Valorisation des ressources génétiques du pommier dans une population d'amélioration élite grâce à la sélection génomique

Xabi Cazenave

DOCTORAT ECOLOGIE BRETAGNE GEOSCIENCES LOIRE AGRONOMIE ALIMENTATION

université angers

THESE DE DOCTORAT DE

L'UNIVERSITE D'ANGERS

ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialité : *Génétique, génomique et bio-informatique*

Par
Xabi Cazenave

Valorisation des ressources génétiques du pommier dans une population d'amélioration élite grâce à la sélection génomique

Thèse présentée et soutenue le 25/05/2022 à l'Université d'Angers

Unité de recherche : UMR 1345 - IRHS, 42 rue Georges Morel 49070 Beaucouzé
Thèse N° : 227713

Rapporteurs avant soutenance :

Alain CHARCOSSET	Directeur de recherches, INRAE
Christèle ROBERT-GRANIE	Directrice de recherches, INRAE

Composition du Jury :

Examinateurs :	**Sophie BOUCHET**	Chargée de recherches, INRAE
	Andrea PATOCCHI	Group Leader, Agroscope
	Didier PELTIER	Professeur, Université d'Angers
Dir. de thèse :	**Charles-Eric DUREL**	Directeur de recherches, INRAE
Co-enc. de thèse :	**Hélène MURANTY**	Chargée de recherches, INRAE

Remerciements

En premier lieu, je tiens à remercier Alain Charcosset et Christèle Robert-Granié d'avoir accepté d'être les rapporteurs de cette thèse, ainsi que les autres membres du jury : Sophie Bouchet, Andrea Patocchi et Didier Peltier. Je voudrais également remercier les membres de mon CSI Coralie Danchin-Burge, Patricia Faivre-Rampant, Laurence Moreau et Etienne Verrier d'avoir suivi l'avancement de mes travaux de thèse et d'avoir su me diriger vers les pistes à privilégier dans ce travail.

Ce travail ne serait pas ce qu'il est sans le binôme de choc qui m'a encadré trois années durant. Je voudrais donc exprimer toute ma gratitude envers mes deux encadrants de thèse Hélène et Charles-Eric pour tous les échanges que nous avons pu avoir, pour leur patience et pour leur bienveillance permanente.

Hélène, tu m'as toujours accueilli avec un grand sourire lorsque j'avais une « petite question » (j'espère que tu auras noté que j'ai abandonné cette expression en deuxième année !) et j'ai pris beaucoup de plaisir à échanger avec toi lors de nos points hebdomadaires. Je te suis également reconnaissant de m'avoir inclus au réseau R2D2, j'y ai beaucoup appris. J'espère maintenant que nous serons deux à pouvoir célébrer un diplôme en 2022 !

Charles-Eric, merci pour tes interrogations et remarques toujours pertinentes sur le travail mené tout au long de la thèse. Merci également pour toutes les anecdotes que tu as partagées lors de nos réunions ou au café concernant le monde des pommes ou celui de la recherche et qui témoignent de la passion pour la science qui t'anime et que tu partages si facilement. Je crois que les jambon-beurre de 13h30 les mercredis pour cause de réunion à rallonge vont me manquer !

Par extension, je voudrais remercier tous les gens que j'ai côtoyés au sein du groupe Carto (et non pas Cartoon Nissia si tu lis ces lignes !), nos retrouvailles en Choupette étaient toujours un moment d'échange agréable. Et vive monsieur de Guerlasse !

Plus généralement, je pense que tout le monde sera d'accord pour dire qu'il fait bon travailler au Bât B, j'aurais bien une histoire à raconter sur chacun d'entre vous pour illustrer la solidarité et la joie de vivre qui unit les membres des différentes équipes mais cela prendrait trop de temps à rédiger. Soyez donc tous remerciés pour les moments agréables passés au café, les repas collectifs ou simplement les échanges dans le couloir. S'il fallait tout de même que je remercie certaines personnes plus en détail, je pense que je citerais :

— **Nicolas** : lorsque je suis arrivé, tout timide et impressionné par la liste de nouvelles têtes à reconnaitre (dans les premiers jours, j'ai quand même parlé 15 minutes avec Emilie en pensant que c'était Mathilde...), tu as été l'un des premiers à me guider (vers la cantine, mais ça compte quand même). Aujourd'hui tu as pris du grade en devenant responsable café, que de chemin accompli ! C'est probablement grâce à la sagesse que te procure le passage de la barre des 40 ans.
— **Jaia** : parce que tu es arrivée très peu avant moi et qu'on a ainsi pu se rassurer mutuellement sur le milieu pas si hostile dans lequel on était arrivés. Merci surtout pour tous les moments en dehors du boulot au Héron, au Blue Monkeys ou au Matt Murphy's (Mate Meurfiz comme on dit chez nous). Je te souhaite à toi et à Djogo tout le meilleur pour la suite et attends impatiemment de pouvoir rencontrer votre bout de chou !
— **Michaela** : even though your stay in Angers was rather short, it was really great to have you with us and I hope we will get to meet again.
— **Bernard** : parce que déjà, tout le monde sait que tu fais les meilleurs gâteaux du bâtiment. Et aussi pour tous ces déplacements en verger à compter et peser des fruits, et aussi à t'amuser à me faire goûter les pires pommes qu'on y trouvait, celles dont Mehdi disait : « Même à un cheval je n'en donnerais pas ».
— **François L.** : parce que malgré ton emploi du temps surchargé, tu as su trouver du temps pour discuter avec moi de l'article et du manuscrit de thèse. Je ne sais pas où ça en est mais j'espère que l'on pourra croiser Thomas Pesquet sur Angers en août prochain !
— **Lydie, Magalie et Nathalie** : parce qu'il en fallait de la patience pour me recevoir au moins une fois par semaine, mais pas une fois le matin. Merci pour toutes ces fois où vous m'avez permis de ne pas me perdre sur le chemin tortueux des tâches administratives. J'ai également une pensée pour Marie-Pierre, qui aura égayé mes trajets lorsque je prenais le bus avec elle et effrayé notre bureau chaque fois que son nom apparaissait sur le téléphone juste après un ordre de mission rempli de façon hasardeuse.
— **Pierre B.** : je me suis demandé si je devais te remercier vu que tu mènes actuellement dans nos confrontations échiquéennes, mais je me suis dit que d'ici à ce que tu lises ces lignes j'avais le temps de te battre 3 ou 4 fois. Hâte de pouvoir refaire le monde avec toi autour d'un verre de Grolleau !

— **Pat** : parce qu'il faisait bon passer dans le bureau que tout le monde envie au Bât B afin d'écouter Radio Free Dom ou pour profiter de vos 25°C quand il faisait 12 ailleurs. Merci pour ton dynamisme qui a relancé les activités du bâtiment quand tu es arrivée, et pour te remercier il est bien possible que j'aie une belette pour toi lors de la soutenance.

— **Amanda** : je me demandais s'il fallait que je te remercie en anglais mais tes progrès en français ont été tellement fulgurants que je suis sûr que tu n'auras aucun mal à lire ces quelques lignes. Merci pour les bons moments et good luck in New Zealand !

— **Aurélien** : pour les relectures qui m'amènent à me demander pourquoi tu as accès au répertoire de Respom, c'est pas normal ça ! Merci pour les échanges zététiques au café, ça aussi c'étaient de bons moments. Maintenant je croise les doigts pour toi pour le concours !

— **Sylvain** (et j'ajoute David aussi tiens) : parce que vous êtes les deux piliers de la troupe, et parce que c'est quand même agréable de pouvoir faire des barbecues à la campagne

— **Wendy** : parce que si je ne t'avais pas mentionnée, tu me l'aurais reproché pendant une semaine et que je ne veux pas me mettre de membre de la BTS Army à dos, ça a l'air dangereux. J'espère que tout roule dans tes vertes prairies !

— **Babacar, Sandra et Nissia** : si j'étais un chercheur d'or, je vous appellerais mes trois pépites. J'ai été honoré d'encadrer vos stages respectifs et de partager de bons moments à vos côtés autour d'un Molkky, de alley-oop en verger ou de trajets de bon matin depuis la gare. Plein de bonnes choses à vous trois pour la suite !

— **Juliette** : parce que tu es la première personne que j'ai vue sur le site de l'INRA ce fameux matin du premier novembre 2018 à 8h, quelle chance d'être tombé sur toi quand même, une autre personne de la diaspora sudiste et qui apprécie tout comme moi l'Ile de Gorée à sa juste valeur ! Et quelle chance d'avoir partagé ton bureau et ta bonne humeur pendant deux ans, aux côtés de la non moins merveilleuse...

— **Marie-Charlotte** : une autre amatrice du domaine de Souleymane parait-il. Merci d'avoir initié les petit-déjeuner (trop) matinaux ? les repas (trop peu nombreux) au Bistrot des Ducs ou au Punjab et les dégustations de Giusepe. Merci surtout pour les bons moments dans la dernière ligne droite, notamment à m'épauler pour martyriser le pauvre...

— **Antoine** : dernier arrivé du bureau, je tiens d'abord à souligner ta patience compte tenu de nos attaques perpétuelles. Pour cela, j'ai l'honneur d'annoncer dans cette section que ta période d'essai est terminée et te déclare donc officiellement membre du bureau 040-118 (que personne n'appelle comme ça d'ailleurs mais tu pourras vérifier, c'est bien son nom). J'espère que cette nouvelle te donnera le sourire et t'empêchera de râler jusqu'à la fin de ta thèse, mais je n'y crois pas trop en fait. Tout le meilleur pour la suite et rendez-vous en 2041 pour Bordeaux-OL en quarts de la LDC

Il y a également deux groupes d'étudiants que je souhaiterais remercier ici pour des raisons assez diamétralement opposées : il s'agit d'une part des étudiants de L3 d'Agrocampus Ouest à qui j'ai eu l'honneur de donner des cours de statistiques mais surtout de karaoké (et ils sont plus doués pour ça) et d'autre part le groupe des étudiants-doctorants qui constituent le Doctoberfest et avec lesquels j'ai pu passer de merveilleuses soirées à essayer de parler d'autre choses que la thèse, sans forcément y arriver. Alex et Alex, Wilfried, Julia, Laure, Martin, Vinciane et tous les autres, merci pour tous ces moments indispensables, on attend la relève désormais !
En plus de ces joyeux trublions, j'aimerais remercier mon fidèle groupe d'amis angevins, qui a rendu les soirs d'hiver moins longs et les soirs d'étés plus ensoleillés. Merci Pierre C. pour les places au 400 Coups, les solos a cappella sur la route de Tours ; le gouda et les backcross. Merci à Aline pour Gérard et Colette, mais surtout Gérard. Merci à Axel pour le wake, les bières de Noël que tu ne m'as toujours pas payées et le squash. J'ai demandé à Philippe, tu es sur 76 défaites consécutives, véridique. Merci à Rafa mon tío pour avoir sauvé le monde des pandémies 92 fois, pour les Inde-Argentine accrochés, pour les buts d'Olivier Giroud et pour les méchants qui se coincent le cou dans une branche en bobsleigh. Merci à Nicolas S. pour son matelas deux places qui traine dans ma chambre depuis plus d'un an, pour le concours qui n'aura pas duré plus de 4 semaines et pour ses discours nuancés et argumentés sur le rôle des sélectionneurs. Merci à Arthur parce qu'il produit Ariane, merci à Sarah pour le violon par-dessus les cymbales et merci à Aurélie pour les pinces croco qui étaient trop courtes au final. Merci enfin aux 12 colocataires différents qui seront passés par la rue Dupetit-Thouars en 3 ans, je ne vous mentionnerai pas tous mais j'ai une pensée pour chacun d'entre vous. Un mot spécial tout de même pour Nico et Margaux qui m'ont accompagné sur cette fin, leur soutien à base d'escape game et de programmes télé de basse qualité aura été salvateur.
Remontons désormais légèrement dans le temps. J'ai rencontré les dernières personnes que je souhaite remercier avant le début de la thèse, à commencer par mes collègues nogentais. Car au final c'est à leur côté que j'ai pris goût à la recherche, moi qui n'envisageais même pas la voie académique à la fin de mes études. Mais l'écosystème bouillonnant du site des Barres m'a captivé et l'amour porté aux coléoptères par mes collègues écologues m'a finalement poussé à me demander « Pourquoi pas moi ? ». Un grand merci donc à Ugoline, Cécile, Ushma, Fabien, Hilaire, Jordan, Pavel, mes voisines Gwendo et Céline (qui pourtant ont passé un an à me dire de ne surtout pas faire de thèse) et tant d'autres d'avoir suscité cette vocation en moi. En parallèle, je salue ici Patrick, Pierre et Carl pour nos expérimentations souvent créatives, même si je ne sais toujours pas ce qu'est une clé de 12.
Les remerciements commençant à devenir un peu longs, je vais être plus concis pour remercier mes amis qui m'ont épaulé pendant la thèse et même avant pour certains. Je voudrais commencer par remercier le groupe des Poètes, pour leur bonne humeur en toute saison mais surtout parce que Pierre Issac. Pierre Issac. Ah ça fait du bien d'écrire ça. Je remercie par ailleurs la

bande corse des nuques protégées : Jujuskateur pour les drops rue des Cordeliers et les coups de casque dans les rucks, Fanch pour les Paris-Melun et Isabelle Ithurburu, Hubert parce que quand tu as conclu ta thèse en mentionnant Anatole et que Didier Gascuel a levé les bras j'ai eu des frissons comme rarement, Niels parce que la carabine et les portes qui claquent, Romain parce qu'on sait bien qu'à tes 30 ans on va appeler Odewan pour l'entedre dire « En Avant Guingamp » et Tilly parce que Valentin Pinson, Tchokounté et que dans les tribunes, Michèle Morgan applaudit. Je remercie aussi la team des sudettes : Marie parce qu'on attend encore le Pruneau Show, Marie-Sophie parce que rgeqghhjhwdfgh (c'est ton père qui m'a demandé d'écrire ça) et Léa parce qu'elle adore les raisins secs et qu'il y a 3 ans elle se baladait avec un sac de raisins et - mais j'avais dit que je serais concis donc restons-en là. Enfin, je remercie tous les amis basques que je vois de moins en moins fréquemment mais que je prends toujours plaisir à croiser. J'ai une pensée spéciale pour les ex-star du rock Hugo Poxta, Kenji Delabaca, Jeannot et les frérots Araspinus avec qui j'ai eu tant de joies sur scène.
Babou, tu as peut-être pensé que je t'avais oublié dans la liste précédente mais pas du tout, je voulais juste terminer par un paragraphe de remerciement pour ma famille : les oncles et tantes que je vois plus ou moins souvent mais qui veulent toujours savoir quelle est la meilleure variété de pomme (la réponse est Akane), les cousins qui écrivent trop de messages pour que j'aie le temps de suivre mais auprès de qui je m'engage à contribuer à une belle cousinade 2022 dans le Sud, mon petit frère Battitta qui ne prend jamais de nouvelles mais que j'aime bien quand même (d'ailleurs je n'en prends jamais non plus, j'espère que tu vas bien) et ma mère qui a bien du courage de devoir gérer deux fils aussi dissipés. Aita, zuri ere pentsatzen dut, zu izan zinen doktoregotzaren aholkua eman zidan lehen pertsona. Ez zaitugu ahanzten.
Evidemment, je ne pouvais pas terminer sans te remercier toi Zourab, pour tout ce que tu m'as apporté jusqu'à maintenant et pour tout ce qui est à venir.

Liste des abréviations

ACP : Analyse en Composantes Principales
AFLP : Amplified Fragment Lenght Polymorphism
AR² : Allelic R^2
BLUP : Best Unbiased Linear Predictor
CBD : Convention on Biological Diversity
CD : Coefficient of Determination
CRB : Centre de Ressources Biologiques
CWR : Crop Wild Relatives
DHS : Distinction Homogénéité Stabilité
ECPGR : European Cooperative Programme for Plant Genetic Resources
FAO : Food and Agriculture Organization of the United States
GBLUP : Genomic Best Unbiased Linear Predictor
GDDH13 : Golden Delicious Doubled Haploid n° 13
GEBV : Genomic Estimated Breeding Value
GIEC : Groupement d'Experts Intergouvernemental sur l'Evolution du Climat
GL : Genotype Likelihood
GRM : Genomic Relationship Matrix
GWAS : Genome-Wide Association Study
HFTH1 : Hanfu derived Trihaploid 1
HiDRAS : High-quality Disease Resistant Apples for a Sustainable Agriculture
IQS : Imputation Quality Score
IRHS : Institut de Recherche en Horticulture et Semences
MAF : Minor Allele Frequency
NIRS : Near Infrared Spectroscopy
OGM : Organisme Génétiquement Modifié
PBA : Pedigree-Based Analysis
PFR : Plant and Food Research Institute
QTL : Quantitative Trait Locus
RAPD : Randomly Amplified Polymorphic DNA
RG : Ressources Génétiques
SAM : Sélection Assistée par Marqueurs
SNP : Single Nucleotide Polymorphism
SSR : Short Sequence Repeats
TBV : True Breeding Value
TS : Training Set
USDA : United States Department of Agriculture
VS : Validation Set
WGD : Whole Genome Duplication
WUE : Water-Use Efficiency

Table des figures

Liste des tableaux

Table des matières

CHAPITRE 1

Introduction générale

1.1 Importance de la diversité génétique en création variétale

1.1.1 Les défis de la création variétale au 21ème siècle

La production alimentaire au niveau mondial depuis 70 ans a connu une évolution spectaculaire (figure 1.1), notamment grâce aux changements des pratiques agricoles et au progrès génétique observé chez les espèces animales et végétales (Miglior et al. 2017 ; Voss-Fels et al. 2019) qui ont permis d'obtenir des rendements plus élevés (Evenson et Gollin 2003 ; Schauberger et al. 2018). Pourtant, le défi de l'alimentation s'annonce de taille au 21ème siècle : la population mondiale devrait atteindre 9 à 10 milliard d'habitants d'ici 2050 (Adam 2021), tandis que la FAO (Food and Agriculture Organization of the United States) indiquait en 2021 que la malnutrition était en progression dans le monde et qu'une personne sur dix souffrait de sous-alimentation (FAO 2021). Dans le même temps, le sixième rapport du GIEC (Groupement d'Experts Intergouvernemental sur l'Evolution du Climat) indique que le changement climatique impactera fortement les écosystèmes agricoles (figure 1.2) du fait de l'élévation des températures, des précipitations accrues et de l'émergence de nouvelles maladies (GIEC 2022). L'objectif majeur de l'agriculture est donc aujourd'hui de produire suffisamment pour nourrir une population grandissante tout en s'adaptant aux nouvelles demandes de la société, notamment en termes de réduction des intrants. A ce titre, la création variétale doit donc être en mesure de proposer des variétés à haut rendement pouvant faire face aux stress biotiques et abiotiques.

Cependant, le progrès génétique annuel semble diminuer chez certaines espèces (Rizzo et al. 2022) et les variétés actuelles ne sont généralement pas adaptées aux futures conditions climatiques (Kahiluoto et al. 2019). Les programmes d'amélioration définissant explicitement l'adaptation au changement climatique comme objectif de sélection sont à ce jour peu nombreux, notamment chez les espèces fruitières et les légumes (GIEC 2022, p. 286). De plus, les programmes d'amélioration actuels n'utilisent généralement qu'une partie réduite de la diversité génétique à leur disposition, ce qui limite les possibilités d'adaptation aux nouveaux environnements tout en desservant potentiellement le progrès génétique à long terme (Swarup et al. 2021). La valorisation des ressources génétiques représente dans ce cas une piste à explorer par les sélectionneurs (Thudi et al. 2021). Nous retenons ici une définition large du terme « ressources génétiques », telles qu'elles sont définies par la Convention sur la Diversité Biologique (CBD), à savoir tout matériel génétique ayant une valeur effective ou potentielle. Il peut par exemple s'agir de variétés de pays, de variétés anciennes ou d'espèces sauvages apparentées.

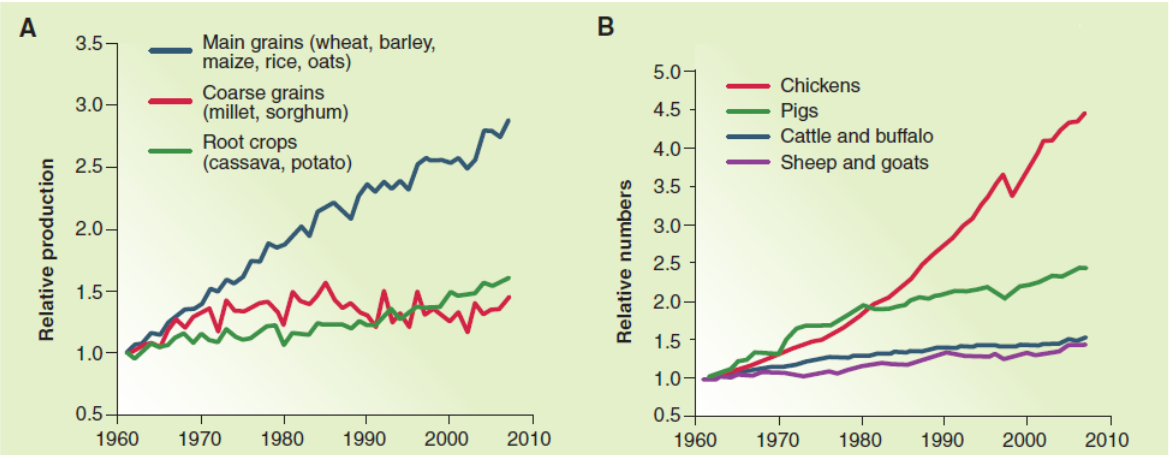


FIGURE 1.1 – Évolution de la production relative au niveau mondial pour (**A**) les productions végétales majeures et (**B**) les espèces animales majeures
D'après Godfray et al. (2010)

1.1.2 Pourquoi s'intéresser aux ressources génétiques ?

Chez de nombreuses espèces, la domestication a entrainé une réduction importante de la diversité génétique par rapport aux espèces sauvages apparentées (Bayer et al. 2022 ; van de Wouw et al. 2010) du fait de la dérive génétique (effet fondateur dû à la sélection d'un petit nombre de génotypes, qui n'a cependant pas été observé chez toutes les espèces (Gross et Olsen 2010 ; Trucchi et al. 2021)) et de la forte sélection de quelques gènes clé. De plus, la sélection variétale a également pu conduire à une réduction de la diversité génétique chez certaines espèces (voir par exemple Yue et al. (2015) en race Holstein, Zeitler et al. (2020) chez le maïs, Scott et al.

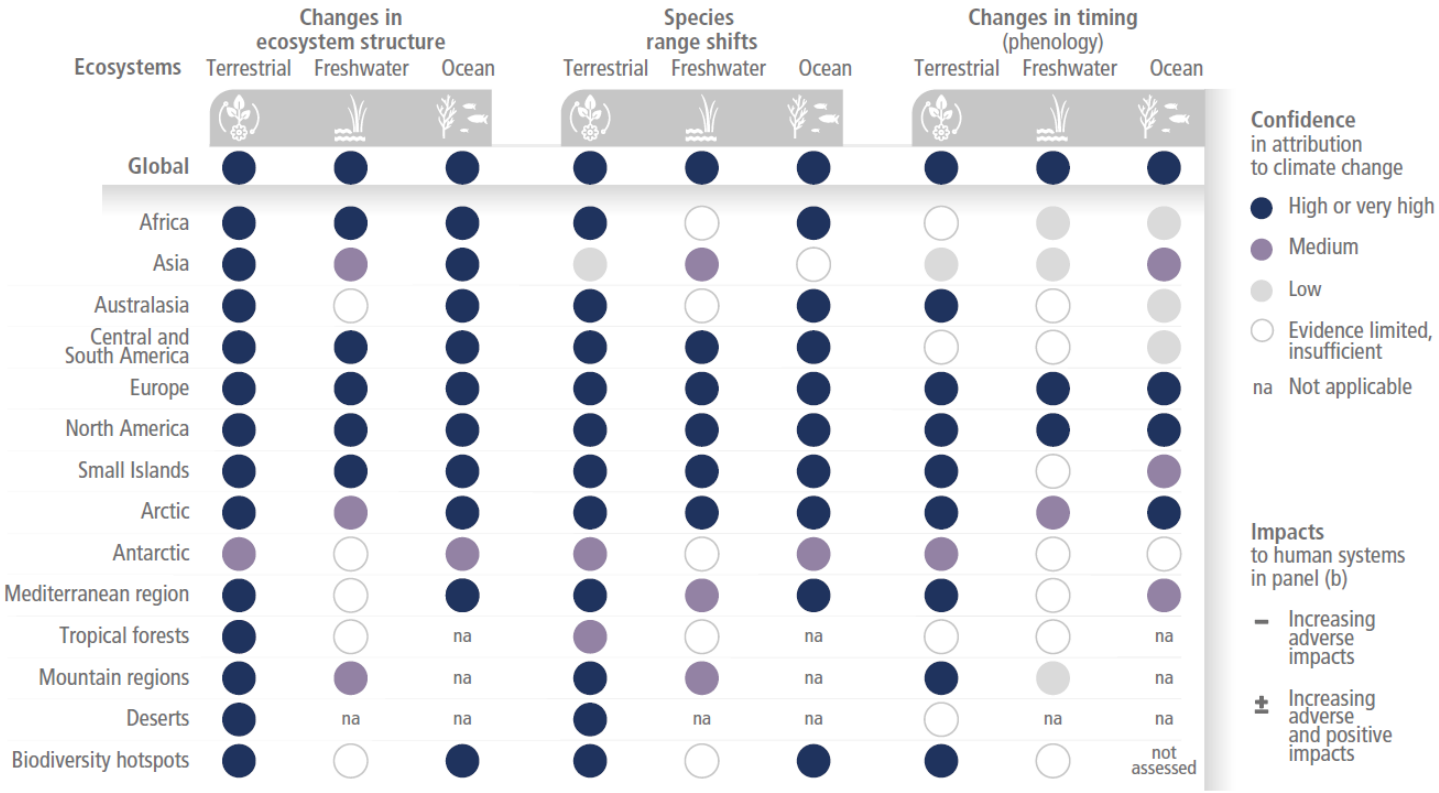


FIGURE 1.2 – Impacts sur les écosystèmes attribués au changement climatique aux niveaux régionaux et globaux. Les résultats présentés reposent sur des revues de la littérature et des méta-analyses effectuées par le GIEC
D'après GIEC (2022)

(2021b) chez le blé), bien que certaines études mettent également en évidence une augmentation de la diversité du fait de l'action humaine (Reif et al. 2005 ; Schouten et al. 2019). Cette « érosion génétique », globalement observée chez la plupart des espèces, a pour conséquence une utilisation limitée de la diversité génétique dans les programmes de sélection actuels (Khoury et al. 2022), qui peut limiter le gain génétique à long terme.
Pour s'en persuader, il est possible de revenir à « l'équation du sélectionneur » (Lush 1937) :

$$\Delta G = \frac{ir\sigma}{L}$$

où ΔG représente le gain génétique par unité de temps, i l'intensité de sélection, r la précision de sélection, L l'intervalle de génération et σ la racine carrée de la variance génétique du caractère d'intérêt dans la population. Une diminution de la variance génétique additive s'accompagne donc d'une limitation de la réponse à la sélection. Compte tenu de la base génétique étroite des programmes d'amélioration constatée chez de nombreuses espèces, l'utilisation de ressources génétiques permettant de réintroduire de la diversité génétique dans les programmes de sélection est indispensable pour assurer un gain génétique à long terme (Allier et al. 2020b)
Par ailleurs, il est probable que les ressources génétiques puissent amener des allèles favorables

absents dans les programmes de sélection, ou tout du moins présents en fréquence faible (Feuillet et al. 2008 ; Tanksley et McCouch 1997). Le transfert de tels allèles vers du matériel élite pourrait par exemple permettre une meilleure tolérance au changement climatique. Singh et al. (2018) ont ainsi mis en évidence un haplotype provenant d'une espèce sauvage apparentée du blé et absent des lignées élite de leur programme d'amélioration, ce qui a permis une meilleure tolérance à la sècheresse des meilleures lignées après introgression. De même, Naz et al. (2019) ont identifié chez le blé plusieurs QTL provenant d'une espèce sauvage apparentée permettant d'obtenir un rendement en grains plus élevé une fois introgressés dans un fond génétique élite et sous contrainte hydrique.

1.1.3 Difficultés d'utilisation des ressources génétiques

A l'heure actuelle, plus de 7 millions d'accessions de ressources génétiques sont conservées dans des banques de gènes (Genetic Resources for Food and Agriculture (2010), figure 1.3) et représentent donc un réservoir de diversité important valorisable en création variétale. Les sélectionneurs sont cependant plutôt réticents à utiliser ce type de matériel dans leurs programmes d'amélioration pour plusieurs raisons (Wang et al. 2017).

Premièrement, la caractérisation phénotypique de ressources génétiques dans les banques de gènes est un travail de longue haleine et qui peut se révéler coûteux si plusieurs caractères sont évalués. De plus, la maladaptation des ressources génétiques peut rendre leur évaluation délicate, par exemple si les conditions environnementales du site d'évaluation ne correspondent pas aux conditions de l'aire d'origine du matériel. Il est aussi possible que certains allèles ne puissent être évalués que dans un fond génétique élite car ils ne s'expriment pas dans les ressources génétiques (Longin et Reif 2014 ; Sommer et al. 2020). Les informations relatives aux performances agronomiques des ressources génétiques sont donc généralement limitées, de telle sorte que Prada (2009) considère ainsi que la recherche d'allèles favorables dans les banques de gènes revient à « chercher une aiguille dans une botte de foin ». L'utilisation de marqueurs moléculaires et le séquençage massif des banques de gènes réalisé depuis quelques années constituent à ce titre une avancée certaine pour l'identification de matériel intéressant, notamment à l'aide de modèles prédictifs (McCouch et al. 2012).

Il existe par ailleurs souvent une différence de performance agronomique entre les ressources génétiques et le matériel élite, ce qui limite leur introduction directe dans les programmes de sélection. Au niveau moléculaire, l'apport d'allèles favorables depuis les ressources génétiques peut conduire à l'introduction simultanée d'allèles défavorables en liaison forte avec les allèles favorables, un phénomène nommé « linkage drag » et pouvant conduire à l'introduction de caractéristiques non désirées (Peng et al. 2014), telles que la détérioration de la qualité gustative (Zhu et al. 2018) ou une pénalisation du rendement (Pasquariello et al. 2020). Il est donc néces-

saire d'avoir recours à plusieurs générations de croisements avec du matériel élite avant d'obtenir des individus présentant potentiellement de bonnes performances agronomiques et pour lesquels les allèles favorables provenant des ressources génétiques se trouvent dans un fond génétique majoritairement élite. Dans le cas de caractères contrôlés par un grand nombre de loci, il est de plus difficile d'identifier les allèles favorables provenant des ressources génétiques et donc de les suivre à l'aide de marqueurs moléculaires. Or la sélection de croisements entre matériel élite et ressources génétiques sur la base d'observations phénotypiques a tendance à favoriser la sélection de génotypes dont le génome provient majoritairement du matériel élite, ce qui peut conduire à contre-sélectionner les allèles favorables provenant des ressources génétiques (Yang et al. 2020)

Récemment, deux approches ont été proposées pour contourner les difficultés liées à l'utilisation des ressources génétiques. La première approche, dénommée « édition du génome » consiste à modifier directement un ou plusieurs nucléotides dans le matériel élite afin d'obtenir les allèles favorables d'intérêt qui auraient pu être obtenus à partir des ressources génétiques, évitant ainsi la « loterie de la sélection » (Schleif et al. 2021). Cependant, les variétés résultant de l'application de cette technique sont actuellement considérées OGM (Organismes Génétiquement Modifiés) au sein de l'Union Européenne, rendant l'utilisation de cette approche incertaine pour les sélectionneurs. La deuxième approche, dénommée sélection génomique, est présentée dans la section suivante et fait l'objet du travail de cette thèse. Nous avons plus particulièrement étudié l'intérêt de son application chez le pommier, une espèce fruitière d'importance économique majeure. Des éléments relatifs à la biologie et à l'amélioration du pommier sont donc présentés dans une troisième section de l'introduction.

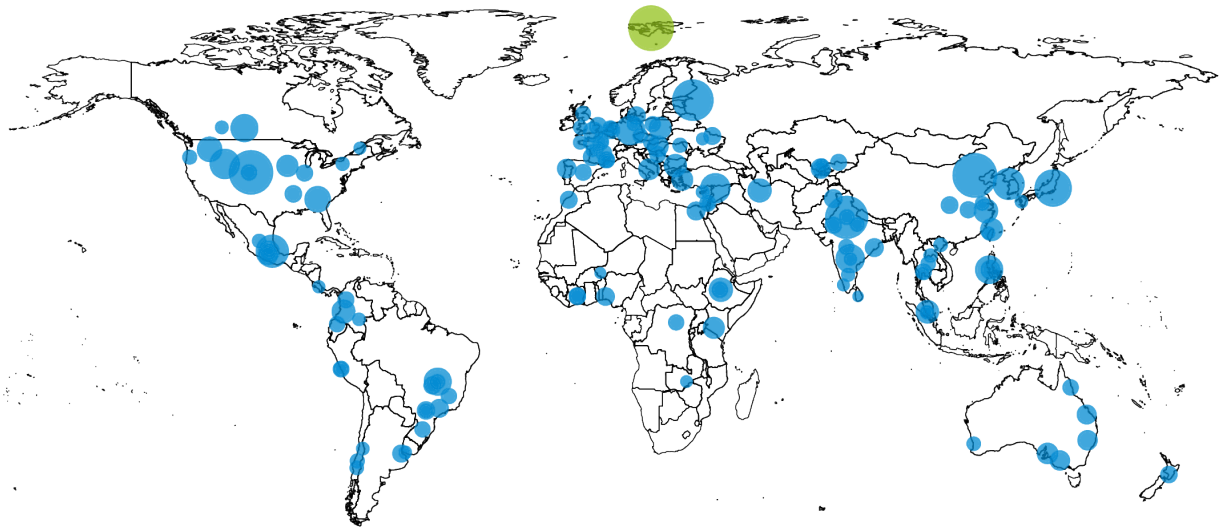


FIGURE 1.3 – **Bleu** : distribution géographique des banques de gènes possédant plus de 10 000 accessions. **Vert** : réserve mondiale de semences Svalbard Global Seed Vault
D'après Yu et al. (2016)

1.2 La sélection génomique : principes et applications

1.2.1 Débuts de la sélection génomique

La sélection génomique (Meuwissen et al. 2001) repose sur l'utilisation d'un grand nombre de marqueurs génétiques couvrant l'intégralité du génome pour prédire la valeur génétique de candidats à la sélection. Bien que les bases théoriques de la sélection génomique aient été posées au début des années 2000 (Bernardo 1994 ; Meuwissen et al. 2001 ; Nejati-Javaremi et al. 1997 ; Whittaker et al. 2000) et que l'intérêt de cette approche ait été rapidement reconnu, les outils nécessaires à sa mise en place n'étaient alors pas disponibles (de Koning 2016). Ce sont donc aussi les progrès spectaculaires dans le domaine de l'étude des génomes et des technologies de génotypage qui ont permis l'application de la sélection génomique. La sélection génomique a d'abord été mise en place chez les bovins laitiers, notamment grâce au développement en 2007 d'une puce de génotypage contenant 54000 marqueurs SNP (Single Nucleotide Polymorphism) (Boichard et al. 2016). L'utilisation de prédictions génomiques a rendu possible la sélection précoce de candidats à la sélection, ce qui a permis des progrès génétiques considérables : pour la race Holstein, l'utilisation de la sélection génomique a par exemple permis de doubler le gain génétique par unité de temps (García-Ruiz et al. 2016). Chez les espèces végétales, les premiers travaux sur la sélection génomique sont apparus un peu plus tard et ont dans un premier temps concerné les espèces de grande culture (Crossa et al. 2010 ; Jannink 2010), puis les espèces pérennes telles que les espèces forestières (Resende et al. 2012 ; Zapata-Valenzuela et al. 2013) et fruitières (Biscarini et al. 2017 ; Kumar et al. 2012a).

1.2.2 Principe de la sélection génomique

La sélection génomique repose sur la construction d'équations de prédiction permettant de calculer la valeur génétique (Genomic Estimated Breeding Value ou GEBV) de candidats à la sélection sur la seule base de leurs données génotypiques. Une partie des candidats peut alors être sélectionnée en se servant de leur GEBV. Une population dite d'entrainement comprenant des individus génotypés et phénotypés est d'abord utilisée pour établir une équation de prédiction à partir des données génotypiques. Les effets estimés des marqueurs sont ensuite utilisés pour prédire la valeur génétique d'une population de candidats (aussi appelée population de validation ou d'application) qui a été génotypée mais pas phénotypée (figure 1.4). Cette approche repose sur l'hypothèse que chaque locus ayant un effet sur le caractère d'intérêt (ou Quantitative Trait Locus, QTL) est en déséquilibre de liaison avec au moins un marqueur utilisé dans le modèle de prédiction.

La sélection génomique présente plusieurs avantages par rapport à d'autres méthodes de sé-

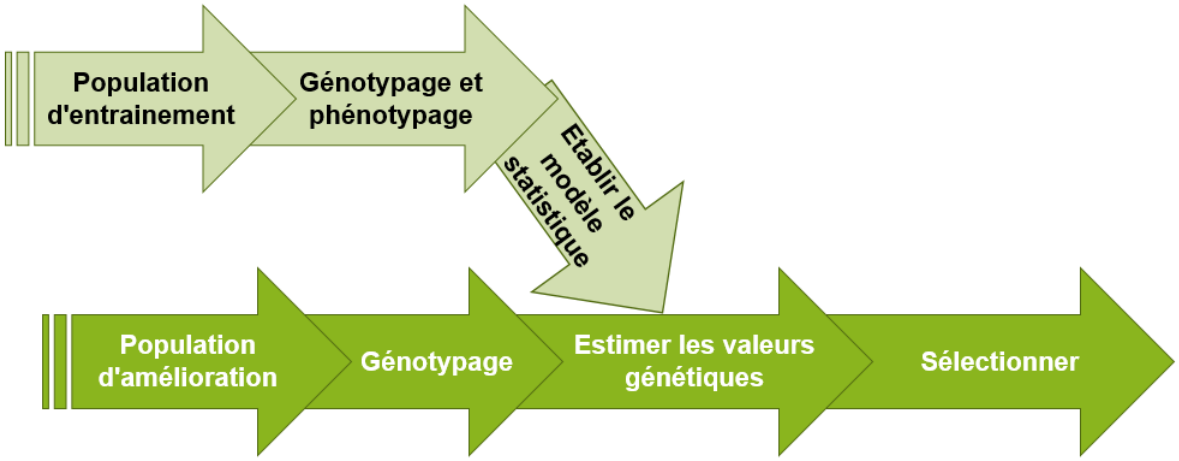


FIGURE 1.4 – Principe de la sélection génomique. D'après Heffner et al. (2009)

lection : premièrement, le calcul des GEBV n'impliquant pas de phénotyper les candidats à la sélection, l'intervalle de générations peut être fortement réduit (Bernardo et Yu 2007). La sélection génomique peut également être bénéfique pour des caractères difficiles ou onéreux à phénotyper (Voss-Fels et al. 2019) et pour des caractères à faible héritabilité (Muir 2007). Enfin, l'utilisation de la sélection génomique peut permettre d'augmenter l'intensité de sélection en testant plus de candidats à la sélection car le génotypage coûte généralement moins cher que le phénotypage.

1.2.3 Modèles de prédiction génomique

Les premiers modèles de prédiction génomique reposaient sur un modèle additif attribuant un effet à chaque marqueur. L'estimation des effets des marqueurs par la méthode classique des moindres carrés est cependant impossible dans le cas de la sélection génomique car le nombre de marqueurs est très grand par rapport à la taille de la population d'entrainement, un problème parfois appelé fléau de la dimensionnalité (de los Campos et al. 2009). De ce fait, de nombreux modèles de prédiction permettant de contourner cette limitation ont été développées (de los Campos et al. 2009 ; Gianola 2013). Ces modèles reposent généralement sur un rétrécissement des effets estimés ou sur la sélection de variables pour pouvoir estimer les effets des marqueurs (Desta et Ortiz 2014). Le modèle RR-BLUP (Ridge-Regression BLUP) fait partie des modèles les plus utilisés et repose sur le modèle mixte suivant :

$$y = X\beta + Wg + e$$

où $\mathbf{y}$ est un vecteur colonne contenant les observations, $\mathbf{X}$ est la matrice d'incidence des effets fixes β, $\mathbf{W}$ est la matrice de génotypage contenant p individus et m marqueurs, $\mathbf{g}$ est un vecteur colonne qui contient l'effet des marqueurs et $\mathbf{e}$ est le vecteur colonne des erreurs résiduelles avec

$e \sim \mathcal{N}(0,\, I\sigma_e^2)$. De plus, on suppose que $g \sim \mathcal{N}(0, I\sigma_g^2)$.
Dans ce cas, on peut écrire :

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + \lambda \end{bmatrix} \begin{bmatrix} \beta \\ g \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

avec $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$ est un paramètre à estimer appelé coefficient de pénalisation. On peut utiliser ces équations pour estimer **g** :

$$\hat{g} = W'(WW' + I\lambda)^{-1}(y - X\hat{\beta})$$

Une fois l'effet des marqueurs estimés, le calcul des GEBV des candidats à la sélection est immédiat et correspond au produit matriciel $W\hat{g}$.

De façon similaire, le modèle GBLUP (pour Genomic BLUP) peut être utilisé pour directement prédire les GEBV des candidats à partir du modèle mixte suivant :

$$y = X\beta + Zu + e$$

où **y**, **X** et β sont définis comme précédemment, **Z** est une matrice identité associée au vecteur colonne **u** contenant les valeurs génétiques. On suppose que $u \sim \mathcal{N}(0, G\sigma_u^2)$, où G est une matrice dite d'apparentement génomique (ou GRM, Genomic Relationship Matrix). On a alors :

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

A partir des équations du modèle mixte, on peut estimer **u** :

$$\hat{u} = GZ'(ZGZ' + I\lambda)^{-1}(y - X\hat{\beta})$$

La matrice d'apparentement génomique utilisée dans le modèle du GBLUP est calculée à partir des informations génotypiques des candidats à la sélection et des individus de la population d'entraînement. La formule proposée par (VanRaden 2008) est souvent utilisée :

$$G = \frac{Z_1 {Z_1}'}{\sum p_i(1-p_i)}$$

où $Z_1 = W - 2p_i$ avec W la matrice de génotypage codée en dosage de l'allèle mineur et p_i la fréquence de l'allèle mineur au locus i.

Les modèles RR-BLUP et GBLUP reposent sur l'hypothèse que le caractère étudié est contrôlé par un très grand nombre de loci à effet faible (modèle infinitésimal) et que la variance génétique à chaque marqueur est identique. Pour s'affranchir de cette hypothèse, il est possible d'utiliser des modèles bayesiens (Gianola 2013) qui permettent d'attribuer des variances différentes à certains loci. Ces modèles sont donc particulièrement adaptés dans le cas où certains QTL à effet fort sont connus pour le caractère d'intérêt, mais sont en contrepartie plus exigeants en temps de calcul que les modèles GBLUP et RR-BLUP. Enfin, des modèles basés sur le machine learning sont également utilisés pour réaliser des prédictions génomiques (Pérez-Enciso et Zingaretti 2019 ; Varshney 2021). La précision de prédiction obtenue à partir de tels modèles est cependant rarement supérieure à la précision de prédiction des modèles bayesiens ou du GBLUP (Azodi et al. 2019).

1.2.4 Précision de prédiction

Le succès de la sélection génomique dépend en grande partie de la précision des prédictions des candidats, qui correspond à la corrélation entre les GEBV des candidats et leur vraie valeur génétique (aussi nommée True Breeding Value ou TBV). Dans la pratique, les TBV ne sont jamais connues et sont souvent remplacées par les valeurs phénotypiques des candidats à la sélection. La précision de prédiction peut alors être mesurée par validation croisée (Isik et al. (2017), figure 1.5). Cette approche consiste à utiliser un sous-ensemble de la population d'entrainement comme population de validation puis à calculer la corrélation entre les valeurs phénotypiques et les GEBV de ce sous-ensemble (figure 1.5). En général, cette opération est répétée plusieurs fois en faisant varier le sous-ensemble de validation à chaque fois, ce qui permet de calculer une précision de prédiction moyenne.

Plusieurs formules ont été proposées afin d'estimer la précision de prédiction qu'il serait possible d'obtenir avec une population d'entrainement donnée. Une des premières formules proposées est celle de Daetwyler et al. (2008) :

$$\hat{r} = \sqrt{\frac{Nh^2}{Nh^2 + M_\mathrm{e}}}$$

où $\hat{r}$ correspond à la précision de prédiction attendue, N correspond à la taille de la population d'entrainement, h^2 à l'héritabilité du caractère étudié et M_e au nombre de segments chromosomiques indépendants ségrégeant dans la population, qui dépend entre autres de la taille efficace de la population. Dans la pratique, M_e est très difficile à évaluer (Grattapaglia et Resende 2011) et la formule présentée ne donne pas toujours des résultats proches de la réalité (Brard et Ricard 2015). Elle permet néanmoins d'identifier les facteurs qui influencent le plus la précision de prédiction. Nous détaillons ci-après deux de ces facteurs et quelques approches

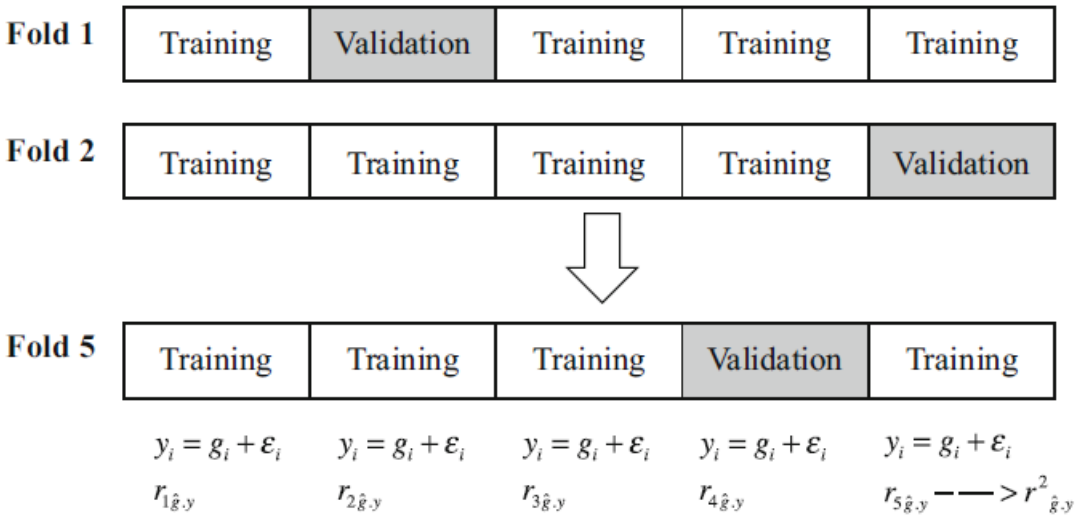


FIGURE 1.5 – Principe de la validation croisée. Ici, la population d'entrainement est divisée en 5 sous-ensembles, qui servent chacun une fois de population de validation. La précision de prédiction moyenne correspond alors à la moyenne des 5 précisions de prédiction mesurées avec les 5 sous-ensembles
D'après Isik et al. (2017)

possibles permettant d'obtenir une meilleure précision de prédiction dans les deux cas. Bien que nous ne les mentionnions pas explicitement, le choix du modèle statistique utilisé pour les prédictions et la qualité du phénotypage de la population d'entrainement sont également des facteurs importants à prendre en compte.

