2597–2606. doi:10.2135/cropsci2011.05.0253

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi:10.2135/cropsci2008.08.0512

Heidaritabar, M., Vereijken, A., Muir, W. M., Meuwissen, T., Cheng, H., Megens, H. J., et al. (2014). Systematic differences in the response of genetic variation to pedigree and genome-based selection methods. *Heredity (Edinb).* 113, 503–513. doi:10.1038/hdy.2014.55

Henderson, C. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31, 423–447

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Sci.* 52, 146. doi:10.2135/cropsci2011.06.0297

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Sci.* 54, 1476–1488. doi:10.2135/cropsci2013.03.0195

Hickey, J. M. and Gorjanc, G. (2012). Simulated Data for Genomic Selection and Genome-Wide Association Studies Using a Combination of Coalescent and Gene Drop Methods. *G3 Genes, Genomes, Genet.* 2, 425–427. doi:10.1534/g3.111.001297

Hill, W. G. and Weir, B. S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet. Res. (Camb).* 93, 47–64. doi:10.1017/S0016672310000480

Howard, D. M., Pong-Wong, R., Knap, P. W., Kremer, V. D., and Woolliams, J. A. (2018). Selective advantage of implementing optimal contributions selection and timescales for the convergence of long-term genetic contributions. *Genet. Sel. Evol.* 50, 24

Howard, R., Carriquiry, A. L., and Beavis, W. D. (2014). Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. *G3: Genes|Genomes|Genetics* 4, 1027–1046. doi:10.1534/g3.114.010298

Howe, G. T., Saruul, P., Davis, J., and Chen, T. H. (2000). Quantitative genetics of bud phenology, frost damage, and winter survival in an F2family of hybrid poplars. *Theor. Appl. Genet.* 101, 632–642. doi:10.1007/s001220051525

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* 5, e1000529. doi:10.1371/journal.pgen.1000529

Huang, W. and Mackay, T. F. (2016). The Genetic Architecture of Quantitative Traits Cannot Be Inferred From Variance Component Analysis. *bioRxiv* , 041434doi:10.1101/041434

Huang, Y., Hickey, J. M., Cleveland, M. A., and Maltecca, C. (2012). Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genet. Sel. Evol.* 44, 25. doi:10.1186/1297-9686-44-25

Isidro, J., Jannink, J. L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 128, 145–158. doi:10.1007/s00122-014-2418-4

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., et al. (2015). Genomic selection in maritime pine. *Plant Sci.* 242, 108–119. doi:10.1016/j.plantsci.2015.08.006

Isik, F. and Toplu, F. (2004). Variation in juvenile traits of natural black poplar (Populus nigra L.) clones in Turkey. *New For.* 27, 175–187

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics* 9, 166–177. doi:10.1093/bfgp/elq001

Jia, Y. and Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192, 1513–1522. doi:10.1534/genetics.112.144246

Jiang, J., Shen, B., O'Connell, J. R., VanRaden, P. M., Cole, J. B., and Ma, L. (2017). Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *BMC Genomics* 18, 1–13. doi:10.1186/s12864-017-3821-4

Johnsson, H. (1942). Cytological studies of triploid progenies of Populus tremula. *Hereditas* 28, 306–312

Johnston, J., Kistemaker, G., and Sullivan, P. G. (2011). Comparison of Different Imputation Methods. *Interbull Bull.* , 25–33

Jorge, V., Dowkiw, A., Faivre-Rampant, P., Bastien, C., Faivre-Rampant, P., and Basrtien, C. (2005). Genetic architecture of qualitative and quantitative Melampsora larici-populina leaf rust resistance in hybrid poplar: genetic mapping and QTL detection. *New Phytol.* 167, 113–127. doi:10.1111/j.1469-8137.2005.01424.x

Kadarmideen, H. N., Thompson, R., Coffey, M. P., and Kossaibati, M. A. (2003). Genetic parameters and evaluations from single-and multiple-trait analysis of dairy cow fertility and milk production. *Livest. Prod. Sci.* 81, 183–195