1.2.4.1 Constitution de la population d'entrainement

Ici, nous définissons la constitution de la population d'entrainement comme la taille de cette population ainsi que les génotypes qui la composent.
D'après l'équation proposée par Daetwyler et al. (2008), plus la taille de la population d'entrainement est grande et plus la précision de prédiction attendue est élevée. De nombreuses études ont effectivement confirmé que la taille de la population d'entrainement est prépondérante pour la calibration des équations de prédiction (Edwards et al. 2019 ; Norman et al. 2018), et ce d'autant plus que l'héritabilité du caractère étudié est faible (figure 1.6) et que les individus de la population d'entrainement sont peu apparentés (Takeda et al. 2021). Dans les faits, la précision de prédiction atteint souvent un plateau malgré l'augmentation de la taille de la population d'entrainement (Liu et al. 2018).

Une façon d'obtenir des populations d'entrainement de grande taille consiste à utiliser des données provenant de cycles de sélection précédents (Asoro et al. 2011 ; Muir 2007) ou des données historiques (Gonzalez et al. 2021 ; Washburn et al. 2021). L'ajout d'individus peu apparentés aux candidats dans la population d'entrainement peut néanmoins dégrader la qualité des prédic-

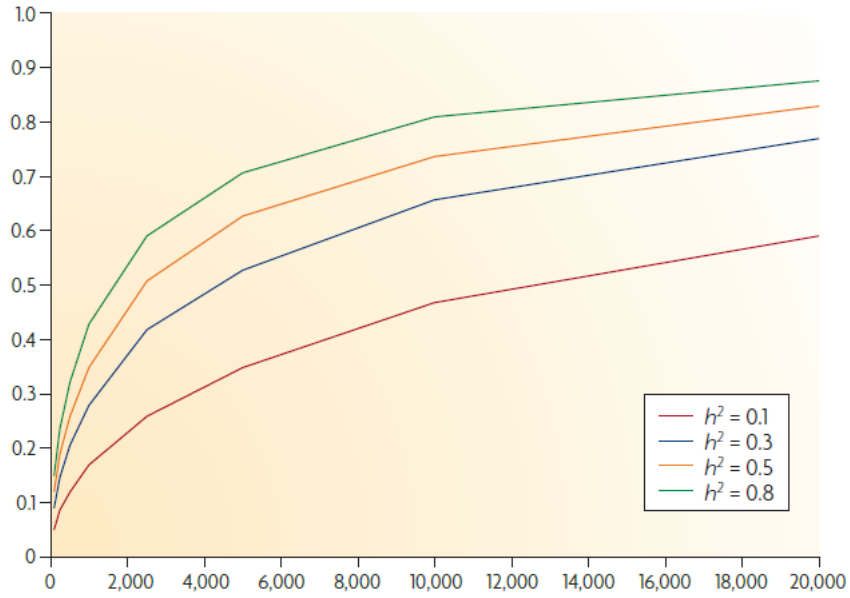


FIGURE 1.6 – Influence de la taille de la population d'entrainement sur la précision de prédiction selon la formule de Daetwyler et al. (2008) pour quatre valeurs d'héritabilité
D'après Goddard et Hayes (2009)

tions génomiques (Lorenz et Smith 2015). De même, il peut être préférable d'ajouter seulement un nombre restreint des générations précédentes à la population d'entrainement plutôt que toutes les générations disponibles : Wolc et al. (2011) et Weng et al. (2016) recommandent ainsi de ne pas utiliser de données au-delà des trois dernières générations, en particulier pour des caractères à faible héritabilité. Il est donc nécessaire de régulièrement mettre à jour la population d'entrainement afin que les fréquences alléliques et les associations marqueur-QTL soient conservées entre la population d'entrainement et les candidats (Eynard et al. 2018 ; Pszczola et Calus 2016).

Lorsque des données antérieures ne peuvent pas être utilisées pour établir les équations de prédiction, des données acquises sur d'autres populations peuvent être utilisées (Olatoye et al. 2020 ; Sverrisdóttir et al. 2018). Chez les animaux domestiques, cette dernière approche a été largement étudiée afin d'évaluer l'intérêt de combiner des données provenant de différentes races dans une même population d'entrainement. Cependant, les gains permis par une telle combinaison de données se sont la plupart du temps révélés marginaux (Hayes et al. 2009 ; Karoui et al. 2012). de Roos et al. (2009) ont montré que la combinaison de populations était d'autant plus efficace que la divergence entre les populations était récente (figure 1.7) et que la densité de marqueurs était élevée, afin de retrouver les mêmes associations marqueur-QTL d'une population à l'autre.

L'apparentement entre les individus de la population d'entrainement et les candidats contribue également grandement à obtenir des prédictions génomiques précises (Habier et al. 2009 ; Pszc-

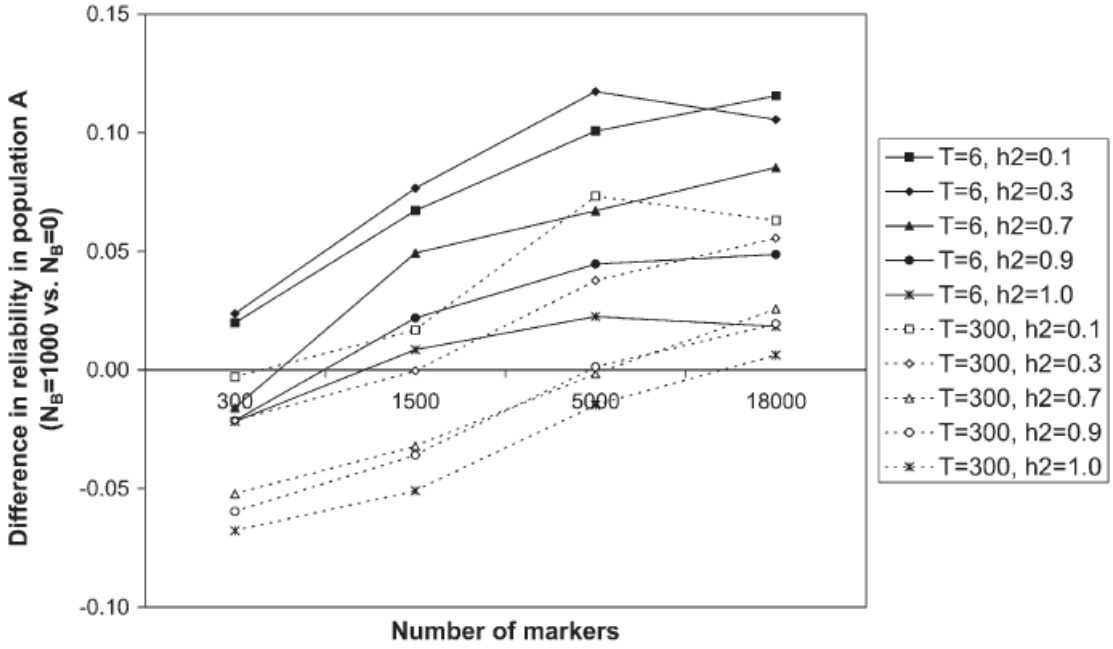


FIGURE 1.7 – Influence de la combinaison de 1000 individus d'une population A et 1000 individus d'une population B dans une même population d'entrainement en fonction du temps T de divergence entre les populations et de la densité de marquage
D'après de Roos et al. (2009)

zola et al. 2012), surtout si la population d'entrainement est de petite taille (Clark et al. 2012 ; Wientjes et al. 2013). Une situation favorable pour obtenir des précisions de prédiction élevées consiste par exemple à utiliser des plein-frères des candidats comme population d'entrainement. Lehermeier et al. (2014) ont ainsi montré en utilisant une population multiparentale de maïs que lorsque la population de validation était composée d'une partie des plein-frères d'une famille donnée, il fallait généralement une population d'entrainement de 375 demi-frères pour obtenir la même précision de prédiction qu'avec une population d'entrainement de 50 plein-frères.
Lorsque la taille de la population d'entrainement doit être limitée, pour des raisons budgétaires par exemple, plusieurs algorithmes ont été proposés pour optimiser le choix des individus constituant la population d'entrainement (Akdemir et al. 2015 ; Bustos-Korts et al. 2016 ; Rincent et al. 2017 ; 2012). Pour davantage de détails concernant ces algorithmes, Isidro y Sánchez et Akdemir (2021) proposent une synthèse concernant l'optimisation de la constitution de la population d'entrainement.

1.2.4.2 Déséquilibre de liaison et densité de marqueurs

Pour que la sélection génomique fonctionne, il est nécessaire que chaque QTL soit en déséquilibre de liaison avec au moins un marqueur. Un fort déséquilibre de liaison est également nécessaire pour que les effets des QTL soient correctement estimés entre populations ou d'une

génération à l'autre (Hayes et al. 2009). Dans l'étude pionnière de Meuwissen et al. (2001) sur la sélection génomique, une valeur de r^2 moyen de 0,2 avait été simulé et avait permis d'obtenir des précisions de prédiction élevées. Calus et al. (2008) ont également confirmé par simulations qu'un r^2 de 0,2 entre loci adjacents permettait d'obtenir des précisions de prédictions élevées mais que la précision de prédiction diminuait significativement pour des valeurs de r^2 plus faibles. Habier et al. (2013) ont par ailleurs montré que le déséquilibre de liaison était le facteur influençant le plus la précision de prédiction lorsque la taille de la population d'entrainement était élevée, mais qu'il avait moins d'importance pour des populations d'entrainement plus petites. Comme mentionné au paragraphe précédent, c'est alors l'apparentement entre population d'entrainement et candidats qui prime (Wientjes et al. 2013).
L'utilisation d'un grand nombre de marqueurs peut permettre d'obtenir un plus fort déséquilibre de liaison entre marqueurs et QTL, entrainant alors théoriquement une précision de prédiction plus élevée (Druet et al. 2014). Dans les faits, la précision de prédiction peut cependant atteindre un plateau avec un nombre limité de marqueurs (Frischknecht et al. 2018 ; Jung et al. 2020), comme illustré figure 1.8.

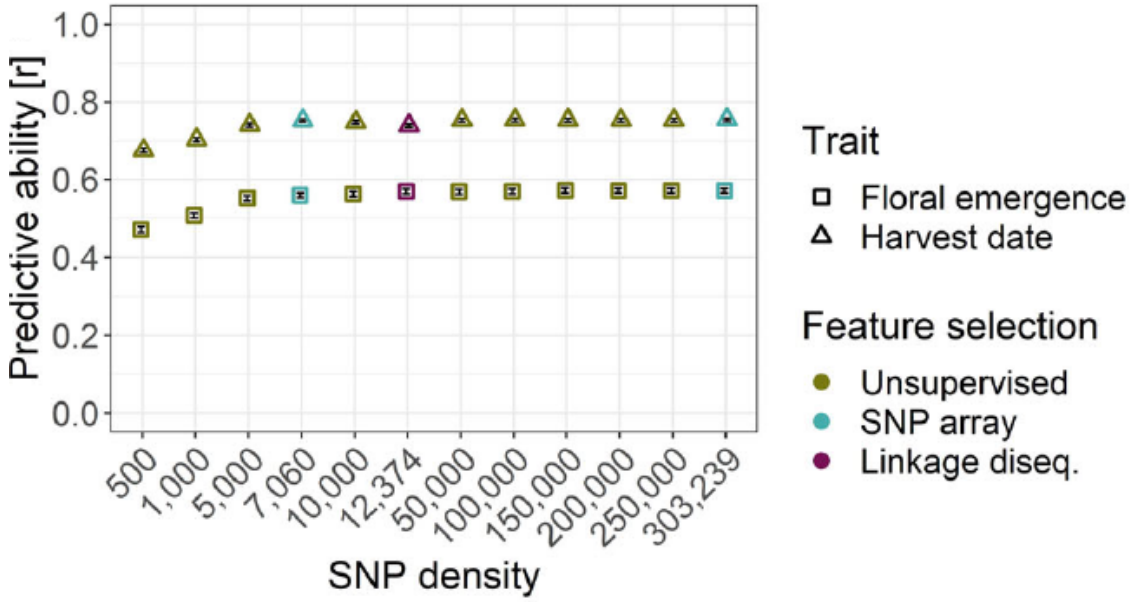


FIGURE 1.8 – Influence du nombre de marqueurs sur la précision de prédiction pour deux caractères : la date de début de floraison et la date de récolte chez le pommier.
D'après Jung et al. (2020)

1.2.5 Principe et utilisation de l'imputation

En pratique, le génotypage à haute densité de marquage des candidats à la sélection n'est pas toujours réalisable, notamment en raison du coût que peut représenter une telle opération. Il est alors possible de génotyper lesdits candidats avec une puce présentant une densité de

marqueurs inférieure et de remonter par la suite à une densité de marqueurs plus importante en se basant sur une population de référence elle-même génotypée avec une puce haute densité, à condition qu'il existe des marqueurs en commun entre les deux puces. Cette approche, nommée « imputation génotypique » ou plus simplement « imputation », a d'abord été utilisée en génétique humaine (Burdick et al. 2006 ; Scheet et Stephens 2006) et rapidement adoptée en sélection animale (Habier et al. 2009) et végétale (Jannink et al. 2009).

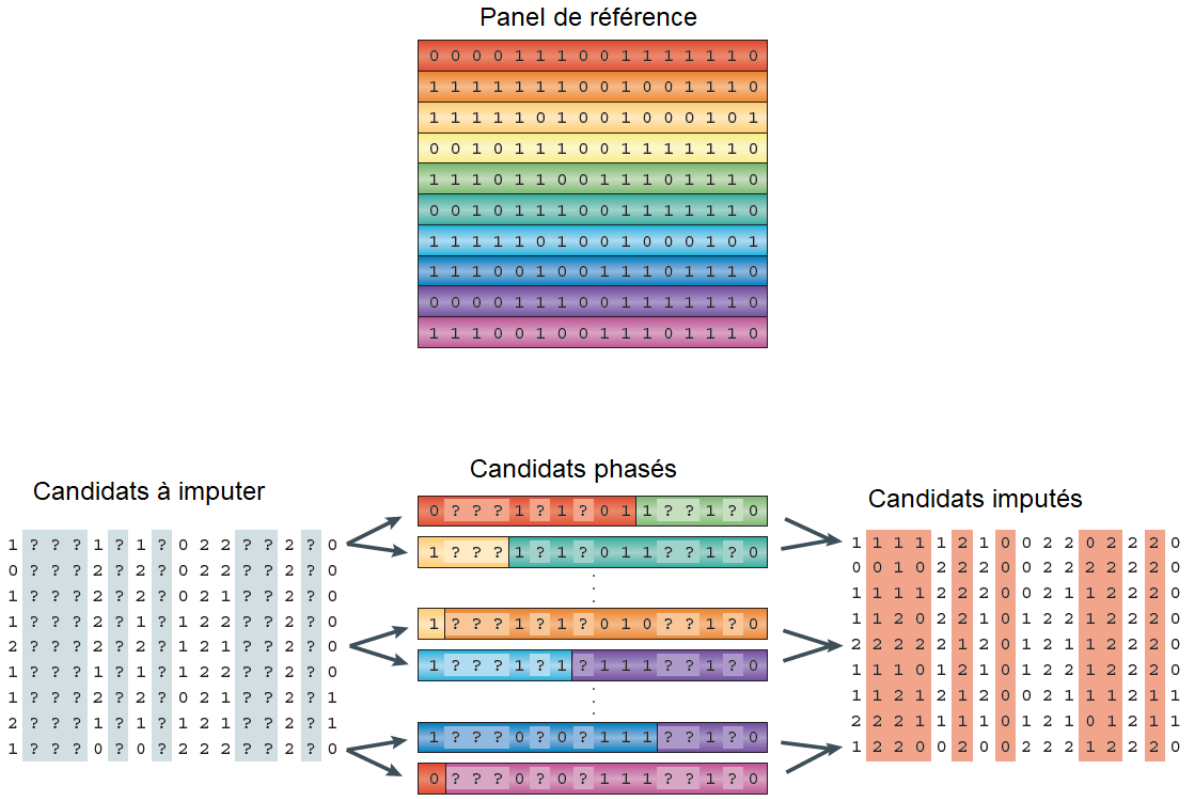


FIGURE 1.9 – Fonctionnement de l'imputation génotypique. D'après Marchini et Howie (2010)

La figure 1.9, issue de l'article de synthèse de Marchini et Howie (2010) illustre la façon dont fonctionne l'imputation génotypique. Cette approche repose sur l'utilisation d'un groupe d'individus qui constituent le panel de référence et sont génotypés à haute densité. Le panel de référence est dans un premier temps phasé (c'est-à-dire que l'origine parentale des allèles des individus du panel est déterminée), ce qui permet de constituer une librairie d'haplotypes apparaissant dans le panel de référence. Les individus à imputer sont ensuite phasés à leur tour et les haplotypes obtenus sont comparés aux haplotypes de la librairie d'haplotypes, en faisant la supposition que les haplotypes des individus à imputer sont une mosaïque des haplotypes du panel de référence. Les données manquantes à moyenne densité sont alors imputées à partir des haplotypes auxquels elles sont associées dans le panel de référence. Compte tenu de l'incertitude dans la correspondance entre les haplotypes phasés des individus à imputer et la librairie d'haplotypes, la majorité des logiciels d'imputation retournent une probabilité d'imputation à

chaque marqueur (par exemple pour un marqueur biallélique, un logiciel d'imputation pourrait retourner les valeurs 0,01/0,98/0,01, ce qui signifie que la probabilité que le marqueur imputé soit hétérozygote est de 0,98). Le logiciel Beagle, largement utilisé en génétique humaine mais également végétale, repose par exemple sur ce principe. Il est également possible d'utiliser les informations du pédigrée des individus à imputer, ce qui peut se révéler bénéfique lorsque les allèles transmis peuvent facilement être inférés sur la base des règles de ségrégation mendélienne (Druet et Georges 2010). Des logiciels tels qu'AlphaImpute ou FImpute utilisent cette approche.

La qualité des imputations obtenues par les approches décrites peut être évaluée en comparant les génotypes imputés aux génotypes réels s'ils sont disponibles. Cette qualité d'imputation, généralement mesurée comme le pourcentage de génotypes correctement imputés ou comme la corrélation entre génotypes réels et imputés, dépend de plusieurs facteurs qu'il est important de prendre en compte avant de réaliser l'imputation. Premièrement, la différence de densité de marqueurs entre la puce utilisée pour génotyper les individus à imputer et la puce ayant servi pour le panel de référence peut avoir un effet majeur sur la qualité d'imputation. Hickey et al. (2012b) ont ainsi observé une forte diminution de la précision d'imputation lorsque plus de 75% des marqueurs SNP d'une puce contenant 35 000 marqueurs étaient masqués. Deuxièmement, la fréquence des allèles des individus à imputer est un facteur connu pour influencer la qualité d'imputation : les allèles rares sont généralement mal imputés car peu d'haplotypes dans le panel de référence portent ces allèles. Li et al. (2011) ont par exemple observé que moins de la moitié des allèles très rares (fréquence inférieure à 0,001) était correctement imputée sur un jeu de données de 550 000 marqueurs SNP. L'utilisation de méthodes utilisant l'information de pédigrée peut permettre une meilleure imputation dans le cas des allèles rares (Sargolzaei et al. 2014). Enfin, la taille et la constitution du panel de référence ont une influence majeure sur la qualité d'imputation : l'augmentation de la taille du panel entraine généralement une meilleure qualité d'imputation, en particulier pour les allèles rares (Zheng et al. 2015). Cependant, l'ajout de génotypes dans le panel de référence ne semble plus avoir d'effet au-delà d'une certaine taille. L'apparentement entre les individus du panel de référence et ceux à imputer joue aussi un rôle dans la précision d'imputation (Hozé et al. 2013).

Lorsque la qualité de l'imputation est bonne, les GEBV obtenues à partir de données imputées ou à partir des vraies données sont généralement très corrélées (Dassonneville et al. 2011), mais les erreurs d'imputation peuvent tout de même détériorer la qualité des prédictions génomiques, surtout si elles concernent uniquement les candidats à la sélection (van den Berg et al. 2017). Des prédictions basées sur des données génotypiques comportant des erreurs peuvent en revanche conduire à sous-estimer les GEBV des meilleurs candidats à la sélection imputés et à surestimer les GEBV des plus mauvais individus (Pimentel et al. 2015).

1.2.6 Impact de la sélection génomique sur la diversité génétique

La mise en place de la sélection génomique chez les espèces animales et végétales étant relativement récente, ses effets sur la diversité génétique ont dans un premier temps été étudiés d'un point de vue théorique ou en utilisant des simulations. Dans le cas des populations animales, certains auteurs avaient ainsi avancé que cette approche permettrait de réduire le taux de consanguinité des individus sélectionnés puisque l'estimation des valeurs génétiques additives des candidats à la sélection inclurait l'aléa de méiose contrairement aux évaluations génétiques classiques, conduisant donc à la sélection d'individus moins apparentés (Daetwyler et al. 2007). Cependant, la réduction de l'intervalle entre générations pouvait également laisser craindre un accroissement plus fort de la consanguinité par unité de temps (Schaeffer 2006). Les études récentes concernant les espèces ayant rapidement mis en place la sélection génomique montrent en fait des résultats contrastés. En effet, une augmentation du taux de consanguinité a effectivement été constatée en race Holstein et Jersey (Doublet et al. 2019 ; Makanjuola et al. 2020 ; Scott et al. 2021a) mais pas en race Montbéliarde, Normande ou Angus (Doublet et al. 2019 ; Lozada-Soto et al. 2021). Bien que la sélection génomique ait permis de proposer un plus grand nombre de taureaux sur le marché, ces taureaux sont tout de même fortement apparentés en Holstein et dérivent d'un petit nombre de taureaux, ce qui peut expliquer ces observations (Le Mézec et al. 2018). Rutkoski et al. (2015) ont aussi observé une augmentation de l'apparentement entre lignées après deux générations de sélection génomique chez le blé. Plusieurs études s'appuyant sur des simulations ont également montré qu'à long terme, la perte de variance génétique additive par génération était plus importante lorsque la sélection génomique était utilisée (Allier et al. 2019a ; Silva et al. 2021 ; Wientjes et al. 2022). Au niveau génomique, Jannink (2010) a montré à l'aide de simulations que la sélection génomique conduisait à la perte d'un nombre d'allèles favorables plus important que la sélection phénotypique, surtout si ces allèles se trouvaient en faible fréquence dans la population d'entrainement. En parallèle, les variants délétères sont également purgés plus rapidement du génome des meilleurs candidats lorsque ces derniers sont choisis sur la base de leur GEBV (Kono et al. 2019).

1.2.7 Utilisation de la sélection génomique dans le cadre de programmes de pré-breeding

Bien que la sélection génomique puisse conduire à une diminution de la variabilité génétique, son utilisation représente également une opportunité d'exploiter des ressources génétiques peu ou pas utilisées dans les programmes d'amélioration.
Les prédictions génomiques peuvent permettre d'identifier des génotypes intéressants sans avoir à phénotyper ce matériel s'il a été génotypé, ce qui représente un avantage certain lorsque le

nombre de génotypes à évaluer est élevé (Martini et al. 2021). Cette approche a été mise en œuvre pour évaluer les génotypes présents dans des banques de gènes chez plusieurs espèces telles que le blé (Crossa et al. 2016), le sorgho (Yu et al. 2016), le soja (Jarquin et al. 2016) ou encore le riz (Tanaka et al. 2021), en utilisant un sous-ensemble des génotypes de la banque de gènes comme population d'entrainement. Les précisions de prédiction généralement élevées obtenues dans ces études, même lorsqu'un faible nombre de marqueurs était utilisé (Muleta et al. 2017) tendent à confirmer l'intérêt des prédictions génomiques pour la valorisation de ressources génétiques.

Dans un programme d'amélioration, les ressources génétiques ne sont généralement pas directement utilisables du fait de leur faible performance agronomique en comparaison au matériel élite. Plusieurs auteurs ont dans ce cas proposé d'avoir recours à la sélection génomique afin de rapidement améliorer du matériel issu de telles ressources génétiques grâce à plusieurs générations de sélection récurrente (Allier et al. 2020b ; Gorjanc et al. 2016) ou de croisements avec du matériel élite (Bernardo 2009 ; 2016) en choisissant les meilleurs candidats sur la base de leur GEBV à chaque génération.

Plutôt que de sélectionner les parents sur la base de leur GEBV, il est également possible de se servir des résultats des prédictions génomiques pour optimiser le choix des parents dans des croisements de type « ressources génétiques x élite » afin de valoriser leur complémentarité (Moeinizade et al. 2020). Il a par exemple été proposé d'utiliser les prédictions génomiques pour prédire la variance dans la descendance d'un croisement afin d'estimer l'intérêt dudit croisement grâce au critère d'utilité, qui dépend de la moyenne et de la variance attendues du croisement (Allier et al. 2019a ; Lado et al. 2017 ; Wolfe et al. 2021). Allier et al. (2019b) ont quant à eux proposé d'estimer les effets d'haplotypes le long du génome afin de sélectionner les donneurs les plus complémentaires d'un parent élite pour certains haplotypes présentant une faible valeur chez l'élite. De nombreuses autres approches permettant d'identifier des donneurs intéressants parmi un panel de ressources génétiques ont été proposées. La plupart de ces approches sont répertoriées par Labroo et al. (2021) et en partie décrites par Civan et al. (2021).

Enfin, les prédictions génomiques peuvent être mises en œuvre dans une optique de gestion de la diversité génétique dans les programmes d'amélioration. Pour limiter la perte des allèles rares favorables dans la population des candidats, plusieurs auteurs ont ainsi proposé de pondérer les effets estimés des marqueurs en fonction de leur fréquence allélique, afin de donner un poids plus fort aux allèles en faible fréquence (Goddard 2009 ; Jannink 2010 ; Liu et al. 2015 ; Sun et VanRaden 2014). Yang et al. (2020) ont quant à eux montré que dans un croisement biparental « exotique x élite », il était possible d'identifier l'origine des allèles afin de donner plus de poids aux allèles favorables provenant du matériel exotique, permettant là aussi d'augmenter leur fréquence. Les marqueurs moléculaires peuvent également être utilisés dans le but de chercher un compromis entre gain génétique et perte de diversité au cours des générations (Fernández

et al. 2021 ; Meuwissen et al. 2020). Des études par simulations ont montré que ces approches permettaient un gain génétique à long terme plus important (Allier et al. 2020b ; Tiret et al. 2021) lorsqu'elles étaient utilisées dans le cadre de prédictions génomiques.

1.3 Le pommier, une espèce fruitière majeure

Le pommier cultivé, *Malus domestica*, est un arbre de la famille des *Rosaceae* (au même titre que d'autres espèces fruitières ou ornementales d'importance économique majeure telles que l'abricotier, le cerisier, le fraisier ou encore le rosier), de la sous-famille des *Maloideae*, de la tribu des *Pyrae* et du genre *Malus*, qui comprend une soixantaine d'espèces botaniques (Korban 2021, chapitre 2). L'importance économique majeure du pommier a conduit au développement rapide d'outils génétiques permettant d'étudier son génome, ce qui fait du pommier un bon modèle d'étude pour l'implémentation de la sélection génomique chez les espèces fruitières.

1.3.1 Importance économique du pommier

La pomme peut être classée en deux grandes catégories : les pommes à couteau, destinées à être consommées telles quelles ou à être transformées pour la production de jus ou de compote, et les pommes à cidre destinées à la production de cidre. Il s'agit d'un des fruits les plus appréciés aux quatre coins du globe : en 2017, la pomme était ainsi la deuxième espèce fruitière la plus cultivée au monde avec une production annuelle de 83Mt et une surface de vergers de 4.6 Mha, essentiellement en climat tempéré (FAOSTAT, 2020). La valeur de la production mondiale en 2017 s'élevait à un peu moins de 37 milliards de dollars (Korban 2021, chapitre 1). En France, elle reste le fruit préféré des Français en 2021 et représentait une production annuelle de 1.6 Mt en 2020, ce qui faisait alors de la France le 10ème producteur mondial, la production étant par ailleurs largement dominée par la Chine (environ 40Mt produites). Bien que plusieurs milliers de variétés à travers le monde aient été recensées, le marché de la pomme est aujourd'hui dominé par un petit nombre de génotypes. En France, les variétés Golden, Gala, Pink Lady®, Granny Smith et Fuji représentaient par exemple à elles seules un peu plus de 70% de la récolte en 2021[1].

1.3.2 Une brève histoire de la domestication du pommier

Le pommier a d'abord été domestiqué dans les montagnes du Tian Shan en Asie centrale à partir de l'espèce sauvage apparentée *Malus sieversii* (Cornille et al. 2014, figure 1.10). Il est intéressant de noter qu'avant toute action humaine, certains animaux de cette région (ours, chevaux) ont vraisemblablement contribué à la sélection de fruits plus gros et plus sucrés (Spengler 2019 ; Yao et al. 2015) et que la sélection humaine s'est ensuite exercée sur ces mêmes caractères au cours de la domestication de *M. domestica* (Khan et al. 2014a).

Les échanges de graines le long de la Route de la Soie ont mené à son introduction en Europe il

1. http://lapomme.org/chiffres/production-en-france-par-varietes

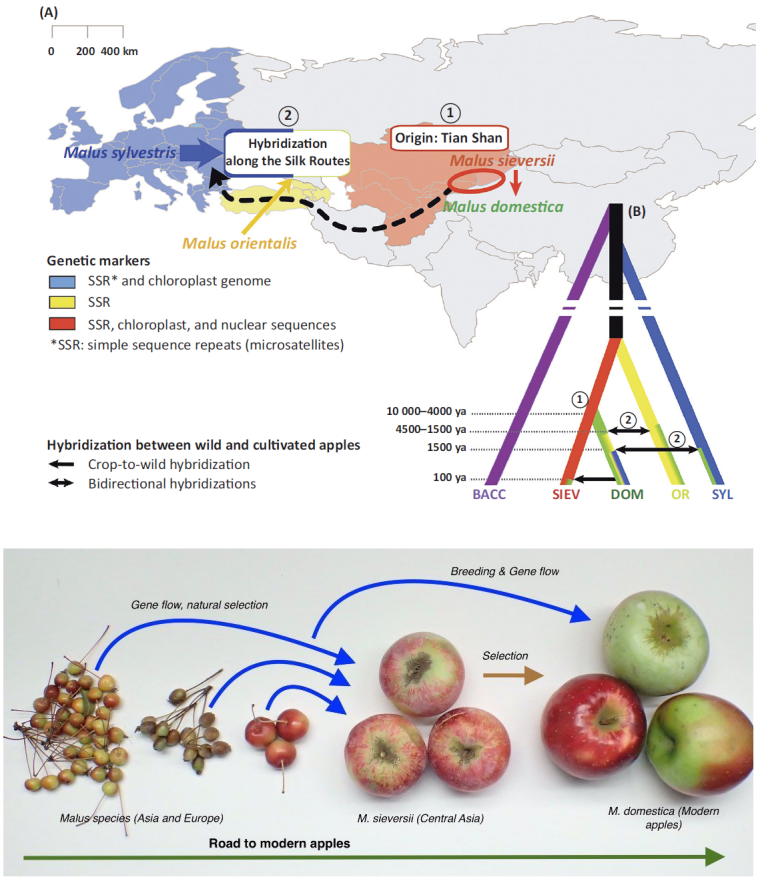


FIGURE 1.10 – (En haut) Histoire évolutive des pommier cultivé. **(A)** (1) Origine dans les montagnes du Tian Shan de *Malus sieversii*, suivie de (2) dispersion de l'Asie vers l'Europe le long de la Route de la Soie, facilitant l'hybridation et l'introgression des pommiers sauvages caucasiens et européens. **(B)** Relations phylogénétiques entre les pommiers sauvages et les pommiers cultivés. Les dates approximatives des événements de domestication et d'hybridation entre les espèces sauvages et cultivées sont détaillées dans la légende. **BACC** : *Malus baccata* ; **DOM** : *Malus domestica* ; **OR** : *Malus orientalis* ; **SIEV** : *Malus sieversii* ; **SYL** : *Malus sylvestris* (d'après Cornille et al. (2014))
(En bas) Evolution de la taille du fruit chez le pommier au cours de la domestication (crédit : Khan Lab, Cornell, USA)

y a environ 3000 à 4000 ans. Des traces archéologiques attestent de l'utilisation de techniques de greffage à cette période, permettant le maintien et la transmission des meilleurs génotypes (Cornille et al. 2014). Des analyses génétiques ont également permis de conclure que plusieurs événements d'hybridation avec des espèces sauvages locales avaient eu lieu le long de la Route de la Soie, notamment au contact des espèces *M. sylvestris* et *M. orientalis* (Cornille et al. 2019 ; 2012). Sun et al. (2020) estiment ainsi que 23% du génome de la variété Gala est d'origine hybride.

1.3.3 Organisation du génome du pommier

Le génome du pommier, d'une taille d'environ 650 Mb (Daccord et al. 2017) est composé de 17 chromosomes. La majorité des variétés est diploïde (2n = 34), bien qu'un nombre important de variétés anciennes soient triploïdes (environ 20% dans la collection française d'après Lassois et al. (2016)). La grande majorité des pommiers présente un mécanisme d'auto-incompatibilité gamétophytique, et de ce fait le génome du pommier est hautement hétérozygote.
Le pommier a été la dixième espèce végétale séquencée (Velasco et al. 2010) et plus récemment, plusieurs génomes de variétés ou génotypes majeurs ont été assemblés, notamment celui d'un haploïde doublé de Golden Delicious (Golden Delicious Doubled Haploid n° 13 ou GDDH13, Daccord et al. (2017)), d'un haploïde triplé de Hanfu (Hanfu derived Trihaploid 1 ou HFTH1, Zhang et al. (2019)) et de la forme diploïde de Gala (Sun et al. 2020), permettant une meilleure compréhension de la structure du génome du pommier. En utilisant la version HFTH1 du génome, 44 667 gènes codant des protéines ont été identifiés (Zhang et al. 2019). L'étude de ces génomes a également permis de montrer que les 17 chromosomes du pommier dérivent d'une duplication complète récente du génome (WGD pour Whole Genome Duplication) ayant eu lieu il y a 13 à 27 millions d'années (Su et al. 2021), ce qui en fait une espèce de choix pour étudier la dynamique des génomes après un tel événement (Han et al. 2011). Les chromosomes ou fragments de chromosomes homéologues résultant de cette duplication complète sont représentés figure 1.11.

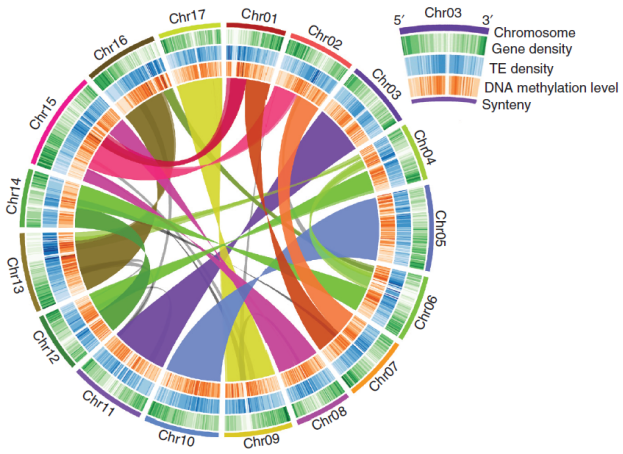


FIGURE 1.11 – Relations de synténie au sein du génome du pommier domestique et distribution des caractéristiques génomiques et épigénomiques du génome du pommier (d'après Daccord et al. (2017))

1.4 Création variétale chez le pommier

1.4.1 Objectifs de sélection

Il existe différents programmes d'amélioration du pommier à travers le monde, principalement portés par le secteur public. Les principaux objectifs de sélection sont globalement les mêmes d'un pays à l'autre (Laurens 1998) et nous pouvons les classer en trois grandes catégories.

— **Résistance aux stress biotiques** : il existe de nombreux bioagresseurs du pommier, parmi lesquels les ascomycètes *Venturia inaequalis*, responsable de la tavelure du pommier et *Podosphaera leucotricha*, responsable de l'oïdium, la bactérie *Erwinia amylovora* à l'origine du feu bactérien ou encore divers insectes tels que le puceron lanigère (*Eriosoma lanigerum*) et le puceron cendré (*Dysaphis plantaginea*). Compte tenu du grand nombre de traitements phytosanitaires actuellement appliqués dans les vergers de pommier (en 2018, ce nombre s'élevait en moyenne à 38 traitements par an en conduite conventionnelle et à 27 en agriculture biologique), la résistance aux stress biotiques est aujourd'hui une priorité dans le processus de création variétale chez le pommier, de manière à réduire drastiquement ce nombre de traitements représentant un risque pour l'environnement et la santé humaine. Tous les programmes d'amélioration cherchent ainsi à obtenir des variétés résistantes à la tavelure, et dans une moindre mesure aux autres bioagresseurs (voir schéma de sélection). Un certain nombre de gènes majeurs et QTL de résistance à ces bioagresseurs sont connus et de nombreux programmes cherchent à pyramider plusieurs QTL afin de proposer une résistance durable en diversifiant les mécanismes de résistance.

— **Qualité commerciale** : certains caractères sont recherchés par les producteurs, à commencer par la productivité, la régularité de production, l'aptitude à la conservation et au stockage des fruits et l'attrait du fruit (notamment lié à la coloration du fruit, à sa forme, à son calibre et à l'absence de défauts cosmétiques).

— **Qualité organoleptique** : la qualité gustative des fruits est un autre élément pris en compte dans les programmes d'amélioration. Les principaux critères retenus pour évaluer la qualité correspondent à la texture du fruit (croquant, absence de farinosité...), son rapport sucre/acide et sa quantité de jus. Sélectionner des fruits de bonne qualité est un processus complexe : d'une part, le terme « qualité » regroupe un grand nombre de paramètres pour un sélectionneur. Par exemple, entre 10 et 20 paramètres associés à la qualité sont notés lors de l'évaluation des hybrides dans les programmes d'amélioration INRAE-NOVADI (voir figure 1.12). D'autre part, le phénotypage des caractères liés à la qualité est complexe : bien qu'il existe des appareils permettant de mesurer un

certain nombre de caractères liés à la qualité [2], il est nécessaire de ramener les fruits en laboratoire pour les évaluer. Le nombre de fruits pouvant être simultanément mesurés est de plus limité, si bien que la qualité est plutôt évaluée en verger par des notateurs entrainés sur la base de dégustations. Enfin, la majorité de ces paramètres sont sous contrôle polygénique (Jenks et Bebeli 2011, chapitre 3), rendant l'identification de loci liés à la qualité difficile. Certains caractères liés à la qualité (notamment le taux de sucre et le taux d'acidité) répondent également à des logiques d'optimum et non pas de maximum, et il est dans ce cas difficile de trouver les meilleures combinaisons alléliques. Pour toutes ces raisons, il n'est pas étonnant que la plupart des programmes de création variétale s'appuient sur un nombre restreint de géniteurs répondant déjà à des critères de qualité.

De nouveaux caractères commencent également à être pris en compte dans les programmes de sélection, à commencer par l'architecture de l'arbre (Segura et al. 2009), qui est parfois corrélée à la régularité de production (Guitton et al. 2012) et peut également permettre de faciliter la conduite de l'arbre par la taille, permettant de limiter les interventions humaines (le coût de la main d'œuvre étant la dépense principale dans le budget des arboriculteurs). L'adaptation au changement climatique est également considérée (Legave 2022, chapitre 11), notamment en cherchant à retarder les dates de floraison afin d'éviter les gelées tardives ou en s'intéressant à des génotypes avec une meilleure efficacité d'utilisation de l'eau (Coupel-Ledru et al. 2022).

1.4.2 Études du déterminisme génétique de caractères d'intérêt

Dès la fin du 20ème siècle, des efforts importants ont été consacrés au développement de cartes génétiques pour comprendre le déterminisme génétique de caractères d'intérêts chez le pommier (Maliepaard et al. 1998 ; Weeden et al. 1994). Ces premières cartes, basées sur des marqueurs RAPD (Randomly Amplified Polymorphic DNA) ont notamment permis d'identifier des zones du génome impliquées dans l'acidité ou la couleur du fruit (Conner et al. 1997). Par la suite, des cartes basées sur l'utilisation de marqueurs SSR (Short Sequence Repeat), AFLP (Amplified Fragment Length Polymorphism) et SNP ont été développées (Kenis et Keulemans 2005 ; Khan et al. 2012 ; Liebhard et al. 2002 ; Silfverberg-Dilworth et al. 2006) et ont été largement utilisées pour cartographier de nombreux caractères, notamment liés à la qualité du fruit (Ben Sadok et al. 2015 ; Costa 2015 ; Liebhard et al. 2003), à la couleur du fruit (Chagné et al. 2007 ; Moriya et al. 2017), à l'architecture de l'arbre (Segura et al. 2009) ou encore à la résistance à certaines maladies (par exemple Calenge et al. (2004) pour la tavelure, Calenge et al. (2005) pour le feu bactérien et Calenge et Durel (2006) pour l'oïdium).

Le séquençage du génome de Golden Delicious en 2010 a également permis de développer

2. Par exemple https://www.setop.eu/fr/produit/pimprenelle

plusieurs puces de génotypage haut-débit contenant 8K SNP (Chagné et al. 2012), 20K SNP (Bianco et al. 2014) et 480K SNP (Bianco et al. 2016). Les deux premières puces ont été développées respectivement en utilisant des SNP détectés à partir de 27 et 13 accessions et en se focalisant sur les régions codantes du génome. La densité de marqueurs n'était de plus vraisemblablement pas suffisante pour effectuer des études d'association compte tenu de la décroissance rapide du déséquilibre de liaison chez le pommier (Kumar et al. 2014). Afin de mieux prendre en compte la diversité génétique existant chez le pommier, la puce haute densité 480K SNP a été développée en utilisant 65 accessions représentatives de la diversité mondiale chez le pommier, mais uniquement issues du pool domestique. Les SNP ont été choisis de façon à couvrir le génome de façon uniforme en ciblant aussi bien les régions codantes (72% des SNP) que non-codantes du génome. Parmi les SNP retenus, 7,5% étaient rares, c'est-à-dire que la fréquence de l'allèle minoritaire n'excédait pas 5% pour ces SNP. Le génotypage par séquençage est également utilisé (Gardner et al. 2014).
Les premières populations utilisées pour la cartographie génétique étaient des famillles biparentales, mais ces familles présentaient l'inconvénient de ne pas exploiter l'ensemble de la diversité allélique rencontrée chez le pommier et d'identifier des QTL qui n'étaient pas détectés dans d'autres familles. L'utilisation de populations multi-parentales (Peace et al. 2014) et l'utilisation d'informations de pédigrée (Pedigree-Based Analysis ou PBA, Bink et al. (2014)) a permis de dépasser ces limitations. La densité de marqueurs de la puce 480K a également permis de réaliser des études de génétique d'association (Genome-Wide Association Study ou GWAS, par exemple Urrestarazu et al. (2017)).
Les études menées depuis une vingtaine d'années sur le déterminisme génétique des caractères d'intérêt chez le pommier montrent que la majorité des caractères est sous contrôle polyénique, en particulier en ce qui concerne la qualité du fruit. Pour certains caractères comme la couleur du fruit, la date de récolte ou encore la résistance à certaines maladies, quelques QTL à effet fort sont connus et des marqueurs moléculaires ont pu être développés pour suivre la présence de ces QTL dans les programmes d'amélioration. Cet aspect est détaillé dans la section 1.4.4.1.

1.4.3 Schémas de sélection chez le pommier

Cette section a pour but de présenter à quoi ressemble un schéma de sélection « classique » chez le pommier en prenant pour exemple le programme de création variétale mis en œuvre sur le site d'Angers d'INRAE. Il existe cependant de fortes similitudes entre les programmes majeurs à travers le monde (voir Evans (2013) pour un exemple aux Etats-Unis et Kumar et al. (2012a) pour un exemple en Nouvelle-Zélande).

Le programme de création variétale pommier mené à l'Institut de Recherche en Horticulture

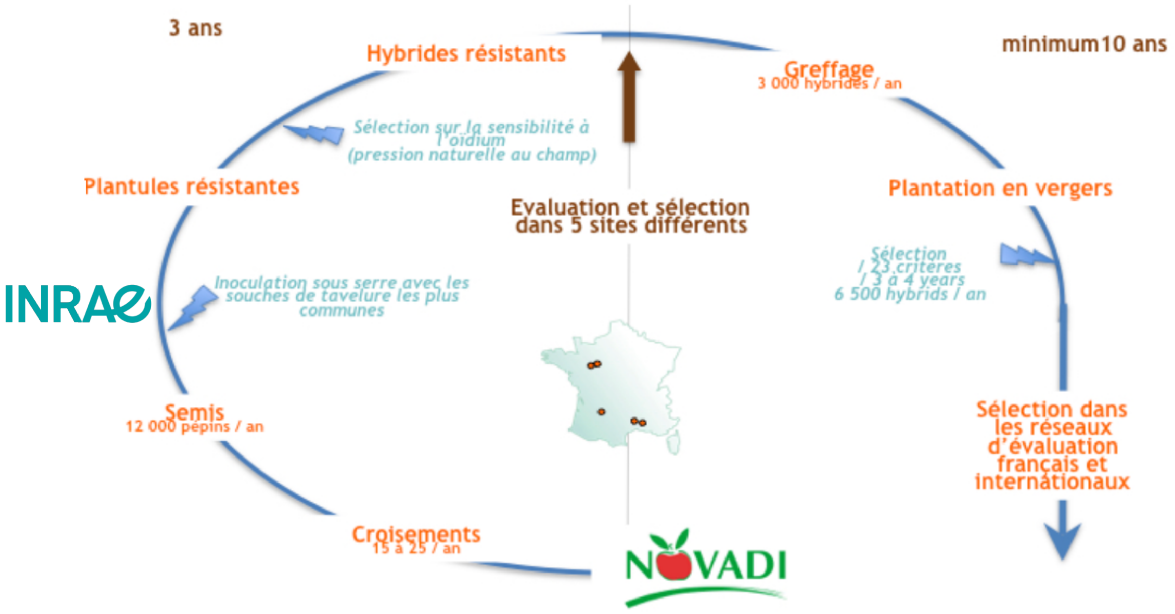


FIGURE 1.12 – Schéma de sélection du pommier tel que développé dans le cadre du partenariat INRAE-NOVADI

et Semences (ou IRHS) à Angers est un programme d'amélioration collaboratif privé-public impliquant des chercheurs INRAE et l'association de pépiniéristes NOVADI, regroupant 13 pépiniéristes français.

Chaque année, 15 à 25 croisements contrôlés sont effectués après concertation entre INRAE et NOVADI et 10 à 15000 pépins résultant de ces croisements sont ensuite semés en plaques afin de tester leur résistance à la tavelure par inoculation en serre. Les plantules résistantes à la tavelure (correspondant en moyenne à 6000 plantules) sont ensuite plantées en pépinière de grossissement et évaluées pour leur tolérance à l'oïdium pendant un an et demi. Ces deux étapes ont lieu sur le site angevin d'INRAE. Les individus tolérants (environ 3000) sont multipliés par greffage et évalués pendant plusieurs années sur 5 sites de NOVADI représentant des environnements contrastés. Les individus les plus prometteurs sont enfin évalués pendant plusieurs années en condition de production et dans un réseau d'expérimentation large au niveau national et international. Le processus de sélection du matériel prometteur représente au minimum 10 années, ce qui signifie que le travail de création variétale prend généralement 15 à 20 ans chez le pommier avant la mise en examen DHS (Distinction Homogénéité Stabilité) pour une inscription au Catalogue Officiel des variétés françaises.

Ce travail est fastidieux car le pommier a une phase juvénile (période sans pouvoir donner de fleurs) dont la durée varie généralement entre 6 et 10 ans, ce qui retarde grandement l'évaluation des caractères de qualité du fruit. Cette phase juvénile peut néanmoins être raccourcie à environ 4 ans grâce à l'utilisation de porte-greffes (Fischer 1994), voire à moins de 4 ans en utilisant des techniques de transgenèse (Flachowsky et al. 2009 ; Yamagishi et al. 2014).