Kainer, D., Stone, E. A., Padovan, A., Foley, W. J., and Külheim, C. (2018). Accuracy of Genomic Prediction for Foliar Terpene Traits in Eucalyptus polybractea. *G3: Genes/Genomes/Genetics* 8, 2573–2583. doi:10.1534/g3.118.200443

Kemperman, J. A. and Barnes, B. V. (1976). Clone size in American aspens. *Can. J. Bot.* 54, 2603–2607

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proc. R. Soc. Lond. B* 143, 103–113

Kizilkaya, K., Fernando, R. L., and Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.* 88, 544–551. doi:10.2527/jas.2009-2064

Lajoie, M. J., Rovner, A. J., Goodman, D. B., Aerni, H.-R., Haimovich, A. D., Kuznetsov, G., et al. (2013). Genomically Recoded Organisms Expand Biological Functions. *Science (80-.).* 342, 357–360. doi:10.1126/science.1241459

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92, 4656–4663. doi:10.3168/jds.2009-2061

Legionnet, A., Muranty, H., and Lefèvre, F. (1999). Genetic variation of the riparian pioneer tree species Populus nigra. II. Variation in susceptibility to the foliar rust Melampsora larici-populina. *Heredity (Edinb).* 82, 318–327. doi:10.1038/sj.hdy.6884880

Lenz, P. R., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M., and 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, 1–17. doi:10.1186/s12864-017-3715-5

Li, C., Weeks, D., and Chakravarti, A. (1993). Similarity of dna fingerprints due to chance and relatedness. *Human heredity* 43, 45–52

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr.*

Lima, B. M. d. (2014). *Bridging genomics and quantitative genetics of Eucalyptus: genome-wide prediction and genetic parameter estimation for growth and wood properties using high-density SNP data.* Ph.D. thesis, Universidade de São Paulo

Lopes, M. S., Silva, F. F., Harlizius, B., Duijvesteijn, N., Lopes, P. S., Guimarães, S. E. F., et al. (2013). Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genet.* 14. doi:10.1186/1471-2156-14-92

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., et al. (2011). *Genomic Selection in Plant Breeding. Knowledge and Prospects.*, vol. 110. doi:10.1016/B978-0-12-385531-2.00002-5

Lund, M. S., De Roos, A. P., De Vries, A. G., Druet, T., Ducrocq, V., Fritz, S., et al. (2011). A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet. Sel. Evol.* 43, 1–8. doi:10.1186/1297-9686-43-43

Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch, H. G., et al. (2013). Relatedness and genotype x environment interaction affect prediction accuracies in genomic selection: a study in cassava. *Crop Sci.* 53, 1312–1325

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* 152, 1753–1766

Mäntysaari, E. A., Liu, Z., and VanRaden, P. (2010). Interbull validation test for genomic evaluations. *Interbull Bull.* , 17

Marchal, A., Legarra, A., Tisné, S., Carasco-Lacombe, C., Manez, A., Suryana, E., et al. (2016). Multivariate genomic model improves analysis of oil palm (Elaeis guineensis Jacq.) progeny tests. *Mol. Breed.* 36, 1–13. doi:10.1007/s11032-015-0423-1

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511. doi:10.1038/nrg2796

Marron, N., Bastien, C., Sabatti, M., Taylor, G., and Ceulemans, R. (2006). Plasticity of growth and sylleptic branchiness in two poplar families grown at three sites across Europe. *Tree Physiol.* 26, 935–946. doi:10.1093/treephys/26.7.935

Marroni, F., Pinosio, S., Zaina, G., Fogolari, F., Felice, N., Cattonaro, F., et al. (2011). Nucleotide diversity and linkage disequilibrium in Populus nigra cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genet. Genomes* 7, 1011–1023. doi:10.1007/s11295-011-0391-5