1.4.4 Utilisation des marqueurs moléculaires

1.4.4.1 Utilisation de la sélection assistée par marqueurs

L'utilisation de marqueurs moléculaires dans les programmes d'amélioration peut permettre une sélection précoce des candidats à la sélection porteurs d'allèles d'intérêt, permettant ainsi de raccourcir la durée des cycles de sélection ou de n'implanter en vergers d'hybrides que les individus présentant les meilleures combinaisons d'allèles aux loci ciblés par les marqueurs. Compte tenu du temps nécessaire pour obtenir une variété chez le pommier, il est normal que de nombreux instituts se soient penchés sur cette possibilité, notamment en développant des marqueurs associés à des gènes majeurs ou QTL détectés dans des études de cartographie génétique (Frey et al. 2004).
Cette approche, que nous désignons dans la suite du texte par « sélection assistée par marqueurs post-QTL » (ou SAM post-QTL) a historiquement été utilisée pour suivre la transmission de gènes de résistance aux maladies (Baumgartner et al. 2016 ; Peil et al. 2008). Aujourd'hui, la SAM post-QTL est utilisée par plusieurs programmes d'amélioration à travers le monde (Ru et al. 2015). A INRAE, les premiers essais de SAM post-QTL remontent à 2003. Cette approche est encore utilisée, surtout pour guider la création de géniteurs résistants à la tavelure, en favorisant le pyramidage de gènes majeurs et de QTLs dans une perspective de durabilité de la résistance (Lasserre-Zuber et al. 2018).
Certains marqueurs liés à la qualité du fruit ont également été développés, notamment pour la texture du fruit (Longhi et al. 2013 ; Migicovsky et al. 2021 ; Zhu et Barritt 2008). Cependant, seul le programme d'amélioration de l'Université de Washington semble utiliser ces marqueurs en routine (K.Evans, communication personnelle, citée dans Kumar et al. (2012a)).
Récemment, les projets FruitBreedomics (Laurens et al. 2018) et RosBREED (Iezzoni et al. 2020), respectivement menés à l'échelle européenne et nord-américaine, ont été menés en collaboration entre plusieurs instituts afin d'accélérer l'utilisation d'outils génomique chez certaines espèces fruitières de la famille des *Rosacées*. Ces deux programmes ont largement contribué au développement et à l'utilisation des marqueurs moléculaires dans les programmes de création variétale chez le pommier.

1.4.4.2 Utilisation de la sélection génomique

Bien que la sélection assistée par marqueurs post-QTL ait démontré son intérêt chez le pommier et soit désormais mise en place dans plusieurs programmes d'amélioration à travers le monde, son utilisation est limitée pour les caractères contrôlés par un grand nombre de QTL à effets faibles (Bernardo 2008). La sélection génomique a donc rapidement été perçue comme une extension pertinente de la SAM post-QTL permettant de s'affranchir de la pré-sélection

de marqueurs moléculaires associés à un caractère donné, tout en représentant une opportunité d'augmenter le gain génétique par unité de temps (Kumar et al. 2012a).
Kumar et al. (2012b) estiment ainsi qu'en utilisant des techniques permettant une initiation rapide de la floraison (van Nocker et Gardiner 2014), la sélection génomique permettrait de ne pas phénotyper les candidats à la sélection, réduisant ainsi la durée entre deux générations de 7 à 4 ans pour un programme d'amélioration évaluant les candidats pendant 3 ans. Cependant, à notre connaissance, seul l'institut Plant and Food Research (PFR) en Nouvelle-Zélande utilise la sélection génomique dans ses programmes d'amélioration à ce jour, bien que d'autres instituts envisagent sérieusement son implémentation dans leur programme de création variétale (Blissett 2019).
Les travaux explorant l'intérêt de la sélection génomique chez le pommier se sont jusqu'à présent surtout focalisés sur la précision de prédiction qu'il était possible d'atteindre pour différents caractères, principalement liés à la qualité du fruit dans des familles biparentales (voir tableau 1.1 pour une synthèse des travaux ayant exploré la question de la prédiction génomique chez le pommier). La première étude visant à mesurer la précision de prédiction chez le pommier (Kumar et al. 2012b) a fourni des résultats encourageants puisque les auteurs ont obtenu des précisions de prédiction élevées (0,68 à 0,86) pour les 5 caractères de qualité du fruit qu'ils avaient étudiés. L'apparentement entre la population d'entrainement et la population des candidats était élevé dans ce cas puisque la précision de prédiction avait été mesurée par validation croisée à partir de 7 familles de plein-frères mais sans considérer cette structure. Des études ultérieures ayant mesuré la précision de prédiction intra-famille avec un population d'entrainement consituée de familles similaires (Kumar et al. 2015 ; Muranty et al. 2015) voire d'un panel de diversité (Roth et al. 2020) ont obtenu des précisions de prédiction moins élevées. De même, les précisions de prédiction obtenues lorsque les candidats étaient choisis parmi un panel de diversité étaient moins élevées que dans le cas de l'utilisation de familles de plein-frères (Kumar et al. 2020 ; McClure et al. 2018 ; 2019 ; Migicovsky et al. 2016). Dans la plupart des études, le nombre de marqueurs utilisés était peu élevé et la taille de la population d'entrainement relativement variable, les travaux menés sur un panel de diversité comportant généralement un nombre limité de génotypes. La grande majorité des études citées a utilisé un modèle de prédiction estimant les effets additifs des marqueurs. L'étude de Kumar et al. (2015) est la seule à avoir pris en compte les effets de dominance dans le modèle mais les auteurs n'ont pas observé d'amélioration de la précision de prédiction dans ce cas. Minamikawa et al. (2021) ont testé différents modèles basés sur l'utilisation d'haplotypes et ont observé pour certains caractères des précisions de prédiction plus élevées que dans le cas de l'utilisation directe des marqueurs SNP.

TABLEAU 1.1 – Synthèse des travaux de prédiction génomique menés chez le pommier à ce jour

Réf.	Caractère(s) étudié(s)	Modèle utilisé	Nombre de marqueurs	Précision de prédiction	Constitution TS	Taille TS	Candidats
[1]	Qualité du fruit	RR-BLUP LASSO	2500	0,68-0,86	7 familles FS	1200	10% TS
[2]	Qualité du fruit	GBLUP	2828	0,15-0,35 0,5-0,83	17 familles FS	248	1 famille FS $1/17^{ème}$ TS
[3]	Productivité Apparence du fruit	BayesCπ	7652	-0,02-0,5	20 familles FS	977	5 familles FS
[4]	Apparence du fruit Qualité du fruit Phénologie	RR-BLUP	4395	0,02-0,57	Panel de diversité	689	20% TS
[5]	Qualité du fruit Résistance tavelure	RR-BLUP	75130	0,08-0,72	Panel de diversité	172	20% TS
[6]	Qualité du fruit	RR-BLUP	98584	0-0,49	Panel de diversité	136	20% TS
[7]	Date début de floraison Date de récolte	RR-BLUP	303239	0,57 0,76	Panel de diversité + 27 familles FS	564	20% TS
[8]	Qualité du fruit	GBLUP	6400	0,22-0,61	Panel de diversité	274	$1/11^{ème}$ TS
[9]	Texture du fruit	RR-BLUP	8294	-0,29-0,73	Panel de diversité	259	6 familles FS
[10]	Qualité du fruit	BayesB	11786	0-0,64	16 familles FS	659	1 famille FS

TS : population d'entrainement (Training Set)
Références utilisées : [1](Kumar et al. 2012b) [2](Kumar et al. 2015) [3](Muranty et al. 2015) [4](Migicovsky et al. 2016) [5](McClure et al. 2018) [6](McClure et al. 2019) [7](Jung et al. 2020) [8](Kumar et al. 2020) [9](Roth et al. 2020) [10](Minamikawa et al. 2021)

1.5 Diversité génétique chez le pommier : gestion et utilisation en création variétale

1.5.1 Diversité disponible chez le pommier domestique

Chez le pommier, aucun goulot d'étranglement lié au processus de domestication n'a pu être mis en évidence (Cornille et al. 2012 ; Gross et al. 2014). Par rapport à des espèces annuelles, un faible nombre de générations sépare de plus le début de la domestication du pommier des variétés actuelles (Spengler (2019) estime ce nombre de générations à une centaine), principalement du fait de la longueur de la phase juvénile chez le pommier ainsi que de l'utilisation de la multiplication végétative (Miller et Gross 2011). De plus, de nombreux contacts secondaires avec les espèces sauvages apparentées ont eu lieu chez le pommier (Cornille et al. 2014). Pour ces raisons, le pommier domestique a conservé une large partie de la diversité génétique observée chez les espèces sauvages fondatrices : en se basant sur des marqueurs SSR, Gross et al. (2014) estiment ainsi que plus de 95% de la diversité génétique présente chez *M. sieversii* est également présente chez *M. domestica*, des résultats confirmés par Duan et al. (2017) en utilisant des données SNP. Chen et al. (2021) avancent même que la diversité nucléotidique serait plus importante chez le pommier domestique que chez *M. sieversii* et *M. sylvestris* sur la base de données SNP, bien que cette conclusion s'appuie sur un échantillon de petite taille. Liao et al. (2021) ont quant à eux observé une diversité nucléotidique significativement plus faible dans les variétés modernes que dans les variétés anciennes, et suggèrent que la phase d'amélioration récente du pommier a donc entrainé un goulot d'étranglement responsable de

la diminution de la diversité génétique dans les variétés actuellement utilisées en production. L'utilisation restreinte de la diversité génétique dans les programmes d'amélioration est un problème identifié de longue date sur la base des pédigrées : Noiton et Alspach (1996) ont étudié 77 variétés modernes et montré que 5 génotypes très utilisés en sélection (McIntosh, Golden Delicious, Jonathan, Cox's Orange Pippin et Red Delicious) se trouvaient dans le pédigrée de 64% de ces variétés. La consanguinité est malgré tout limitée chez le pommier cultivé (seulement 2,3% des génotypes étudiés dans l'étude de Muranty et al. (2020) présentaient de la consanguinité sur la base de leur pédigrée, même s'il s'agissait principalement de variétés anciennes), vraisemblablement du fait de son système d'auto-incompatibilité et de la forte diversité génétique conservée. Les effets de la dépression de consanguinité chez le pommier sont par ailleurs mal connus : une seule étude ancienne détaille de tels effets (Brown (1971), cité par Noiton et Alspach (1996)) qui semblent se traduire par une réduction sévère de la vigueur de l'arbre entraînant un allongement considérable de la période juvénile. Dans les vergers d'INRAE, le port chétif de certains haploïdes doublés obtenus après sélection drastique sur semis en serre a pu être observé (Lespinasse, communication personnelle).

1.5.2 Caractérisation des ressources génétiques domestiques

Le développement des marqueurs moléculaires chez le pommier a permis une meilleure connaissance du matériel conservé dans les collections. Urrestarazu et al. (2016) a ainsi montré que les variétés anciennes européennes pouvaient être réparties en trois groupes qui correspondent préférentiellement au Sud, à l'Ouest et au Nord + Est de l'Europe (figure 1.13) mais que la différenciation au niveau du génome entre ces groupes était tout de même faible ($F_{ST} = 0{,}031$) traduisant le grand nombre d'échanges de matériel à l'échelle européenne. Il existe cependant des portions du génome fortement différenciées entre ces groupes, comme cela a par exemple pu être démontré pour certains loci contrôlant la date de maturité (Urrestarazu et al. 2017).

Au moment de l'élaboration du projet de cette thèse en 2016, les relations de parenté entre les variétés anciennes étaient par ailleurs peu connues et limitées aux relations présentées dans la littérature pomologique. L'utilisation de marqueurs moléculaires a permis de mettre en évidence de nombreuses relations d'apparentement entre les variétés anciennes au niveau européen (figure 1.14) qui n'avaient jusque là pas été documentées. Certaines variétés se retrouvent ainsi dans le pédigrée de nombreuses variétés anciennes : Muranty et al. (2020) ont par exemple montré que 18% d'un ensemble de 1325 variétés représentatives de la diversité génétique européenne descendait de Reinette Franche, variété normande du début du 16ème siècle. D'autres fondateurs majeurs en Europe sont les variétés Cox's Orange Pippin', Alexander ou encore Borowitsky.

Ces travaux de reconstruction des pédigrées sont actuellement poursuivis dans le cadre d'un projet collaboratif (Howard et al. 2018b) basé sur l'étude d'haplotypes partagés entre variétés

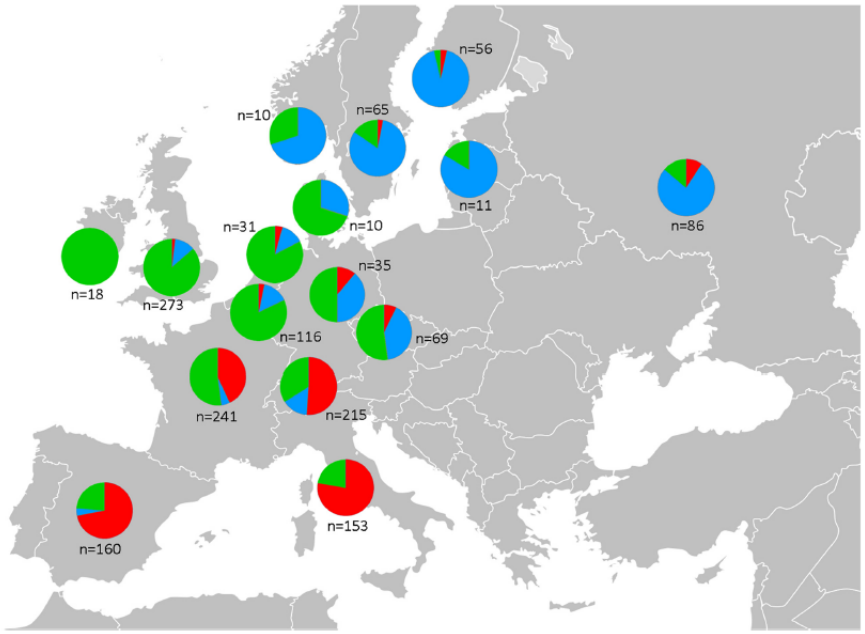


FIGURE 1.13 – Composition génétique de cultivars à l'échelle européenne en fonction de la collection d'origine. **Rouge** : origine Sud Europe. **Vert** : origine Ouest Europe **Bleu** : origine Nord + Est Europe
D'après Urrestarazu et al. (2016)

(Howard et al. 2021) afin notamment de prendre en compte un plus grand nombre de variétés à l'échelle européenne et américaine (Luby et al. 2022) ainsi que pour détecter des relations plus lointaines que celles mises en évidence par Muranty et al. (2020).

1.5.3 Gestion des ressources génétiques

Le maintien dans le temps des ressources génétiques de pommier est un enjeu capital pour leur utilisation en amélioration variétale, d'autant plus que certaines espèces sauvages sont menacées du fait des activités humaines (Shan et al. 2021) et des flux de gènes depuis le pommier domestique vers ces espèces (Feurtey et al. 2020 ; Omasheva et al. 2017). Par conséquent, de nombreux pays ont entrepris la gestion de ressources génétiques depuis la fin du 20ème siècle, participant à la sauvegarde des accessions sauvages, locales et internationales. La majorité des pays gère ses ressources génétiques uniquement sous forme de vergers (Bramel et Volk 2019). La plus grosse collection mondiale de ressources génétiques de pommiers est celle du United States Department of Agriculture (USDA), qui est largement constituée d'accessions de pommiers sauvages en verger : près de 4000 des 7000 accessions maintenues proviennent d'espèces sauvages (Fazio et al. 2008). En France, la gestion ex situ des ressources génétiques de pommier implique de nombreux acteurs : associations d'amateurs, parcs naturels régionaux, jardins botaniques, vergers conservatoires ou encore centres de recherche (Durel 2016 ; Lassois et al. 2016). Ces acteurs sont regroupés dans le Réseau Fruits à pépins dont l'animation est confiée

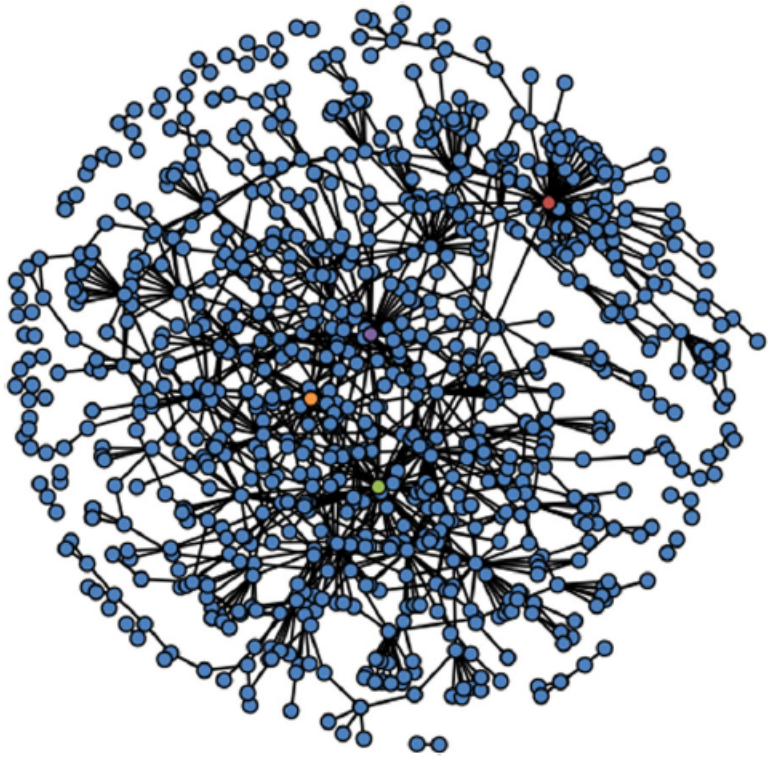


FIGURE 1.14 – Réseau d'apparentement de 832 variétés anciennes. Les variétés présentant une relation d'apparentement de premier degré sont connectées par une ligne noire et chaque variété est représentée par un point bleu, mis à part les variétés suivantes : **Rouge** : Reinette Franche. **Vert** : Cox's Orange Pippin' **Violet** : Alexander **Orange** : Borowitsky
D'après Muranty et al. (2020)

à INRAE. A INRAE, un Centre de Ressources Biologiques (ou CRB), nommé RosePom (car il rassemble à la fois des ressources biologiques de rosier et de pomoïdées fruitières) est coordonné par l'IRHS à Angers et l'Unité Expérimentale Horti. A l'heure actuelle, le CRB préserve près de 5000 accessions de pommier en verger et plusieurs milliers d'accessions sous la forme d'échantillons d'ADN (Roux-Cuvelier et al. 2021). Environ un millier de ces accessions sont des variétés locales et un petit nombre de pommiers d'espèces sauvages apparentées est également conservé. Ces ressources génétiques sont régulièrement phénotypées pour des caractères agronomiques, morphologiques et organoleptiques. Une partie des ressources génétiques a par ailleurs été génotypée à l'aide de marqueurs SSR et plus récemment SNP, permettant de mettre en évidence un certain nombre de redondances, c'est-à-dire de clones ayant des noms différents au sein des ressources génétiques mais le même profil moléculaire (Lassois et al. 2016). L'utilisation de marqueurs moléculaires a en outre permis d'établir une core-collection représentative de la diversité des ressources génétiques du CRB qui a notamment été utilisée dans des travaux des génétique d'association (Coupel-Ledru et al. 2022 ; Leforestier et al. 2015 ; Urrestarazu et al. 2017).

1.5.4 Utilisation des ressources génétiques en sélection

Les actions de gestion de la diversité évoquées dans la section précédente ont une importance capitale pour l'amélioration du pommier puisque les accessions présentes dans les collections comportent très certainement des allèles favorables en fréquence faible voire absents dans les variétés modernes.
Jusqu'à présent, les ressources génétiques de pommier ont très majoritairement été utilisées pour introduire des allèles de résistance à des maladies dans les programmes d'amélioration. L'exemple emblématique d'une telle utilisation est l'introduction dans les programmes de sélection du gène de résistance à la tavelure *Rvi6* provenant de l'espèce sauvage apparentée *M. floribunda*. Les premiers croisements impliquant *M. floribunda* datent du début du 20ème siècle, et aujourd'hui plus de 80 variétés modernes sont porteuses du gène *Rvi6* (Gessler et al. 2006). Des allèles de résistance à l'oïdium et au feu bactérien ont aussi été détectés dans du matériel sauvage (Bus et al. 2010 ; Durel et al. 2009) et ont été utilisés dans certains programmes d'amélioration à travers le monde (Flachowsky et al. 2011 ; Luo et al. 2020a).
Les projets FruitBreedomics et RosBREED ont aussi contribué à la création de géniteurs combinant plusieurs QTL de résistance à la tavelure et au feu bactérien, utilisables pour de futurs croisements et certains instituts commencent à utiliser des géniteurs issus de variétés anciennes et cumulant plusieurs gènes majeurs ou QTL de résistance aux principales maladies rencontrées chez le pommier (Kellerhalls et al. 2018 ; Testolin et al. 2021).
Cependant, peu de publications décrivent l'utilisation de ressources génétiques pour améliorer des caractères sous contrôle polygénique, probablement du fait de la difficulté d'identifier des allèles favorables liés à de tels caractères et de les suivre dans les programmes de sélection. L'utilisation de ressources génétiques peut de plus entrainer des défauts qui sont actuellement absents du matériel élite, tels que l'alternance, une mauvaise conservation des fruits, voire une détérioration de la qualité des fruits dans le cas de l'utilisation de matériel exotique. Dans le cas de l'introgression d'un gène majeur depuis les ressources génétiques vers le matériel élite, ces défauts peuvent être éliminés après quelques générations de croisements en revenant à un fond génétique majoritairement élite (Luo et al. 2020b), mais cette approche est plus délicate dans le cas de caractères contrôlés par un grand nombre de gènes. Le programme de création variétale de l'institut PFR est un des rares exemples documentés de l'utilisation de ressources génétiques dans un programme d'amélioration variétale ayant pour objectif d'élargir la base génétique du matériel élite (Kumar et al. 2010). En 1990, l'institut a constitué une population très diverse composée de descendances obtenues en pollinisation libre d'environ 500 génotypes comprenant des variétés élite, des variétés anciennes ainsi que d'espèces sauvages apparentées (Noiton et Shelbourne 1992). Les individus les plus prometteurs ont ensuite été sélectionnés sur la base de la taille du fruit et de la résistance en première génération et sur ces critères ainsi

que sur la qualité du fruit en deuxième génération (Kumar et al. 2010). La troisième génération de cette population est actuellement en cours d'évaluation en verger et des essais de prédictions génomiques ont notamment été proposés pour choisir les candidats de la génération suivante.

1.6 Objectifs de la thèse et questions de recherche

Les programmes d'amélioration actuels chez le pommier reposent sur l'utilisation récurrente d'un faible nombre de géniteurs plus ou moins apparentés entre eux et n'utilisent qu'une faible partie de la diversité génétique disponible au sein du pool domestique. Or il est probable qu'il existe des allèles favorables dans les ressources génétiques qui sont absents ou en tout cas présents en faible fréquence dans le matériel élite et qui pourraient être introduits dans les programmes de sélection.
Les ressources génétiques ont généralement une valeur agronomique moindre par rapport aux variétés récentes et peuvent de plus présenter des défauts qui ont été éliminés des variétés modernes et que les sélectionneurs ne souhaitent pas réintroduire dans leur matériel avancé. En effet, l'élimination de tels défauts peut nécessiter plusieurs générations de croisements avec des variétés élite. De tels croisements représentent un investissement conséquent chez le pommier compte tenu de la longueur de sa phase juvénile. L'utilisation de marqueurs moléculaires pourrait alors permettre d'identifier rapidement les candidats à la sélection les plus prometteurs sans attendre leur mise à fruit. La sélection génomique en particulier apparaît comme une opportunité dans les programmes de pré-breeding à deux égards : premièrement, son utilisation permettrait d'identifier des génotypes intéressants parmi les ressources génétiques à disposition d'un sélectionneur, même si ces ressources génétiques n'ont pas été évaluées (ce qui est parfois le cas pour certains caractères liés à la qualité du fruit par exemple). De plus, l'utilisation de la sélection génomique pourrait permettre de guider le choix des croisements à réaliser entre ressources génétiques et matériel élite afin de générer des progéniteurs porteurs d'allèles favorables rares ou absents dans les variétés modernes. Dans les deux cas, il est impératif d'obtenir des précisions de prédiction élevées afin que les choix qui en découlent permettent d'obtenir des géniteurs de bonne qualité. Dans le cadre de cette thèse, j'ai donc étudié trois facteurs affectant la précision des prédictions génomiques dans le cadre de l'utilisation de ressources génétiques en pré-breeding, à savoir la densité de marquage utilisée et la taille ainsi que la composition de la population d'entrainement utilisée pour calibrer les modèles de prédiction. Le travail mené peut se décliner en trois questions de recherche :

— Quelle précision d'imputation est-il possible d'obtenir dans des familles biparentales chez le pommier ?
— Peut-on identifier des génotypes intéressants issus des ressources génétiques à utiliser dans des programmes de pré-breeding grâce à la prédiction génomique ? Quels facteurs

peuvent permettre d'améliorer la qualité des prédictions ?

— Comment intégrer la sélection génomique dans un programme de pré-breeding chez le pommier ?

Je tente de répondre à chacune de ces questions de recherche dans les chapitres 3, 4 et 5 de la thèse. Dans le chapitre 3, j'ai simulé les données génotypiques haute densité de 27 familles de plein-frères correspondant à du matériel élite et ai mesuré la précision d'imputation qu'il était possible d'obtenir pour ces familles, en les ramenant artificiellement à une densité de marqueurs moins élevée et en utilisant un panel de référence composé de ressources génétiques plus ou moins apparentées aux familles simulées. Dans le chapitre suivant, j'ai utilisé deux jeux de données comprenant à la fois des génotypes provenant de ressources génétiques et de matériel élite afin de mesurer la précision de prédiction qu'il était possible d'obtenir pour différents caractères selon la constitution et la taille de la population d'entrainement, et ce en utilisant des données de marquage à moyenne ou haute densité. Enfin, dans le chapitre 5, j'ai simulé deux schémas de sélection classiquement utilisés dans des programmes de pré-breeding et ai comparé l'intérêt de la sélection génomique par rapport à la sélection phénotypique en termes de gain génétique et de valorisation des allèles rares.

CHAPITRE 2

Données utilisées au cours de la thèse

Pour répondre aux objectifs de la thèse, je me suis appuyé sur des données génotypiques et phénotypiques de trois panels provenant de différents projets de recherche passés ou en cours. Le premier panel, nommé « panel FBo-Hi » est issu de deux projets européens achevés tandis que le second panel, nommé « panel REFPOP » provient d'un projet européen toujours en cours. Les génotypes de ces deux panels correspondent soit à des variétés anciennes (que j'appellerai « ressources génétiques » dans le reste du manuscrit) soit à des descendances de familles biparentales issues de divers programmes d'amélioration du pommier à l'échelle de l'Europe qui correspondent au matériel élite des panels. Les génotypes du troisième panel proviennent d'une initiative de pré-breeding menée à INRAE Angers. Je retiens ici et dans le reste du manuscrit une définition large du terme « pré-breeding », à savoir toute action visant à la valorisation de ressources génétiques en vue d'une potentielle utilisation future dans des programmes d'amélioration.

2.1 Jeux de données utilisés

2.1.1 Jeu de données FBo-Hi

Le panel FBo-Hi contient à la fois des données provenant d'un panel de ressources génétiques exploité durant le projet EU-FP7 FruitBreedomics (Laurens et al. 2018) et d'un panel de matériel élite provenant du projet européen HiDRAS (High-quality Disease Resistant Apples for a Sustainable Agriculture, Gianfranceschi et Soglio (2004)). Les ressources génétiques correspondent à 1194 génotypes uniques (principalement des variétés anciennes de pommes à couteau,

voir figure 2.1) ayant été génotypés et phénotypés pour au moins un caractère (voir plus bas) et sont supposées représenter la diversité génétique à l'échelle des variétés européennes. Ces variétés proviennent de six core-collections européennes situées en Belgique, France, Italie, République Tchèque, Royaume-Uni et Suède. Le matériel élite est constitué de génotypes issus de six programmes d'amélioration du pommier menés dans quatre pays européens et correspond aux descendants de 23 combinaisons biparentales, représentées chacune par 18 à 115 génotypes (tableau 2.1). Au total, 1018 génotypes ont été génotypés et phénotypés pour au moins un caractère.

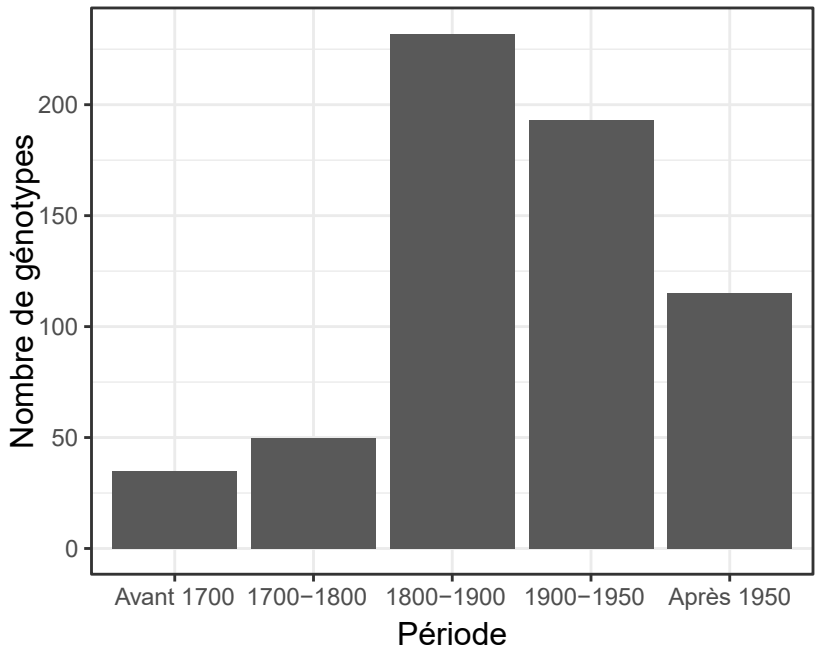


FIGURE 2.1 – Période à laquelle les variétés des ressources génétiques du panel FBo-Hi sont apparues[2](d'après les données de Muranty et al. (2020))

Les génotypes du jeu de données FBo-Hi ont uniquement été évalués sur le site de l'institut dont ils provenaient, excepté pour un petit nombre de génotypes présents sur tous les sites afin de pouvoir ajuster les données phénotypiques.

2.1.2 Jeu de données REFPOP

Les données ont été acquises à partir d'une « population de référence » du pommier décrite par Jung et al. (2020), population que je nommerai « REFPOP » dans le reste du manuscrit. Ce panel est constitué d'une part de 269 variétés anciennes qui représentent la diversité mondiale du pommier et que je dénommerai « ressources génétiques de la REFPOP » et d'autre part de

2. Pour de nombreuses variétés du panel, la période d'apparition n'a pu être établie ou était trop incertaine, elles ne sont donc pas représentées sur cette figure

265 descendants de 27 familles biparentales provenant de divers programmes d'amélioration en Europe, que je nommerai « matériel élite de la REFPOP » (tableau 2.1). Les 265 descendants peuvent être plus ou moins apparentés du fait de l'utilisation de certains parents dans plusieurs familles (voir tableau 2.1 et figure 2.2). Certains génotypes de la REFPOP font également partie du jeu de données FBo-Hi : 188 variétés anciennes et 155 descendants des familles biparentales sont ainsi communs aux deux jeux de données. Le réseau REFPOP est constitué de 5 instituts de recherche européens situés en Espagne, France, Italie, Pologne et Suisse et d'une entreprise privée belge. La REFPOP est implantée en verger par chacun des partenaires du réseau, ce qui représente des environnements contrastés. Chaque génotype au sein de la REFPOP est présent en au moins deux copies. Sur chaque site, les ressources génétiques et le matériel élite sont plantés dans le même dispositif.

2.1.3 Jeu de données Hybrides RG x E

Le panel « Hybrides RG x E » comporte 348 génotypes issus de dix combinaisons biparentales qui ont été créées à INRAE Angers dans le cadre d'une initiative de pré-breeding. Chaque combinaison biparentale résulte du croisement contrôlé entre une variété ancienne et un géniteur moderne, d'où l'appellation « Hybrides RG x E », pour « Ressources Génétiques x Elite ». Le plan de croisement du panel est présenté dans le tableau 4.1 du chapitre 4. Chaque génotype du panel est représenté par un unique arbre en verger.

2.2 Génotypage des panels

Dans le cas des panels FBo-Hi et REFPOP, les ressources génétiques ont été génotypées à haute densité en utilisant la puce Affymetrix Axiom® 480K du pommier (Bianco et al. 2016) et le matériel élite ainsi que les hybrides RG x E ont été génotypés à moyenne densité à l'aide de la puce de génotypage 20K du pommier (Bianco et al. 2014). Après l'application d'un certain nombre de filtres (voir Jung et al. (2020) pour la puce 480K et Howard et al. (2021) pour la puce 20K), 303 239 marqueurs SNP de la puce 480K ont été conservés et 10 295 marqueurs SNP dans le cas de la puce 20K. Parmi ceux-ci, 7 060 marqueurs SNP faisaient également partie de l'ensemble de marqueurs SNP retenus à partir de la puce 480K.

2.3 Phénotypage des panels

Plusieurs caractères ont été phénotypés dans les différents panels utilisés mais nous ne mentionnerons ici que ceux qui ont été analysés dans la suite de la thèse. Ces caractères correspondent

à des caractères ayant à la fois été phénotypés dans les ressources génétiques et le matériel élite d'un panel donné (tableau 2.2). Ce choix sera justifié au chapitre 4.

2.3.1 Panel FBo-Hi

Les ressources génétiques du panel ont été phénotypées entre 2012 et 2014 dans le cadre du projet FruitBreedomics par les six instituts précédemment évoqués. Plusieurs caractères liés à la qualité du fruit et à la phénologie de l'arbre ont été mesurés selon les recommandations de l'European Cooperative Programme for Plant Genetic Resources (ECPGR) et les notations ont été harmonisées entre instituts. Lorsque des données antérieures à 2012 étaient disponibles pour les caractères mesurés, elles ont également été partagées par les instituts concernés.
Le matériel élite du panel a été phénotypé de 2003 à 2005 dans le cadre du projet HiDRAS et plusieurs caractères liés à la productivité de l'arbre et à la qualité du fruit ont été mesurés. Des détails quant aux caractères mesurés et aux protocoles de phénotypage sont disponibles dans Kouassi et al. (2009) et Muranty et al. (2015).
Au total, cinq caractères ont été phénotypés à la fois dans les ressources génétiques et le matériel élite du panel FBo-Hi : la date de récolte, la couleur du fruit, l'acidité du fruit, le caractère croquant du fruit et la jutosité du fruit.

2.3.2 Panel REFPOP

Le panel a été phénotypé sur chaque site entre 2018 et 2020 par les membres du réseau REFPOP en utilisant un protocole commun présenté par Jung et al. (2022). La date de récolte et la couleur du fruit étaient les deux seuls caractères également mesurés dans le panel FBo-Hi.

2.3.3 Panel « Hybrides RG x E »

Les hybrides RG x E ont été mesurés en 2019 et 2020 sur le site d'Angers en suivant le même protocole que pour le panel REFPOP. Dans ce cas, seuls le poids de 20 fruits, le nombre de fruits par arbre, la date de récolte et la couleur du fruit ont été mesurés (voir S2.5).

2.4 Pédigrée des différents panels

Muranty et al. (2020) ont mis en évidence de nombreuses relations d'apparentement au sein des panels FBo-Hi et REFPOP, ce qui a permis de reconstruire le pédigrée des ressources génétiques de ces panels sur plusieurs générations (voir par exemple figures 3.1 et 3.2 du chapitre 3). Au début de ma thèse, une première version du pédigrée des génotypes des différents panels était

disponible et a été utilisée dans le chapitre 3. Cette première version était basée en partie sur des relations connues dans la littérature et en partie sur les pédigrées reconstruits par Muranty *et al.* Dans le cadre d'un projet plus large de reconstruction des pédigrées chez le pommier mené par Howard et al. (2018a), une deuxième version du pédigrée a été mise à ma disposition en cours de thèse. Cette version du pédigrée a été utilisée dans les chapitres 4 et 5 du manuscrit. Le tableau 2.3 indique pour chaque famille de la REFPOP le nombre d'individus dans les la première et deuxième version des pédigrées ainsi que le nombre de générations dans les deux versions. Notons que ces nombres n'augmentent pas forcément dans la version 2 du pédigrée, notamment du fait de la correction d'erreurs de relations parent-enfant par rapport à la première version du pédigrée.

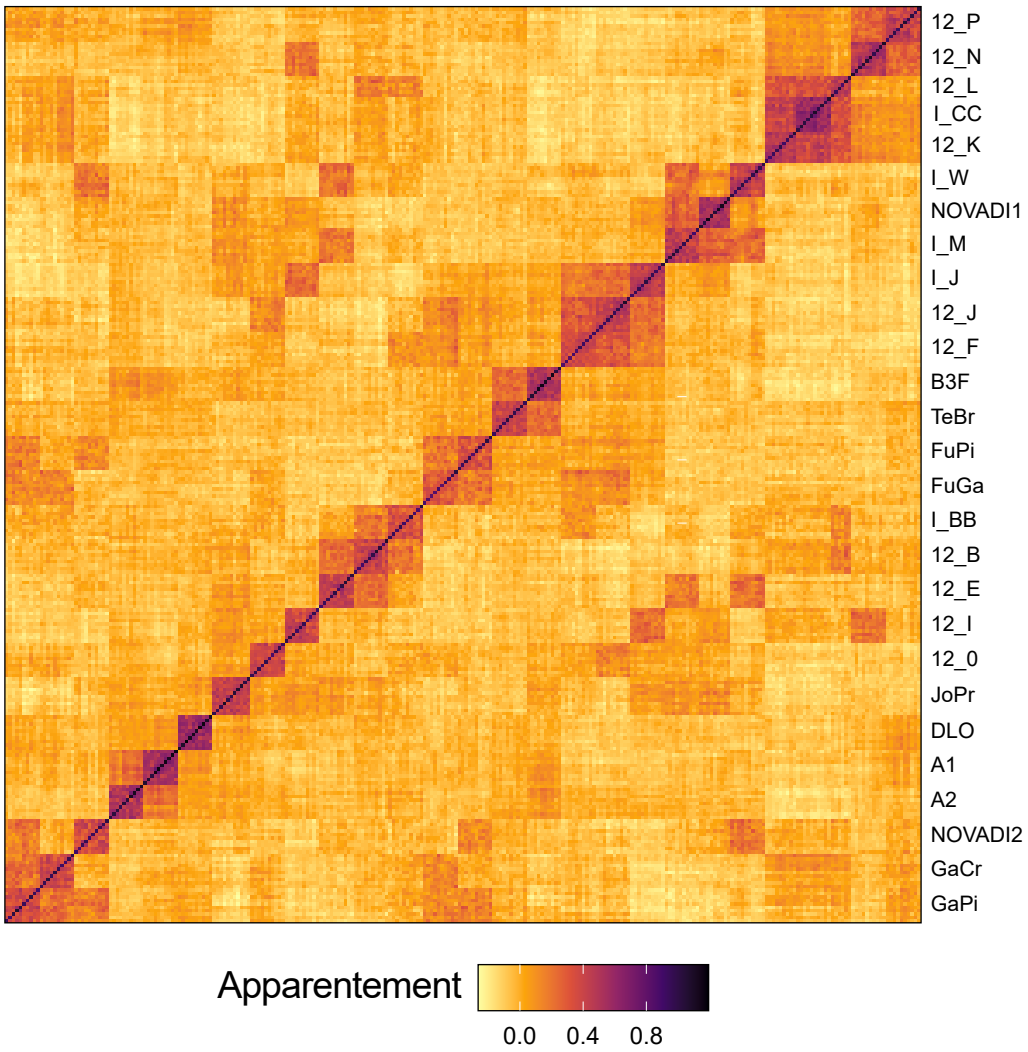


FIGURE 2.2 – Carte de chaleur représentant la matrice d'apparentement génomique du matériel élite de la REFPOP

TABLEAU 2.1 – Taille des combinaisons biparentales des panels REFPOP et FBo-Hi et parents utilisés dans les croisements

Famille	Parent 1	Parent 2	n_1	n_2
GaPi	Gala	Pinova	10	43
GaCr	Gala	Cripps Lady	10	31
NOVADI2	Pinova	X-6396	10	-
A2	11567	13652	10	-
A1	18522	18303	10	-
DLO-12	1980-15-25	1973-01-41	10	-
JoPr	Jonathan	Prima	11	-
12_O	RW x X03177	Galarina	10	-
12_I	X03259	X03263	10	46
12_E	Generos	X06683	10	-
12_B	Generos	X06417	10	-
I_BB	X06564	X06417	10	39
FuGa	Fuji	Gala	10	115
FuPi	Fuji	Pinova	10	85
TeBr	Telamon	Braeburn	10	10
B3F	12-040	Braeburn	10	-
12_F	X03318	X06564	10	44
12_J	X03318	Galarina	10	23
I_J	X03318	X03263	10	47
I_M	X06683	X06681	10	36
NOVADI1	Choupette	X06681	9	-
I_W	X06683	X06398	10	36
12_K	X06679	X06808	9	45
I_CC	X06679	Doriane	10	49
12_L	X06679	X06417	6	25
12_N	X03305	X03259	10	45
12_P	X03305	Rubinette	10	46
AlPi	Alwa	Pinova	-	18
AlSz	Alwa	Sampion	-	29
DiPr	Discovery	Prima	-	74
MeLi	Melrose	Liberty	-	41
PiGa	Pinova	Gala	-	30
PiRe	Pinova	Reanda	-	30
RePi	Rewena	Pirol	-	31
Total			265	1018

n_1 : nombre d'individus provenant d'une famille donnée dans le jeu de données REFPOP
n_2 : nombre d'individus provenant d'une famille donnée dans le jeu de données FBo-Hi

TABLEAU 2.2 – Synthèse des caractères phénotypés par panel et analysés dans le travail de thèse

	Fbo-Hi		REFPOP		Hybrides
	RG	E	RG	E	RG x E
Date de récolte	X	X	X	X	X
Couleur du fruit	X	X	X	X	X
Nombre de fruits			X	X	X
Poids de 20 fruits			X	X	X
Acidité	X	X			
Croquant	X	X			
Jutosité	X	X			

TABLEAU 2.3 – Nombre d'individus et de générations dans la première et la seconde version des pédigrées de la REFPOP

Famille	n_{ped1}	n_{ped2}	Gen_{ped1}	Gen_{ped2}
12_B	14	27	7	9
12_E	36	29	10	8
12_F	25	38	7	9
12_I	35	30	10	8
A1	47	67	8	11
A2	40	50	7	10
B3F	23	15	7	7
DLO-12	33	37	10	11
Novadi1	32	22	10	8
Novadi2	36	23	10	8
12_J	25	38	7	9
12_K	22	14	7	5
12_L	22	27	7	9
12_N	25	20	7	6
12_O	35	27	10	9
12_P	15	28	7	7
FuGa	11	26	5	8
FuPi	12	31	4	8
GaCr	13	26	5	8
GaPi	12	27	5	8
I_BB	24	39	7	9
I_CC	22	14	7	5
I_J	35	37	10	9
I_M	40	34	10	8
I_W	40	27	10	8
JoPr	20	11	8	5
TeBr	10	19	4	7

n_{ped1} : nombre d'individus dans la première version du pédigrée d'une famille donnée
n_{ped2} : nombre d'individus dans la deuxième version du pédigrée d'une famille donnée
Gen_{ped1} : nombre de générations dans la première version du pédigrée d'une famille donnée
Gen_{ped2} : nombre de générations dans la deuxième version du pédigrée d'une famille donnée

S2.5 Protocole de phénotypage de la REFPOP

Nous présentons ici le protocole retenu pour noter la date de récolte, la coloration du fruit, le poids de 20 fruits et le nombre de fruits par arbre dans la population dite REFPOP et par la suite utilisé pour phénotyper les hybrides RG x E.

Date de récolte

Date à laquelle 50% des fruits ont atteint leur maturité physiologique, déterminée par coloration au Lugol ou par l'expertise du notateur.

Couleur du fruit

Pourcentage moyen de la surface de couleur rouge des fruits d'un arbre, noté de 0 à 5 selon l'échelle présentée ci-dessous :

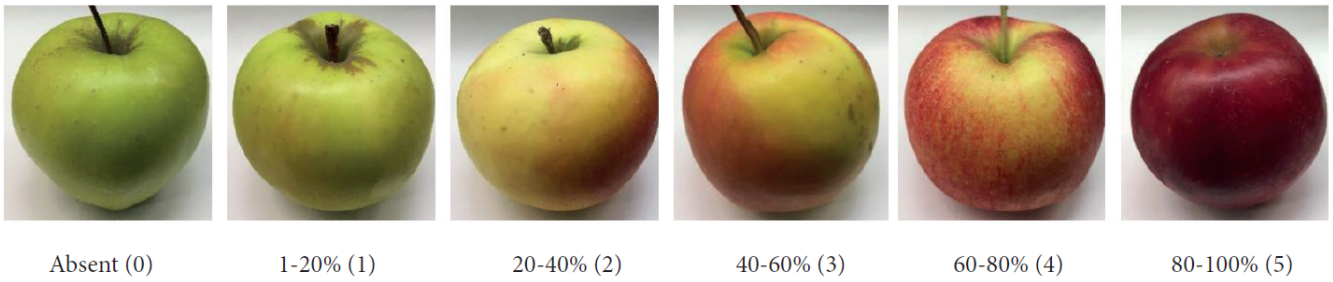


FIGURE S2.3 – Echelle de notation de la couleur du fruit (d'après Jung et al. (2022))

Nombre de fruits

Nombre total de fruits par arbre au moment de la récolte. Seuls les fruits parfaitement développés sont comptabilisés

Poids des fruits

Poids de 20 fruits choisis aléatoirement au moment de la récolte et pesés à l'aide d'une balance

CHAPITRE 3

Analyse de la qualité de l'imputation par simulations de familles biparentales

3.1 Introduction

Chez le pommier, le déséquilibre de liaison décroît rapidement (par exemple, Leforestier a montré dans sa thèse que le r^2 moyen décroît en-dessous de 0,2 au-delà de 10kb dans une core-collection INRAE), ce qui peut se révéler problématique pour appliquer des modèles de prédiction génomique avec des données obtenues à basse ou moyenne densité. Il existe chez le pommier une puce à haute densité générant environ 300 000 marqueurs SNP robustes, qui semble appropriée pour l'application de la sélection génomique (Bianco et al. 2016). Malgré la baisse continue des coûts de génotypage, génotyper à haute densité un grand nombre de candidats à la sélection peut cependant s'avérer délicat d'un point de vue économique chez une espèce comme le pommier. Une stratégie alternative consiste à génotyper certains ancêtres des candidats à haute densité et les candidats à moyenne ou basse densité, puis à imputer les données génomiques de la moyenne vers la haute densité. De nombreuses études ont montré l'intérêt de l'imputation combinée à l'utilisation de la sélection génomique, dans le but d'améliorer la précision des modèles de prédiction génomique (Habier et al. 2009 ; Nyine et al. 2019), d'augmenter l'intensité de sélection (Gorjanc et al. 2017) ou de réduire les coûts de génotypage (Herry et al. 2018 ; Tsairidou et al. 2020). Avant le début de la thèse, environ 1300 variétés anciennes de pomme représentatives de la diversité génétique européenne avaient été génotypées à haute densité dans le cadre d'un projet de génétique d'association (voir section 2.2 ainsi que Bianco et al. (2016) et Urrestarazu et al. (2017)). Ces données ont aussi servi à la reconstruc-

tion de pédigrées (voir section 2.4 et Muranty et al. (2020)), et un grand nombre des ancêtres connus des variétés actuelles possédaient donc des données à haute densité, rendant possible la mise en place de la stratégie d'imputation évoquée plus haut. La connaissance de plus en plus précise des pédigrées (Howard et al. 2021) et la bonne qualité des données de séquençage génomique (Daccord et al. 2017 ; Zhang et al. 2019) devraient permettre d'obtenir une bonne précision d'imputation chez le pommier. Cependant, au début de ma thèse, les deux seules études s'étant intéressées à la précision d'imputation chez le pommier étaient basées sur des données GBS (Money et al. 2015 ; Zheng et al. 2018), pour lesquelles les données manquantes sont réparties aléatoirement et aucune étude n'avait étudié la précision d'imputation qu'il était possible d'atteindre en imputant des données moyenne densité vers de la haute densité. Dans ce chapitre, nous nous appuyons sur des données réelles et des simulations pour évaluer la précision d'imputation qu'il est envisageable d'obtenir lorsque les individus à imputer représentent du matériel élite ou des hybrides entre matériel élite et ressources génétiques. Nous examinons également la répartition des erreurs d'imputation et discutons de stratégies permettant d'améliorer la précision d'imputation chez le pommier.

3.2 Matériel et méthodes

3.2.1 Processus de l'imputation

3.2.1.1 Choix du logiciel d'imputation

Les différents travaux d'imputation présentés dans ce chapitre ont été menés à l'aide du logiciel Beagle (Browning et Browning 2007), qui est un outil largement utilisé en génétique humaine, animale et végétale notamment car il permet d'obtenir des précisions d'imputation élevées de manière relativement rapide (Pook et al. 2020 ; Swarts et al. 2014 ; Whalen et al. 2018a). De plus, Beagle peut utiliser un panel de référence constitué d'individus phasés et génotypés à haute densité afin d'imputer des données basse ou moyenne densité. La version 4.0 de Beagle permet en outre à l'utilisateur de renseigner des informations de pédigrée lors du phasage du panel de référence, afin d'indiquer les apparentements entre individus du panel et ainsi améliorer la qualité du phasage, et par conséquent l'imputation des candidats par la suite. Dans la suite du manuscrit, toute mention à l'outil Beagle fait donc référence à Beagle v4.0, sauf si précisé autrement.

3.2.1.2 Jeux de données utilisés

Données réelles

Note : ce travail préliminaire sur données réelles a été réalisé par Michaela Jung, alors docto-

rante à Agroscope (Suisse), lors d'un séjour à Angers fin 2018. L'approche qu'elle a employée et ses résultats sont succinctement décrits ici car ils ont constitué une base de travail pour les analyses menées par simulations. Michaela a dans un premier temps mesuré la précision d'imputation obtenue avec Beagle en utilisant les données réelles des familles FuPi et GDR. Pour ce faire, elle a créé un jeu de données constitué des deux familles, pour lesquelles les données des marqueurs qui n'apparaissent pas sur la puce 20K ont été masquées. Les données à moyenne densité ainsi obtenues ont été imputées à l'aide de Beagle de cinq façons : (1) en utilisant Beagle 5.0 et de ce fait sans information de pédigrée, ou en utilisant Beagle 4.0 et en incluant dans le pédigrée (2) les parents des deux familles, (3) les ancêtres de deux familles, (4) tous les individus apparentés aux deux familles connus ou (5) tous les individus génotypés à haute densité (Jung et al. 2020). Dans chacun des cas, la précision d'imputation a été mesurée comme la corrélation entre les vraies données génotypiques et les données génotypiques imputées.

Données simulées

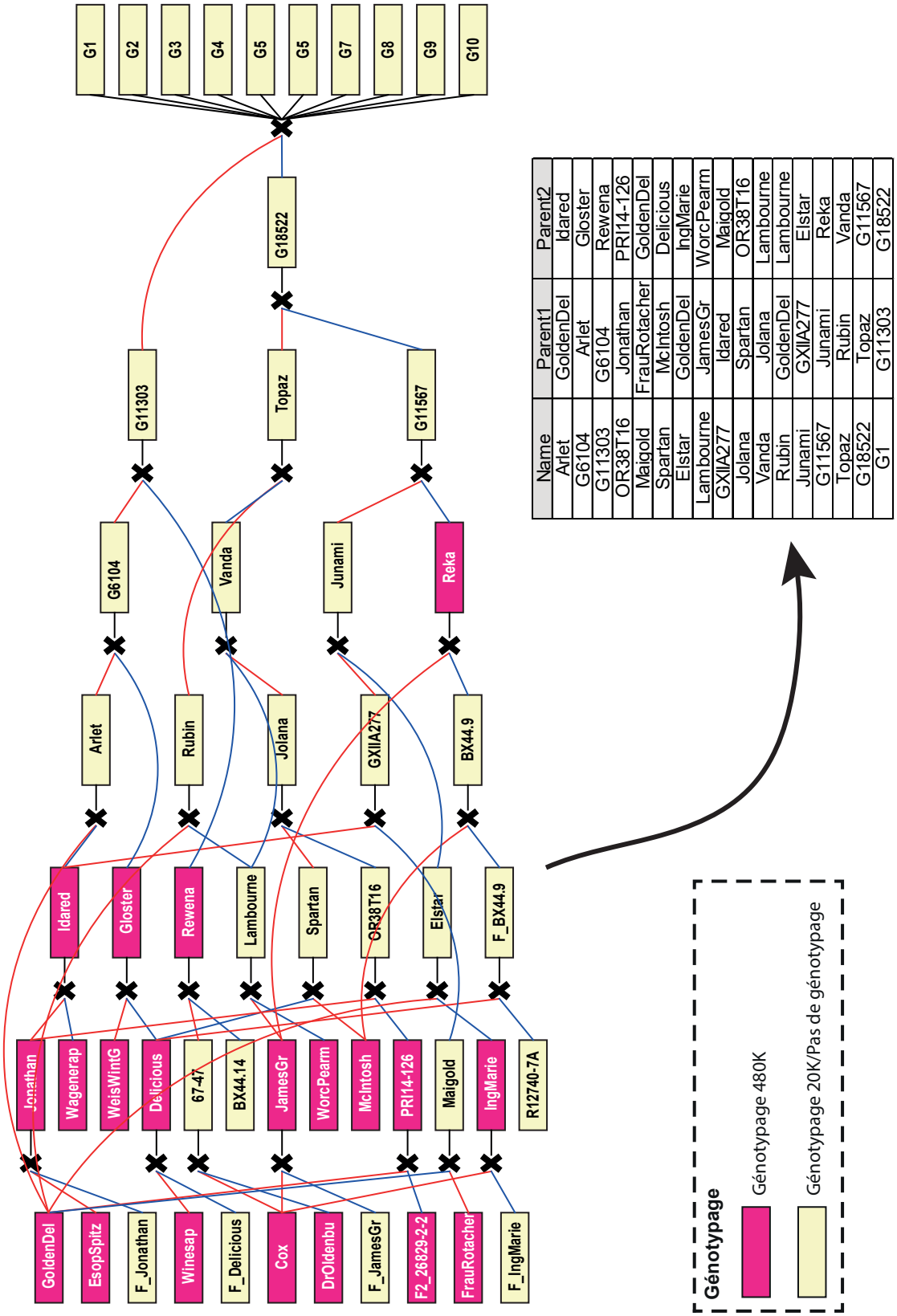
Afin de généraliser les résultats obtenus sur deux familles à partir de données réelles, nous avons simulé des données génotypiques à haute densité pour les 27 familles de la REFPOP, avec pour but d'appliquer l'approche mise en œuvre sur les données réelles aux familles simulées. Nous avons simulé 20 plein-frères par famille en nous servant des pédigrées connus de ces familles et des données génotypiques haute densité au sein de chaque famille. Plus précisément, la démarche mise en œuvre lors des simulations peut se décliner comme suit.

3.2.2 Utilisation des pédigrées et chemin de simulation par famille

Pour une famille donnée, nous identifions au sein du pédigrée quels individus sont génotypés et à quelle densité (haute-densité, moyenne densité ou pas de génotypage). Lorsque les deux parents (notés P_1 et P_2) des plein-frères terminaux du pédigrée sont génotypés à haute densité, nous simulons 20 plein-frères résultant d'un croisement entre P_1 et P_2. Imaginons maintenant que P_1 ait été génotypé à haute densité, mais que P_2 ait été génotypé à moyenne densité. Dans ce cas, si les parents de P_2 (notés GP_{21} et GP_{22}) sont génotypés à haute densité, nous simulons un marquage à haute densité pour P_2, résultant d'un croisement entre GP_{21} et GP_{22}. A partir des données génotypiques haute densité simulées pour P_2, il devient alors possible de simuler un croisement entre P_1 et P_2.

3.2.2.1 Simulation d'un croisement donné

Les données haute densité utilisées lors des simulations correspondent à des données phasées obtenues lors du phasage du panel de référence à l'aide de Beagle. Lors de la simulation d'un croisement, nous simulons donc un gamète pour chaque parent et l'individu simulé porte donc



Name	Parent1	Parent2
Arlet	GoldenDel	Idared
G6104	Arlet	Gloster
G11303	G6104	Rewena
OR38T16	Jonathan	PRI14-126
Maigold	FrauRotacher	GoldenDel
Spartan	McIntosh	Delicious
Elstar	GoldenDel	IngMarie
Lambourne	JamesGr	WorcPearm
GXIIA277	Idared	Maigold
Jolana	Spartan	OR38T16
Vanda	Jolana	Lambourne
Rubin	GoldenDel	Lambourne
Junami	GXIIA277	Elstar
G11567	Junami	Reka
Topaz	Rubin	Vanda
G18522	Topaz	G11567
G1	G11303	G18522

FIGURE 3.1 – Pédigrée de la famille A1 de la REFPOP. La connaissance du pédigrée et des individus génotypés à haute densité (colorés en rouge) permet d'obtenir le "chemin de simulation" nécessaire pour simuler les 20 plein-frères

un haplotype de chaque parent pour chacun des 17 chromosomes. Le nombre de crossing-over lors de la simulation des gamètes suit une loi de Poisson dont l'espérance dépend de la longueur génétique du chromosome. La distance génétique utilisée lors des simulations est basée sur la distance physique, en prenant l'approximation qu'un centiMorgan équivaut à 500 kb chez le pommier. La position physique des points de recombinaison est tirée aléatoirement, sans simuler d'interférence. Nous avons appliqué cette approche aux 27 familles, en écrivant une fonction R permettant de déterminer les croisements à effectuer au sein de chaque famille afin de pouvoir simuler les plein-frères. La figure 3.1 présente un exemple de pédigrée associé à une famille de la REFPOP, ainsi que le chemin de simulation découlant de ce pédigrée et permettant de simuler les plein-frères de cette famille. Le chemin de simulation est d'autant plus complexe qu'il faut remonter dans le pédigrée pour pouvoir simuler des données à haute densité. Pour une des familles de la REFPOP (famille A2, voir sous-section 3.2.2.2), il n'existe d'ailleurs aucun chemin permettant de simuler les plein-frères en suivant la méthode décrite ci-dessus. Le tableau 3.2 indique quels ascendants des plein-frères terminaux ont été génotypés à haute densité au sein de chaque famille et le tableau 3.3 indique les caractéristiques des chromosomes simulés.

3.2.2.2 Cas de la famille A2

Le pédigrée de la famille A2 est présenté figure 3.2. Comme indiqué précédemment, la simulation de 20 plein-frères n'est pas possible pour cette famille car les individus 163-42 et 368-25 n'ont aucun ascendant génotypé à haute densité. Nous avons profité de cette situation particulière pour tester si la présence de génotypes peu apparentés au matériel élite dans le pédigrée avait un impact sur la qualité d'imputation (comme suggéré dans le paragraphe 3.2.5). Pour ce faire, nous avons dans un premier temps défini deux groupes de variétés génotypées à haute densité, constitués soit de matériel exotique (groupe « exo »), soit de variétés modernes couramment utilisées dans les programmes d'amélioration (groupe « commun »). Ces deux groupes ont été constitués sur la base de travaux antérieurs à ma thèse : une analyse de partage d'haplotypes avait permis d'identifier les haplotypes partagés de chaque variété génotypée à haute densité avec un groupe de 42 cultivars présents dans le pédigrée de nombreux génotypes et de compter avec combien de ces cultivars chaque variété partageait au moins un haplotype de plus de 8 Mb. A partir de cette information, nous avons défini le groupe « exotique » comme contenant l'ensemble des génotypes ne partageant aucun haplotype avec au moins 17 des 42 variétés (ce qui est le cas pour 55 génotypes) et le groupe « commun » comme contenant l'ensemble des génotypes partageant au moins un haplotype avec au moins 35 des 42 variétés (ce qui correspond à 462 génotypes). Nous avons ensuite attribué des données génomiques aux individus 163-42 et 368-25 selon trois scenarios :

— Les données génomiques des deux individus proviennent du groupe « exotique » : scenario

exo_exo

— Les données génomiques des deux individus proviennent du groupe « commun » : scenario commun_commun

— Les données génomiques de 163-42 proviennent d'un des deux groupes et celles de 368-25 proviennent du groupe opposé : scenario exo_commun

Nous avons réalisé 10 simulations par scénario en faisant varier les données génomiques attribuées aux individus 163-42 et 368-25 d'une simulation à l'autre et en tirant ces données aléatoirement et sans remise parmi les données génomiques disponibles pour les groupes « exo » ou « commun » selon le scénario. Nous avons d'autre part simulé les données génomiques à haute densité du parent G11567 une seule fois et avons attribué ces données à G11567 dans chaque simulation, de telle sorte que les variations de précision d'imputation observées d'une simulation à l'autre ne sont dues qu'aux données génomiques utilisées pour les individus 163-42 et 368-25.

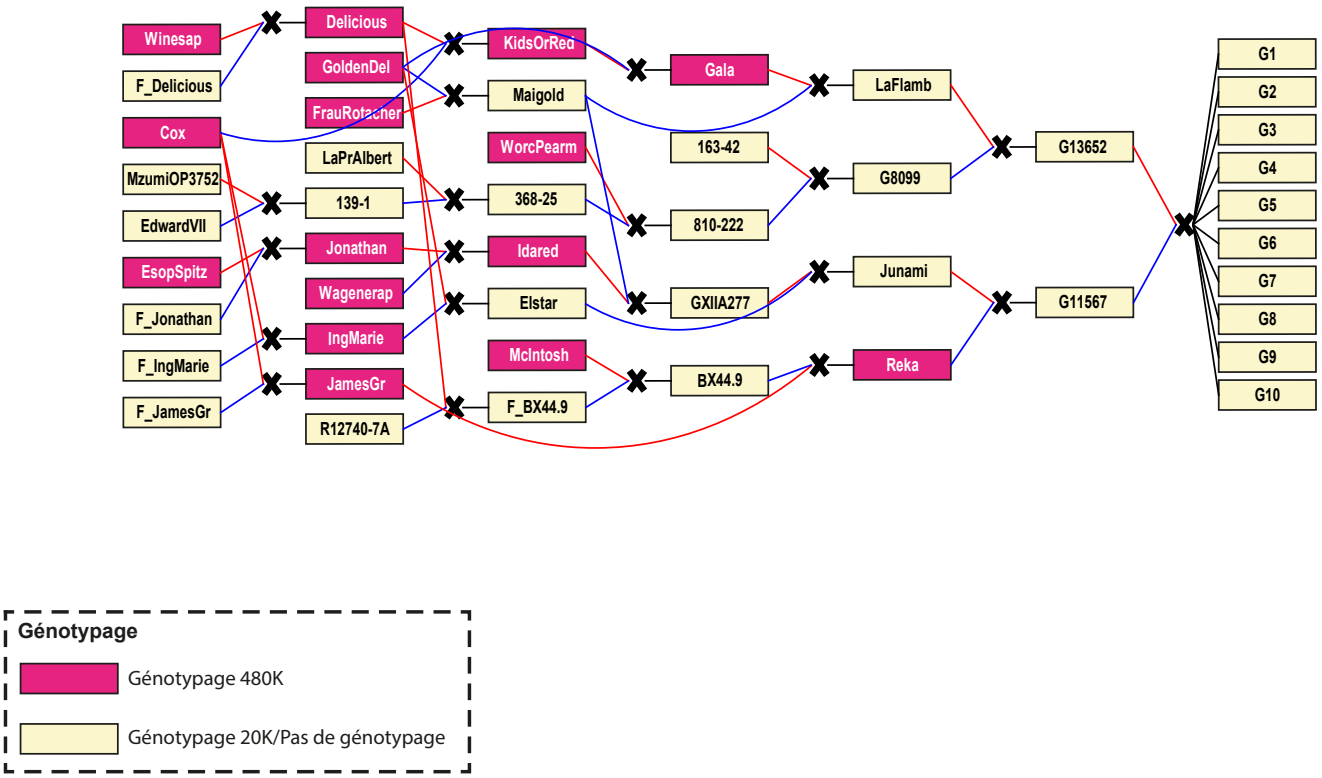


FIGURE 3.2 – Pédigrée de la famille A2 de la REFPOP. Dans cette situation, il n'est pas possible de simuler les 20 plein-frères car les individus 368-25 et 163-42 n'ont pas d'ancêtre génotypés à haute densité

3.2.3 Imputation des données simulées

Comme dans le cas des données réelles, nous avons créé un jeu de données constitué des familles simulées pour lesquelles nous avons masqué les marqueurs qui se trouvaient uniquement à haute densité afin d'obtenir artificiellement des données moyenne densité pour les 26 familles (famille A2 exclue). Nous avons ensuite imputé les données génomiques moyenne densité vers des données génomiques haute densité. Suite aux résultats obtenus par M. Jung sur données réelles (voir ci-après), nous avons uniquement comparé le cas de l'imputation avec Beagle 4.0 en utilisant un panel de référence constitué de tous les individus génotypés à haute densité et phasé en utilisant ou non les informations de pédigrée.

3.2.4 Mesures de la qualité d'imputation des données simulées

Nous avons dans un premier temps mesuré la qualité de l'imputation comme la corrélation entre les génotypes imputés et les génotypes réels, tel que préconisé par Calus et al. (2014). Nous appellerons désormais cette mesure « précision d'imputation » par analogie avec la précision de prédiction rencontrée dans la littérature concernant la sélection génomique. Contrairement au taux de concordance (défini comme le pourcentage de génotypes correctement imputés), cette approche a l'avantage d'être indépendante des fréquences alléliques, qui peuvent entrainer un taux de concordance d'autant plus élevé que la fréquence de l'allèle minoritaire (Minor Allele Frequency ou MAF) est faible (Hickey et al. 2012b). Lorsqu'un marqueur est monomorphe (avant ou après imputation), une mesure basée sur la corrélation est cependant impossible. Nous avons donc mesuré la qualité de l'imputation en utilisant deux autres critères :

— L'AR^2 (Allelic R^2) défini par Browning et Browning (2009) : il s'agit du carré de la corrélation entre les doses alléliques imputées et les vraies doses alléliques. Dans le cas où les vraies doses alléliques ne sont pas connues, les probabilités postérieures peuvent être utilisées à la place. C'est l'approche mise en place dans Beagle, qui renvoie une valeur d'AR^2 pour chaque individu et chaque marqueur. Cet indicateur a l'avantage de pouvoir être exploité pour estimer la qualité d'imputation, même lorsque les vrais génotypes ne sont pas connus.

— L'IQS (Imputation Quality Score) défini par Lin et al. (2010) : il s'agit d'un score qui correspond à une mesure de qualité d'imputation qui prend en compte les marqueurs correctement imputés du fait du hasard. La valeur IQS maximum est de 1 et il n'existe pas de minimum théorique. Une valeur de 0 indique que l'imputation aléatoire basée sur la seule fréquence allélique donnerait autant de génotypes correctement imputés que ce qui est observé. Pour chaque marqueur, nous pouvons représenter le nombre de génotypes imputés en fonction des génotypes réellement observés en utilisant le tableau 3.2.4.

L'IQS du marqueur est alors calculé de la façon suivante :

TABLEAU 3.1 – Tableau de classes génotypiques des données réelles et imputées permettant d'illustrer le calcul de l'IQS

	Vrais génotypes			
Imputés	0	1	2	Total
0	n_{11}	n_{12}	n_{13}	$n_{1.}$
1	n_{21}	n_{22}	n_{23}	$n_{2.}$
2	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

$$IQS = \frac{P_0 - P_c}{1 - P_c}$$

avec $P_0 = \frac{\sum n_{ii}}{n}$ et $P_c = \frac{\sum n_{i.} n_{.i}}{n^2}$

Afin de prendre en compte l'incertitude liée à l'imputation, il est aussi possible de remplacer le nombre de génotypes correctement imputés par la somme des probabilités postérieures pour chaque génotype. Afin de distinguer les deux cas, nous appellerons les valeurs d'IQS calculées de cette manière IQS_{imp} et les valeurs d'IQS basées sur la comparaison entre données imputées et réelle IQS_{true}.

Nous avons évalué la précision d'imputation par marqueur puis par chromosome, et par individu puis par famille. Nous avons également évalué la qualité d'imputation par marqueur en utilisant les autres indicateurs et examiné leur relation avec la fréquence allélique. Pour cela, nous avons étudié la distribution de la qualité d'imputation en répartissant les marqueurs dans les classes de MAF suivantes : [0 ; 0,01],]0,01 ; 0,05],]0,05 ; 0,15],]0,15 ; 0,25],]0,25 ; 0,35],]0,35 ; 0,45] et MAF $>$ 0,45. Nous nous sommes ensuite intéressés plus spécifiquement aux allèles rares définis comme ceux dont la fréquence de l'allèle minoritaire est inférieure à 0,05 et avons mesuré l'IQS de ces allèles par famille et par chromosome.

3.2.5 Analyse exploratoire de la précision d'imputation pour deux familles

Les précisions d'imputation mesurées par simulation ayant révélé des différences entre chromosomes et familles, nous nous sommes intéressés aux familles 12_B et 12_E, pour lesquelles les chromosomes 11 et 17 sont moins bien imputés que chez les autres familles. Nous avons alors cherché à comprendre les raisons de telles différences. Les plein-frères des deux familles possèdent un parent en commun, Generos, dont les ancêtres n'étaient pas connus au moment de l'étude mais pour lequel des données d'archive soulignent la présence de pommier sauvage apparenté (*M. kaido*) dans le pédigrée (Crosby et al. 1992), qui pourrait être à l'origine d'haplotypes rares dans ces deux familles. Des résultats de l'équipe ont en outre confirmé la présence d'un

haplotype partagé de 27.5 cM entre Generos et *M. kaido* sur le chromosome 17, en utilisant des données de génotypage moyenne densité et la méthode développée par Howard et al. (Howard et al. 2021). Pour tester l'influence de Generos dans les familles 12_B et 12_E, nous avons de nouveau simulé 20 plein-frères pour chacune des deux familles mais en remplaçant les données génotypiques de Generos par celles de Jonathan, une variété largement utilisée dans les programmes d'amélioration et qui est apparentée à de nombreux individus du panel de référence. Nous avons ensuite cherché à savoir si les erreurs d'imputation étaient observées pour les mêmes marqueurs entre les individus de chacune des deux familles. Nous avons pour cela représenté la moyenne de la précision d'imputation à l'aide de fenêtres glissantes de 1 Mb avec un pas de 100 kb le long des chromosomes 11 et 17. Nous nous sommes enfin intéressés à la précision d'imputation des allèles rares spécifiquement dans ces deux familles, en faisant l'hypothèse que des allèles rares provenant de Generos pourraient ségréger uniquement dans ces deux familles tout en étant également rares dans le panel de référence, expliquant ainsi la mauvaise qualité d'imputation pour les chromosomes 11 et 17. Nous avons dans un premier temps identifié les allèles rares définis sur la base des fréquences alléliques des 26 familles simulées et calculé l'IQS de ces allèles lorsqu'ils ségrégeaient dans la famille 12_B ou la famille 12_E. L'IQS a été calculé au sein de fenêtres de 100kb le long des 17 chromosomes pour les deux familles. Nous avons réalisé la même mesure mais en nous limitant aux allèles rares ségrégeant dans la famille 12_B ou la famille 12_E qui étaient également rares dans le panel de référence.

3.2.6 Qualité d'imputation pour les jeux de données d'application

L'objectif des travaux menés dans ce chapitre était avant tout d'estimer la qualité de l'imputation qu'il était possible d'obtenir chez le pommier afin de montrer que l'imputation des jeux de données utilisés par la suite dans la thèse pouvait être jugée de bonne qualité. Nous rapportons donc ici à titre informatif les valeurs d'AR^2 et de probabilités postérieures obtenues lorsque nous avons imputé le matériel élite et les hybrides RG x E (matériel utilisé pour les prédictions génomiques, voir section 2.1) en utilisant la méthodologie proposée dans ce chapitre et les comparons aux valeurs obtenues sur les simulations.

3.3 Résultats

3.3.1 Précision d'imputation pour les 26 familles simulées

Sur données réelles

Comme illustré sur la figure 3.3, la meilleure précision d'imputation moyenne pour les deux familles FuPi et GDR (corrélation de 0,98) est obtenue lorsque tous les individus génotypés à haute densité sont inclus dans le panel de référence et que l'imputation utilise un panel de référence phasé à l'aide des informations de pédigrée. En comparaison, la précision d'imputation est de 0,9 lorsque le même panel de référence est utilisé mais que les informations de pédigrée n'ont pas été utilisées pour le phaser. La précision d'imputation avec pédigrée sur un panel de référence réduit est également plus élevée que lorsque tous les individus sont utilisés sans pédigrée (précision d'imputation moyenne de 0,91 en incluant seulement les ancêtres des deux familles dans le panel de référence et de 0,95 en incluant tous les individus apparentés aux plein-frères des deux familles), sauf lorsque le panel de référence ne contient que les parents des deux familles (précision d'imputation moyenne de 0,75), probablement du fait de la taille trop limitée du panel dans ce cas. Au vu de ces résultats préliminaires obtenus sur données réelles par M. Jung, nous avons donc choisi pour la suite des travaux d'utiliser un panel de référence contenant tous les individus génotypés à haute densité et ayant été phasés à l'aide des informations de pédigrée pour réaliser les imputations.

Sur données simulées

La précision d'imputation moyenne par chromosome est globalement élevée, allant de 0,92 (chromosomes 7, 9 et 16) à 0,97 (chromosome 12, voir tableau 3.6). Il existe cependant une variabilité de précision d'imputation en fonction des individus et des familles pour chaque chromosome (tableau 3.4 et figure 3.4). Ainsi, quelques individus sont mal imputés sur certains chromosomes (en particulier les individus des familles 12_B et 12_E sur les chromosome 11 et 17 ainsi que les individus de la famille Novadi1 sur le chromosome 16, voir figure 3.4), ce qui se traduit par des valeurs minimales plus faibles et un écart-type de précision d'imputation plus important pour ces familles (tableau 3.4). La précision d'imputation moyenne par individu est malgré tout de 0,95 et est supérieure à 0,9 pour les 520 individus simulés. La qualité d'imputation par marqueur est également bonne quel que soit l'indicateur de mesure de qualité utilisé et surtout dans le cas où les vrais génotypes sont utilisés à la place de probabilités postérieures (corrélation moyenne de 0,91 sur l'ensemble du génome, IQS_{true} moyen de 0,87, IQS_{imp} moyen de 0,79, AR^2 moyen de 0,79). La précision d'imputation moyenne par marqueur est de 0,91 et est supérieure à 0,7 pour 93% des marqueurs (voir tableau 3.5).
Dans le cas de la famille A2, la précision d'imputation au niveau du génome ne diffère quasiment pas entre scénarios (précision d'imputation de 0,92 pour le scénario « commun_commun » et

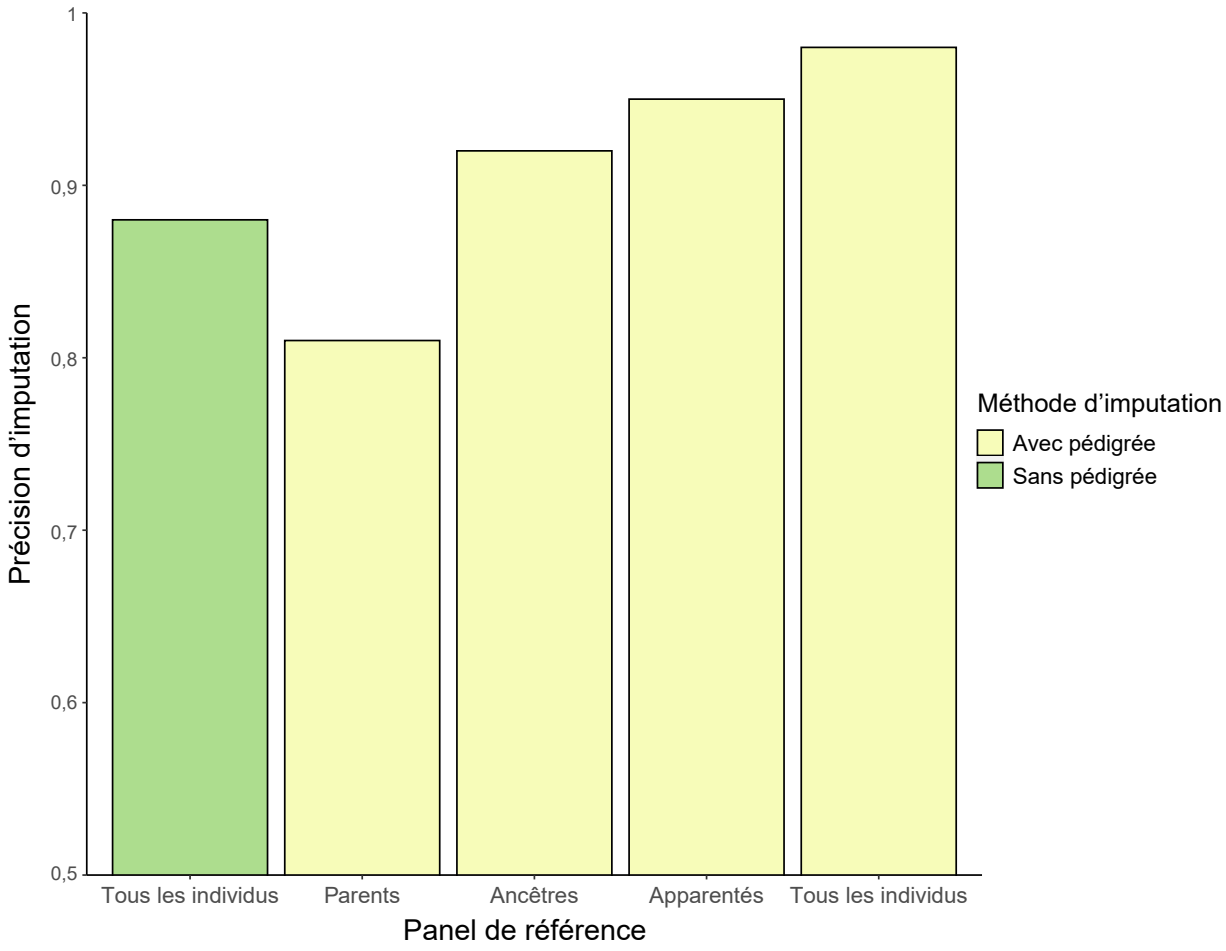


FIGURE 3.3 – Influence des individus présents dans le panel de référence sur la précision d'imputation. La comparaison entre données réelles et imputées concerne deux familles de 46 plein-frères génotypés à haute densité (d'après les travaux de M. Jung).

0,91 pour les autres scénarios), avec une variabilité très faible entre les dix simulations d'un scénario donné (tableau 3.7).

3.3.2 Comparaison des indicateurs de la qualité d'imputation utilisés

Comme illustré sur la figure S3.5, les valeurs obtenues en utilisant les différents indicateurs de qualité d'imputation sont fortement corrélées, sauf pour le taux de concordance, qui est dépendant des fréquences alléliques (contrairement aux autres indicateurs). Nous pouvons observer sur la figure S3.6 que les différences entre les quatre autres méthodes concernent majoritairement les marqueurs dont la MAF est faible, l'IQS étant l'indicateur pénalisant le plus les erreurs pour les allèles rares (Rowan et al. 2019). L'AR^2 étant le seul indicateur n'utilisant pas les vrais génotypes, il s'agit d'un indicateur d'intérêt pour évaluer la qualité d'imputation sur données réelles. Les valeurs d'AR^2 sont d'autant plus fiables que les valeurs de probabilités postérieures

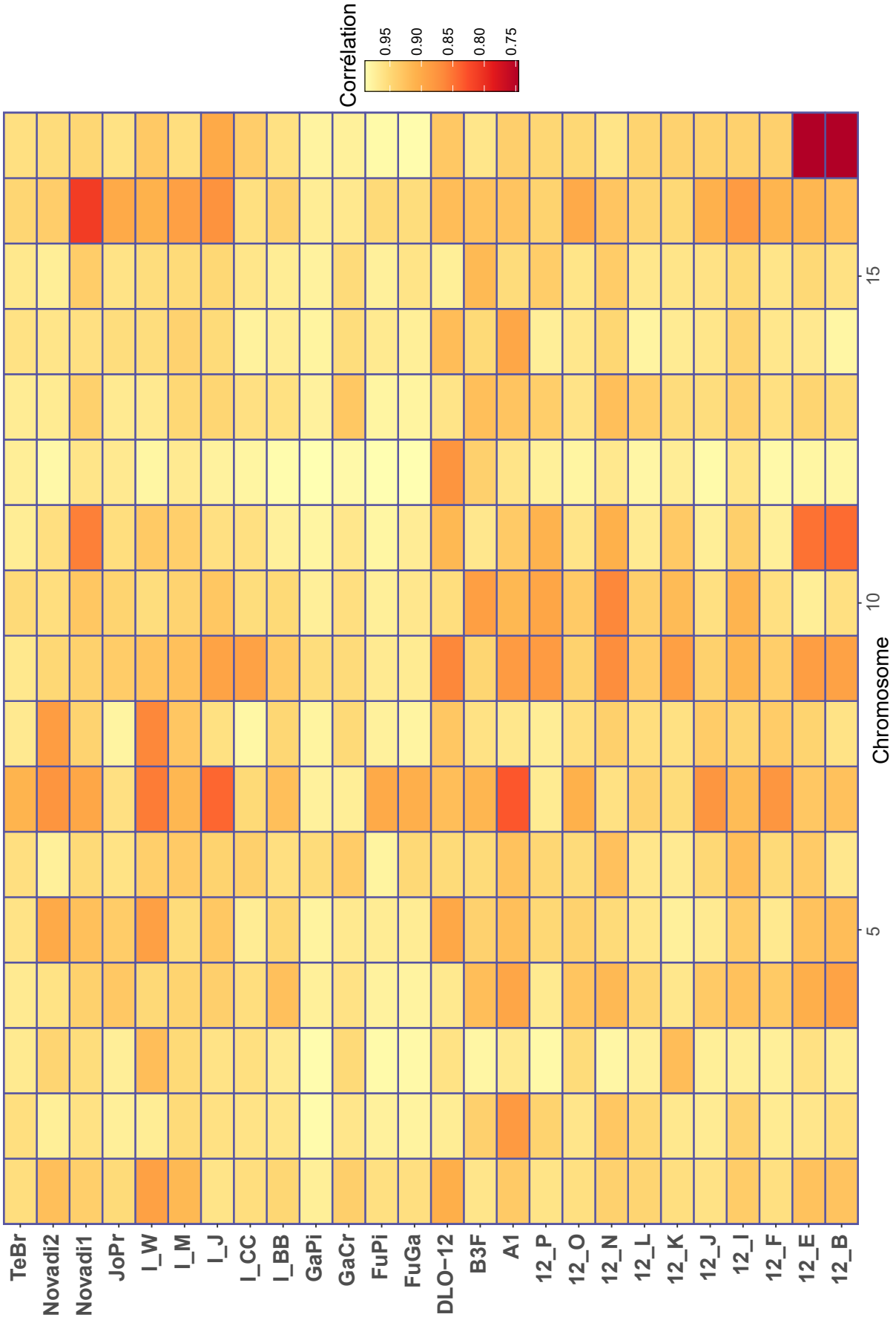


FIGURE 3.4 – Précision d'imputation par chromosome et par famille lorsque le panel de référence a été phasé en utilisant les informations de pédigrée

sont correctement évaluées (Browning et Browning 2007) mais peuvent être erronées lorsque les probabilités postérieures traduisent mal le lien entre génotype réel et imputé (par exemple dans le cas de probabilités postérieures élevées pour des imputations erronées ou dans le cas opposé de probabilités postérieures faibles pour des imputations correctes). Dans notre cas, les valeurs de probabilité postérieure sont supérieures à 0,9 dans 87% des cas, et dans ce cas, les génotypes imputés et réels diffèrent dans seulement 0,6% des cas.

3.3.3 Précision d'imputation des familles 12_B et 12_E

La précision d'imputation des familles 12_B et 12_E augmente pour presque tous les chromosomes lorsque Generos est remplacé par Jonathan dans les simulations, et la variabilité de la précision d'imputation par chromosome est réduite, surtout pour les chromosomes qui étaient initialement mal imputés (figure S3.7). Le gain de précision d'imputation est le plus marqué pour les chromosomes 11 (passage respectivement de 0,85 à 0,98 et de 0,85 à 0,94 dans les famille 12_B et 12_E) et 17 (passage respectivement de 0,75 à 0,94 et de 0,75 à 0,97 dans les famille 12_B et 12_E). La précision d'imputation mesurée à l'aide de fenêtres glissantes sur le chromosome 17 permet d'identifier deux zones qui sont mal imputées chez tous les plein-frères dans les deux familles (voir par exemple figure S3.8 pour la famille 12_B et figures S3.9 et S3.10), situées autour des fenêtres 100 et 220 (soit entre 10 et 23 Mb). D'autres fragments de chromosomes sont mal imputés sur le chromosome 17 mais diffèrent d'un individu à l'autre. De même, différents fragments chromosomiques apparaissent mal imputés sur le chromosome 11 mais leur localisation diffère généralement d'un plein-frère à l'autre et il ne semble pas y avoir de portion chromosomique qui soit mal imputée chez tous les plein-frères. La majorité des allèles rares est bien imputée dans les deux familles (89% des allèles rares ont un IQS supérieur à 0,9) mais les allèles rares également rares dans le panel de référence sont très mal imputés sur les chromosomes 11, 17 et dans une moindre mesure 5 (figures S3.11 et S3.12).

3.3.4 Application de l'imputation à du matériel élite et des hybrides RG x E

La qualité de l'imputation sur les données réelles ne peut être évaluée qu'au travers des probabilités postérieures et des valeurs d'AR^2, en supposant que l'AR^2 soit fortement corrélé à la précision d'imputation comme cela a pu être observé sur les données simulées. La distribution des valeurs d'AR^2 est globalement très proche entre les familles simulées et les jeux de données d'application (figure S3.13). Nous observons davantage de marqueurs ayant une valeur d'AR^2 élevée chez le matériel élite et les hybrides que chez les familles simulées (79% des marqueurs

du matériel élite et des hybrides ont une valeur d'AR^2 supérieure à 0,7, contre 66% des marqueurs dans le cas des familles simulées). Ces valeurs d'AR^2 élevées sont liées à des valeurs de probabilité postérieure maximales également élevées : chez le matériel élite et les hybrides, 89% des génotypes ont une probabilité postérieure maximale supérieure à 0,9 (87% pour les données simulées).

3.4 Discussion

3.4.1 Des résultats prometteurs quant à la qualité d'imputation

Les résultats présentés dans ce chapitre montrent que l'imputation basée sur l'utilisation d'un panel de référence contenant certains ancêtres des individus à imputer est une stratégie pertinente chez le pommier. L'utilisation préalable des informations de pédigrée pour phaser correctement le panel de référence est également un élément important de la stratégie d'imputation mise en place dans ce chapitre. Cette stratégie a permis d'obtenir des précisions d'imputation élevées tant sur données simulées que réelles (précision d'imputation moyenne de 0,98 sur données réelles et précision d'imputation supérieure à 0,9 pour chaque chromosome dans le cas des données simulées). Dans le cas des simulations, il convient tout de même de garder à l'esprit que nous avons utilisé les données phasées du panel de référence pour simuler les 26 familles de plein-frères. De ce fait, les haplotypes des 26 familles sont forcément présents dans le panel de référence, ce qui peut conduire à une surestimation de la précision d'imputation par rapport à une situation réelle pour laquelle une partie des haplotypes du panel de référence serait inférée de manière incorrecte. Dans le cas de l'imputation du matériel élite et des hybrides ressources génétiques x matériel élite, les vrais génotypes à haute densité ne sont pas disponibles et la qualité d'imputation ne peut donc pas être mesurée avec des indicateurs tels que la corrélation entre génotypes imputés et réels ou l'IQS. Les seules informations permettant de présumer de la qualité d'imputation sont alors les valeurs des probabilités postérieures et la distribution de l'AR^2, qui découle directement de ces probabilités et qui est apparu fortement corrélé aux autres indicateurs basés sur les données simulées. Dans notre cas, les valeurs élevées d'AR^2 par marqueur semblent indiquer une bonne qualité d'imputation des deux jeux de données d'application (figure S3.13). Les valeurs d'AR^2 sont même légèrement plus élevées sur les données imputées du matériel élite que sur les données simulées, pour lesquelles nous venons de souligner une probable surestimation du fait de la conception des simulations. Une première explication réside possiblement dans le nombre d'individus à imputer : dans le cas des simulations, les familles 12_B et 12_E contiennent environ 8% (2 familles sur 26) des individus à imputer, alors que les individus de ces familles représentent 1,5% des individus élite à imputer pour les données du matériel élite (deux familles de 10 plein-frères sur 1018 individus élite à imputer, voir tableau 2.1). En supposant que les valeurs d'AR^2 soient majoritairement influencées par ces deux familles dans le cas des données simulées, les valeurs d'AR^2 peuvent être plus élevées dans le matériel élite puisque ces deux familles y sont représentées en fréquence plus faible. Il faut cependant garder à l'esprit que l'AR^2 est une mesure pouvant surestimer la qualité d'imputation pour des marqueurs présentant des allèles rares dans le panel de référence et chez les individus à imputer (figure S3.6). La présence d'allèles rares ne ségrégeant que dans

certaines familles est un cas de figure qui peut typiquement se présenter dans un programme d'amélioration cherchant à intégrer des allèles rares dans les variétés élite, auquel cas il peut être nécessaire de génotyper certains candidats à haute densité pour s'assurer de la qualité du phasage puis de l'imputation et pour suivre le devenir de ces allèles rares potentiellement mal imputés.

3.4.2 Qualité d'imputation en présence d'haplotypes rares

Malgré la bonne qualité générale de l'imputation, nous avons obtenu des précisions d'imputation basses pour les familles 12_B et 12_E sur les chromosomes 11 et 17, probablement du fait de la présence d'haplotypes comprenant des variants rares hérités de leur parent commun Generos. En effet, la répartition des erreurs d'imputation le long du chromosome 17 pour ces deux familles montre qu'il y a davantage d'erreurs d'imputation dans deux régions chromosomiques situées vers 10 Mb et 23 Mb chez tous les plein-frères simulés. Or des travaux menés dans l'équipe ont montré que *M. kaido* et Generos partagent un haplotype de plusieurs Mb sur une partie du chromosome, ce qui pourrait expliquer les erreurs d'imputation observées. De nombreux variants ne ségrégeant que dans les familles 12_B et 12_E se trouvent sur cette portion chromosomique, et plusieurs de ces variants sont également rares dans le panel de référence, conduisant à des difficultés pour faire correspondre les haplotypes inférés dans les deux familles et les haplotypes du panel de référence. Nous avons obtenu des précisions d'imputation nettement plus élevées (surtout pour les chromosomes 11 et 17) lorsque les données génomiques de Generos étaient remplacées par celles de Jonathan dans les simulations, Jonathan étant une variété ayant de nombreux descendants dans le panel de référence et donc de nombreux haplotypes communs. Ceci confirme ainsi que les erreurs d'imputation étaient liées aux données génomiques de Generos et probablement aux haplotypes rares hérités de *M. kaido*. Il n'est pas exclu qu'une telle situation se retrouve dans d'autres programmes d'amélioration. En effet, chez le pommier, il est courant d'utiliser des espèces sauvages apparentées (ou CWR pour Crop Wild Relatives) comme géniteurs pour introgresser des gènes de résistance (Papp et al. 2020a ; Patocchi et al. 2020). Parmi les 27 familles de la REFPOP, nous pouvons citer au moins deux cas relativement récents d'utilisation de matériel exotique pour introgresser des gènes de résistance à la tavelure. Une partie des familles (familles commençant par I_, familles NOVADI et famille DLO) comporte le gène *Rvi6* hérité de l'espèce sauvage *M. floribunda* et situé sur le chromosome 1 (Gessler et al. 2006), tandis que les plein-frères des familles A1 et A2 ont hérité de gènes de résistance introgressés depuis l'espèce sauvage *M. sieversii* (Bus et al. 2005) via l'accession R12740-7A (figures 3.1 et 3.2) et situés sur le chromosome 2. Nous n'avons cependant observé de différence majeure de précision d'imputation ni entre ces familles et le reste de la population, ni entre les scénarios « exo_exo » et « commun_commun » de la famille A2. Dans les

deux cas, au moins 5 générations séparent l'utilisation du matériel exotique et les plein-frères de la REFPOP, conduisant probablement à la réduction de la longueur des haplotypes sauvages du fait des recombinaisons. Cette réduction a d'ailleurs été constatée en utilisant des marqueurs moléculaires dans le cas du gène *Rvi6* (King et al. 1999). Le cas des familles 12_B et 12_E, pour lesquelles l'utilisation de matériel exotique est plus récente (*M. kaido* étant un grand-parent ou un arrière grand-parent de Generos), est donc un cas à part parmi les 27 familles que nous avons étudiées. Compte tenu de la longueur des cycles de sélection chez le pommier, d'autres croisements récents non étudiés ici peuvent encore porter une part non négligeable du génome de l'espèce sauvage, particulièrement autour des régions introgressées. Il convient donc de s'interroger sur la manière d'améliorer la qualité de l'imputation de tels génotypes, pour lesquels la fréquence des allèles rares dans un panel de référence peut être une contrainte.