Martini, J. W., Gao, N., Cardoso, D. F., Wimmer, V., Erbe, M., Cantet, R. J., et al. (2017). Genomic prediction with epistasis models: On the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinformatics* 18, 1–16. doi:10.1186/s12859-016-1439-1

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

Meuwissen, T. and Goddard, M. (2010). The Use of Family Relationships and Linkage Disequilibrium to Impute Phase and Missing Genotypes in Up to Whole-Genome Sequence Density Genotypic Data. *Genetics* 185, 1441–1449. doi:10.1534/genetics.110.113936

Meuwissen, T. H., Solberg, T. R., Shepherd, R., and Woolliams, J. A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol.* 41, 1–10. doi:10.1186/1297-9686-41-2

Meuwissen, T. H. E. and Goddard, M. E. (1997). Estimation of effects of quantitative trait loci in large complex pedigrees. *Genetics* 146, 409–416

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi:11290733

Miller, A. J. and Gross, B. L. (2011). From forest to field: Perennial fruit crop domestication. *Am. J. Bot.* 98, 1389–1414. doi:10.3732/ajb.1000522

Moghaddar, N. and van der Werf, J. H. (2017). Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. *J. Anim. Breed. Genet.* 134, 453–462. doi:10.1111/jbg.12287

Montesinos-Lopez, O. A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C. M., and Martín-Vallejo, J. (2018). Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. *G3: Genes, Genomes, Genetics* , g3–200728

Moore, J. K., Manmathan, H. K., Anderson, V. A., Poland, J. A., Morris, C. F., and Haley, S. D. (2017). Improving genomic prediction for pre-harvest sprouting tolerance in wheat by weighting large-effect quantitative trait loci. *Crop Sci.* 57, 1315–1324. doi:10.2135/cropsci2016.06.0453

Muir, W. (2007). Comparison of genomic and traditional blup-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics* 124, 342–355

[Dataset] Muñoz, F. and Sanchez, L. (2018). breedR: Statistical Methods for Forest Genetic Resources Analysts

Muñoz, P. R., Resende, M. F., Gezan, S. A., Resende, M. D. V., de los Campos, G., Kirst, M., et al. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* 198, 1759–1768. doi:10.1534/genetics.114.171322

Nakaya, A. and Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Ann. Bot.* 110, 1303–1316. doi:10.1093/aob/mcs109

Nepveu, G., Keller, R., and Others (1978). Sélection juvénile pour la qualité du bois chez certains peupliers noirs. In *Ann. des Sci. For.* (EDP Sciences), vol. 35, 69–92

Neves, H. H. R., Carvalheiro, R., and Queiroz, S. A. (2012). A comparison of statistical methods for genomic selection in a mice population. *BMC Genet.* 13, 100

Nishio, M. and Satoh, M. (2014). Parameters affecting genome simulation for evaluating genomic selection method. *Anim. Sci. J.* 85, 879–887. doi:10.1111/asj.12224

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3: Genes|Genomes|Genetics* , g3.200311.2018doi:10.1534/g3.118.200311

Novaes, E., Osorio, L., Drost, D. R., Miles, B. L., Boaventura-Novaes, C. R. D., Benedict, C., et al. (2009). Quantitative genetic analysis of biomass and wood chemistry of Populus under different nitrogen levels. *New Phytol.* 182, 878–890

Ornella, L., Pérez, P., Tapia, E., González-Camacho, J. M., Burgueño, J., Zhang, X., et al. (2014). Genomic-enabled prediction with classification algorithms. *Heredity (Edinb).* 112, 616

Paffetti, D., Travaglini, D., Labriola, M., Buonamici, A., Bottalico, F., Materassi, A., et al. (2018). Land use and wind direction drive hybridization between cultivated poplar and native species in a Mediterranean floodplain environment. *Sci. Total Environ.* 610, 1400–1412. doi:10.1016/j.scitotenv.2017.08.238

Patry, C. and Ducrocq, V. (2011a). Accounting for genomic pre-selection in national BLUP evaluations in dairy cattle. *Genet. Sel. Evol.* 43, 30. doi:10.1186/1297-9686-43-30

Patry, C. and Ducrocq, V. (2011b). Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci.* 94, 1011–1020

Pegard, M., Rogier, O., Bérard, A., Faivre-Rampant, P., Le Paslier, M.-C., Bastien, C., et al. (2018). Sequence Imputation from Low Density Single Nucleotide Polymorphism Panel in a Black Poplar Breeding population. *bioRxiv* Submitted

Pérez, P., de los Campos, G., Crossa, J., and Gianola, D. (2010). Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. *Plant Genome J.* 3, 106. doi:10.3835/plantgenome2010.04.0005

Perron, M., DeBlois, J., and Desponts, M. (2013). Use of resampling to assess optimal subgroup composition for estimating genetic parameters from progeny trials. *Tree Genet. genomes* 9, 129–143

Pichot, C. and Tessier Du Cros, E. (1988). Estimation of genetic parameters in the European black poplar (Populus nigra L.). Consequence on the breeding strategy. *Ann. des Sci. For.* 45, 223–237

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., et al. (2016). Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Mol. Biol. Evol.* 33, 2706–2719. doi:10.1093/molbev/msw161

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news* 6, 7–11

Pospíšková, M. and Šálková, I. (2006). Population structure and parentage analysis of black poplar along the Morava River. *Can. J. For. Res.* 36, 1067–1076

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of ibd and ibs in complex trait studies. *Nature Reviews Genetics* 11, 800

Pszczola, M., Strabel, T., Mulder, H., and Calus, M. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95, 389–400. doi:10.3168/jds.2011-4338

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Rae, A. M., Street, N. R., Robinson, K. M., Harris, N., and Taylor, G. (2009). Five QTL hotspots for yield in short rotation coppice bioenergy poplar: the poplar biomass loci. *BMC Plant Biol.* 9, 23

Rambolarimanana, T., Ramamonjisoa, L., Verhaegen, D., and Tsy, J.-m. L. P. (2018). Performance of multi-trait genomic selection for Eucalyptus robusta breeding program doi:10.1007/s11295-018-1286-5

Rameau, J.-C., Mansion, D., and Dumé, G. (1989). *Flore forestière française: Région méditerranéenne*, vol. 3 (Forêt privée française)

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

Rathmacher, G., Niggemann, M., Köhnen, M., Ziegenhagen, B., and Bialozyt, R. (2010). Short-distance gene flow in Populus nigra L. accounts for small-scale spatial genetic structures: implications for in situ conservation measures. *Conserv. Genet.* 11, 1327–1338

Resende, M. D., Resende Jr, M. F., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., et al. (2012a). Genomic selection for growth and wood quality in Eucalyptus: Capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol.* 194, 116–128. doi:10.1111/j.1469-8137.2011.04038.x

Resende, M. F. R. (2012). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments (vol 193, 617, 2012). *New Phytol.* 193, 1099. doi:DOI10.1111/j.1469-8137.2011.04048.x

Resende, M. F. R. D. V., Munoz, P., Resende, M. F. R. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., et al. (2012b). Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). *Genetics* 190, 1503–1510. doi:10.1534/genetics.111.137026

Resende, R., Resende, M., Silva, F., Azevedo, C., Takahashi, E., Silva-Junior, O., et al. (2017). Assessing the expected response to genomic selection of individuals and families in eucalyptus breeding with an additive-dominant model. *Heredity* 119, 245

Ricodeau, N., Bastien, C., Fabre, B., Baubet, O., Et, P. B., Bourlon, V., et al. (2018). Caractéristiques générales des peupliers , 1–7

Riedelsheimer, C. and Melchinger, A. E. (2013). Optimizing the allocation of resources for genomic selection in one breeding cycle. *Theor. Appl. Genet.* 126, 2835–2848. doi: 10.1007/s00122-013-2175-9