3.4.3 Comment améliorer davantage la qualité de l'imputation chez le pommier ?

Une première possibilité pour améliorer la qualité de l'imputation dans notre cas pourrait résider dans l'utilisation des informations de pédigrée non seulement lors de l'étape de phasage du panel de référence mais également lors de l'étape d'imputation en elle-même. En effet, nous avons vu dans ce chapitre que même dans le cas où les haplotypes des candidats se retrouvent dans le panel de référence, l'imputation peut être de mauvaise qualité si ces haplotypes sont présents en faible fréquence dans le panel. L'utilisation d'information de pédigrée pour indiquer quels haplotypes du panel de référence se retrouvent chez les candidats, indépendamment de leur fréquence, est donc une solution particulièrement intéressante, d'autant plus que les méthodes basées sur le pédigrée ont plusieurs fois montré leur supériorité par rapport aux méthodes basées sur le seul déséquilibre de liaison dans le cas où les candidats sont structurés en famille (Hickey et al. 2012a ; Miar et al. 2017), comme c'est souvent le cas dans les programmes d'amélioration. La prise en compte des informations de pédigrée permet en outre une meilleure précision d'imputation pour les allèles rares (Sargolzaei et al. 2014). Au début de la thèse, deux logiciels d'imputation basés sur l'utilisation de pédigrée étaient fréquemment utilisés : FImpute (Sargolzaei et al. 2014) et AlphaImpute (Antolín et al. 2017 ; Hickey et al. 2012a). FImpute n'a pas pu être testé du fait de la structure des données à renseigner au logiciel. En effet, l'information de pédigrée donnée au logiciel doit comporter l'indication du sexe (mâle ou femelle) des parents des individus à imputer, ce qui ne peut être fait chez le pommier puisque ses fleurs sont hermaphrodites. Il n'est donc pas possible de connaitre le sexe des parents dans un pédigrée, d'autant plus qu'un même individu peut être utilisé comme mâle dans certains croisements et comme femelle dans d'autres croisements. Dans le cas d'AlphaImpute, des premiers essais peu concluants en termes de précision d'imputation et le temps de calcul important (phasage

et imputation en 2 jours pour Beagle 4.0 mais en une semaine pour AlphaImpute) nous ont poussé à abandonner l'étude de cet outil. Il existe cependant de nouveaux outils développés par les créateurs d'AlphaImpute, permettant par exemple de prendre en compte la structure en familles biparentales souvent présente dans les programmes d'amélioration chez les végétaux (Gonen et al. 2018 ; Whalen et al. 2018a) ou basés sur une optimisation des algorithmes d'imputation (méthode dite « hybrid peeling », Whalen et al. (2018b)). A terme, il serait intéressant d'évaluer la performance de ces outils chez le pommier, tant en termes de précision d'imputation que de temps de calcul. Soulignons par ailleurs que nous avons utilisé les paramètres par défaut de Beagle lors de l'imputation. Or Beagle est à la base un logiciel conçu pour imputer des données de génétique humaine et certains paramètres peuvent donc ne pas être adaptés au cas du pommier. Un ajustement de ces paramètres pourrait donc permettre d'obtenir une meilleure précision d'imputation (Pook et al. 2020). Une autre approche que nous avons explorée pour les familles 12_B et 12_E consiste à phaser simultanément les individus à imputer et leurs apparentés en utilisant les informations de pédigrée afin d'obtenir des phases de façon plus précise, puis d'imputer les individus en utilisant leurs haplotypes plutôt que leurs données génotypiques. Cette approche, nommée pré-phasage, a été initialement proposée en génétique humaine afin d'éviter de phaser les individus à imputer à chaque fois qu'un nouveau panel de référence est disponible, permettant ainsi d'optimiser le temps de calcul de l'imputation (Howie et al. 2012 ; 2011). Dans le cas des familles 12_B et 12_E, le pré-phasage pourrait permettre de déterminer précisément les haplotypes des plein-frères et donc de retrouver plus facilement ces haplotypes dans le panel de référence même sans utilisation des informations de pédigrée. Cependant, la méthode de pré-phasage n'a eu quasiment aucun impact sur la précision d'imputation dans notre étude, quels que soient les chromosomes et les familles étudiés. Le gain de précision le plus important était de 0,01 et a été observé pour le chromosome 17 de la famille 12_B. L'amélioration quasi-nulle de la précision d'imputation en utilisant la méthode de pré-phasage suggère donc que les individus à imputer étaient déjà correctement phasés même sans utilisation des informations de pédigrée, et que la faible précision d'imputation observée pour les familles 12_B et 12_E provient plutôt de la faible fréquence de certains haplotypes dans le panel de référence. Ainsi, lorsqu'il est connu que certains individus sont porteurs d'allèles rares (par exemple dans le cas où le pédigrée indique qu'un individu apparenté provient de matériel exotique), il peut être nécessaire d'élargir le panel de référence utilisé lors de l'imputation afin que les haplotypes propres à cet individu y soient plus fréquemment représentés. Cela peut passer par le génotypage à haute densité d'individus apparentés à l'individu en question, ou par l'utilisation d'un panel de référence spécialement conçu pour optimiser l'imputation (Rowan et al. 2019). L'identification d'individus pouvant impacter la qualité d'imputation n'est cependant pas aisée si l'on se fie uniquement aux pédigrées. D'une part, les pédigrées peuvent être relativement superficiels. Au début de la thèse, le pédigrée utilisé n'indiquait d'ailleurs pas que Generos

descendait de *M. kaido* et aucun élément n'aurait donc pu permettre de préjuger la qualité d'imputation des familles 12_B et 12_E. D'autre part, les pédigrées historiques peuvent être erronés chez une espèce allogame comme le pommier du fait des multiples flux de pollen (Evans et al. 2011). Les erreurs de pédigrée peuvent avoir un impact non négligeable sur la précision d'imputation (Ros-Freixedes et al. 2020) et il est donc préférable d'utiliser des informations de pédigrée validées sur la base de marqueurs moléculaires afin d'améliorer la qualité des travaux d'imputation. Chez le pommier, l'utilisation de marqueurs moléculaires a récemment permis de vérifier ou de découvrir des relations d'apparentement entre variétés et ces relations sont régulièrement mises à jour sur la base de nouvelles données génomiques disponibles (Howard et al. 2021). Si l'imputation est basée sur un panel de référence, il est alors recommandé de régulièrement mettre à jour le phasage du panel avec les informations les plus récentes. Nous avons d'ailleurs bénéficié d'une mise à jour des pédigrées entre l'imputation des familles simulées et l'imputation des données réelles, ce qui peut aussi expliquer la bonne qualité apparente de l'imputation pour ces données. Une fois les résultats d'imputation obtenus, il peut aussi être intéressant de ne conserver que les marqueurs pour lesquels la probabilité d'imputation correcte est élevée, par exemple sur la base de l'AR^2. Un seuil d'AR^2 de 0,7 a été proposé par Tiplady et Ric (2015) et la figure S3.14 montre que la suppression des marqueurs en-dessous de ce seuil améliore effectivement la précision d'imputation dans notre cas. Cependant, ce seuil conduit à supprimer environ 20% des marqueurs et nous avons montré dans le prochain chapitre que la suppression de ces marqueurs n'améliorait pas la précision de prédiction dans les modèles de prédiction génomique pour les caractères étudiés (voir la sous-section 4.2.1). Nous avons donc conservé l'intégralité des marqueurs dans la suite des travaux menés.
Enfin, la qualité de l'imputation est influencée par la qualité des données de séquence du génome utilisées. Les positions physiques des marqueurs, renseignées lors de l'imputation, sont basées sur le génome dit GDDH13 (Daccord et al. 2017), qui est un génome récent de bonne qualité par rapport à la première version du génome disponible chez le pommier (Velasco et al. 2010). Cependant, une version plus récente du génome (Zhang et al. 2019) a montré qu'il subsistait des erreurs dans la version GDDH13 du génome, notamment 14 larges inversions dont une de plusieurs Mb sur le chromosome 1. L'utilisation de données génomiques de haute qualité et donc de positions physiques précises devrait permettre une meilleure qualité d'imputation dans les futurs travaux menés chez le pommier.

3.5 Conclusion

Nous avons montré dans ce chapitre qu'une précision d'imputation élevée pouvait être obtenue chez le pommier pour des données simulées représentatives du matériel élite ou des hybrides RG x E présentés dans le chapitre précédent. Le présent chapitre met également en évidence

la nécessité de prendre en compte la présence éventuelle d'haplotypes rares dans les méthodes d'imputation basées sur l'utilisation du seul déséquilibre de liaison. Malgré cette possible difficulté, les travaux présentés dans ce chapitre suggèrent que la précision d'imputation des données de la thèse est élevée et que ces données imputées peuvent donc être utilisées sans souci dans des modèles de prédiction génomique

Dans le prochain chapitre de la thèse, nous explorons l'intérêt de combiner du matériel élite et des ressources génétiques au sein d'une même population d'entrainement. A ce titre, l'effet de l'utilisation de données imputées ou non est étudié.

Figures supplémentaires du chapitre

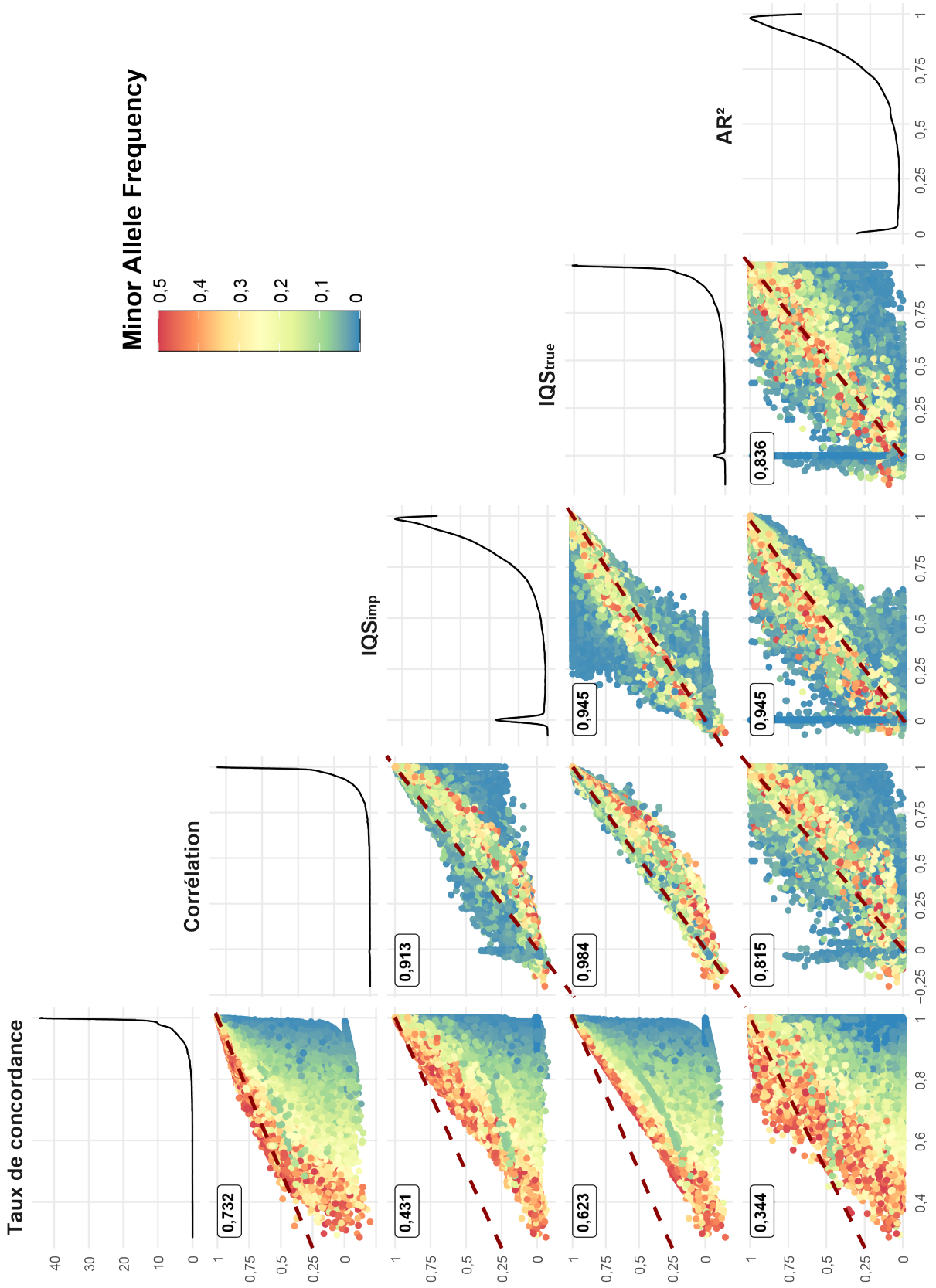


FIGURE S3.5 – Distributions des précisions d'imputation par marqueurs (sur la diagonale) et graphes de corrélation entre les différents indicateurs utilisés pour mesurer la qualité d'imputation des données simulées (la ligne rouge en pointillés représentent la diagonale de chaque graphe, certains n'ayant pas la même échelle en abscisse que d'autres).

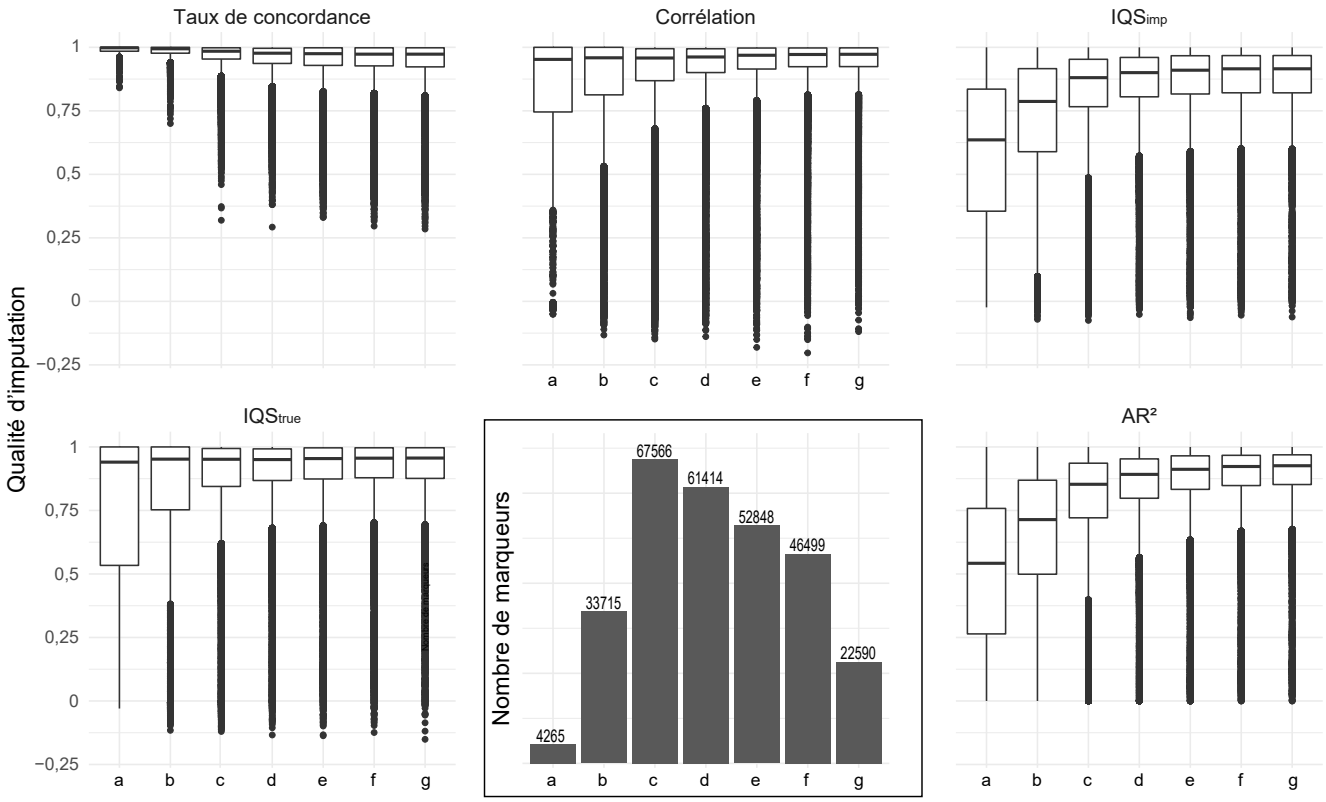


FIGURE S3.6 – Qualité de l'imputation mesurée par différents indicateurs pour les classes de MAF étudiées et nombre de marqueurs par classe de MAF (en bas au milieu).
Les lettres utilisées sur la figure correspondent aux classes de MAF suivantes : **a**) [0 ; 0,01] **b**)]0,01 ; 0,05] **c**)]0,05 ; 0,15] **d**)]0,15 ; 0,25] **e**)]0,25 ; 0,35] **f**)]0,35 ; 0,45] **g**) MAF $> 0{,}45$

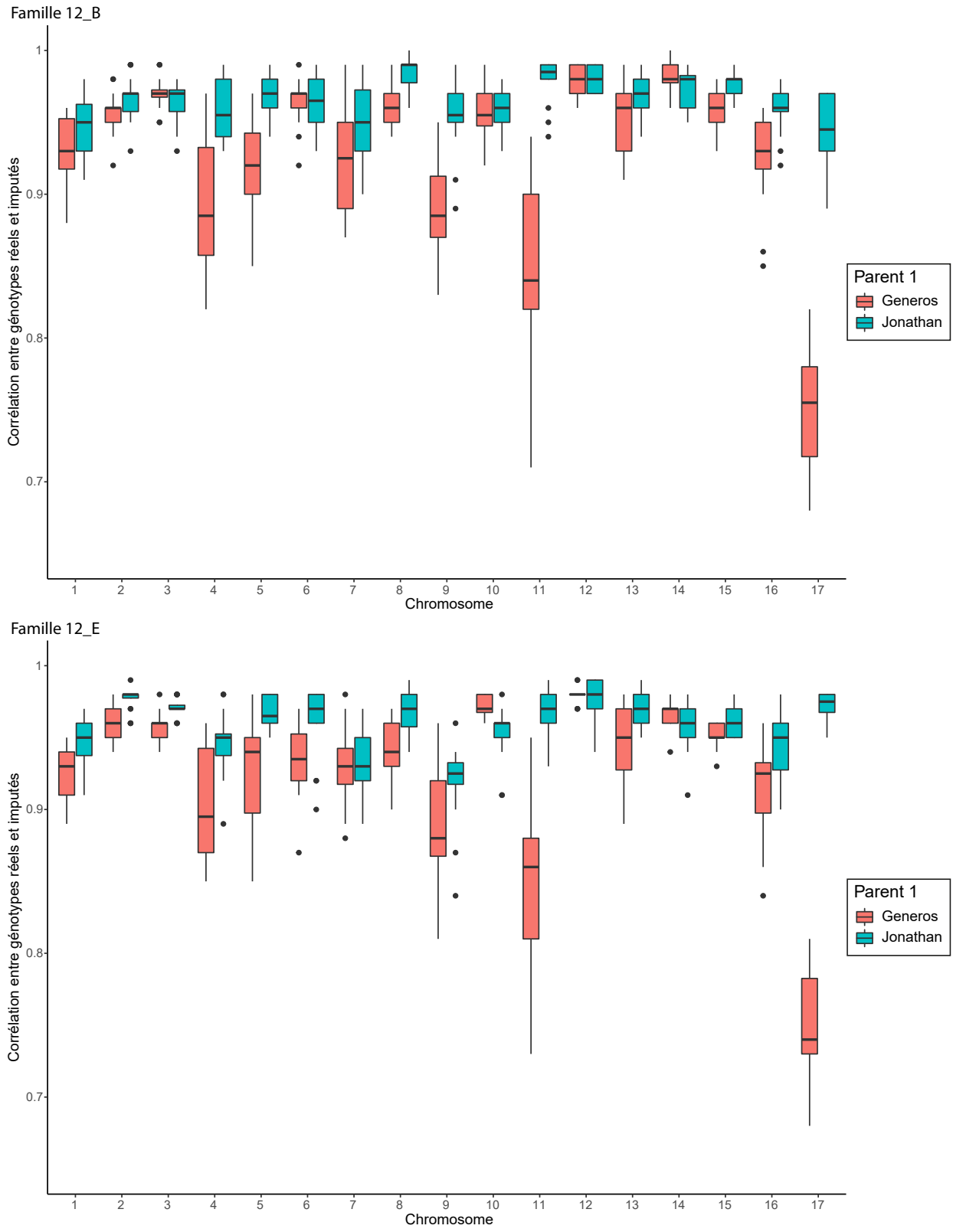


FIGURE S3.7 – Précision d'imputation des familles 12_B et 12_E lorsque Generos est utilisé comme parent dans les simulations (en rouge) ou lorsque Generos est remplacé par Jonathan (en bleu)

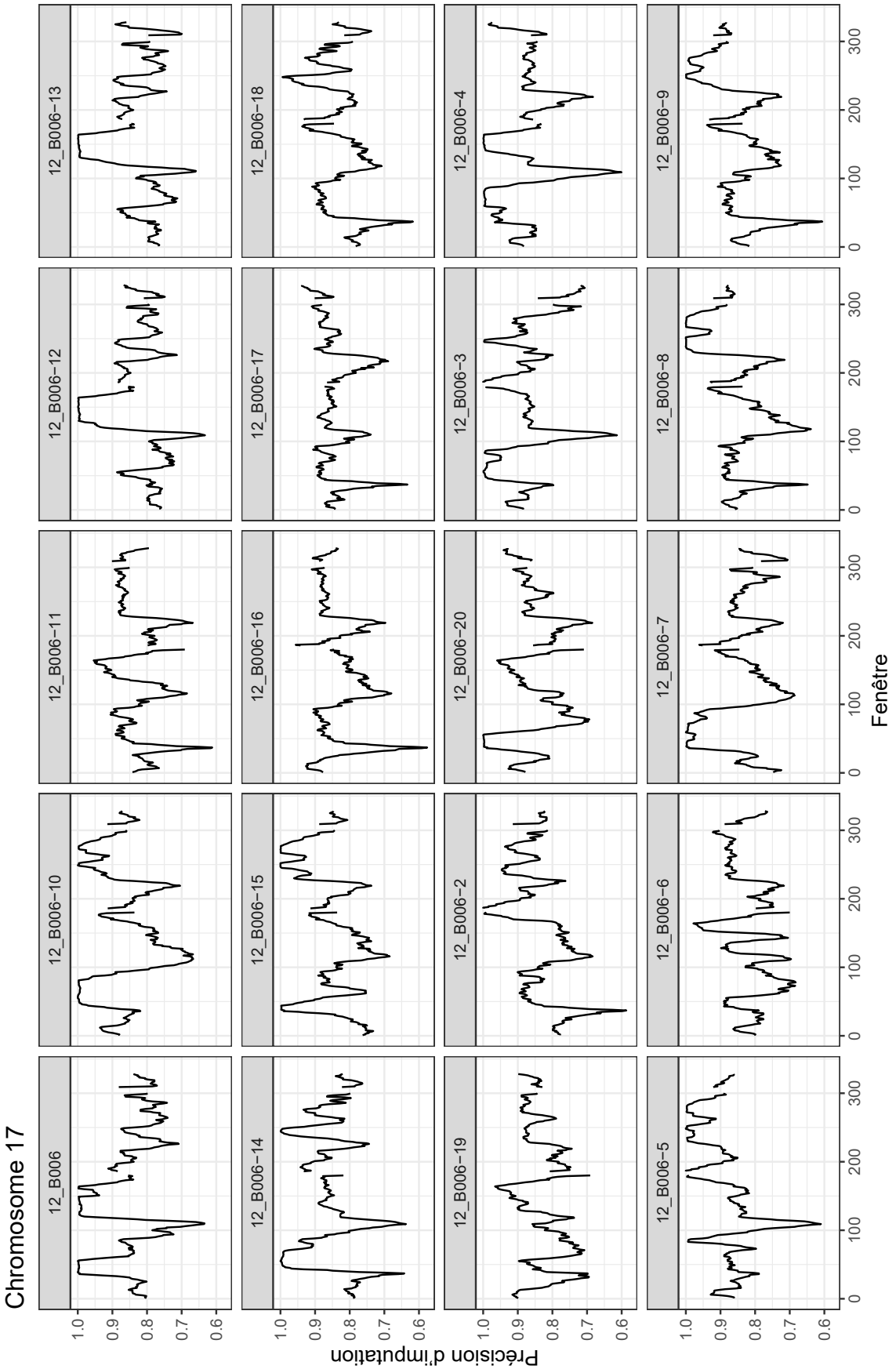


FIGURE S3.8 – Précision d'imputation le long du chromosome 17 de la famille 12_B. La précision est calculée au sein de fenêtres glissantes de 1Mb avec un pas de 100kb

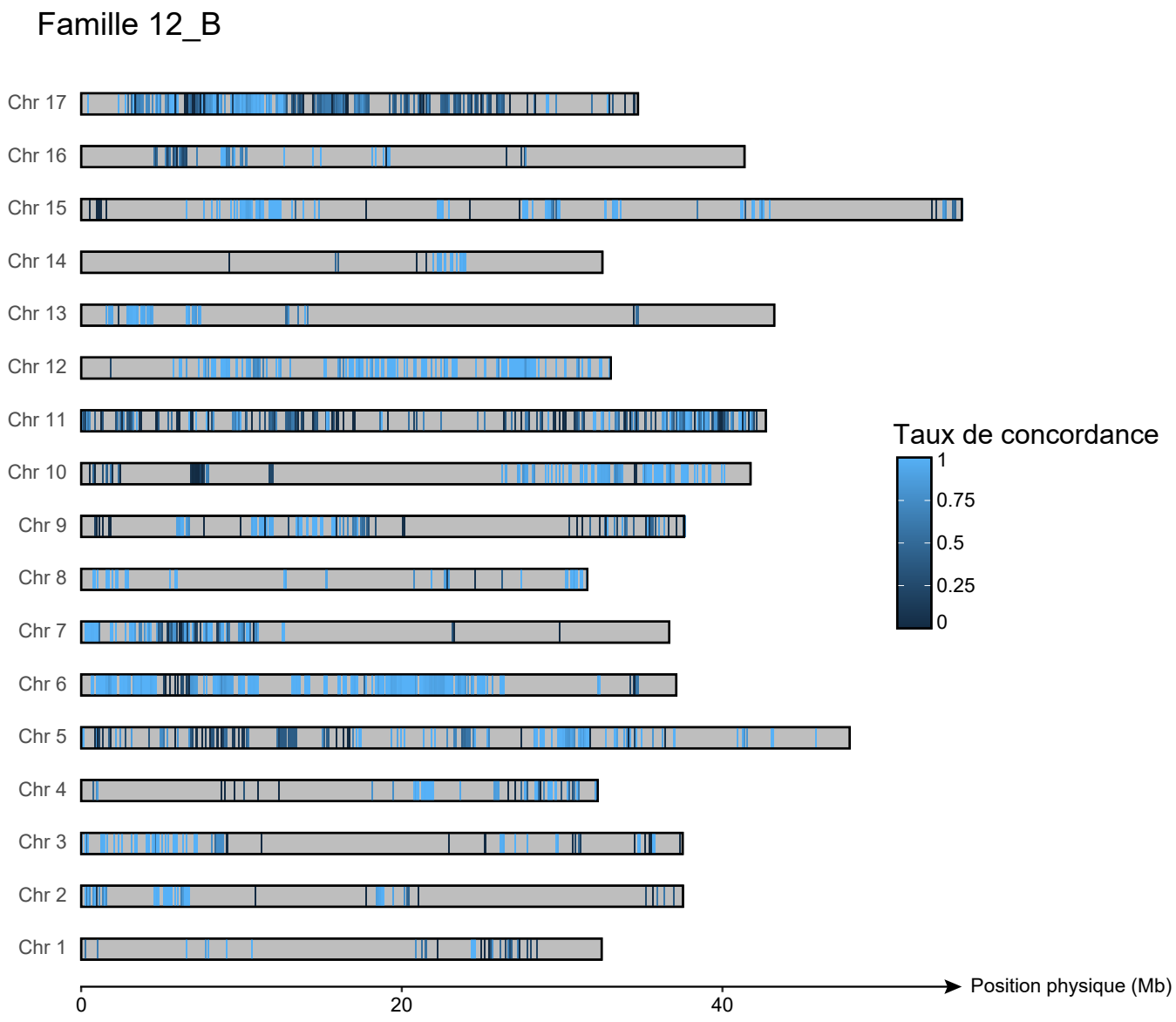


FIGURE S3.9 – Précision d'imputation des allèles rares dans les 26 familles simulées et ségrégeant dans la famille 12_B. La précision est calculée au sein de fenêtres glissantes de 1Mb avec un pas de 100kb. Les zones en gris correspondent aux régions génomiques pour lesquelles aucun allèle rare n'est présent

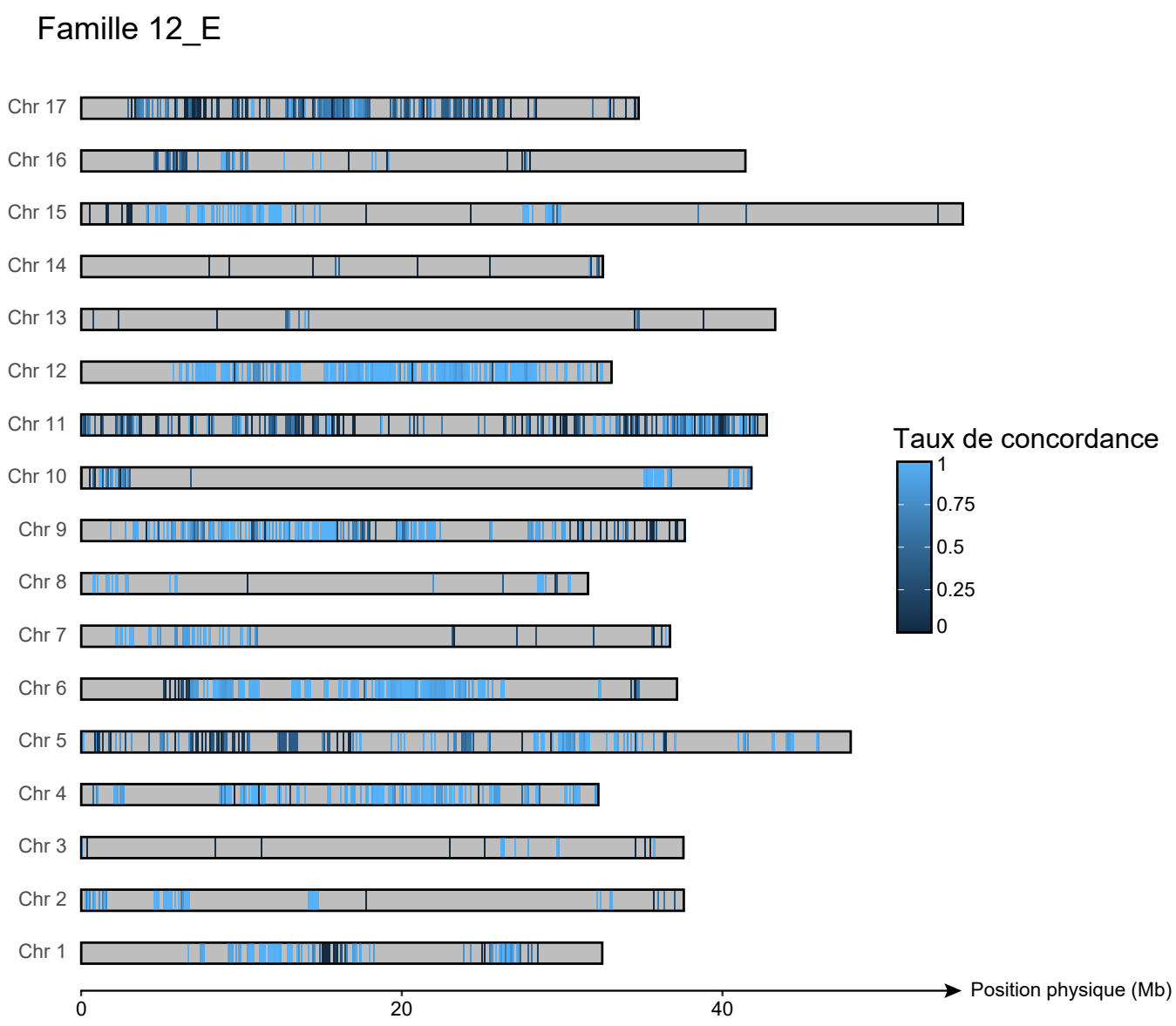


FIGURE S3.10 – Précision d'imputation des allèles rares dans les 26 familles simulées et ségrégeant dans la famille 12_E. La précision est calculée au sein de fenêtres glissantes de 1Mb avec un pas de 100kb. Les zones en gris correspondent aux régions génomiques pour lesquelles aucun allèle rare n'est présent

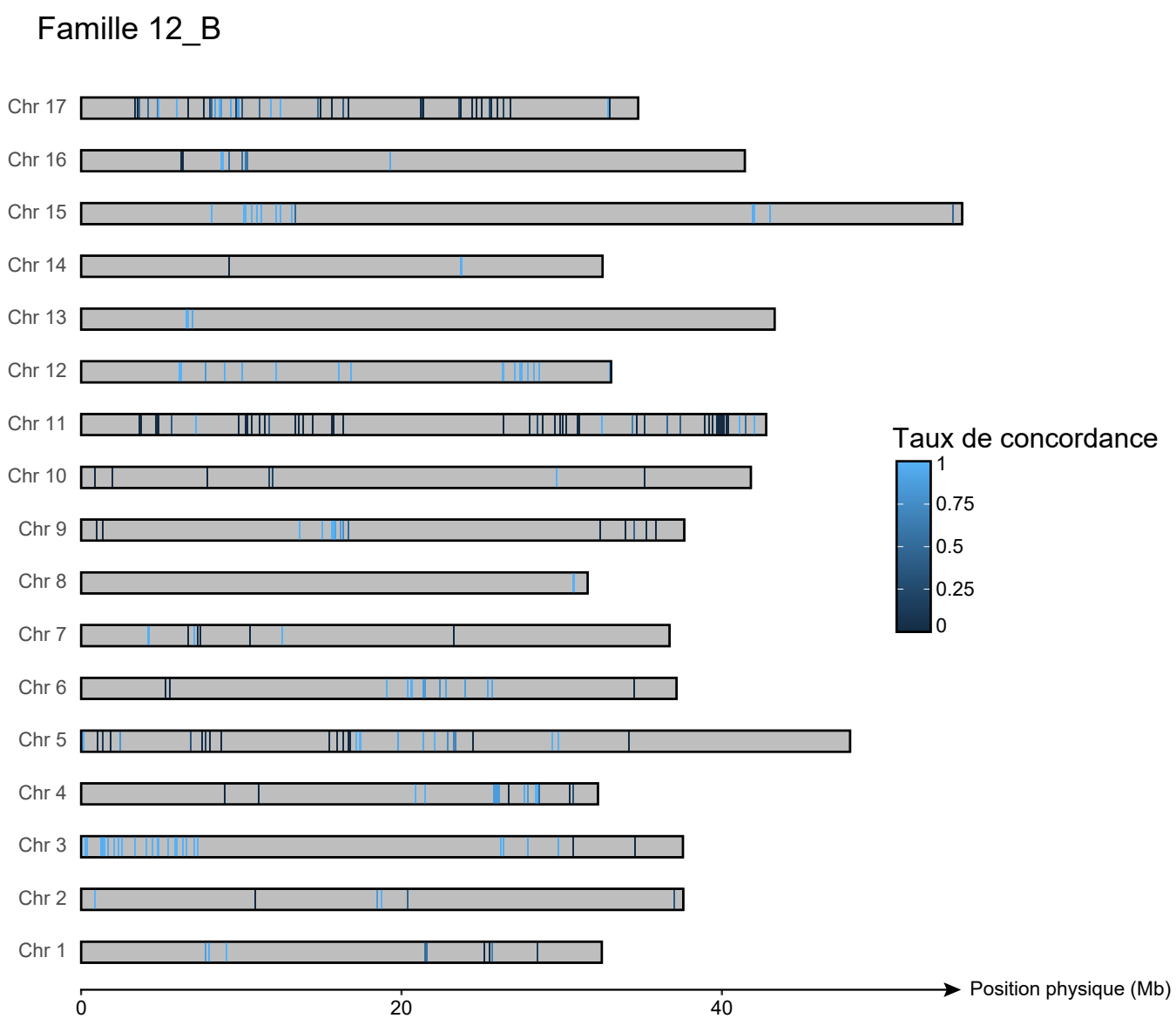


FIGURE S3.11 – Précision d'imputation des allèles rares dans le panel de référence ainsi que dans les 26 familles simulées et ségrégeant dans la famille 12_B. La précision est calculée au sein de fenêtres glissantes de 1Mb avec un pas de 100kb. Les zones en gris correspondent aux régions génomiques pour lesquelles aucun allèle rare n'est présent

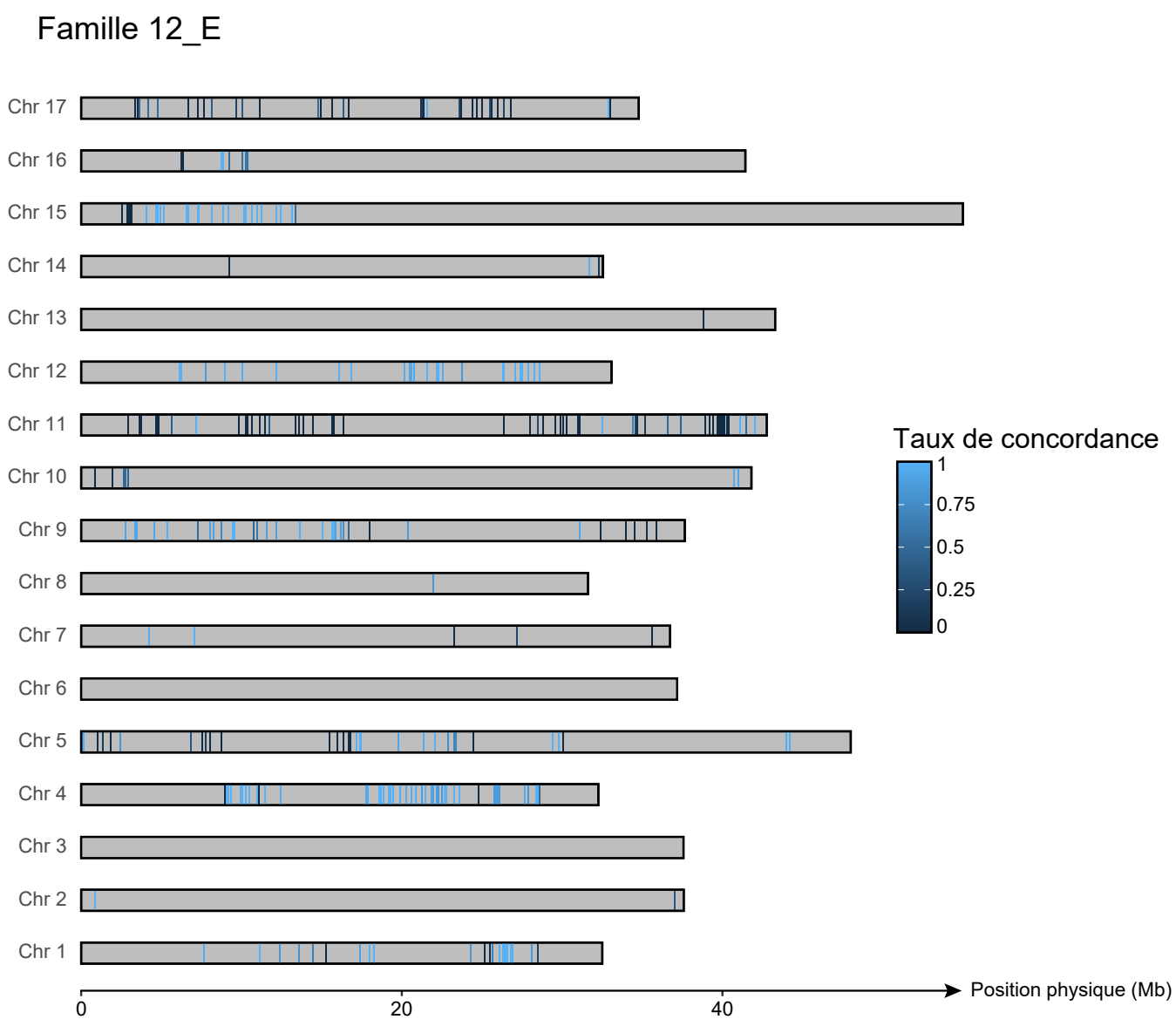


FIGURE S3.12 – Précision d'imputation des allèles rares dans le panel de référence ainsi que dans les 26 familles simulées et ségrégeant dans la famille 12_E. La précision est calculée au sein de fenêtres glissantes de 1Mb avec un pas de 100kb. Les zones en gris correspondent aux régions génomiques pour lesquelles aucun allèle rare n'est présent

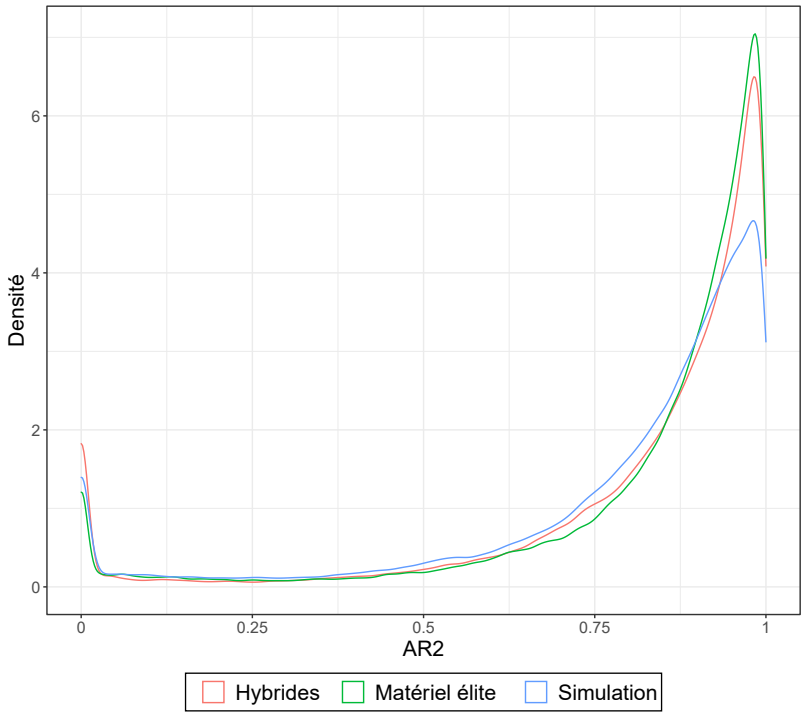


FIGURE S3.13 – Distribution des valeurs d'AR^2 selon le jeu de données utilisé lors de l'imputation

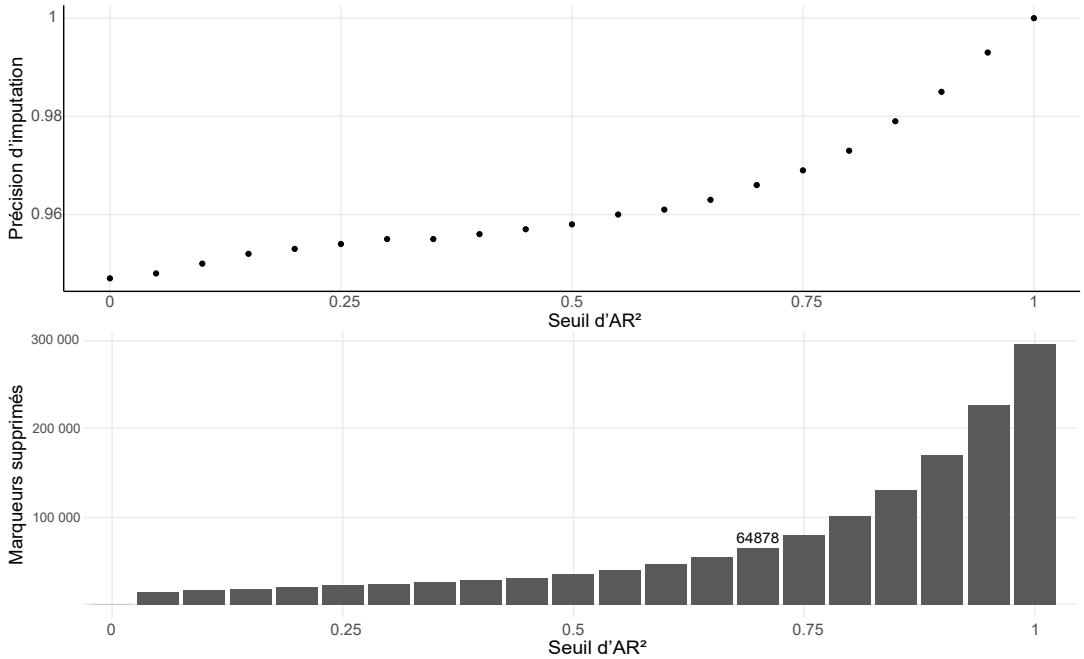


FIGURE S3.14 – Évolution de la précision de prédiction lorsque des marqueurs ne sont pas retenus sur la base de l'AR^2 et nombre de marqueurs supprimés selon le seuil d'AR^2 choisi

Tableaux supplémentaires du chapitre

TABLEAU 3.2 – Nombre d'individus et densité de génotypage au sein des familles utilisées lors des simulations

Famille	n_{ped}	n_{HD}	% HD	2P	1P 2GP	4GP	1P Autre	Autre	n_{sim}
12_B	10	8	80		X				2
12_E	29	18	62,1	X					1
12_F	21	15	71,4		X				2
12_I	29	15	51,7					X	5
A1	41	20	48,8					X	18
A2	34	16	47,1					X	-
B3F	17	12	70,6				X		3
DLO	25	12	48				X		4
Novadi1	30	14	46,7	X					1
Novadi2	28	17	60,7	X					1
12_J	21	15	71,4		X				2
12_K	18	9	50					X	4
12_L	17	10	58,8					X	4
12_N	21	12	57,1				X		3
12_O	29	19	65,5					X	4
12_P	13	8	61,5		X				2
FuGa	8	8	100	X					1
FuPi	9	8	88,9	X					1
GaCr	9	8	88,9		X				2
GaPi	9	8	88,9	X					1
I_BB	19	15	78,9			X			3
I_CC	18	9	50					X	4
I_J	29	17	58,6				X		3
I_M	33	20	60,6	X					1
I_W	32	18	56,3	X					1
JoPr	15	9	60	X					1
TeBr	6	5	83,3		X				2
Total				9	6	1	3	8	

n_{ped} : nombre d'individus composant le pédigrée ; **n_{HD}** : nombre d'individus génotypés à haute densité dans le pédigrée ; **%HD** : pourcentage d'individus génotypés à haute densité dans le pédigrée ; **2P** : les deux parents des plein-frères terminaux sont génotypés à haute densité ; **1P 2GP** : un parent et les deux parents de l'autre parents des plein-frères terminaux sont génotypés en HD ; **4 GP** : les quatre grand-parents des plein-frères terminaux sont génotypés en HD ; **1P Autre** : un parent des plein-frères terminaux est génotypé en HD et au moins un des grand-parents n'est pas génotypé en HD ; **Autre** : les parents des plein-frères terminaux ne sont pas génotypés en HD ; **n_{sim}** : nombre d'individus du pédigrée à simuler avant de pouvoir simuler les individus terminaux

TABLEAU 3.3 – Caractéristiques des chromosomes simulés

Chr.	Taille		Nombre de marqueurs	Mk/Mb	Marqueurs monomorphes*
	cM	Mb			
1	63	32,6	12405	380	458
2	78,4	37,6	17879	476	904
3	73,9	37,5	17558	468	801
4	65,5	32,3	15629	484	466
5	77,8	47,9	23089	482	973
6	75,3	37,1	15333	413	1436
7	82,4	36,7	16597	452	1153
8	68,5	31,6	14936	473	407
9	67	37,6	16117	429	520
10	81,3	41,8	22779	546	1446
11	80,9	43	20254	471	888
12	65,4	33	16817	509	695
13	71,4	44,3	16489	372	658
14	64,4	32,5	17955	552	802
15	112,1	54,9	22467	409	863
16	67,5	41,4	18843	455	1012
17	71,8	34,7	18092	521	860

* Nombre de marqueurs monomorphes parmi les 26 familles après simulation. Les positions physiques renseignées correspondent à la position physique du dernier marqueur de chaque chromosome

TABLEAU 3.4 – Précision d'imputation moyenne par chromosome pour les 26 familles simulées lorsque le panel a été phasé en utilisant les informations de pédigrée

Famille	Précision d'imputation		
	Minimum	Moyenne	Maximum
12_B	0,75	0,93 (0,06)	0,98
12_E	0,75	0,92 (0,06)	0,98
12_F	0,88	0,95 (0,03)	0,98
12_I	0,88	0,93 (0,02)	0,97
12_J	0,88	0,95 (0,03)	0,98
12_K	0,89	0,95 (0,02)	0,97
12_L	0,93	0,95 (0,01)	0,98
12_N	0,86	0,93 (0,03)	0,98
12_O	0,9	0,95 (0,02)	0,98
12_P	0,88	0,95 (0,03)	0,98
A1	0,83	0,92 (0,04)	0,97
B3F	0,89	0,94 (0,02)	0,98
DLO-12	0,86	0,93 (0,03)	0,97
FuGa	0,9	0,97 (0,02)	0,99
FuPi	0,9	0,97 (0,02)	0,99
GaCr	0,93	0,96 (0,01)	0,98
GaPi	0,95	0,97 (0,01)	0,99
I_BB	0,92	0,95 (0,02)	0,99
I_CC	0,89	0,95 (0,02)	0,98
I_J	0,84	0,93 (0,04)	0,98
I_M	0,89	0,94 (0,02)	0,97
I_W	0,86	0,93 (0,04)	0,98
JoPr	0,9	0,95 (0,02)	0,98
Novadi1	0,81	0,93 (0,04)	0,96
Novadi2	0,88	0,94 (0,03)	0,98
TeBr	0,91	0,96 (0,01)	0,97

L'écart-type des précisions d'imputation par chromosome est indiqué entre parenthèses