Rincent, R., Charpentier, J.-P., Faivre-Rampant, P., Paux, E., Le Gouis, J., Bastien, C., et al. (2018). Phenomic selection: a low-cost and high-throughput alternative to genomic selection. *bioRxiv* , 302117

Rincent, R., Kuhn, E., Monod, H., Oury, F. X., Rousset, M., Allard, V., et al. (2017). Optimization of multi-environment trials for genomic selection based on crop models. *Theor. Appl. Genet.* 130, 1735–1752. doi:10.1007/s00122-017-2922-4

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (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, 715–728. doi:10.1534/genetics.112.141473

Rodrigues, S., Bréhéret, J.-G., Macaire, J.-J., Greulich, S., and Villar, M. (2007). In-channel woody vegetation controls on sedimentary processes and the sedimentary record within alluvial environments: a modern example of an anabranch of the River Loire, France. *Sedimentology* 54, 223–242

Rohde, A., Bastien, C., and Boerjan, W. (2011). Temperature signals contribute to the timing of photoperiodic growth cessation and bud set in poplar. *Tree Physiol.* 31, 472–482. doi:10.1093/treephys/tpr038

Romero Navarro, J. A., Phillips-Mora, W., Arciniegas-Leal, A., Mata-Quirós, A., Haiminen, N., Mustiga, G., et al. (2017). Application of Genome Wide Association and Genomic Prediction for Improvement of Cacao Productivity and Resistance to Black and Frosty Pod Diseases. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.01905

Rood, S. B., Goater, L. A., McCaffrey, D., Montgomery, J. S., Hopkinson, C., and Pearce, D. W. (2017). Growth of riparian cottonwoods: heterosis in some intersectional Populus hybrids and clonal expansion of females. *Trees* 31, 1069–1081

Roshyara, N. R., Kirsten, H., Horn, K., Ahnert, P., and Scholz, M. (2014). Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genet.* 15, 88. doi:10.1186/s12863-014-0088-5

Ruffinoni, C., Pautou, G., and Piégeay, H. (2003). Les forêts riveraines des cours d'eau. *Paris, Inst. pour le Développement For.*

Russell, J. C. and Fewster, R. M. (2009). Evaluation of the linkage disequilibrium method for estimating effective population size. In *Model. Demogr. Process. Mark. Popul.* (Springer). 291–320

Santangelo, E., Scarfone, A., Giudice, A. D., Acampora, A., Alfano, V., Suardi, A., et al. (2015). Harvesting systems for poplar short rotation coppice. *Ind. Crops Prod.* 75, 85–92. doi:10.1016/j.indcrop.2015.07.013

Santure, A. W., Stapley, J., Ball, A. D., Birkhead, T. R., Burke, T., and Slate, J. (2010). On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 snps. *Molecular Ecology* 19, 1439–1451

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15, 478. doi: 10.1186/1471-2164-15-478

Scheet, P. and Stephens, M. (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *Am. J. Hum. Genet.* 78, 629–644. doi:10.1086/502802

Schefers, J. M. and Weigel, K. A. (2012). Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. *Anim. Front.* 2, 4–9

Schulthess, A. W., Wang, Y., Miedaner, T., Wilde, P., Reif, J. C., and Zhao, Y. (2016). Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. *Theor. Appl. Genet.* 129, 273–287. doi:10.1007/s00122-015-2626-6

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477

Slavov, G. T., Difazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., et al. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytol.* 196, 713–725. doi:10.1111/j.1469-8137.2012.04258.x

Smulders, M. J. M., Beringen, R., Volosyanchuk, R., Broeck, A. V., Van der Schoot, J., Arens, P., et al. (2008). Natural hybridisation between Populus nigra L. and P. x canadensis Moench. Hybrid offspring competes for niches along the Rhine river in the Netherlands. *Tree Genet. Genomes* 4, 663–675

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *J. Anim. Sci.* 86, 2447–2454

Sorensen, D. and Kennedy, B. (1984). Estimation of genetic variances from unselected and selected populations. *Journal of Animal Science* 59, 1213–1223

Sow, M. D., Segura, V., Chamaillard, S., Jorge, V., Delaunay, A., Lafon-Placette, C., et al. (2018). Narrow-sense heritability and P ST estimates of DNA methylation in three Populus nigra L. populations under contrasting water availability. *Tree Genet. Genomes* 14, 78

Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., and Balding, D. J. (2017). Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* 49, 986–992. doi:10.1038/ng.3865

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* 91, 1011–1021. doi:10.1016/j.ajhg.2012.10.010

Stanton, B. J., Serapiglia, M. J., and Smart, L. B. (2014). The domestication and conservation of Populus and Salix genetic resources. *Poplars willows trees Soc. Environ. Wallingford, UK CAB Int.* , 124–199

Stebbins, G. L. (1959). The role of hybridization in evolution. *Proc. Am. Philos. Soc.* 103, 231–251

Stebbins, G. L. (1985). Polyploidy, hybridization, and the invasion of new habitats. *Ann. Missouri Bot. Gard.* , 824–832

Stelkens, R. and Seehausen, O. (2009). Genetic distance between species predicts novel trait expression in their hybrids. *Evolution (N. Y).* 63, 884–897. doi:10.1111/j.1558-5646.2008.00599.x

Stettler, R., Bradshaw, T., Heilman, P., and Hinckley, T. (1996). *Biology of Populus and its implications for management and conservation* (NRC Research Press)

Stettler, R. F., Fenn, R. C., Heilman, P. E., and Stanton, B. J. (1988). Populus trichocarpa x Populus deltoides hybrids for short rotation culture: variation patterns and 4-year field performance. *Can. J. For. Res.* 18, 745–753

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

Su, G., Madsen, P., Nielsen, U. S., Mäntysaari, E. A., Aamand, G. P., Christensen, O. F., et al. (2012). Genomic prediction for Nordic Red Cattle using one-step and selection index blending. *J. Dairy Sci.* 95, 909–917

Sun, C., VanRaden, P., O'Connell, J., Weigel, K., and Gianola, D. (2013). Mating programs including genomic relationships and dominance effects1. *Journal of dairy science* 96, 8014–8023

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., and 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 Biol.* 17, 110. doi:10.1186/s12870-017-1059-6

Tan, B., Grattapaglia, D., Wu, H. X., and Ingvarsson, P. K. (2018). Genomic relationships reveal significant dominance effects for growth in hybrid Eucalyptus. *Plant Sci.* 267, 84–93. doi:10.1016/j.plantsci.2017.11.011

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., et al. (2015). Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. *Front. Plant Sci.* 6, 1–11. doi:10.3389/fpls.2015.00941

Teissier, M., Larroque, H., and Robert-Granié, C. (2018). Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: A quantitative trait influenced by a major gene. *Genet. Sel. Evol.* 50, 1–12. doi:10.1186/s12711-018-0400-3

TEISSIER du CROS, E. (1977). Aperçu de la transmission héréditaire de quelques caractères juvéniles chez Populus nigra L. *Ann. Sei. For.* 34, 311–322

Tier, B. and Sölkner, J. (1993). Analysing gametic variation with an animal model. *Theor. Appl. Genet.* 85, 868–872. doi:10.1007/BF00225031

Toghiani, S., Aggrey, S. E., and Rekaya, R. (2016). Multi-generational imputation of single nucleotide polymorphism marker genotypes and accuracy of genomic selection. *animal* 10, 1077–1085. doi:10.1017/S1751731115002906

Toro, M. A. and Varona, L. (2010). A note on mate allocation for dominance handling in genomic selection. *Genet. Sel. Evol.* 42, 1–9. doi:10.1186/1297-9686-42-33