TABLEAU 3.5 – Précision d'imputation mesurée par marqueur et par chromosome

Chr.	Imp. mono	Imp_poly & true_mono	Imp_mono & true_poly	Précision d'imputation			
				Corrélation	IQS_{true}	IQS_{imp}	AR^2
1	372	124	38	0,91 (0,15)	0,87 (0,19)	0,78 (0,23)	0,77 (0,24)
2	839	130	65	0,94 (0,12)	0,91 (0,17)	0,83 (0,24)	0,83 (0,23)
3	736	185	120	0,94 (0,13)	0,91 (0,19)	0,84 (0,24)	0,83 (0,24)
4	350	180	64	0,92 (0,13)	0,88 (0,19)	0,81 (0,22)	0,79 (0,22)
5	886	269	182	0,89 (0,17)	0,85 (0,22)	0,77 (0,26)	0,77 (0,26)
6	1154	362	80	0,92 (0,14)	0,87 (0,23)	0,75 (0,3)	0,75 (0,29)
7	1009	449	305	0,87 (0,18)	0,79 (0,27)	0,68 (0,29)	0,68 (0,29)
8	374	92	59	0,93 (0,12)	0,9 (0,17)	0,83 (0,21)	0,84 (0,2)
9	450	160	90	0,88 (0,16)	0,84 (0,22)	0,76 (0,24)	0,75 (0,24)
10	1350	267	171	0,91 (0,14)	0,87 (0,2)	0,77 (0,27)	0,77 (0,26)
11	804	324	240	0,9 (0,16)	0,86 (0,22)	0,79 (0,25)	0,79 (0,24)
12	653	111	69	0,95 (0,12)	0,93 (0,16)	0,86 (0,23)	0,86 (0,23)
13	491	213	46	0,93 (0,13)	0,89 (0,19)	0,82 (0,23)	0,82 (0,23)
14	573	295	66	0,94 (0,12)	0,9 (0,19)	0,83 (0,24)	0,81 (0,23)
15	783	191	111	0,93 (0,15)	0,9 (0,2)	0,83 (0,25)	0,83 (0,24)
16	745	383	116	0,87 (0,19)	0,82 (0,26)	0,74 (0,29)	0,73 (0,29)
17	722	222	84	0,9 (0,14)	0,86 (0,2)	0,77 (0,25)	0,76 (0,24)

Imp. mono : marqueurs monomorphes après imputation **Imp_poly & true_mono** : marqueurs polymorphes après imputation mais monomorphes en réalité **Imp_mono & true_poly** : marqueurs monomorphes après imputation mais polymorphes en réalité

TABLEAU 3.6 – Précision d'imputation mesurée par individu simulé et par chromosome lorsque le panel a été phasé en utilisant les informations de pédigrée

Chr.	Précision d'imputation		
	Minimum	Moyenne	Maximum
1	0,83	0,94 (0,03)	0,99
2	0,81	0,96 (0,03)	0,99
3	0,85	0,96 (0,02)	0,99
4	0,82	0,94 (0,04)	1
5	0,81	0,94 (0,03)	1
6	0,83	0,95 (0,02)	1
7	0,75	0,92 (0,05)	1
8	0,81	0,95 (0,03)	1
9	0,75	0,92 (0,04)	0,99
10	0,79	0,94 (0,04)	1
11	0,71	0,94 (0,04)	1
12	0,82	0,97 (0,03)	1
13	0,82	0,95 (0,03)	1
14	0,84	0,96 (0,03)	1
15	0,88	0,96 (0,02)	0,99
16	0,75	0,92 (0,04)	0,99
17	0,67	0,94 (0,06)	0,99

TABLEAU 3.7 – Précision d'imputation par chromosome pour la famille A2 dans les scénarios exo_exo, common_exo et common_common

	exo_exo			common_exo			common_common		
Chr.	Minimum	Moyenne	Maximum	Minimum	Moyenne	Maximum	Minimum	Moyenne	Maximum
1	0,75	0,9	0,99	0,79	0,9	0,99	0,76	0,89	0,99
2	0,73	0,88	0,98	0,75	0,88	0,98	0,72	0,87	0,98
3	0,87	0,95	0,99	0,83	0,95	0,99	0,88	0,95	0,99
4	0,82	0,93	0,98	0,77	0,92	0,99	0,75	0,91	0,99
5	0,84	0,93	0,98	0,79	0,91	0,98	0,8	0,92	0,98
6	0,82	0,93	0,99	0,85	0,92	0,97	0,81	0,91	0,98
7	0,73	0,88	0,96	0,71	0,88	0,97	0,75	0,89	0,97
8	0,81	0,92	0,98	0,82	0,92	0,99	0,86	0,93	0,98
9	0,82	0,92	0,98	0,81	0,91	0,98	0,82	0,93	0,98
10	0,77	0,88	0,99	0,68	0,88	0,99	0,76	0,88	0,97
11	0,85	0,93	0,98	0,83	0,93	0,98	0,77	0,92	0,98
12	0,84	0,92	0,98	0,75	0,9	0,98	0,75	0,91	0,98
13	0,77	0,89	0,97	0,78	0,89	0,97	0,77	0,89	0,97
14	0,8	0,93	0,99	0,8	0,93	0,99	0,82	0,94	0,99
15	0,84	0,92	0,98	0,83	0,91	0,98	0,82	0,91	0,99
16	0,81	0,89	0,95	0,76	0,89	0,95	0,76	0,89	0,95
17	0,84	0,93	0,99	0,79	0,92	0,98	0,78	0,92	0,99

CHAPITRE 4

Combinaison d'un panel de ressources génétiques et de matériel élite en vue d'améliorer la précision de prédiction chez le pommier

4.1 Introduction

Comme rappelé dans la partie introductive du manuscrit, les programmes d'amélioration chez le pommier s'appuient aujourd'hui sur l'utilisation récurrente d'un petit nombre de variétés élite plus ou moins apparentées les unes aux autres. De ce fait, l'élargissement de la base génétique est une problématique actuellement abordée par plusieurs équipes de recherche à travers le monde. Dans ce contexte, la sélection génomique apparait comme une approche intéressante afin d'identifier des ressources génétiques actuellement pas ou peu utilisées dans les programmes d'amélioration et donc afin d'accélérer le transfert d'allèles favorables absents des variétés élite actuelles. Une des principales difficultés liées à l'utilisation de la sélection génomique dans le cadre d'actions de pré-breeding chez le pommier réside dans l'établissement de la population d'entrainement servant à calibrer les équations de prédiction. En effet, générer des populations de grande taille est difficile du fait de contraintes d'espace et du coût de phénotypage et d'entretien de telles populations. Dans ce chapitre, structuré autour d'un article publié dans la revue G3 :Genes|Genomes|Genetics, nous avons étudié l'intérêt de regrouper des génotypes provenant de programmes de sélection actuels et de variétés anciennes (que nous nommons « populations élite » et « population des ressources génétiques ») afin de constituer une population d'entrainement diverse qui pourrait être utilisée dans des programmes de pré-breeding. Dans un premier temps, nous avons mesuré les précisions de prédiction intra et inter-population pour différents caractères afin d'étudier l'intérêt d'une combinaison de ces populations dans une même

population d'entrainement. Nous avons ensuite comparé les précisions de prédiction obtenues intra-population et en combinant les deux populations pour différents caractères. Nous nous sommes enfin intéressés à l'intérêt de combiner les deux populations pour prédire des hybrides résultant d'un croisement entre variétés modernes et anciennes, en testant dans ce cas différentes proportions [1] des deux populations et différentes tailles de population d'entrainement.

1. Sur les figures 4, S4.14 et S4.15, les proportions sont notées sont la forme $\text{Proportion}_{\text{Elite}}/\text{Proportion}_{\text{RG}}$. Par exemple, une proportion notée 0.6/0.4 signifie que 60% des génotypes de la population d'entrainement proviennent de la population élite et 40% de la population des ressources génétiques

G3, 2021, jkab420
https://doi.org/10.1093/g3journal/jkab420
Advance Access Publication Date: 10 December 2021
Genomic Prediction

Combining genetic resources and elite material populations to improve the accuracy of genomic prediction in apple

Xabi Cazenave,[1] Bernard Petit,[1] Marc Lateur,[2] Hilde Nybom ,[3] Jiri Sedlak ,[4] Stefano Tartarini ,[5] François Laurens ,[1] Charles-Eric Durel ,[1] and Hélène Muranty [1]*

[1]Univ Angers, INRAE, Institut Agro, IRHS, SFR QuaSaV, F-49000 Angers, France,
[2]Plant Breeding and Biodiversity, Centre Wallon de Recherches Agronomiques, Gembloux, Belgium,
[3]Department of Plant Breeding, Swedish University of Agricultural Sciences, Kristianstad, Sweden,
[4]Výzkumný a Šlechtitelský ústav Ovocnářský Holovousy s.r.o, Holovousy, Czech Republic, and
[5]Department of Agricultural Sciences, University of Bologna, Bologna, Italy

*Corresponding author: Email: helene.muranty@inrae.fr

Abstract

Genomic selection is an attractive strategy for apple breeding that could reduce the length of breeding cycles. A possible limitation to the practical implementation of this approach lies in the creation of a training set large and diverse enough to ensure accurate predictions. In this study, we investigated the potential of combining two available populations, *i.e.*, genetic resources and elite material, in order to obtain a large training set with a high genetic diversity. We compared the predictive ability of genomic predictions within-population, across-population or when combining both populations, and tested a model accounting for population-specific marker effects in this last case. The obtained predictive abilities were moderate to high according to the studied trait and small increases in predictive ability could be obtained for some traits when the two populations were combined into a unique training set. We also investigated the potential of such a training set to predict hybrids resulting from crosses between the two populations, with a focus on the method to design the training set and the best proportion of each population to optimize predictions. The measured predictive abilities were very similar for all the proportions, except for the extreme cases where only one of the two populations was used in the training set, in which case predictive abilities could be lower than when using both populations. Using an optimization algorithm to choose the genotypes in the training set also led to higher predictive abilities than when the genotypes were chosen at random. Our results provide guidelines to initiate breeding programs that use genomic selection when the implementation of the training set is a limitation.

Keywords: genomic selection; training set design; population combination; germplasm; *Malus domestica*; Genomic Prediction; GenPred; Shared Data Resource

Introduction

Breeding programs in outbred fruit tree crops can take many years before new varieties are released, in part because of the long juvenile phase of the trees. Shortening the breeding cycle length for these crops could help increase genetic gain (van Nocker and Gardiner 2014). The length of the breeding cycle is also a constraint when breeders aim to introgress new traits from distant relatives or genetic resources, because several generations are usually needed before genotypes with the desired traits can be released as varieties.

In both cases, the use of molecular markers is an attractive strategy for the early identification of the most promising selection candidates or genetic resources (Myles 2013). Since most apple breeding programs around the world rely on a limited number of genotypes that are frequently used as parents for variety development (Noiton and Alspach 1996), mating strategies based on molecular markers could be used to broaden the genetic base of elite material (Yu *et al.* 2016) by identifying without phenotyping promising genetic resources that could be used as novel parents in pre-breeding programs (Crossa *et al.* 2016).

Recently, an approach referred to as genomic selection has particularly gained popularity among plant breeders (Voss-Fels *et al.* 2019), as genetic gain is expected to be higher with genomic selection compared to marker-assisted breeding, especially for complex traits (Xu *et al.* 2020). Genomic selection uses a training population of individuals that have been genotyped and phenotyped in order to estimate marker effects, then allowing the estimation of genomic breeding values of a candidate population that has only been genotyped (Meuwissen *et al.* 2001). The success of genomic selection depends on the accuracy of the predicted breeding values. The size of the training set (Zhang *et al.* 2017; Edwards *et al.* 2019), the relatedness between the training set and the candidates (Clark *et al.* 2012; Lehermeier *et al.* 2014) or the genetic architecture of the trait (Daetwyler *et al.* 2010; Wimmer *et al.* 2013) have been reported as factors affecting prediction accuracy. In fruit tree crops, the potential of genomic

Received: August 29, 2021. **Accepted:** November 29, 2021

selection has been outlined (Kumar *et al.* 2012a; Nsibi *et al.* 2020) and its implementation could help in efficiently using genetic resources in apple (Kumar *et al.* 2020). However, initiating pre-breeding programs based on genomic predictions can be challenging, in part because maintaining and phenotyping populations of large size is time-consuming and costly. The establishment of a large training set could thus limit the implementation of genomic selection. A way to overcome this limitation could be to combine data coming from several breeding populations, like historical data or data from other countries. This strategy has been explored in animal breeding (Hayes *et al.* 2009; Ibáñez-Escriche *et al.*, 2009; Wientjes *et al.* 2015) and more recently in plant breeding (Technow and Totir 2015; Sverrisdóttir *et al.* 2018; Olatoye *et al.* 2020) generally showing little to no gain in prediction accuracy. This observation may result from differences in marker effects between the combined populations, as well as from the differences in relatedness between these populations and the candidates to predict. Prediction models that allow the estimation of population-specific marker effects have been proposed (Karoui *et al.* 2012; Schulz-Streeck *et al.* 2012; Lehermeier *et al.* 2015) but few studies have used such models in plant breeding (Ramstein and Casler 2019; Rio *et al.* 2019). Differences in LD patterns between the combined populations may also explain the low observed gains from combination. In this case, it has been suggested to use high-density marker data in order to ensure consistency in SNP-QTL linkage disequilibrium between populations (de Roos *et al.* 2008).

Multi-population training sets can also be interesting when the genotypes to predict results from crosses between the combined populations. If no phenotypic records are available for the progenies of such crosses (as can be the case when initiating a new breeding program or when crossing selection material with exotic germplasm), one could consider combining populations in order to increase the diversity of the training set (Brandariz and Bernardo 2019). In this context, an optimum proportion of each combined population may exist. If this occurs and the sizes of the populations differ, it can be relevant to use a subset of the populations rather than to combine all the genotypes into a unique training set. When the size of the training set has to be fixed in advance, different algorithms exist to optimize the composition of the training set (Rincent *et al.* 2012; Akdemir *et al.* 2015; Mangin *et al.* 2019; Ou and Liao 2019) and they could be used to choose the genotypes to include in a training set in each of the combined populations (Isidro *et al.* 2015).

In this study, we considered the two abovementioned scenarios to investigate the interest of combining different populations into a unique training set for genomic prediction in apple. First, we assessed the potential increase in prediction accuracy when combining two populations, namely genetic resources and elite material, instead of using only one of the two, either using a standard GBLUP model or a model that allows the marker effects to differ between populations. We then predicted the GEBV of the progenies from crosses between elite material and genetic resources and evaluated the prediction accuracy when using different proportions of genetic resources and elite material in the training set and different strategies to choose the genotypes of this training set.

The objectives of our study were to (1) compare the predictive ability of genomic prediction models using simple or combined training sets within and across populations, (2) investigate the effect of different proportions of the two populations used in a training set and (3) assess the impact of high and medium marker density in these cases.

Materials and methods

In this study, three datasets were used for genomic predictions: the first one (hereafter referred to as the FBo-Hi dataset) regroups data coming from two past European projects while the second dataset (from here on referred to as the REFPOP dataset) contains data from an ongoing European project. The genotypes in both the FBo-Hi and REFPOP datasets are either old varieties (called genetic resources from now on) or progenies from breeding programs (called elite material in this study). Due to the different experimental design between the panels of the two datasets (see below), the FBo-Hi and REFPOP datasets provide an opportunity to investigate the effect of genomic predictions with or without the presence of genotype by environment interactions.

The third dataset contains data from crosses between old and modern varieties, which are part of a pre-breeding initiative engaged at IRHS, INRAE Angers. In this study, this dataset was only used as a validation set (VS).

Plant material

FBo-Hi panel

The panel regroups two different apple populations, from here on referred to as the genetic resources and elite material of the FBo-Hi dataset, for which phenotypic and genotypic data are available. The genetic resources represent European apple germplasm preserved in six European core-collections along with data coming from the EU-FP7 FruitBreedomics (FBo) project described in Laurens *et al.* (2018), while the elite material consists of progenies originating from biparental combinations from six European breeding programs that were gathered for pedigree-based QTL analysis during the HiDRAS (Hi) project (Kouassi *et al.* 2009). A total of 1194 unique genotypes (mainly old dessert apple cultivars) from the genetic resources were genotyped and phenotyped for at least one trait (see below). Similarly, 1018 progenies from 23 biparental combinations were genotyped and phenotyped for at least one trait. The genotypes of the FBo-Hi panel were not replicated across sites, except for some genotypes that were used to adjust phenotypic data.

REFPOP panel

The apple REFPOP is described in detail in Jung *et al.* (2020). This panel consists of 269 cultivars (hereafter referred to as genetic resources of the REFPOP) that are representative of the worldwide apple genetic diversity and of 265 progenies (hereafter referred to as elite material of the REFPOP) originating from 27 biparental combinations from various European breeding programs. Some of the genotypes of the REFPOP panel are also part of the FBo-Hi panel, as 189 accessions and 155 progenies are found in both panels. The panel is replicated across six locations in six European countries with contrasting environments and each genotype is replicated at least twice in each environment. At each site, the genetic resources and elite material are planted in the same orchard.

Panel of hybrids

The panel consists of 348 so-called "hybrids" originating from 10 biparental combinations of approximately the same size. Each combination involved a controlled cross between an old cultivar and an elite cultivar. The mating design is presented in Supplementary Table S1 and involved five old cultivars and five elite cultivars. Each hybrid genotype was represented by only one tree in the orchard in Angers, France.

Genotypic data

Genotyping of the plant material

For both the FBo-Hi and REFPOP panels, the genetic resources were genotyped using the Affymetrix Axiom Apple 480K SNP genotyping array (Bianco et al. 2016) and the elite material using the Illumina Infinium 20K SNP genotyping array (Bianco et al. 2014). The hybrids were also genotyped with the 20K genotyping array. After filtering, 303,239 SNP markers were retained from the 480K array and 10,295 SNP markers from the 20K array, of which 7060 were common to the 480K array. More information about the filtering and quality check procedure can be found in Jung et al. (2020).

Genotype imputation

For the elite and hybrids genotypes, the 20K SNPs were completed to reach the 480K genotyping array density by imputation with BEAGLE 4.0 software (Browning and Browning 2007), which can use a reference set of phased marker data along with pedigree information to improve the imputation quality. The reference panel proposed by Jung et al. (2020) was used in this study. Missing SNP marker data in the reference panel were first imputed and haplotypes were phased using the default parameters of BEAGLE 4.0. The phased marker data along with pedigree information inferred in Muranty et al. (2020) and updated in an ongoing apple pedigree project (Howard et al. 2018) were then provided to the software for the imputation from medium to high-density. Since high-density genotypes were not available for the elite material in the FBo-Hi and hybrids datasets, we could not directly assess the imputation accuracy by comparing imputed and true marker data. Imputation accuracy was estimated by computing the mean of the markers Allelic R^2 (AR^2) values provided by Beagle 4.0 after imputation, where the AR^2 value of a marker is an approximation of the correlation between true and imputed genotypic values based on the posterior probabilities of imputation (Browning and Browning 2007). We considered that a marker was well imputed if its AR^2 value was larger than 0.7.

Phenotypic data

FBo-Hi dataset

The genetic resources were phenotyped between 2012 and 2014 in six European research institutes. Several fruit quality and phenology traits were measured by assessor pairs according to the recommendations of the European Cooperative Programme for Plant Genetic Resources (Watkins and Smith 1982) and the notation for each trait was harmonized between institutes. When available, each institute also provided FBo-Hi notations for the measured traits (Supplementary Table S2).

The elite material was phenotyped between 2003 and 2005 in six European countries. Several traits linked to the productivity and fruit quality were measured. More details about the phenotyping procedure can be found in Muranty et al. (2015) and Kouassi et al. (2009).

Five traits were phenotyped in both populations: harvest date, fruit over-color, fruit juiciness, fruit acidity, and fruit crispness. As we were interested in combining phenotypic information from both populations, only these traits were used for the genomic predictions. When the traits in the two populations were evaluated using different ordinal scales, a correspondence table between the two scales was created in order to use phenotypic data that could be combined.

REFPOP dataset

Traits related to yield, fruit quality and phenology were measured between 2018 and 2020 using a common protocol in each location. The detailed protocol is described in Jung et al. (2021). Harvest date and fruit over-color were the only two traits that were measured in both the FBo-Hi and the REFPOP panels.

Dataset of hybrids

The hybrids were phenotyped in 2019 and 2020. Fruit weight, fruit number, fruit over-color, and harvest date were measured for each tree following the protocol proposed for the REFPOP dataset.

Data adjustment

The phenotypic data of the REFPOP dataset and the genetic resources of the FBo-Hi dataset were adjusted for year and site effects and those of the hybrids were adjusted for year effect, as they were evaluated in only one site. To do so, the estimated marginal means of the raw phenotypic data were computed using the emmeans function of the emmeans package (Lenth 2021). Prior to this adjustment, the phenotypic data of the REFPOP dataset were also corrected for spatial heterogeneity as in Jung et al. (2020) using the P-spline ANOVA approach with the PSANOVA function of the SpATS package (Rodríguez-Álvarez et al. 2018).

Heritability of genotypic means was estimated in both cases as $H^2 = \frac{\sigma^2_G}{\sigma^2_G + \frac{\sigma^2_E}{n}}$, where σ^2_G is the variance of genotype effects, σ^2_E is the residual variance and n is the mean number of observations per genotype.

Genomic heritability was estimated as $h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e}$, where σ^2_g and σ^2_e are the posterior means of the genomic and residual variances estimated by the MG-GBLUP model (see below). The standard deviation of these genomic heritability values was also computed from the posterior genomic and residual variances as in Lehermeier et al. (2015).In the case of the elite material of the FBo-Hi dataset, the best linear unbiased predictors (BLUPs) of clonal values computed in Kouassi et al. (2009) were used as phenotypic values.

Genetic characterization of the populations

Structure and genetic diversity across the populations

The structure of the populations was investigated through a principal component analysis (PCA) performed on the SNP marker data of the hybrids, the genetic resources, and the elite material. A PCA was thus performed for the FBo-Hi dataset and another one for the REFPOP dataset. To reduce the computational time of the analysis, the marker data were first pruned based on linkage disequilibrium using the indep-pairwise function in the PLINK 1.9 program (Purcell et al. 2007), by pruning the markers with a pairwise $r^2 > 0.1$ in a 50-kb window. In total, 12,363 SNP markers were retained after pruning from the FBo-Hi dataset and 12,290 markers from the REFPOP dataset. The PCA was carried out with the prcomp function from the R *stats* package. To further assess the differentiation between the elite material, the genetic resources, and the hybrids, pairwise F_{ST} values between the populations were computed using the pairwise. WCfst function from the R *hierfstat* package (Goudet 2005).

The allele frequencies along the genome of the elite material and genetic resources were compared using sliding windows for both datasets. For each chromosome, windows of 2 Mb with a shift of 400 kb were built and the average minor allele frequency of the SNPs included in the windows was computed for the genetic

resources. The frequency of the minor allele in the genetic resources was then computed in the elite material following the same procedure. When the number of SNPs of a window was less than 200, the mean minor allele frequency was set to missing.

Linkage disequilibrium

For each population of the FBo-Hi and REFPOP datasets, the linkage disequilibrium (r^2) was computed as the square correlation coefficient between pairs of markers within a 500 kb distance using the r^2 function of the PLINK 1.9 program. Marker pairs were placed into bins of 500 bp according to their pairwise distance and the LD decay was plotted as the mean r^2 of each bin.

Evaluated scenarios for genomic predictions

We evaluated the interest of population combination for two purposes: in the first case, we predicted a population (genetic resources or elite material) using either the same population (within-population prediction), the other population (across-population prediction) or a training set (TS for short) including genotypes from both populations (combination prediction). In the second case, we predicted the "genetic resources × elite" hybrids previously described with a training set combining genotypes from the genetic resources and elite material populations (Prop_hybrids scenario). In the latter case, we investigated the effect of different proportions of the two combined populations on the predictive ability given different TS sizes.

Genomic prediction models

Two different prediction models were evaluated in this study. For both models, the genomic estimated breeding values (GEBV) of the candidates were calculated using the following mixed model:

$$y = X\beta + Zu + e,$$

where y is a vector of phenotypic values, X is an incidence matrix relating fixed effects to the observations, β is the grand mean, Z is an incidence matrix linking the observations to the breeding values, u is a vector of breeding values for each individual and e is the vector of residuals.

Standard GBLUP model

We used the GBLUP model which is derived from the mixed model presented above with $u \sim N\left(0,\ G\sigma_g^2\right)$ and $e \sim N(0, \sigma_e^2)$ σ_g^2 and σ_e^2 are respectively the genetic and residual variances and G is the genomic relationship matrix (GRM, see below). The variance components and the breeding values were estimated with the kin.blup function of the R *rrBLUP* package (Endelman 2011).

Multigroup GBLUP model

A second model that takes the genomic correlation between the populations into account was used. In contrast to the standard GBLUP, this model allows the marker effects to be different (but correlated) between the two populations. Following Lehermeier *et al.* (2015), the model can be written as:

$$\begin{pmatrix} y_{GR} \\ y_{Elite} \end{pmatrix} = \begin{pmatrix} X_{GR} & 0 \\ 0 & X_{Elite} \end{pmatrix} \begin{pmatrix} \beta_{GR} \\ \beta_{Elite} \end{pmatrix} + \begin{pmatrix} Z_{GR} & 0 \\ 0 & Z_{Elite} \end{pmatrix} \begin{pmatrix} u_{GR} \\ u_{Elite} \end{pmatrix} + \begin{pmatrix} e_{GR} \\ e_{Elite} \end{pmatrix}$$

where:

$$\begin{pmatrix} u_{GR} \\ u_{Elite} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{g_{GR}} & \sigma_{g_{GR,Elite}} \\ \sigma_{g_{Elite,GR}} & \sigma^2_{g_{Elite}} \end{pmatrix} \otimes G \right).$$

Here, X_{GR} and X_{Elite} are incidence matrices relating fixed effects to the observations in each population, β_{GR} and β_{Elite} represent the mean of the two populations, Z_{GR} and Z_{Elite} are the incidence matrices relating the breeding values to the observations in each population, u_{GR} and u_{Elite} are the vectors of breeding values of the genetic resources and the elite material, σ_g^2GRand σ_g^2*Elite* are the genomic variances of the genetic resources and the elite material, $\sigma_{g_{GR,Elite}}$ is the genomic covariance between the two populations and e_{GR} and e_{Elite} are vectors of residuals for each population. The genomic correlation between the two populations, defined as the correlation of the marker effects of each population, was computed from the estimated parameters as:

$$r_{GR,\ Elite} = \frac{\sigma_{g_{GR,Elite}}}{\sqrt{\sigma_g^2 GR\ \sigma_g^2 Elite}}.$$

The genomic variances and covariances were estimated using a Gibbs sampler implemented in the *MTM* package. The Gibbs sampler was run for 20,000 iterations, the first 4000 being discarded as burn-in. One in every two samples was kept and the genetic parameters were then estimated by computing the posterior means of the remaining samples.

Genomic relationship matrix

For both models, the genomic relationship matrix (GRM) G was estimated as in VanRaden (2008):

$$G = \frac{ZZ'}{2\Sigma p_i(1-p_i)},$$

where $Z = M - 2p_i$ is the centered matrix of the marker data, M is the matrix of marker data and p_i is the minor allele frequency at locus i. Here, Z corresponds to the marker data of both genetic resources and elite material combined into one dataset, allowing the estimation of the relationships within and across populations. The GRM was calculated with the *A.mat* function of the rrBLUP package.

The GRM used in the models was computed according to the populations used in the training and VSs as presented in Table 1. In the case of within-population prediction, the marker data of the predicted population was used to compute the GRM, while the marker data of the two populations was used to compute the GRM for the across-population prediction and combination prediction cases. To evaluate the influence of the marker density on

Table 1 Composition of the training and validation set in the different scenarios

Method	Training set	Validation set
WP	Genetic resources	Genetic resources
	Elite	Elite
AP	Elite	Genetic resources
	Genetic resources	Elite
Comb	Genetic resources + Elite	Genetic resources
	Genetic resources + Elite	Elite
Prop_hybrids	Genetic resources + Elite	Hybrids

the predictive ability of the models, the GRM was computed using the high-density marker data and a subset corresponding to the SNP markers of the 20K genotyping array in each case.

Predictive ability

The predictive ability of the different models was evaluated as the Pearson correlation coefficient between the GEBV and the phenotypic values of the individuals in the VS. In a first step, the elite material and genetic resources were considered separately and a fivefold cross-validation scheme was used to randomly split the genotypes between training and VSs within each population, allowing within-population genomic predictions (from here on referred to as WP method). The same VSs were also predicted using all the genotypes in the complementary population (across-population prediction, from here on referred to as AP method) or by combining the genotypes of the two populations into a unique training set (combined-populations prediction, Comb and MG-Comb methods, see below). In this case, both the standard GBLUP (Comb method) and multi-group GBLUP (MG-Comb method) models described above were used. This procedure was replicated 20 times for both the FBo-Hi and REFPOP datasets. The approach is summarized in Figure 1A and the different TS and VS compositions are shown in Table 1. For the Comb method, we also studied the effect of the TS size by combining an increasingly large number of genotypes from the same population as the candidates with all the genotypes from the complementary population. This approach is described in more detail in Supplementary Materials S1.

Proportion of GR/elite in the training set

When predicting the hybrids, the influence of the elite/GR proportion in the TS was also investigated (Prop_hybrids scenario in Table 1). To do so, the total size of the TS was first fixed and the TS was then built with increasing proportions of elite individuals, from 0% to 100% by steps of 20%, and correspondingly decreasing proportions of genetic resources individuals (Figure 1B). From here on, the tested proportions will be noted as "$prop_{elite}$/$prop_{GR}$ proportion," where $prop_{elite}$ and $prop_{GR}$ are values between 0 and 1 that refer to the tested proportions (for example the 0.2/0.8 proportion means that the TS contains 20% of elite genotypes and 80% of genotypes from the genetic resources). For all the proportions, different TS sizes were evaluated: the TS size ranged from 50 to 500 by steps of 50 individuals when the genotypes in the TS were chosen from the FBo-Hi dataset and from 50 to 250 by steps of 50 individuals when using the REFPOP dataset. In this last case, the maximum TS size was smaller than for the FBo-Hi datasets because all the proportions could not be tested otherwise given the smaller number of genotypes in each population of the REFPOP. The individuals in the TS were chosen with 3 different methods, as described below. These methods were applied separately to the elite and GR populations by sampling n_1 genotypes from the elite population and n_2 genotypes from the genetic resources, where n_1 and n_2 depend on the studied proportion and $n_1 + n_2 = n$ is a predefined TS size as described above. The chosen individuals in the two populations were then combined into a unique TS that was used for the prediction of the hybrids The GEBV of the hybrids were predicted with the standard GBLUP

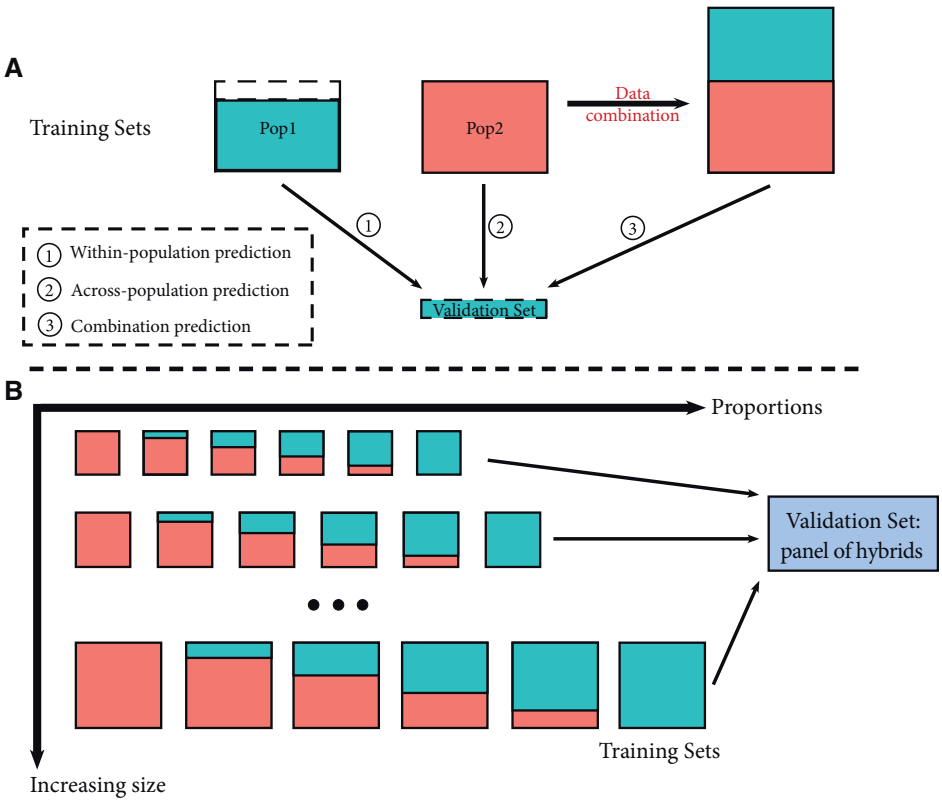


Figure 1 Schematic representation of the tested scenarios in this study. (A) The VS is composed of 20% of one of the two populations (elite material or genetic resources) and is predicted either by the remaining 80% of the population (within-population prediction), by the other population (across-population prediction) or by combining both populations (combination prediction). (B) The panel of hybrids is used as VS and is predicted by combining elite material and genetic resources into a unique population. Several training sets with increasing sizes and different proportions of the two populations are tested.

model described above. In order to compare the predictive abilities obtained with defined training set sizes or with larger training set sizes, the hybrids were also predicted using all the genetic resources and elite material genotypes in the FBo-Hi dataset or the REFPOP dataset.

Method 1: choice of the TS based on the CDpop criterion

The CDpop criterion proposed by Rincent *et al.* (2017) was used to choose the individuals for the TS. CDpop is derived from the coefficient of determination or CD (Rincent *et al.* 2012), which is a measure of the expected reliability of the predictions. The optimization based on CDpop aims to maximize the mean CD of the contrast matrix of the VS to predict, as defined in Rincent *et al.* (2017). For the two populations, an initial number n_1 (respectively n_2) of genotypes was first sampled. An exchange algorithm was then used to replace an individual within each group by an individual from the same population that had not been included yet, and this latter was kept in the group if the mean CD of the group increased after its inclusion. With our data, we found that the mean CD reached a plateau after 1000 to 5000 iterations. Since the choice of the TS based on CDpop is based on the initial sampling, the final TS composition can vary. The method was thus repeated 20 times with 3000 iterations.

Method 2: choice of the TS based on relatedness

For each population, the genotypes were chosen based on their relationship with the VS. To do so, we calculated the mean relationship coefficient between an individual and the candidates from the GRM and chose the genotypes with the highest mean relationship to be part of the TS.

Method 3: choice of the TS by random sampling

In the "Random_strat" method, the TS was constituted by combining n_1 genotypes sampled from the elite population and n_2 genotypes sampled from the genetic resources. The "Random" method is similar except that n genotypes were sampled to constitute the TS, regardless of the population to which they belonged. The two methods were applied 100 times.

Results

Imputation accuracy

We used the AR^2 values provided by Beagle 4.0 as a proxy of imputation accuracy. The distribution of the AR^2 values for the imputed datasets is shown in Supplementary Figure S1. Respectively 81% and 79% of the markers had an AR^2 value higher than 0.7 in the elite material of the FBo-Hi dataset and hybrids dataset. In the case of the REFPOP dataset, Jung *et al.* (2020) reported an imputation accuracy of 0.95 using Beagle 4.0 and the same reference panel as in this study.

Characterization of the populations

In order to investigate the potential of combining genetic resources and elite material into a training set, we first investigated the genetic and phenotypic differences between the two populations. Since the results were very similar between the two datasets, we only present the results obtained with the FBo-Hi dataset. See Supplementary Data for the same analyses with the REFPOP dataset.

Figure 2 shows the first two principal components from the PCA obtained using the pruned SNP marker data. A clear distinction between the two populations can be observed, although some genetic resources (for example Golden Delicious, Jonathan or Cox's Orange Pippin) appear to be closer from the elite material than from the other genetic resources. As expected, the hybrids were plotted between the two populations. The F_{ST} value computed for the marker data indicated a low differentiation between the genetic resources and elite material (F_{ST} = 0.023, see Supplementary Table S3). The linkage disequilibrium decay was rapid in both genetic resources and elite material (Supplementary Figure S2), the decay being faster in the genetic resources with an average r^2 calculated at 1, 5, 100, and 250 kb of 0.24, 0.2, 0.18, and 0.15 in the genetic resources and 0.29, 0.26, 0.22, and 0.2 in the elite material in the FBo-Hi dataset. Very similar values were observed in the REFPOP dataset. In the two panels, allelic frequencies were similar between elite material and genetic resources for the largest part of the genome (Supplementary Figures S3 and S4), although some major differences could be observed for some genomic regions (for example in chromosomes 3 and 15).

Supplementary Figures S5–S7 show that the phenotypic distribution is different for some traits between the two populations. Fruits phenotyped in the elite material were juicier and crispier than in the genetic resources, and there were more fruits per tree in the elite material. The mean fruit over-color was also slightly higher in the elite material, although the same bimodal distribution could be observed in the two populations. The mean fruit acidity, fruit weight, and harvest date were similar for both populations, but a wider range in acidity score was measured in the elite material compared to genetic resources, while the range in harvest date was larger in the genetic resources. The fruit weight distribution was extremely similar in both populations. In both datasets, the heritability values were generally high (>0.7) for all the traits (Supplementary Figure S8 and Table S2).

The genomic correlations between the two populations were moderate to high, ranging from 0.42 to 1 (Table 2). The lowest genomic correlations were measured for fruit weight, fruit crispness, and fruit juiciness. The correlation was high for fruit over-color and acidity (around 0.7 in both cases) and was highest for harvest date (in both datasets) and fruit number, with a correlation of almost 1.

Predictive abilities when combining populations

Table 3 presents the predictive abilities for within and across-population predictions, as well as for predictions obtained when combining elite material and genetic resources into a unique training set for the FBo-Hi and REFPOP datasets.

For the within-populations predictions, the predictive abilities varied between 0.33 and 0.82 for the elite material of the FBo-Hi dataset (respectively between 0.46 and 0.79 for the REFPOP dataset), and between 0.34 and 0.83 for the genetic resources of the FBo-Hi dataset (respectively between 0.5 and 0.79 for the REFPOP dataset). In both datasets, predictive abilities were high for harvest date (between 0.79 and 0.83 for FBo-Hi and between 0.6 and 0.79 for REFPOP, see Figure 3A) and fruit over-color (between 0.59 and 0.72 for FBo-Hi and between 0.64 and 0.79 for REFPOP, see Figure 3B), while they were moderately high for the other traits (0.33 to 0.58, Supplementary Figures S9–S13). The predictive abilities were higher when predicting harvest date (with a difference up to 0.19 and smaller standard errors in the REFPOP dataset), fruit weight (difference up to 0.06) and fruit number (difference up to 0.05) in the genetic resources than in the elite material, while fruit over-color (difference up to 0.14) and juiciness (difference up to 0.09) were better predicted when using elite material in the training and VS.

The across-population predictions always resulted in lower predictive abilities than the within-population predictions, with values ranging from 0.13 (juiciness) to 0.64 (harvest date) for the FBo-Hi dataset and from 0.19 (fruit weight) to 0.73 (fruit over-color) for the REFPOP dataset. The standard errors of predictive abilities were also generally higher than for within-population predictions. Across-

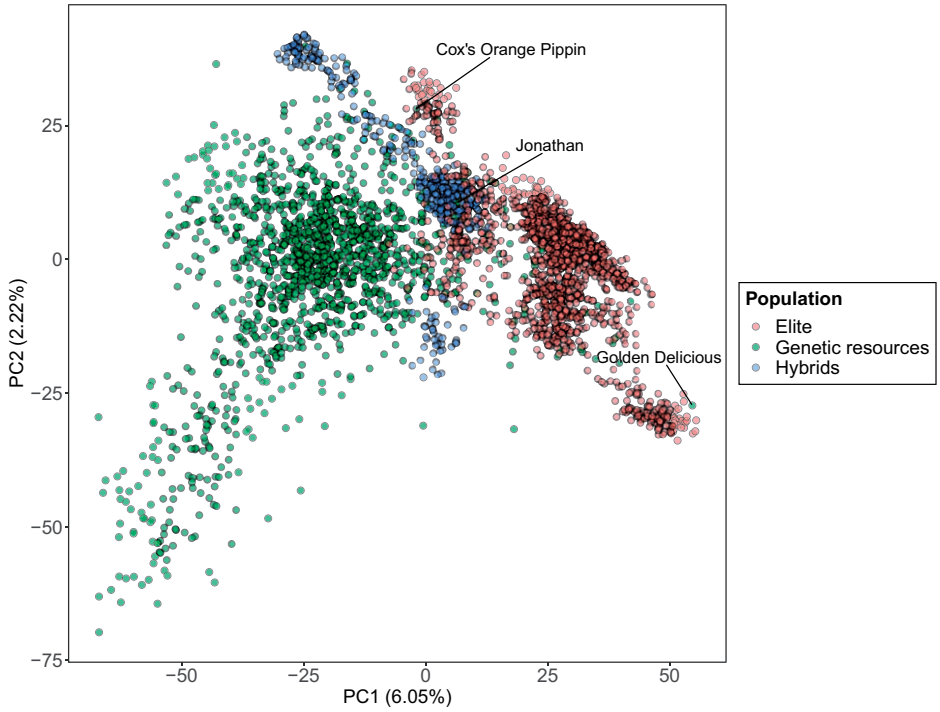


Figure 2 PCA performed using pruned marker data of the FBo-Hi and hybrids panel.

Table 2 Genomic and environmental variances and genomic correlations estimated from the Gibbs sampler for each trait in the two population of the FBo-Hi and REFPOP datasets

	Elite			Genetic Resources			
Trait	σ^2_g	σ^2_e	h^2	σ^2_g	σ^2_e	h^2	$r_{\text{Elite, GR}}$
FBo-Hi dataset							
Harvest date	194.2 (20.58)	85.81 (5.92)	0.69 (0.03)	314.5 (25.47)	108.91 (10.37)	0.74 (0.03)	0.99 (0.02)
Fruit over-color	0.6 (0.08)	0.53 (0.03)	0.52 (0.04)	0.8 (0.09)	0.43 (0.04)	0.65 (0.04)	0.65 (0.07)
Acidity	1.5 (0.24)	1.63 (0.1)	0.48 (0.05)	1.1 (0.15)	1.01 (0.08)	0.53 (0.05)	0.74 (0.07)
Juiciness	0.8 (0.15)	1.09 (0.07)	0.42 (0.05)	0.5 (0.07)	0.7 (0.05)	0.43 (0.04)	0.52 (0.12)
Crispness	0.4 (0.1)	1.16 (0.07)	0.27 (0.05)	0.5 (0.08)	0.85 (0.06)	0.37 (0.05)	0.45 (0.13)
REFPOP dataset							
Harvest date	245.8 (46.47)	52.34 (14.25)	0.82 (0.06)	341.2 (38.37)	36.33 (13.84)	0.9 (0.04)	1 (0)
Fruit over-color	1 (0.16)	0.34 (0.06)	0.75 (0.05)	1.3 (0.15)	0.25 (0.06)	0.83 (0.05)	0.72 (0.08)
Fruit number	69.6 (23.01)	112.77 (15.64)	0.38 (0.1)	70.6 (16.03)	47.21 (10.4)	0.59 (0.1)	0.98 (0.03)
Fruit weight	0.4 (0.09)	0.21 (0.04)	0.66 (0.08)	0.6 (0.1)	0.23 (0.05)	0.74 (0.06)	0.42 (0.15)

The standard deviation of the genomic and environmental variances, of the heritability values, and of the genomic correlations is shown between brackets.

population predictive abilities could be moderately high for some traits, such as harvest date (between 0.47 and 0.64 for the FBo-Hi dataset and between 0.41 and 0.47 for the REFPOP dataset) or fruit over-color (between 0.35 and 0.47 for the FBo-Hi dataset and between 0.52 and 0.73 for the REFPOP dataset). All the traits were better predicted when using genetic resources in the training set to predict elite material than when using elite material to predict genetic resources, with the exception of fruit number and crispness that were slightly better predicted in the latter case. The decrease in predictive ability between within and across-population prediction was also more important when genetic resources constituted the VS, especially for fruit over-color (for instance, with a decrease from 0.71 to 0.45 for the FBo-Hi dataset at high marker density), fruit weight (from 0.57 to 0.19) and juiciness (from 0.49 to 0.18).

Combining populations into a unique training set allowed to obtain predictive abilities that were slightly higher than with the corresponding within-population prediction in most cases (exceptions are some predictions for harvest date, fruit over-color, and juiciness, for which the combination led to the same mean predictive abilitiy). The highest increases in predictive ability were observed for the fruit over-color in the genetic resources in the REFPOP (from 0.69 to 0.73 with medium density). Allowing marker effects to differ between populations (MG-GBLUP model) rarely led to improvements of the predictive ability compared to a prediction using a standard GBLUP model, and when it did, the increase in predictive ability was limited, the highest gain being for juiciness when predicting genetic resources with an increase of 0.02 at both medium and high marker density.

Using a high marker density allowed higher predictive abilities compared to medium density only for harvest date and fruit weight, especially when predicting across populations (for instance, an increase from 0.60 to 0.64 for the FBo-Hi was observed for harvest date when predicting elite material with a training set of genetic resources). For some traits, measured predictive

Table 3 Predictive abilities of the measured traits in the within-population (WP), across-population (AP), combined populations (Comb), and MG-GBLUP method

		Elite material		Genetic resources	
Trait	**Method**	**Medium density**	**High density**	**Medium density**	**High density**
FBo-Hi dataset					
Harvest date	WPP	**0.79 (0.03)**	0.79 (0.03)	**0.82 (0.02)**	**0.83 (0.02)**
	APP	0.6 (0.05)	0.64 (0.04)	0.47 (0.05)	0.51 (0.05)
	Comb	0.78 (0.03)	0.79 (0.03)	**0.82 (0.02)**	**0.83 (0.02)**
	MG-GBLUP	**0.79 (0.03)**	**0.8 (0.03)**	**0.82 (0.02)**	**0.83 (0.02)**
Fruit over-color	WPP	**0.72 (0.03)**	0.71 (0.03)	0.59 (0.04)	0.57 (0.05)
	APP	0.47 (0.05)	0.45 (0.05)	0.39 (0.06)	0.35 (0.06)
	Comb	**0.72 (0.03)**	**0.72 (0.03)**	**0.61 (0.05)**	0.58 (0.05)
	MG-GBLUP	**0.72 (0.03)**	**0.72 (0.03)**	**0.61 (0.04)**	**0.59 (0.04)**
Crispness	WPP	0.34 (0.06)	0.33 (0.06)	0.34 (0.05)	0.34 (0.05)
	APP	0.15 (0.07)	0.17 (0.06)	0.21 (0.05)	0.22 (0.05)
	Comb	**0.36 (0.06)**	**0.36 (0.06)**	**0.35 (0.05)**	**0.36 (0.05)**
	MG-GBLUP	0.35 (0.06)	0.35 (0.06)	**0.35 (0.04)**	0.35 (0.05)
Juiciness	WPP	**0.49 (0.05)**	0.48 (0.05)	**0.4 (0.05)**	0.41 (0.05)
	APP	0.18 (0.07)	0.16 (0.07)	0.13 (0.07)	0.14 (0.07)
	Comb	**0.49 (0.05)**	**0.49 (0.05)**	0.38 (0.05)	0.4 (0.05)
	MG-GBLUP	**0.49 (0.05)**	**0.49 (0.05)**	**0.4 (0.05)**	**0.42 (0.05)**
Acidity	WPP	0.48 (0.05)	0.47 (0.05)	0.45 (0.05)	0.43 (0.05)
	APP	0.39 (0.05)	0.39 (0.05)	0.28 (0.06)	0.23 (0.06)
	Comb	**0.51 (0.04)**	**0.5 (0.04)**	**0.47 (0.06)**	**0.46 (0.06)**
	MG-GBLUP	**0.51 (0.04)**	0.49 (0.04)	**0.47 (0.05)**	0.45 (0.05)
REFPOP dataset					
Harvest date	WPP	**0.6 (0.09)**	0.6 (0.09)	0.79 (0.04)	0.78 (0.04)
	APP	0.44 (0.12)	0.45 (0.12)	0.41 (0.08)	0.47 (0.08)
	Comb	**0.6 (0.10)**	**0.63 (0.10)**	**0.8 (0.04)**	**0.8 (0.04)**
	MG-GBLUP	**0.6 (0.10)**	0.62 (0.10)	0.79 (0.04)	0.79 (0.04)
Fruit over-color	WPP	0.79 (0.05)	0.78 (0.05)	0.69 (0.06)	0.64 (0.06)
	APP	0.73 (0.05)	0.69 (0.06)	0.56 (0.09)	0.52 (0.09)
	Comb	0.81 (0.05)	0.80 (0.05)	**0.73 (0.05)**	**0.68 (0.06)**
	MG-GBLUP	**0.82 (0.05)**	**0.81 (0.05)**	**0.73 (0.05)**	**0.68 (0.06)**
Fruit weight	WPP	0.52 (0.09)	0.52 (0.09)	0.57 (0.08)	0.58 (0.08)
	APP	0.29 (0.11)	0.3 (0.11)	0.19 (0.10)	0.2 (0.10)
	Comb	**0.54 (0.10)**	**0.55 (0.09)**	**0.59 (0.08)**	**0.61 (0.07)**
	MG-GBLUP	**0.54 (0.09)**	0.54 (0.09)	**0.59 (0.08)**	0.6 (0.08)
Fruit number	WPP	0.48 (0.08)	0.46 (0.08)	0.5 (0.11)	0.51 (0.11)
	APP	0.27 (0.12)	0.24 (0.03)	0.27 (0.11)	0.29 (0.11)
	Comb	**0.49 (0.08)**	**0.47 (0.09)**	**0.52 (0.12)**	**0.54 (0.11)**
	MG-GBLUP	0.47 (0.08)	0.45 (0.09)	0.51 (0.11)	0.52 (0.11)

The standard deviation of the predictive abilities is shown between brackets. Bold values represent the highest predictive ability value for a given trait at a given marker density.

abilities were higher with medium marker density than with high marker density, as can be observed for over-color in both datasets and for fruit acidity. The decrease in predictive ability was also more important for across-population predictions for these two traits, especially when predicting genetic resources (decrease of 0.04 for fruit over-color in both datasets and of 0.05 for acidity). For the remaining traits, the influence of the marker density depended on the population used in the VS: predictive abilities for fruit number, juiciness, and crispness decreased when predicting elite material using a high marker density but increased in the genetic resources.

Proportion of the combined populations

When predicting the hybrids, we also investigated the influence of the proportion of elite material and genetic resources used in the training set. Four methods were compared to define the training set: the CDpop algorithm (CDstrat method), a choice based on kinship between the training set and VS (Kinship method), or a random choice of genotypes either within each population (Random_strat method) or regardless of the populations (Random method). The results for harvest date and fruit over-color are presented in Figure 4 and the predictive abilities for fruit weight and fruit number are presented in Supplementary Figures S14 and S15. The predictive ability values obtained with the different methods for all the traits are also included in the Supplementary File "Prop_predAbi".

We observed that the predictions based on the CDstrat method generally outperformed the Kinship, Random_strat and Random methods, with the exception of the prediction of the fruit number where the Random_strat method allowed to reach slightly higher predictive abilities. The standard errors of the predictive abilities were also lower when using the CDstrat method than when using the Random_strat or Random method. This observation holds true for all the tested training population sizes and at medium and high marker density in the case of harvest date predicted with the FBo-Hi dataset and for fruit weight (except for one tested proportion, see below). When predicting harvest date with the REFPOP dataset, the CDstrat and Kinship methods gave very similar results for all the tested proportions and training set sizes, but CDstrat allowed to reach higher predictive abilities for the 1/0 proportion, especially for small training set sizes (with a maximum predictive ability difference of 0.46 between the two methods for a TS of 50 genotypes). For fruit over-color, the best method depended on the tested proportion: with

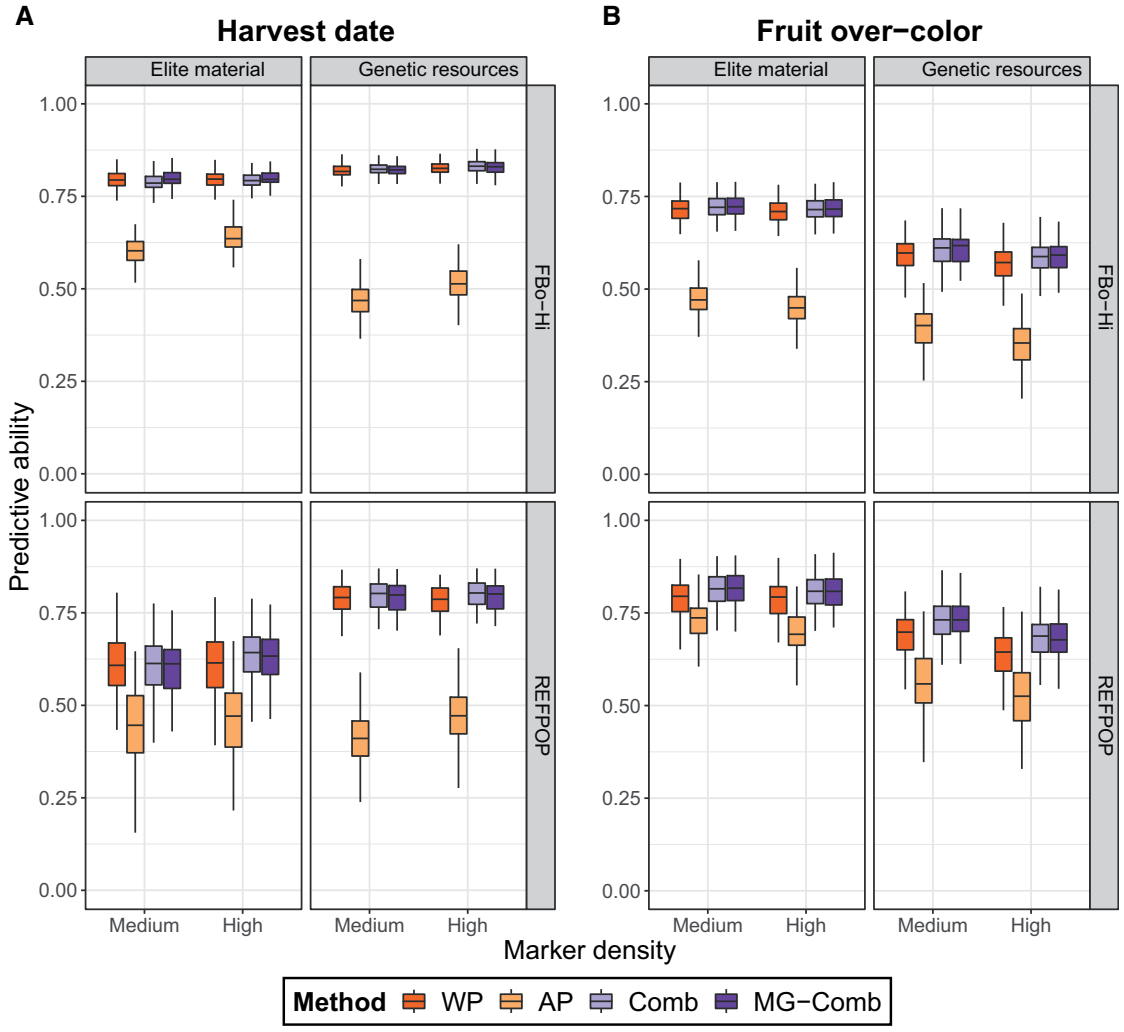


Figure 3 Predictive abilities for (A) harvest date and (B) fruit over-color in the FBo-Hi or REFPOP dataset with medium and high marker density. WP: within-population prediction AP: across-population prediction Comb: combination prediction MG-Comb: combination prediction with the MG-GBLUP method.

the FBo-Hi dataset, CDstrat outperformed Kinship except for the extreme proportions (*i.e.*, TS constituted of only one of the two populations), the highest increases in predictive ability being again measured for small training set sizes. With the REFPOP dataset, CDstrat outperformed Kinship only until a TS size of 100 genotypes, and the two methods gave similar results for larger TS sizes. For all the tested TS sizes, most genotypes included in the TS with the CDstrat method were selected at least 15 times out of the 20 repetitions (data not shown).

The predictive abilities based on the CDstrat method rapidly reached a plateau for harvest date. The maximum predictive ability of 0.85 could be reached with 100 genotypes from the FBo-Hi dataset at medium density, and the maximum predictive ability was reached with 50 genotypes from the REFPOP dataset for both densities. For fruit over-color, a bigger TS size resulted in higher predictive abilities with the FBo-Hi dataset for all but the 0/1 and 1/0 proportions. The increase in TS size also led to an increase in predictive ability only for some proportions for fruit over-color with the REFPOP dataset: increasing the TS from 50 to 250 genotypes allowed an increase in predictive ability for proportions 0.4/0.6 to 1/0 (with increase between 0.02 for the 0.4/0.6 proportion and 0.43 for the 0.8/0.2 proportion) but the predictive ability decreased for the 0/1 and 0.2/0.8 proportions. A similar pattern was observed for fruit weight: when increasing the TS size, predictive abilities also increased for proportions 0.6/0.4 to 1/0 but decreased for proportion 0.2/0.8 and 0/1 beyond respectively 150 and 100 genotypes in the TS. The predictive abilities for fruit number increased when the TS size increased for all the tested proportions, but the gains were smaller than for the other traits (maximum increase from 0.16 to 0.26 when increasing the TS size from 50 to 250 genotypes). In all the studied traits, the standard errors of the predictive abilities also decreased when the TS size increased.

With the Kinship method, the largest increases in predictive ability were measured for harvest date predicted with the FBo-Hi dataset, especially for the 0.8/0.2 and 1/0 proportions at medium marker density (respectively from −0.01 and −0.05 with a TS size of 50 genotypes to 0.84 and 0.72 with a TS size of 500 genotypes) but not in the REFPOP dataset: the predictions with 50 genotypes in the TS led to better predictive abilities than a TS with 250 genotypes, except for the 0.8/0.2 and 0/1 proportions. The increase in TS size also led to higher predictive abilities for all the tested proportions for fruit number, with larger gains than for the CDstrat method (with a maximum increase from 0.13 to 0.29 when increasing the TS size from 50 to 250 genotypes). The predictive ability also increased for fruit weight and fruit over-color for all

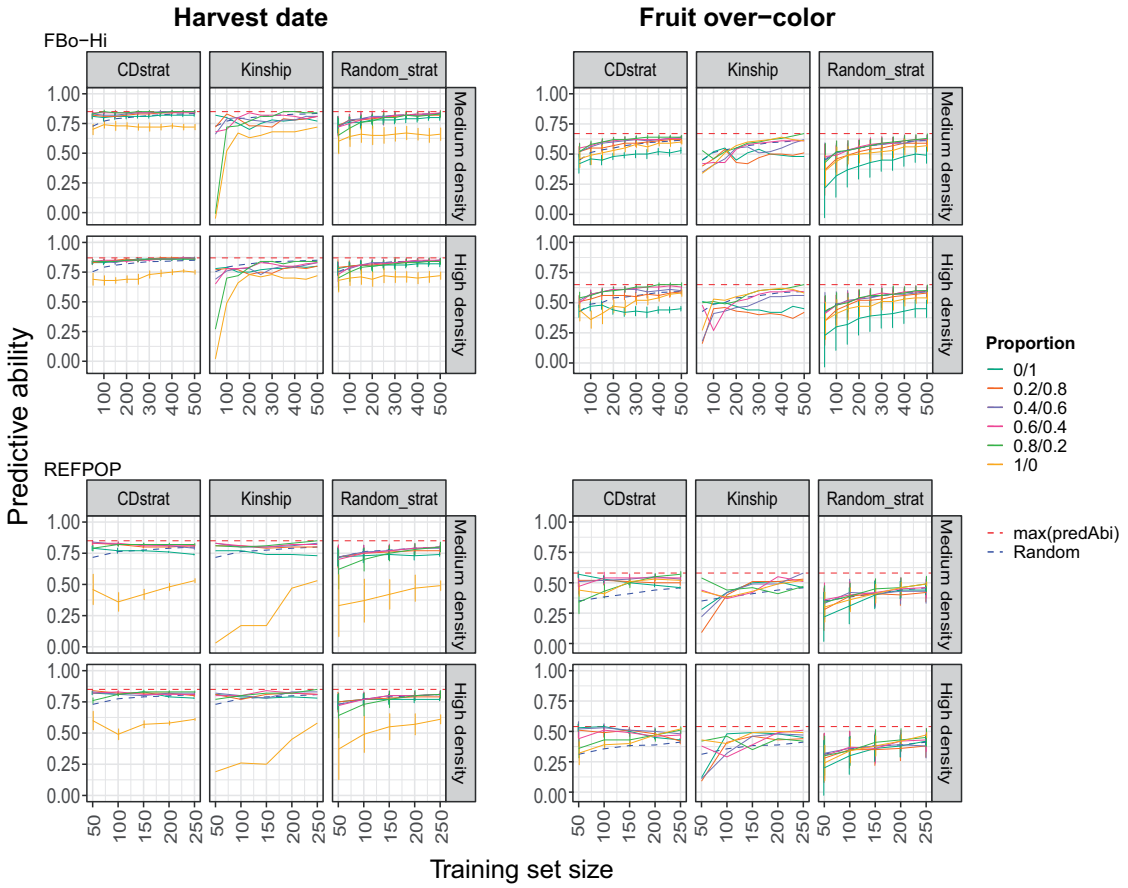


Figure 4 Predictive abilities for harvest date and fruit over-color in the dataset of hybrids with medium and high marker density when the training set is composed of varying proportions of elite material and genetic resources of the FbBo-Hi or REFPOP dataset. max(predAbi): maximum predictive ability obtained for a given marker density with one of the three tested methods regardless of the training set size Random: predictive ability obtained when randomly choosing the genotypes included in the training set.

the tested proportions, except for the 0/1 proportion for fruit weight and the 0.8/0.2 proportion for fruit over-color when predicted with the REFPOP dataset, the maximum predictive ability being observed with a TS size of 50 genotypes in both cases.

For each trait, the proportion allowing the highest predictive ability depended on the TS size and the method used to select the genotypes of the TS. For the three methods, the worst predictive abilities were obtained when only one population was present in the TS, with predictive abilities lower than the ones obtained from random sampling in the two populations: the 0/1 proportion led to the lowest predictive abilities for fruit weight, fruit number and fruit over-color when predicted with the FBo-Hi dataset while for harvest date, the predictions with the 1/0 proportion gave the lowest predictive abilities. Note that for fruit weight and fruit over-color, the 1/0 proportion also led to predictive abilities lower than the predictive abilities obtained with the Random method.

As was observed in the within and across-population predictions, the effect of the marker density on predictive ability depended on the trait: harvest date generally benefited from the higher marker density in the two datasets, especially when the TS size increased. Interestingly, predictive abilities at high marker density with the 1/0 proportion were lower than with medium marker density for small TS sizes when predicting with the FBo-Hi dataset, while the increase in predictive ability due to higher marker density was the highest for the 1/0 proportion when predicting with the REFPOP dataset, with increases ranging from 0.08 to 0.15 with the CDstrat method. Predictive abilities for fruit weight were higher at high density only for TS sizes up to 100 genotypes and medium density led to higher predictive abilities for larger TS sizes with the CDstrat method, except for the 0/1 proportion, in which case high density always gave better results (with increases in predictive ability ranging from 0.04 and 0.13). Predictive abilities measured from high marker density data for fruit over-color and fruit number were slightly lower than when using medium marker density for every method and every tested proportion, except for the 1/0 proportion for fruit over-color predicted with the REFPOP dataset with the Kinship method which allowed slightly higher predictive abilities until a TS size of 200 genotypes.

Discussion

One possible limitation to the implementation of genomic selection in fruit tree crops is the establishment of a training set large enough to allow accurate prediction of selection candidates because of the possible costs of maintaining and phenotyping trees. In this study, we first evaluated the potential of combining two different apple populations to increase the size of the training set by estimating the predictive ability in this case. We then investigated the effect of using subsets of the two populations to predict hybrids resulting from crosses between genetic resource and elite material. In each case, we compared the predictive abilities obtained using either medium or high SNP marker density.

Predictive abilities within and across populations

When predicting within-population, the measured predictive abilities were moderate to high, ranging from 0.33 for fruit crispness to 0.83 for harvest date. These values were overall in line with previous studies in apple that predicted the same traits. For example, harvest date was well predicted in other studies performed on apple (Migicovsky *et al.* 2016; McClure *et al.* 2018; Jung *et al.* 2020; Minamikawa *et al.* 2021) with predictive ability values between 0.6 and 0.8, while prediction accuracies for fruit weight and fruit quality traits were generally low to moderate (Kumar *et al.* 2015; Migicovsky *et al.* 2016; McClure *et al.* 2018; Minamikawa *et al.* 2021), with predictive abilities not exceeding 0.5. An exception was the study from Kumar *et al.* (2012b) that reported predictive ability values over 0.7 for traits related to fruit quality, in a case where genotypes from seven full-sib families derived from a 4 × 2 factorial design were randomly used for cross-validation, leading to a higher relatedness between training and population set than in our study.

Across-population predictions always resulted in predictive abilities lower than within-population predictions. This is an expected result as linkage disequilibrium, allele effects, and allele frequencies are generally different between populations. In our case, the LD decay was indeed more rapid in the genetic resources than in the elite material. The allele frequencies were similar along the 17 chromosomes but differed for some genomic regions, which could be due to selection, thus leading to different allele effects estimates as discussed in further. For some traits like fruit crispness or juiciness, the different phenotypic distribution in the two populations could also explain the poor predictive abilities in across-population prediction (Supplementary Figures S4–S6). Roth *et al.* (2020) reported a similar result when predicting apple fruit texture of full-sib families using a germplasm collection. The high genomic correlations between the two populations could have suggested that across-population predictions could perform as well as within-population prediction, which was not the case in our study. Similar results were observed by Lyra *et al.* (2018) and Rio *et al.* (2019), who hypothesized that this apparently contradictory observation could be explained by population-specific allele frequencies at the causal QTL. Such differences are expected in our case because of selection (see below). The moderate to high predictive abilities observed for some traits like harvest date and fruit over-color and the high genomic correlations between the two populations nevertheless suggest that marker effects are conserved to a certain extent between genetic resources and elite material. In the case of the FBo-Hi dataset, the lower predictive abilities observed in the across-population predictions could also be due to G×E interactions since the trees of the two populations were phenotyped by different assessors in different sites and years. Combining genotypes that were phenotyped in environments that differ from the environments of the VS is not always detrimental for genomic predictions (Jarquin *et al.* 2016) but when updating the training set, such genotypes should be discarded and replaced by genotypes from the same environment when it becomes possible to do so.

Potential of combining populations

Combining datasets can help increase predictive abilities in genomic prediction by allowing larger training populations, which is a key parameter for accurate predictions. However, the marker effects may be different between the combined populations, in which case the addition of new genotypes to the training set is not expected to improve predictive abilities (Lund *et al.* 2016). For instance, several studies in livestock found no increase in predictive ability when combining breeds (Hayes *et al.* 2009; Erbe *et al.* 2012) and adding genetically distinct individuals to the training population can even result in lower predictive abilities (Lorenz and Smith 2015). Multi-population genomic prediction models combining populations that have diverged for a few generations are then expected to lead to higher predictive abilities (de Roos *et al.* 2009; Wientjes *et al.* 2016) and the choice of the populations to combine should be based on knowledge about the genetic distance between the populations.

Several studies showed evidence of selection during apple domestication, *e.g.*, for fruit size (Yao *et al.* 2015; Duan *et al.* 2017), fruit quality traits (Wedger *et al.* 2021), metabolite content (Khan *et al.* 2014) or disease-related genes (Singh *et al.* 2019). However, little information is available about the genomic consequences of the more recent apple improvement and the resulting genetic differences between genetic resources and modern cultivars. Supplementary Figure S4 shows that the genotypes from the elite material are less acid, juicier and crispier than genetic resources, which can be a consequence of breeders' choices since these fruit quality traits have been targeted for decades in apple breeding programs (Janick and Moore 1975; Liao *et al.* 2021). On average, elite material is also harvested later than genetic resources, most probably because fruits with a later ripening are more firm (Migicovsky *et al.* 2016) and can be stored longer (Nybom *et al.* 2008). Modern varieties also have a lower phenolic content than genetic resources in general (Ceci *et al.* 2021; Watts *et al.* 2021). A recent pedigree reconstruction study showed that the two populations are separated by approximately 5–10 generations (Muranty *et al.* 2020), which could result in the persistence of marker phase and effects across populations.

In this study, combining elite material and genetic resources generally led to small increases in predictive ability compared to within-population predictions. For some traits, the combination allowed to obtain slightly higher predictive abilities in the FBo-Hi dataset, in spite of the genotypes from the two populations being evaluated at different sites and years. This suggests that the potential adverse effect of GxE interactions on predictive ability was counterbalanced by the large increase in population size. The addition of genotypes from the complementary population did not decrease the predictive abilities when combining data, but we can hypothesize that selecting subsets of the two populations to combine closer genotypes into a training set could be advantageous. The approaches proposed in the Prop_hybrids scenario could thus be used for the Comb scenario. Supplementary Figures S16–20 show that the small increase in predictive ability when combining populations could be due to the initial large population sizes for some of the studied traits, in which case the marker effects would already be accurately estimated. Combining populations may thus be of particular interest when one of the two populations is small and the complementary population is large. To further increase the size of the training set in such a case, historical data from trees that only have pedigree information could also be used in a single-step GBLUP model that uses both genotyped and nongenotyped individuals at the same time (Sood *et al.* 2020).

Toward improvements of the predictive abilities

For all the studied traits, the MG-GBLUP model, which allows different but correlated effects between populations, did not perform better than the standard GBLUP model. This result can be explained in the light of the genomic correlations between the two populations: when the genomic correlation between the

combined populations is equal to one, the MG-GBLUP is equivalent to the standard GBLUP. Therefore, the MG-GBLUP and standard GBLUP were expected to yield similar results since the genomic correlation between the genetic resources and elite material was high for most traits. The GEBV obtained from the GBLUP and MG-GBLUP model were almost the same even when the correlation between elite material and genetic resources was moderate (Supplementary Figures S21 and S22). Lehermeier *et al.* (2015) point out that MG-GBLUP is expected to perform better than GBLUP when population used for within-population prediction is small because borrowing information from the combined population could improve marker effect estimation in this case. In this study, the populations used for genomic prediction were large, which could explain why the GBLUP and MG-GBLUP performed similarly even when the genomic correlation between the combined populations was moderate. In addition, the MG-GBLUP shows limitations when dealing with genotypes that cannot clearly be assigned to a given population (Lehermeier *et al.* 2015), which is sometimes the case with the elite material and genetic resources (Figure 2 and Supplementary Figure S23). In this case, genomic prediction models that account for admixture could be better adapted (Rio *et al.* 2020).

Ibánez-Escriche *et al.* (2009) suggested that models that fit population-specific marker effects may not be necessary at high marker density because a density high enough could allow the marker-QTL association to be the same in the combined populations. In this study, we observed limited increases in predictive abilities when using high-density marker data. Other studies also reported that predictive abilities could reach a plateau after a given number of markers (Hickey *et al.* 2014; Jung *et al.* 2020) and that high marker density was not always needed to obtain high predictive abilities. Considering the rapid linkage disequilibrium decay in the genetic resources, we could have expected the predictive abilities to be lower at medium than at high density when the marker effects are estimated from such a population. However, the increases in predictive ability when using high-density marker data were generally small for all the datasets. This could imply that even at medium density, some markers are in high linkage disequilibrium with the most important causal QTLs and that increasing marker density does not better capture their variation in this case. In the REFPOP dataset, Jung *et al.* found that predictive ability was the same for harvest date and floral emergence at medium or high density and that as few as 5000 markers could be sufficient to reach a similar predictive ability. In the case of the elite material, imputation errors could also explain the absence of predictive ability improvement when using high-density marker data. However, we obtained very similar predictive abilities for all the traits when we removed markers with an AR^2 value under 0.7 (data not shown).

For some traits, the predictive abilities were even lower with high than with medium marker density. Two of these traits, namely fruit over-color and acidity, are controlled by an oligogenic determinism with a few known major genes (Chagné *et al.* 2016; Verma *et al.* 2019). As the GBLUP and MG-GBLUP models make the hypothesis that all the marker effects are drawn from the same normal distribution and do not allow marker effects to be null, a high marker density could overshrink the marker effects and poorly capture the effect of the large QTLs, as suggested by Daetwyler *et al.* (2010). Similar to our case, Erbe *et al.*(2012) found that predictive abilities were lower with a GBLUP model and 800K markers than with the same model with 50K markers and suggested that the number of effects to estimate was too important compared to the number of phenotypic records. In their case, models that allow marker effects to be set to zero or models resulting in a selection of a subset of markers led to higher predictive abilities. Bayesian models could thus result in higher predictive abilities for traits with an oligogenic genetic architecture (Hayes *et al.* 2009). If major QTLs are known, they could also be considered as fixed effects (Bernardo, 2014; Sarinelli *et al.* 2019) or be weighted accordingly in the GRM (Tiezzi and Maltecca 2015; Raymond *et al.* 2018). Since genomic predictions are generally used for traits with a polygenic determinism, high marker density should give better results than lower marker densities in general. Moreover, DoVale *et al.* (2021) showed that, in outbred crops, genomic prediction models should be updated more regularly when high marker densities are not used. Using high marker density for genomic selection could thus still prove useful when combining populations, especially if a high-quality imputation step can be implemented to reduce the costs of genotyping (Song *et al.* 2019).

Optimization of the training set composition

When predicting the hybrids, we never observed a particular proportion of the two combined populations that would allow higher predictive abilities. However, using only one of the two populations was detrimental for each studied trait, regardless of the training set size or the method used to choose the genotypes. This point highlights the need to use at least some genotypes from both populations in the training set in order to achieve high enough predictive abilities. This observation can have at least two explanations. First, we observed in our data that the phenotypic variation in the hybrids is larger than the variation of the two populations taken separately. When using only one of the two populations in the training set, the range of phenotypic values of the chosen population can by consequence not be large enough to accurately estimate the marker effects. Second, if some alleles linked to the desired trait are in low frequency in one population and segregate in the other population, using only the population with the low-frequency alleles as the training set will lead to incorrect marker effect estimations. For example, Migicovsky *et al.* (2021) showed that the *NAC18.1* gene marker, which is linked to harvest date and fruit firmness, is homozygous for the favorable allele in the nine most marketed varieties in the United States, and several marker alleles detected by GWAS in the REFPOP panel are fixed in the elite population whereas they segregate in the genetic resources (Jung *et al.*, 2021). As expected in this case, the prediction for harvest date in the hybrids when using only the elite material in the training set led to predictive abilities lower than when using both populations. Another such example would be the *Rvi6* gene that confers resistance to apple scab and all QTLs in the associated introgressed segment: the favorable allele at *Rvi6* segregates in elite material (Jung *et al.* 2021) but is absent in old varieties, because the gene introgression from the wild relative *M. floribunda* is recent (Gessler and Pertot 2012).

When the training set size is a limitation for the implementation of genomic selection, methods to optimize the choice of the genotypes in the TS when genomic data are already available can be advantageous. In this study, we observed that for three out of four traits the hybrids were better predicted when the TS was chosen based on the CDpop criterion compared to a choice based on kinship or genotypes sampled randomly, especially for small training set sizes. The CD algorithm has been implemented in order to better sample the genetic diversity than algorithms based on kinship alone (Rincent *et al.* 2012), which could explain the better performance of the CDstrat method, since it is probably

necessary to use a training set with a large diversity to predict the hybrids, as discussed above.

Note that we applied the CD algorithm to each population separately in order to simultaneously study the effect of the proportion of the two populations on predictive abilities. One way to further optimize the composition of the training set could be to use the CDpop criterion with genotypes of the elite material and genetic resources in a single step, letting the algorithm choose the proportion of the two populations to be used. However, using the optimization procedure based on the CDpop criterion is computationally demanding for large populations, as the computation of the criterion requires to calculate the inverse of the GRM of the genotypes in the training and VSs (Rincent *et al.* 2017), which could not be achieved in our case when considering both populations at the same time. It would also be interesting to evaluate the effect of the optimization methods proposed in this study in a case where the hybrids would be predicted by a training set also composed of hybrids. This situation should lead to the highest predictive abilities, and while the combination of populations proposed in this study is a suitable approach to initiate the prediction of hybrids, a training set composed of hybrids only should be envisioned as soon as enough phenotypic and genotypic data for the hybrids are available. Such a training set can be built gradually, by replacing a part of the genotypes of the two populations by newly phenotyped and genotyped hybrids (Fritsche-Neto *et al.* 2021).

Conclusion

We showed in this study that combining genetic resources and elite material into a unique training set could be beneficial for genomic predictions. First, larger training populations can be obtained with this approach, leading to slightly higher predictive abilities in return. Second, using both populations in the training set appeared necessary to predict "genetic resources × elite" hybrids. Combining populations could thus be an effective way to initiate pre-breeding programs that incorporate genomic prediction when a large training population is too costly. Genotypic and phenotypic data from ongoing breeding programs can be used to generate the training set, or historical data can be used to further reduce the costs of phenotyping. The training set composition can be further optimized to reduce the number of genotypes to include, which can be of particular interest for traits that are hard or long to phenotype, like biennial fruit bearing or abiotic stress tolerance. The proposed approach can be used in other fruit tree species provided that the genetic differences of the combined populations are taken into account when necessary.

Data availability

All SNP genotypic data generated with the 480K array used in this study have been deposited in the INRAe dataset archive (https://data.inrae.fr/) at https://doi.org/10.15454/IOPGYF. The SNP genotypic data of the REFPOP elite material generated with the 20K array and used in this study have been deposited in the INRAe dataset archive at https://doi.org/10.15454/1ERHGX. The SNP genotypic data of the elite material of the FBo-Hi panel generated using the 20K array and used in this study have been deposited at https://doi.org/10.15454/PMVCFI. The SNP genotypic data of the hybrids generated using the 20K array in this study have been deposited at https://doi.org/10.15454/SB2JSB. The raw phenotypic data of the REFPOP panel have been deposited at https://doi.org/10.15454/VARJYJ. The raw phenotypic data of the genetic resources of the FBo-Hi dataset have been deposited at https://doi.org/10.15454/KNECMS, the BLUPs of clonal values of the elite material of the FBo-Hi dataset have been deposited at https://doi.org/10.15454/UJPCOV, the raw phenotypic data of the hybrids have been deposited at https://doi.org/10.15454/H6PXRX. The scripts used to obtain the results presented in this manuscript are available at https://sourcesup.renater.fr/projects/apple-gensel.

Supplementary material is available at G3 online.

Acknowledgments

The authors would like to thank the past and present staff and field technicians of UE HORTI, Horticulture Experimental Facility (https://doi.org/10.15454/1.5573931618268674E12), INRAE, Angers-Beaucouzé, for maintaining the apple collections used in the FBo-Hi panel and the "genetic resources × elite" hybrids, as well as the past and present curational staff and field technicians from CRA-W, RBIPH, AUB-UNIBO, NFC-Brogdale (UK), and SLU, and acknowledge Defra (UK) for supporting the characterization of the collections. They also thank the Biological Resource Center "Pome Fruits and Roses" (https://www6.angers-nantes.inrae.fr/irhs/Ressources-mutualisees/Ressources-genetiques/CRB-Fruits-a-pepins-et-rosier) and associated staff for maintaining the plant material and associated datasets used in the present study. We thank Sandra Gabard for her help with the phenotyping of the hybrids in Angers and the collaborators of the REFPOP network for evaluating trees and reporting phenotypic data of the REFPOP panel. The authors in particular thank Michaela Jung for collecting and cleaning the raw data from the different partners of the project. We would like to thank the INRAE EPGV Unit, Evry, France for the genotyping of the panel of hybrids with the Illumina Infinium 20K SNP genotyping array. We acknowledge CNRGH Illumina Platform and DNA bank Teams for technical assistance.

C-E.D. and H.M. conceived and coordinated the study. Selection of germplasm and acquiring phenotypic data were performed by B.P., M.L., H.N., J.S., S.T., and F.L. X.C. performed the imputation of genomic data and carried out the statistical analyses under the supervision of C-ED and H.M. X.C. wrote the manuscript with decisive contributions of C-E.D. and H.M. All authors read and approved the final manuscript.

Funding

This study received financial support from the INRAE metaprogram SelGen and more specifically from the GdivSelgen project. This research was conducted in the framework of the regional programme "Objectif Végétal, Research, Education and Innovation in Pays de la Loire," supported by the French Region Pays de la Loire, Angers Loire Métropole and the European Regional Development Fund. The genotyping and part of the phenotyping of the FBo-Hi panel was funded with the financial support from the Commission of the European Communities (contract No. QLK5-CT-2002-01492), Directorate—General Research—Quality of Life and Management of Living Resources Programme and the EU seventh Framework Programme by the FruitBreedomics project No. 265582: Integrated Approach for increasing breeding efficiency in fruit tree crops (http://www.fruitbreedomics.com/). Phenotypic data collection for the REFPOP panel was partially supported by the Horizon 2020

Framework Program of the European Union under grant agreement No. 817970 (project INVITE: "Innovations in plant variety testing in Europe to foster the introduction of new varieties better adapted to varying biotic and abiotic conditions and to more sustainable crop management practices").

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Akdemir D, Sanchez JI, Jannink J-L. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genet Sel Evol. 47:10.1186/s12711-015-0116-6

Bernardo R. 2014. Genomewide Selection when Major Genes Are Known. Crop Sci. 54:68–75. doi: 10.2135/cropsci2013.05.0315.

Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S, Di Guardo M, Salvi S, Jansen J, Viola R, Gut I, *et al.* 2014. Development and Validation of a 20K Single Nucleotide Polymorphism (SNP) Whole Genome Genotyping Array for Apple (Malus × domestica Borkh). PLoS ONE. 9:e110377 10.1371/journal.pone.0110377

Bianco L, Cestaro A, Linsmith G, Muranty H, Denancé C, *et al.* 2016. Development and validation of the Axiom ® Apple480K SNP genotyping array. Plant J. 86:62–74. doi: 10.1111/tpj.13145.

Brandariz SP, Bernardo R. 2019. Small ad hoc versus large general training populations for genomewide selection in maize biparental crosses. Theor Appl Genet. 132:347–353. 10.1007/s00122-018-3222-3

Browning SR, Browning BL. 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics. 81:1084–1097. 10.1086/521987

Ceci AT, Bassi M, Guerra W, Oberhuber M, Robatscher P, *et al.* 2021. Metabolomic characterization of commercial, old, and Red-Fleshed apple varieties. Metabolites. 11:378.doi:10.3390/metabo11060378.

Chagné D, Kirk C, How N, Whitworth C, Fontic C, *et al.* 2016. A functional genetic marker for apple red skin coloration across different environments. Tree Genetics & Genomes. 12:Doi:10.1007/s11295-016-1025-8.

Clark SA, Hickey JM, Daetwyler HD, van der Werf JHJ. 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.

Crossa J, Jarquín D, Franco J, Pérez-Rodríguez P, Burgueño J, *et al.* 2016. Genomic prediction of gene bank wheat landraces. G3 (Bethesda). 6:1819–1834. doi:10.1534/g3.116.029637.

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 185:1021–1031. doi:10.1534/genetics.110.116855.

DoVale JC, Carvalho HF, Sabadin F, Fritsche-Neto R. 2021. Reduction of genotyping marker density for genomic selection is not an affordable approach to long-term breeding in cross-pollinated crops. biorXiv.doi: 10.1101/2021.03.05.434084

Duan N, Bai Y, Sun H, Wang N, Ma Y, *et al.* 2017. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. Nat Commun. 8:249. doi: 10.1038/s41467-017-00336-7.

Edwards SM, Buntjer JB, Jackson R, Bentley AR, Lage J, *et al.* 2019. The effects of training population design on genomic prediction accuracy in wheat. Theor Appl Genet. 132:1943–1952. doi: 10.1007/s00122-019-03327-y.

Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 4:250–255. doi: 10.3835/plantgenome2011.08.0024.

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, *et al.* 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 95:4114–4129. doi: 10.3168/jds.2011–5019.

Fritsche-Neto R, Galli G, Borges KLR, Costa-Neto G, Alves FC, *et al.* 2021. Optimizing genomic-enabled prediction in small-scale maize hybrid breeding programs: a roadmap review. Front Plant Sci. 12:658267.doi:10.3389/fpls.2021.658267.

Gessler C, Pertot I. 2012. Vf scab resistance of Malus. Trees. 26: 95–108. doi:10.1007/s00468-011-0618-y.

Goudet J. 2005. hierfstat, a package for r to compute and test hierarchical F-statistics. Mol Ecol Notes. 5:184–186. doi: 10.1111/j.1471-8286.2004.00828.x.

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel. Evol. 41:51.

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, *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.

Howard NP, Albach DC, Lubby JJ. 2018. The identification of apple pedigree information on a large diverse set of apple germplasm and its application in apple breeding using new genetic tools. In: Foerdergemeinschaft Oekologischer Obstbau e. V. (FOEKO). Hohenheim: Germany.

Ibánez-Escriche N, Fernando RL, Toosi A, Dekkers JCM. 2009. Genomic selection of purebreds for crossbred performance. Genet Sel Evol. 41:12.doi:10.1186/1297–9686-41-12.

Ibánēz-Escriche N, Fernando RL, Toosi A, Dekkers JC.Genomic selection of purebreds for crossbred performance. Genet Sel Evol. 2009;41(1):10.1186/1297-9686-41-12

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

Janick J, Moore JN, editors. 1975. Advances in Fruit Breeding. West Lafayette, IN: Purdue University Press.

Jarquin D, Specht J, Lorenz A. 2016. Prospects of genomic prediction in the USDA soybean germplasm collection: historical data creates robust models for enhancing selection of accessions. G3 (Bethesda). 6:2329–2341. doi:10.1534/g3.116.031443.

Jung M, Keller B, Roth M, Aranzana MJ, Auwerkerken A, *et al.* 2021. Genetic architecture and genomic prediction accuracy of apple quantitative traits across environments. biorXiv. doi: 10.1101/2021.11.29.470309.

Jung M, Roth M, Aranzana MJ, Auwerkerken A, Bink M, *et al.* 2020. The apple REFPOP-a reference population for genomics-assisted breeding in apple. Hortic Res. 7:189. doi:10.1038/s41438-020–00408-8.

Karoui S, Carabaño MJ, Díaz C, Legarra A. 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet Sel Evol. 44:39.doi:10.1186/1297-9686-44-39.

Khan SA, Tikunov Y, Chibon P-Y, Maliepaard C, Beekwilder J, *et al.* 2014. Metabolic diversity in apple germplasm. Plant Breed. 133: 281–290. doi:10.1111/pbr.12134.

Kouassi AB, Durel C-E, Costa F, Tartarini S, van de Weg E, *et al.* 2009. Estimation of genetic parameters and prediction of breeding values for apple fruit-quality traits using pedigreed plant material in Europe. Tree Genet Genomes. 5:659–672. doi: 10.1007/s11295-009-0217-x.

Kumar S, Bink MCAM, Volz RK, Bus VGM, Chagné D. 2012a. Towards genomic selection in apple (Malus × domestica Borkh.) breeding programmes: prospects, challenges and strategies. Tree Genet Genomes. 8:1–14. doi:10.1007/s11295-011-0425-z.

Kumar S, Chagné D, Bink MCAM, Volz RK, Whitworth C, *et al.* 2012b. Genomic selection for fruit quality traits in apple (Malus×domestica Borkh.), (T. Zhang, Ed.). PLoS One. 7: e36674.doi:10.1371/journal.pone.0036674.

Kumar S, Hilario E, Deng CH, Molloy C. 2020. Turbocharging introgression breeding of perennial fruit crops: a case study on apple. Hortic Res. 7:47.doi:10.1038/s41438-020-0270-z.

Kumar S, Molloy C, Muñoz P, Daetwyler H, Chagné D, *et al.* 2015. Genome-enabled estimates of additive and nonadditive genetic variances and prediction of apple phenotypes across environments. G3 (Bethesda). 5:2711–2718. doi:10.1534/g3.115.021105.

Laurens F, Aranzana MJ, Arus P, Bassi D, Bink M, *et al.* 2018. An integrated approach for increasing breeding efficiency in apple and peach in Europe. Hortic Res. 5:11. doi:10.1038/s41438-018-0016-3.

Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, *et al.* 2014. Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. Genetics. 198:3–16. doi:10.1534/genetics.114.161943.

Lehermeier C, Schön C-C, de los Campos G. 2015. Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. Genetics. 201:323–337. doi:10.1534/genetics.115.177394.

Lenth RV. 2021. emmeans: estimated Marginal Means, aka Least-Squares Means. R package version 1.6.0.

Liao L, Zhang W, Zhang B, Fang T, Wang X-F, *et al.* 2021. Unraveling a genetic roadmap for improved taste in the domesticated apple. Mol Plant. 14:1454–1471. doi:10.1016/j.molp.2021.05.018.

Lorenz AJ, Smith KP. 2015. Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. Crop Sci. 55:2657–2667. doi:10.2135/cropsci2014.12.0827.

Lund MS, van den Berg I, Ma P, Brøndum RF, Su G. 2016. Review: how to improve genomic predictions in small dairy cattle populations. Animal. 10:1042–1049. doi:10.1017/S1751731115003031.

Lyra DH, Granato ÍSC, Morais PPP, Alves FC, dos Santos ARM, *et al.* 2018. Controlling population structure in the genomic prediction of tropical maize hybrids. Mol Breed. 38. doi:10.1007/s11032-018–0882-2.

Mangin B, Rincent R, Rabier C-E, Moreau L, Goudemand-Dugue E. 2019. Training set optimization of genomic prediction by means of EthAcc. PLoS One. 14:e0205629.doi:10.1371/journal.pone.0205629.

McClure KA, Gardner KM, Douglas GM, Song J, Forney CF, *et al.* 2018. A genome-wide association study of apple quality and scab resistance. Plant Genome. 11:170075.doi:10.3835/plantgenome2017.08.0075.

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829.

Migicovsky Z, Gardner KM, Money D, Sawler J, Bloom JS, *et al.* 2016. Genome to phenome mapping in apple using historical data. Plant Genome. 9. doi:10.3835/plantgenome2015.11.0113.

Migicovsky Z, Yeats TH, Watts S, Song J, Forney CF, *et al.* 2021. Apple ripening is controlled by a NAC transcription factor. Front Genet. 12:671300.doi:10.3389/fgene.2021.671300.

Minamikawa MF, Kunihisa M, Noshita K, Moriya S, Abe K, *et al.* 2021. Tracing founder haplotypes of Japanese apple varieties: application in genomic prediction and genome-wide association study. Hortic Res. 8:49. doi:10.1038/s41438-021-00485-3.

Muranty H, Denancé C, Feugey L, Crépin J-L, Barbier Y, *et al.* 2020. Using whole-genome SNP data to reconstruct a large multi-generation pedigree in apple germplasm. BMC Plant Biol. 20:2.doi:10.1186/s12870-019-2171-6.

Muranty H, Troggio M, Sadok IB, Rifaï MA, Auwerkerken A, *et al.* 2015. Accuracy and responses of genomic selection on key traits in apple breeding. Hortic Res. 2:15060. doi:10.1038/hortres.2015.60.

Myles S. 2013. Improving fruit and wine: what does genomics have to offer? Trends Genet. 29:190–196. doi:10.1016/j.tig.2013.01.006.

Noiton DAM, Alspach PA. 1996. Founding clones, inbreeding, coancestry, and status number of modern apple cultivars. J Am Soc Hortic Sci. 121:773–782. doi:10.21273/JASHS.121.5.773.

Nsibi M, Gouble B, Bureau S, Flutre T, Sauvage C, *et al.* 2020. Adoption and optimization of genomic selection to sustain breeding for apricot fruit quality. G3 (Bethesda). 10:4513–4529. doi:10.1534/g3.120.401452.

Nybom H, Sehic J, Garkava-Gustavsson L. 2008. Modern apple breeding is associated with a significant change in the allelic ratio of the ethylene production gene *Md-ACS1*. J Hortic Sci Biotechnol. 83:673–677. doi:10.1080/14620316.2008.11512442.

Olatoye MO, Clark LV, Labonte NR, Dong H, Dwiyanti MS, *et al.* 2020. Training population optimization for genomic selection in *Miscanthus*. G3 (Bethesda). 10:2465–2476. doi:10.1534/g3.120.401402.

Ou J-H, Liao C-T. 2019. Training set determination for genomic selection. Theor Appl Genet. 132:2781–2792. doi:10.1007/s00122-019-03387-0.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, *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.

Ramstein GP, Casler MD. 2019. Extensions of BLUP models for genomic prediction in heterogeneous populations: application in a diverse switchgrass sample. G3 (Bethesda). 9:789–805. doi:10.1534/g3.118.200969.

Raymond B, Bouwman AC, Wientjes YCJ, Schrooten C, Houwing-Duistermaat J, *et al.* 2018. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. Genet Sel Evol. 50:49.doi:10.1186/s12711-018-0419-5.

Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, *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.

Rincent R, Charcosset A, Moreau L. 2017. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. Theor Appl Genet. 130:2231–2247. doi:10.1007/s00122-017-2956-7.

Rio S, Moreau L, Charcosset A, Mary-Huard T. 2020. Accounting for Group-Specific Allele Effects and Admixture in Genomic Predictions: Theory and Experimental Evaluation in Maize. Genetics. 216:27–41. doi: 10.1534/genetics.120.303278.

Rio S, Mary-Huard T, Moreau L, Charcosset A. 2019. Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. Theor Appl Genet. 132:81–96. doi:10.1007/s00122-018-3196-1.

Rodríguez-Álvarez MX, Boer MP, van Eeuwijk FA, Eilers PHC. 2018. Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spat Stat. 23:52–71. doi:10.1016/j.spasta.2017.10.003.

de Roos APW, Hayes, BJ, de Goddard ME. 2009. Reliability of genomic predictions across multiple populations. Genetics. 183:1545–1553. doi:10.1534/genetics.109.104935.

de Roos APW, de Hayes BJ, Spelman RJ, Goddard ME. 2008. Linkage disequilibrium and persistence of phase in Holstein–Friesian,

Jersey and Angus cattle. Genetics. 179:1503–1512. doi:10.1534/genetics.107.084301.

Roth M, Muranty H, Guardo MD, Guerra W, Patocchi A, *et al.* 2020. Genomic prediction of fruit texture and training population optimization towards the application of genomic selection in apple. Hortic Res. 7:148. doi:10.1038/s41438-020-00370-5.

Sarinelli JM, Murphy JP, Tyagi P, Holland JB, Johnson JW, *et al.* 2019. Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. Theor Appl Genet. 132:1247–1261. doi: 10.1007/s00122-019-03276-6.

Schulz-Streeck T, Ogutu JO, Karaman Z, Knaak C, Piepho HP. 2012. Genomic Selection using Multiple Populations. Crop Science. 52:2453–2461. doi: 10.2135/cropsci2012.03.0160.

Singh A, Dubey PK, Chaurasia R, Dubey RK, Pandey KK, *et al.* 2019. Domesticating the undomesticated for global food and nutritional security: four steps. Agronomy. 9:491.doi:10.3390/agronomy9090491.

Song H, Ye S, Jiang Y, Zhang Z, Zhang Q, *et al.* 2019. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. Genet Sel Evol. 51:58.doi:10.1186/s12711-019-0500-8.

Sood S, Lin Z, Caruana B, Slater AT, Daetwyler HD. 2020. Making the most of all data: combining non-genotyped and genotyped potato individuals with HBLUP. Plant Genome. 13:e20056.doi:10.1002/tpg2.20056.