Tsai, H.-Y., Matika, O., Edwards, S. M., Antolín–Sánchez, R., Hamilton, A., Guy, D. R., et al. (2017). Genotype Imputation To Improve the Cost-Efficiency of Genomic Selection in Farmed Atlantic Salmon. *G3 Genes, Genomes, Genet.* 7, 1377–1383. doi:10.1534/g3.117.040717

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science (80-.).* 313, 1596–1604. doi:10.1126/science.1128691

VanRaden, P., Van Tassell, C., Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J., et al. (2009). Invited review: Reliability of genomic predictions for north american holstein bulls. *Journal of dairy science* 92, 16–24

VanRaden, P. M. (2007). Genomic Measures of Relationship and Inbreeding. *Interbull Bull.* 25, 111–114. doi:10.1007/s13398-014-0173-7.2

Varona, L., Legarra, A., Toro, M. A., and Vitezica, Z. G. (2018). Non-additive effects in genomic selection. *Front. Genet.* 9, 1–12. doi:10.3389/fgene.2018.00078

Villar, M. and Forestier, O. (2009). Le peuplier noir en France: Pourquoi conserver ses ressources génétiques et comment les valoriser? *Rev. For. Fr.* 61, 457–466

Villar, M. and Forestier, O. (2017). La France à la sauvegarde du Peuplier noir : état actuel du programme de conservation et de valorisation des ressources génétiques. *Rev. For. Fr.* LXIX, 195–204

Villar, M., Lefevre, F., Bradshaw, H. D., and du Cros, E. T. (1996). Molecular genetics of rust resistance in poplars (Melampsora larici-populina Kleb/Populus sp.) by bulked segregant analysis in a 2 x 2 factorial mating design. *Genetics* 143, 531–536

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics* 206, 1297–1307. doi:10.1534/genetics.116.199406

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195, 1223–1230. doi:10.1534/genetics.113.155176

WANG, H., MISZTAL, I., Aguilar, I., Legarra, A., and MUIR, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res. (Camb).* 94, 73–83. doi:10.1017/S0016672312000274

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164–e164. doi:10.1093/nar/gkq603

Weir, B. S. (1979). Inferences about linkage disequilibrium. *Biometrics* , 235–254

Weir, B. S. (1996). *Genetic data analysis IImethods for discrete population genetic data.* 575.1072 W4

Westhues, M., Schrag, T. A., Heuer, C., Thaller, G., Utz, H. F., Schipprack, W., et al. (2017). Omics-based hybrid prediction in maize. *Theoretical and Applied Genetics* 130, 1927–1939

Whittaker, J. (2000). Marker-assisted selection using ridge regression. *Genet. ...* , 351–367

Wigginton, J. E., Cutler, D. J., and Abecasis, G. R. (2005). A Note on Exact Tests of Hardy-Weinberg Equilibrium. *Am. J. Hum. Genet.* 76, 887–893. doi:10.1086/429864

Wolak, M. E. (2012). nadiv : an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. *Methods Ecol. Evol.* 3, 792–796. doi:10.1111/j.2041-210X.2012.00213.x

Wright, S. (1921). Systems of mating. I. The biometric relations between parent and offspring. *Genetics* 6, 111

Wyckoff, G. W. and Zasada, J. C. (2002). Populus L. Woody plant seed manual. *Website http//www. nsl. fs. fed. us/wpsm/Populus. pdf [accessed 15 Oct. 2009]*

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of heritability for human height. *Nat. Genet.* 42, 565–569. doi:10.1038/ng.608.Common

Ye, S., Yuan, X., Lin, X., Gao, N., Luo, Y., Chen, Z., et al. (2018). Imputation from SNP chip to sequence: a case study in a Chinese indigenous chicken population. *J. Anim. Sci. Biotechnol.* 9, 30. doi:10.1186/s40104-018-0241-5

Zapata-Valenzuela, J., Isik, F., Maltecca, C., Wegrzyn, J., Neale, D., McKeand, S., et al. (2012). SNP markers trace familial linkages in a cloned population of Pinus taeda-prospects for genomic selection. *Tree Genet. Genomes* 8, 1307–1318. doi:10.1007/s11295-012-0516-5

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

Zhang, C., Kemp, R. A. R., Stothard, P., Wang, Z., Boddicker, N., Krivushin, K., et al. (2018). Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. *Genet. Sel. Evol.* 50, 14. doi:10.1186/s12711-018-0387-9

Zhang, M., Zhou, L., Bawa, R., Suren, H., and Holliday, J. A. (2016). Recombination Rate Variation, Hitchhiking, and Demographic History Shape Deleterious Load in Poplar. *Mol. Biol. Evol.* 33, 2899–2910. doi:10.1093/molbev/msw169

Ziegenhagen, B., Gneuss, S., Rathmacher, G., Leyer, I., Bialozyt, R., Heinze, B., et al. (2008). A fast and simple genetic survey reveals the spread of poplar hybrids at a natural Elbe river site. *Conserv. Genet.* 9, 373–379

Marie PEGARD

Nouveaux modèles pour la mise en œuvre de l'évaluation pan-génomique dans le programme d'amélioration du peuplier

Résumé :
Les espèces forestières sont particulières à bien des égards par rapport aux autres espèces domestiquées. Les arbres forestiers ont de longues phases juvéniles, entrainant de long et couteux cycles de sélection et nécesitant une sélection en plusieurs étapes indpendantes. Bien que cette méthode soit efficace du point de vue opérationnel, elle reste couteuse en temps et en ressources, entrainant une dillution de l'intensité et de la précision de sélection. Au vu de ces contraintes, les arbres sont de bons candidats pour la mise en œuvre de l'évaluation génomique. La sélection génomique (SG) repose sur le classement et la sélection d'individus à partir de l'information contenu dans leur génome sans utilisé une étape d'évaluation phénotypique et ainsi accélérer le processus de sélection.Ce travail visait à identifier les situations, les critères et les facteurs dans lesquelles la SG pourrait être une option réalisable pour le peuplier. Notre étude a montré que les avantages de l'évaluation génomique dépendent du contexte. C'est dans des situations les moins avantageuse que l'évaluation génomique se montre la plus performante, elle profite également de la densification de l'inforamtion génétique de faible à moyenne suite à une étape d'imputation de haute qualité. La sélection génomique pourrait être une option intéressante à stade précoce, où la précision de la sélection est généralement faible et la variabilité génétique abondante. Notre travail a également montré qu'il est important d'évaluer les performances avec des critères alternatifs, comme ceux liés au classement, notamment lorsque ces critères répondent au contexte opérationnel du programme d'élevage étudié.

Mots clés : *Populus nigra*, Selection génomique, Imputation vers la séquence

New models for implementation of Genome Wide Evaluation in poplar breeding program

Abstract :
Forest species are unique in many ways compared to other domesticated species. Forest trees have long juvenile phases, leading to long and costly selection cycles and requiring selection in several independent stages. Even if this method is operationally effective, it remains costly in terms of time and resources, resulting in a diluted intensity and accuracy of selection. In view of these constraints, trees are good candidates for the implementation of genomic evaluation. Genomic selection (SG) is based on the classification and selection of individuals from the information contained in their genome without using a phenotypic evaluation step and thus accelerating the selection process, in order to identify the situations, criteria and factors in which SG could be a feasible option for poplar. Our study showed that the benefits of genomic evaluation are context-dependent. Genomic evaluation is most effective in the less-advantageous situations, it also benefits from low to medium density genetic information following a high-quality imputation step. Genomic selection could be an interesting option at an early stage, when the accuracy of selection is generally low and genetic variability is abundant. Our work has also shown that it is important to evaluate performance with alternative criteria, such as those related to ranking, especially when these criteria fit the operational context of the breeding programme under study.

Keywords : populus nigra, genomic selection, sequence imputation, proof of concept

INRA SCIENCE & IMPACT

INRA Centre Val de Loire BioForA

BioForA