Sverrisdóttir E, Sundmark EHR, Johnsen HØ, Kirk HG, Asp T, *et al.* 2018. The value of expanding the training population to improve genomic selection models in tetraploid potato. Front Plant Sci. 9:1118.doi:10.3389/fpls.2018.01118.

Technow F, Totir LR. 2015. Using Bayesian multilevel whole genome regression models for partial pooling of training sets in genomic prediction. G3 (Bethesda). 5:1603–1612. doi:10.1534/g3.115.019299.

Tiezzi F, Maltecca C. 2015. Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. Genet Sel Evol. 47:24.doi:10.1186/s12711-015-0100-1.

van Nocker S, Gardiner SE. 2014. Breeding better cultivars, faster: applications of new technologies for the rapid deployment of superior horticultural tree crops. Hortic Res. 1:14022.doi:10.1038/hortres.2014.22.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980.

Verma S, Evans K, Guan Y, Luby JJ, Rosyara UR, *et al.* 2019. Two large-effect QTLs, Ma and Ma3, determine genetic potential for acidity in apple fruit: breeding insights from a multi-family study. Tree Genetics & Genomes. 15:Doi:10.1007/s11295-019-1324-y.

Voss-Fels KP, Cooper M, Hayes BJ. 2019. Accelerating crop genetic gains with genomic selection. Theor Appl Genet. 132:669–686. doi:10.1007/s00122-018-3270-8.

Watkins R, Smith RA. 1982. International board for plant genetic resources, commission of the European communities, and committee on disease resistance breeding and use of Genebanks. In: Descriptor List for Apple (Malus.). Brussels; Rome: CEC Secretariat; IBPGR Secretariat.

Watts S, Migicovsky Z, McClure KA, Amyotte B, *et al.* 2021. Quantifying apple diversity: a phenomic characterization of Canada's apple biodiversity collection. Plants People Planet. 10211. doi:10.1002/ppp3.10211.

Wedger MJ, Schumann AC, Gross BL. 2021. Candidate genes and signatures of directional selection on fruit quality traits during apple domestication. Am J Bot. 108:616–627. doi:10.1002/ajb2.1636.

Wientjes Y, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, *et al.* 2015. Empirical and deterministic accuracies of across-population genomic prediction. Genet Sel Evol. 47:5.doi:10.1186/s12711-014-0086-0.

Wientjes YCJ, Bijma P, Veerkamp RF, Calus MPL. 2016. An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. Genetics. 202:799–823. doi:10.1534/genetics.115.183269.

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

Xu Y, Liu X, Fu J, Wang H, Wang J, *et al.* 2020. Enhancing genetic gain through genomic selection: from livestock to plants. Plant Commun. 1:100005.doi:10.1016/j.xplc.2019.100005.

Yao J-L, Xu J, Cornille A, Tomes S, Karunairetnam S, *et al.* 2015. A *microRNA* allele that emerged prior to apple domestication may underlie fruit size evolution. Plant J. 84:417–427. doi:10.1111/tpj.13021.

Yu X, Li X, Guo T, Zhu C, Wu Y, *et al.* 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat Plants. 2:16150. doi:10.1038/nplants.2016.150.

Zhang A, Wang H, Beyene Y, Semagn K, Liu Y, *et al.* 2017. Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. Front Plant Sci. 8:1916. doi:10.3389/fpls.2017.01916.

Communicating editor: P. Brown

Cette section correspond au document **Supplementary Materials S1** de l'article de ce chapitre.

Effect of the training set size on predictive ability

For the elite material and genetic resources of the FBo-Hi dataset, we evaluated the impact of the number of genotypes in the TS on predictive ability for the within-population prediction scenario. To do so, we used the 100 sets of candidates used in the WP scenario (coming from 20 replications of a fivefold cross-validation scheme) and predicted each set with a training set of randomly chosen x individuals from the remaining genotypes of the same population, where x starts at 50 genotypes and is increased by steps of 50 genotypes until all the available genotypes are part of the training set. For each training set constituted this way, the predictive ability was evaluated as the Pearson correlation coefficient between the GEBV and the phenotypic values of the individuals in the validation set. In a second step, we added all the genotypes from the complementary population to the x genotypes of the training set. We call the first scenario WP_{inc} and the second one $Comb_{inc}$. Note that the WP and Comb scenarios correspond to the WP_{inc} and $Comb_{inc}$ scenarios when the value of x corresponds to the maximum number of genotypes in the training set.
For the $Comb_{inc}$ scenario, the MG-GBLUP model was not evaluated because of the high computational time it would require to apply the model for each value of x.

Figures supplémentaires de l'article

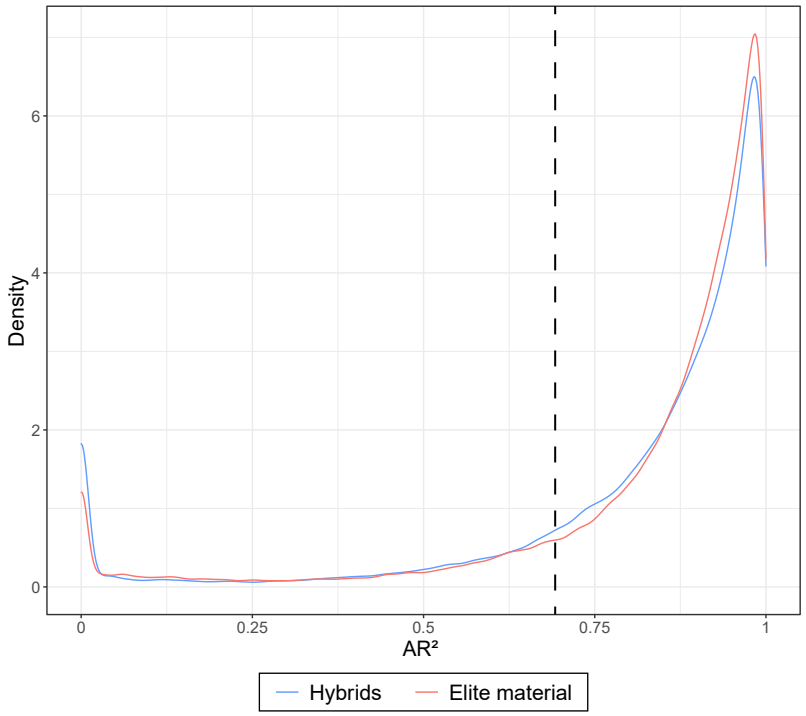


FIGURE S4.1 – Distribution des valeurs d'AR² obtenues après imputation du matériel élite et des hybrides RG x E

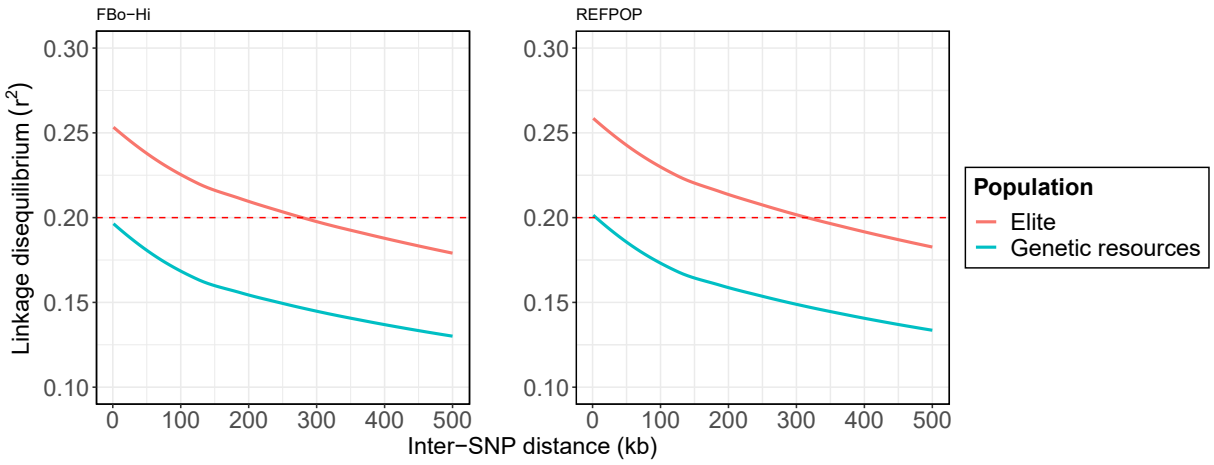


FIGURE S4.2 – Décroissance du déséquilibre de liaison entre SNP compris dans une fenêtre de 500kb dans le matériel élite et les ressources génétiques des jeux de données **FBo-Hi** et **REFPOP**

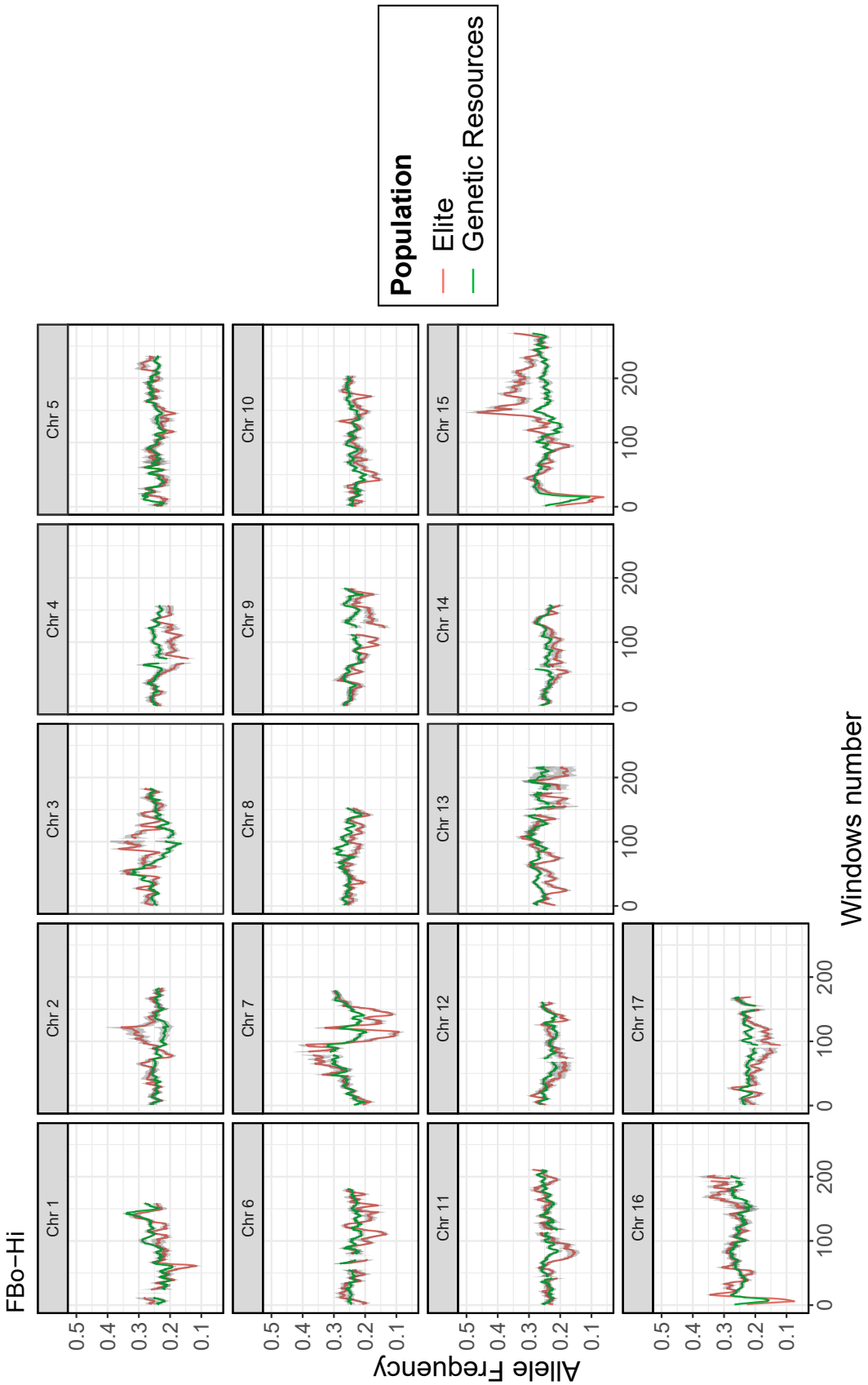


FIGURE S4.3 – Fréquence de l'allèle mineur dans les ressources génétiques et fréquence du même allème dans le matériel élite le long des 17 chromosomes du génome du pommier dans le jeu de données **FBo-Hi**. Les fréquences alléliques ont été calculées dans des fenêtres glissantes de 2Mb avec un pas de 400kb. L'intervalle de confiance des fréquences alléliques au sein des fenêtres glissantes est représenté en gris

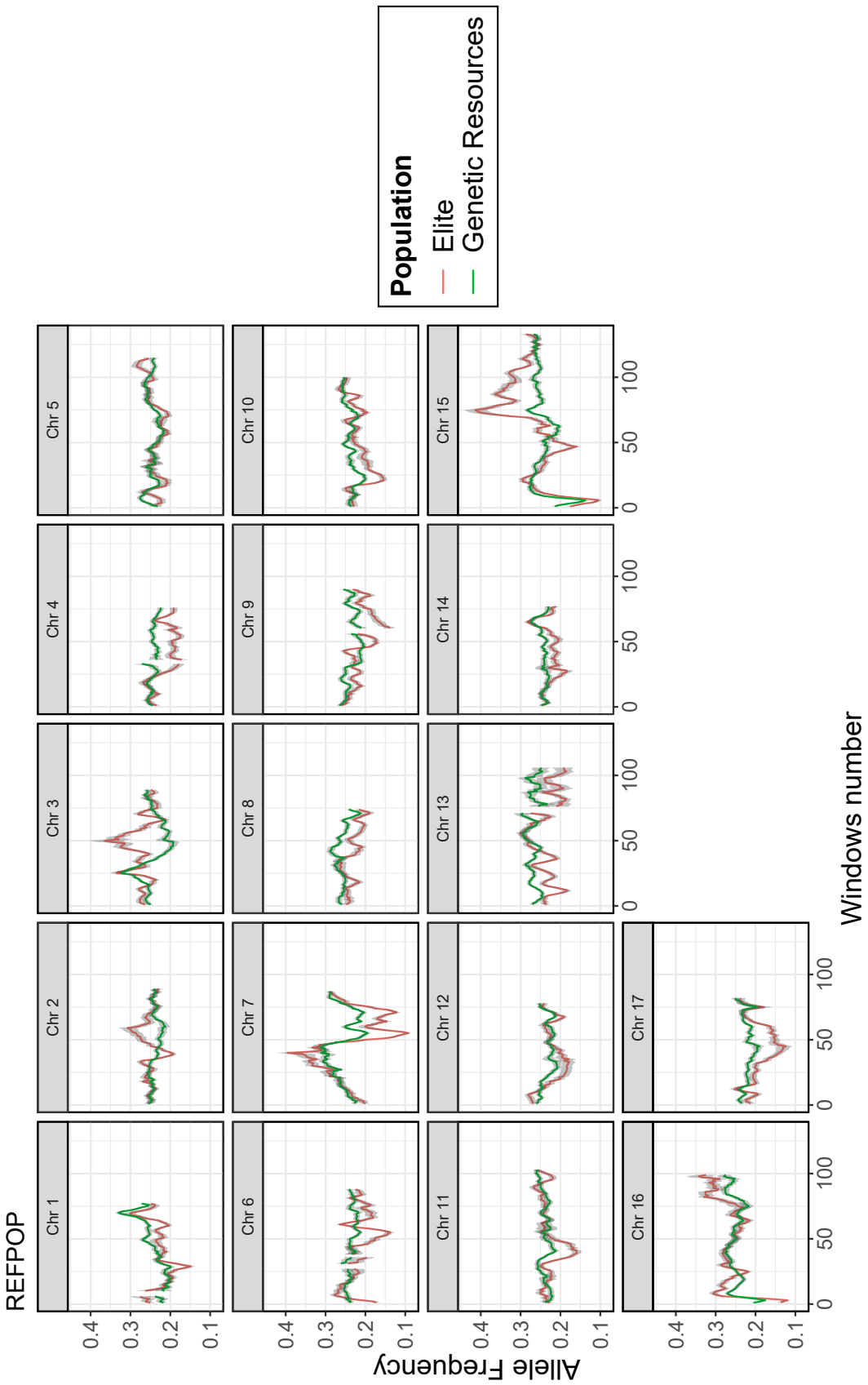


FIGURE S4.4 – Fréquence de l'allèle mineur dans les ressources génétiques et fréquence du même allème dans le matériel élite le long des 17 chromosomes du génome du pommier dans le jeu de données **REFPOP**. Les fréquences alléliques ont été calculées dans des fenêtres glissantes de 2Mb avec un pas de 400kb. L'intervalle de confiance des fréquences alléliques au sein des fenêtres glissantes est représenté en gris

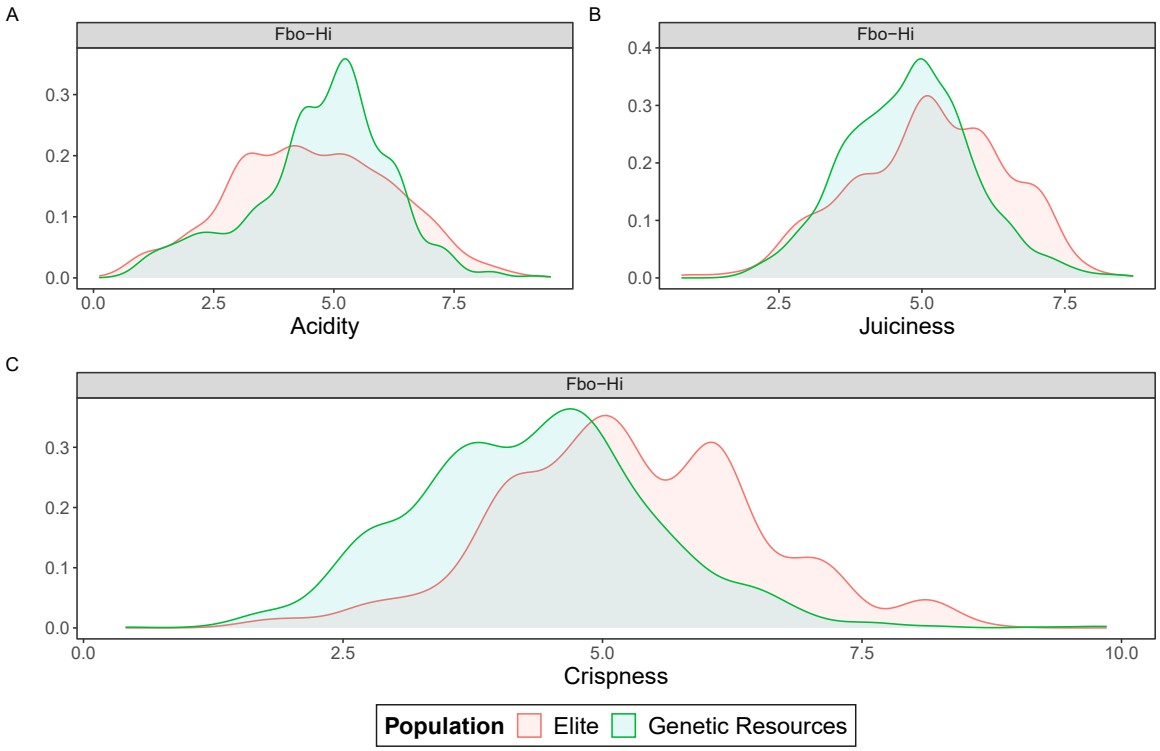


FIGURE S4.5 – Distribution phénotypique des caractères uniquement mesurés dans le jeu de données **FBo-Hi**. **A**. Acidité **B**. Jutosité **C**. Croquant

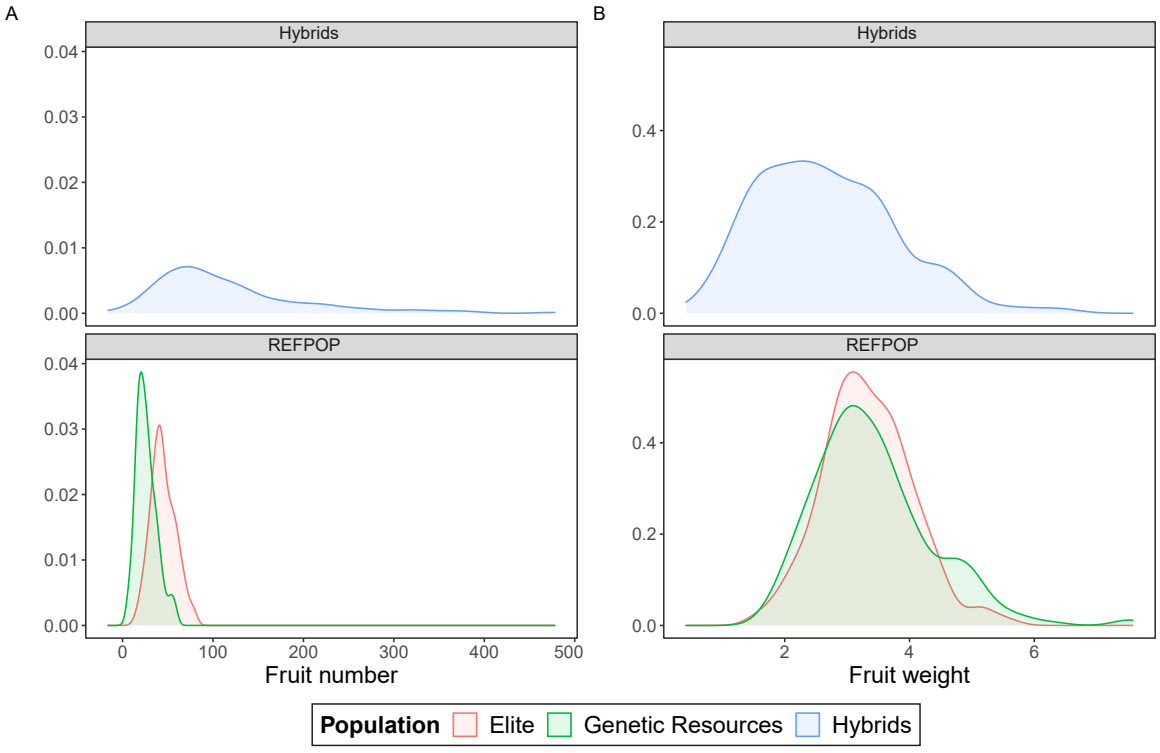


FIGURE S4.6 – Distribution phénotypique des caractères mesurés dans le jeu de données **REFPOP** et sur les hybrides RG x E. **A**. Nombre de fruits **B**. Poids de 20 fruits

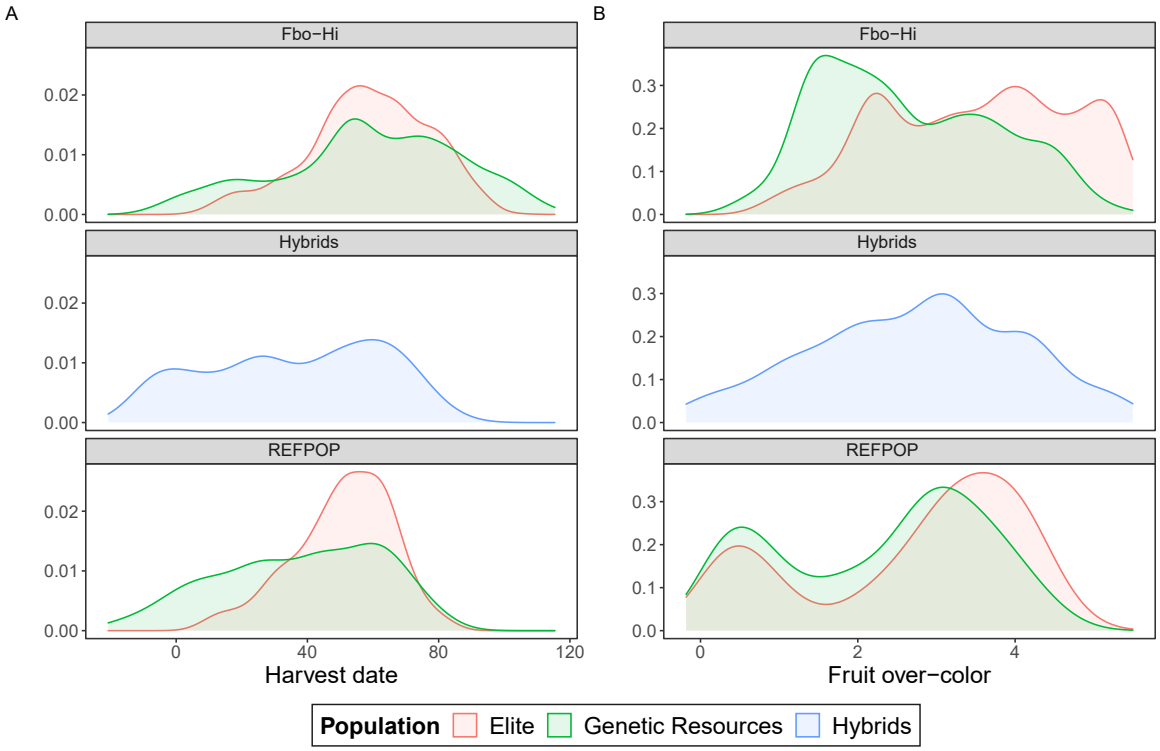


FIGURE S4.7 – Distribution phénotypique des caractères mesurés dans les trois jeux de données. **A**. Date de récolte **B**. Couleur du fruit

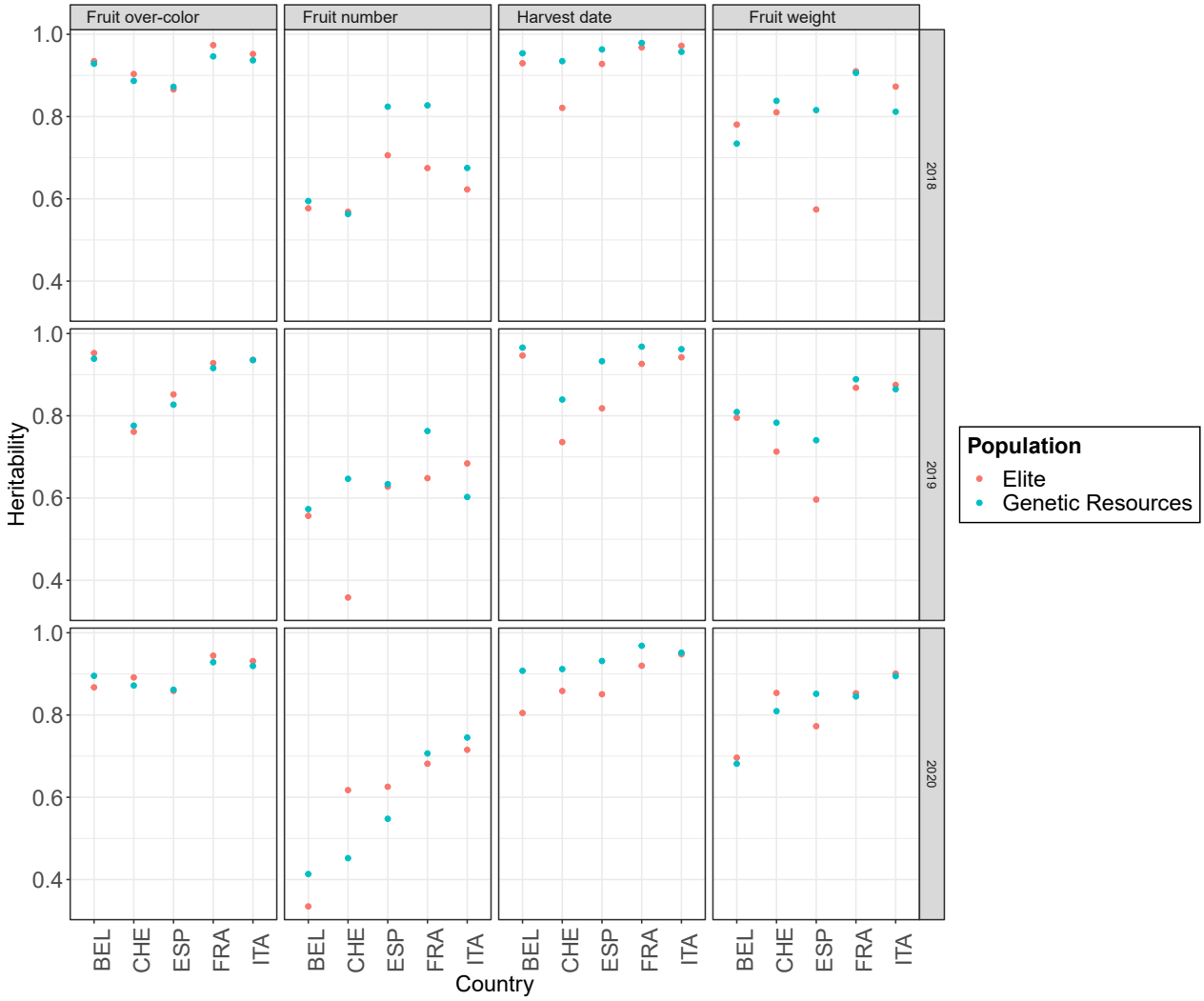


FIGURE S4.8 – Héritabilité au sens large par site et par année pour le jeu de données **REFPOP**. **BEL** Belgique **CHE** Suisse **ESP** Espagne **FRA** France **ITA** Italie

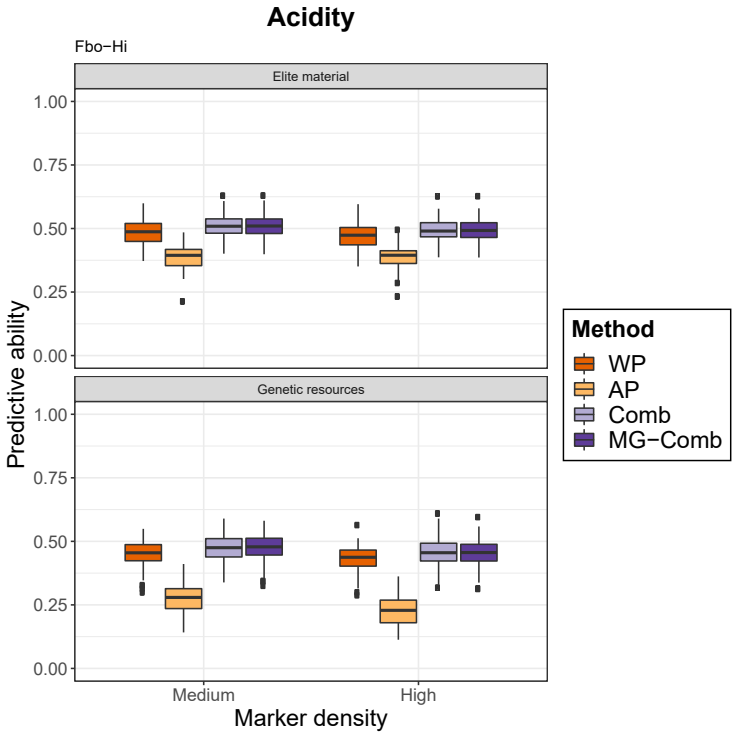


FIGURE S4.9 – Précision de prédiction de l'acidité dans le jeu de données **FBo-Hi** en utilisant des données moyenne ou haute densité. **WP** : within-population prediction **AP** : across-population prediction **Comb** : prédiction en combinaison **MG-Comb** : prédiction en combinaison avec le modèle MG-GBLUP

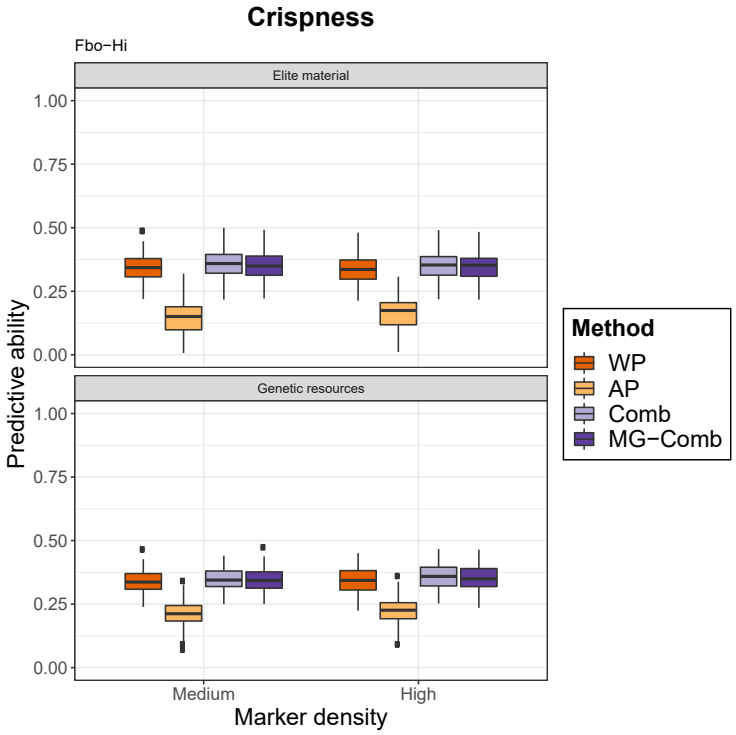


FIGURE S4.10 – Précision de prédiction du caractère croquant dans le jeu de données **FBo-Hi** en utilisant des données moyenne ou haute densité. **WP** : within-population prediction **AP** : across-population prediction **Comb** : prédiction en combinaison **MG-Comb** : prédiction en combinaison avec le modèle MG-GBLUP

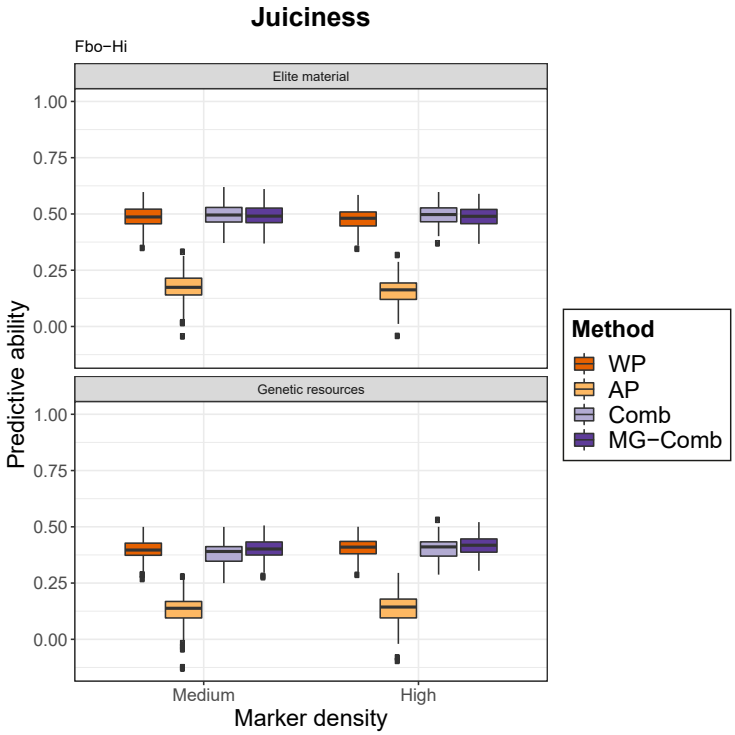


FIGURE S4.11 – Précision de prédiction de la jutosité dans le jeu de données **FBo-Hi** en utilisant des données moyenne ou haute densité. **WP** : within-population prediction **AP** : across-population prediction **Comb** : prédiction en combinaison **MG-Comb** : prédiction en combinaison avec le modèle MG-GBLUP

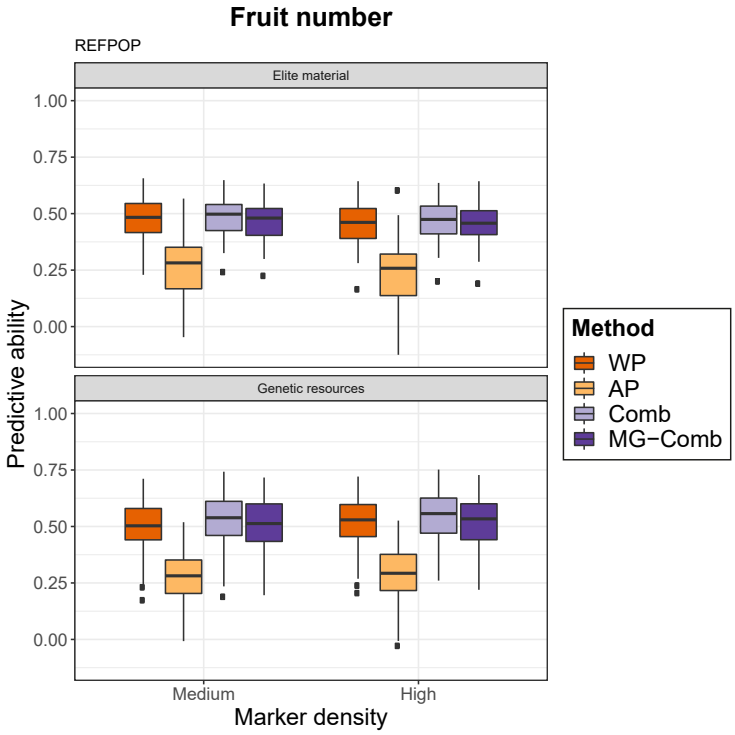


FIGURE S4.12 – Précision de prédiction du nombre de fruits dans le jeu de données **REFPOP** en utilisant des données moyenne ou haute densité. **WP** : within-population prediction **AP** : across-population prediction **Comb** : prédiction en combinaison **MG-Comb** : prédiction en combinaison avec le modèle MG-GBLUP

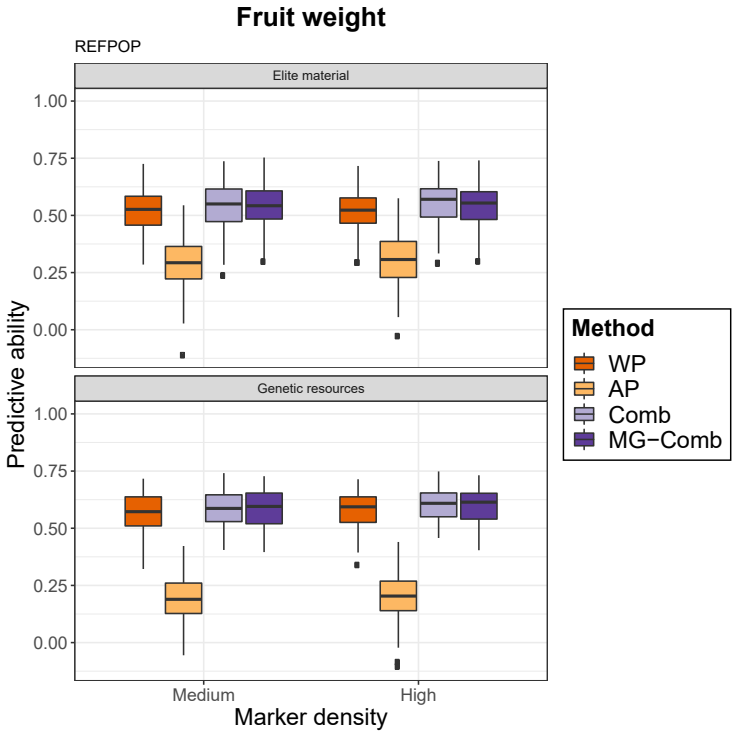


FIGURE S4.13 – Précision de prédiction du poids de 20 fruits dans le jeu de données **REFPOP** en utilisant des données moyenne ou haute densité. **WP** : within-population prediction **AP** : across-population prediction **Comb** : prédiction en combinaison **MG-Comb** : prédiction en combinaison avec le modèle MG-GBLUP

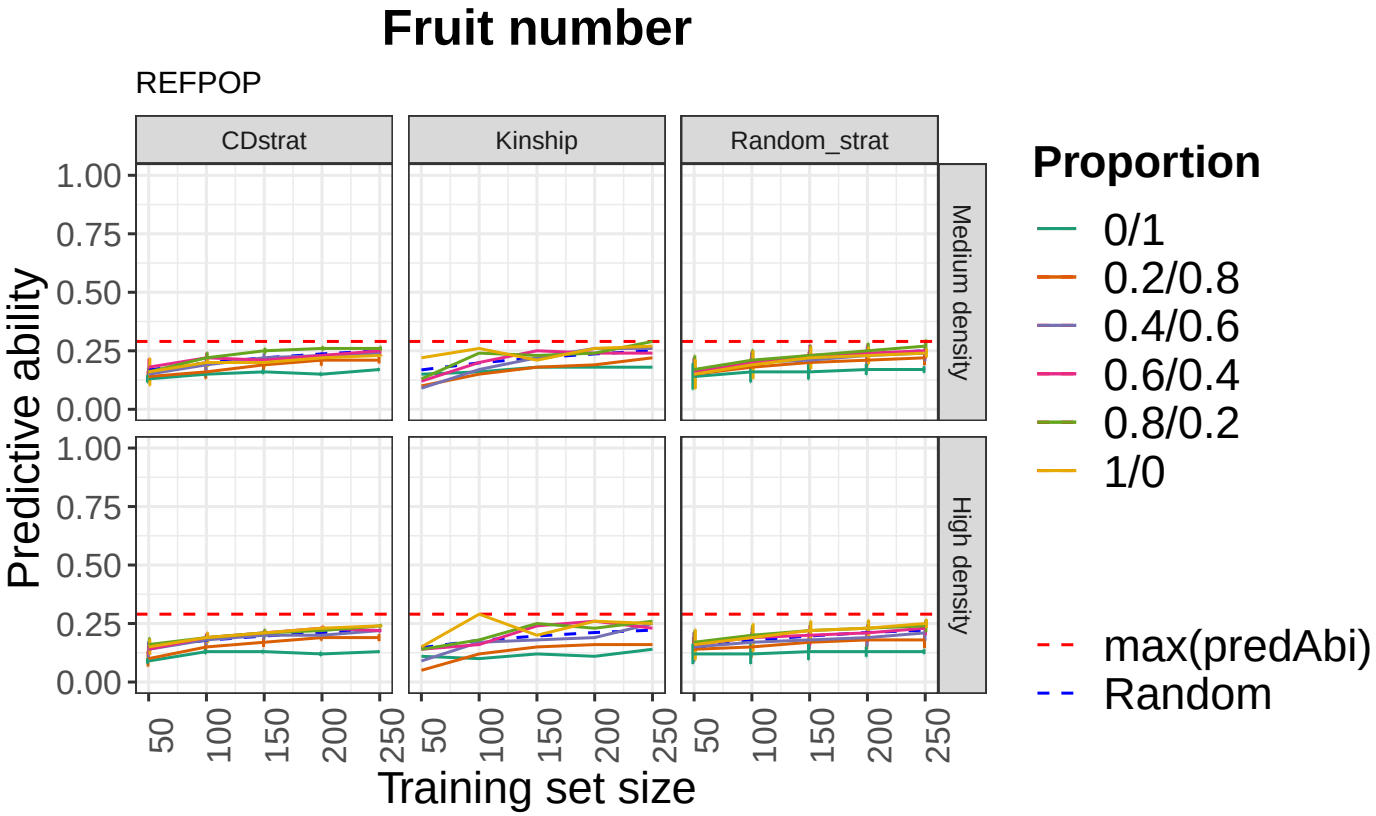


FIGURE S4.14 – Précision de prédiction du nombre de fruits dans le jeu de données des hybrides en utilisant des données moyenne ou haute densité pour une population d'entrainement composée d'une proportion variable de génotypes provenant du matériel élite et des ressources génétiques du jeu de données **REFPOP**. **max(predAbi)** : précision de prédiction maximale obtenue pour une densité de marqueurs donnée en utilisant une des trois méthodes étudiées, indépendamment de la taille de la population d'entrainement. **Random** : précision de prédiction obtenue en choissant aléatoirement les génotypes de la population d'entrainement

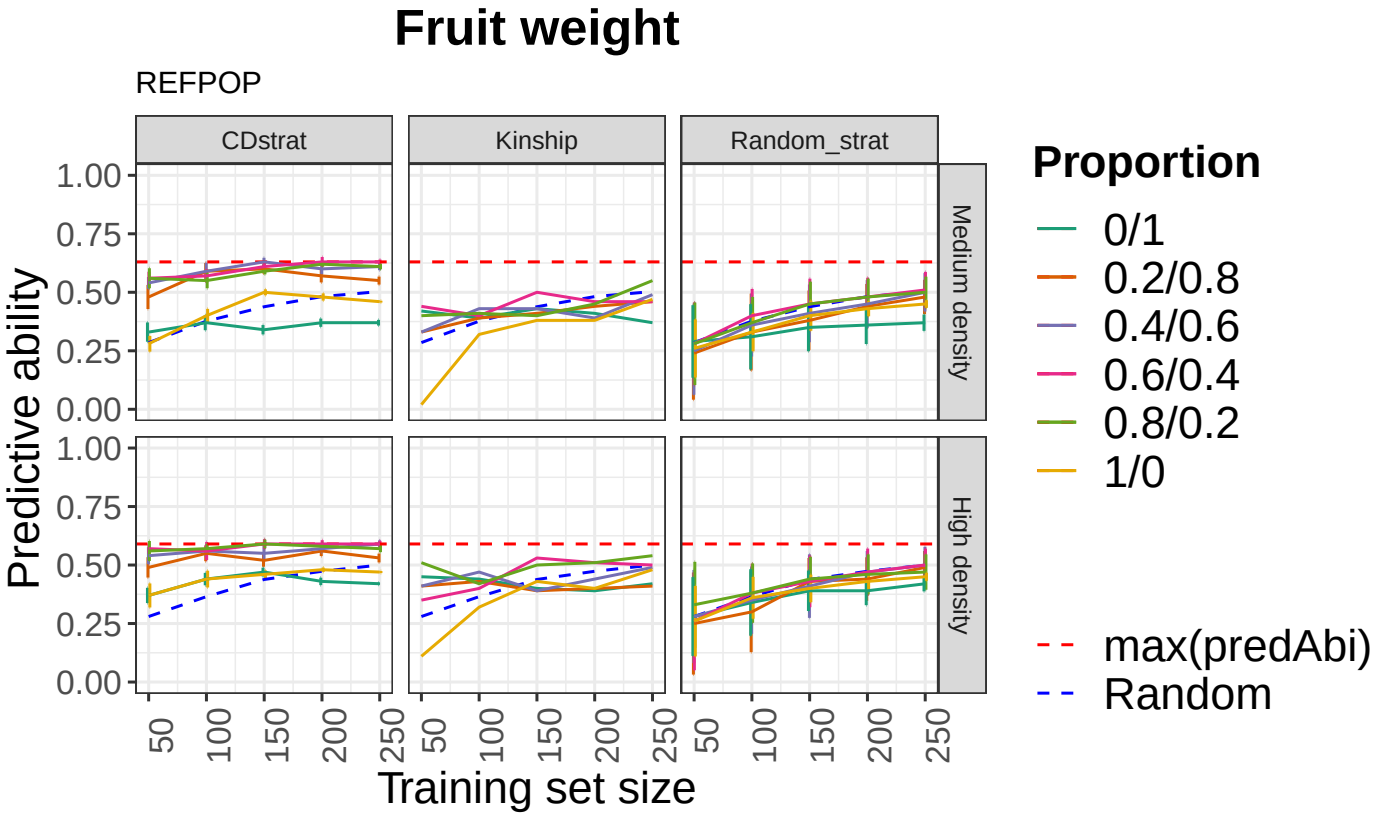


FIGURE S4.15 – Précision de prédiction du poids de 20 fruits dans le jeu de données des hybrides en utilisant des données moyenne ou haute densité pour une population d'entrainement composée d'une proportion variable de génotypes provenant du matériel élite et des ressources génétiques du jeu de données **REFPOP**. **max(predAbi)** : précision de prédiction maximale obtenue pour une densité de marqueurs donnée en utilisant une des trois méthodes étudiées, indépendamment de la taille de la population d'entrainement. **Random** : précision de prédiction obtenue en choissant aléatoirement les génotypes de la population d'entrainement

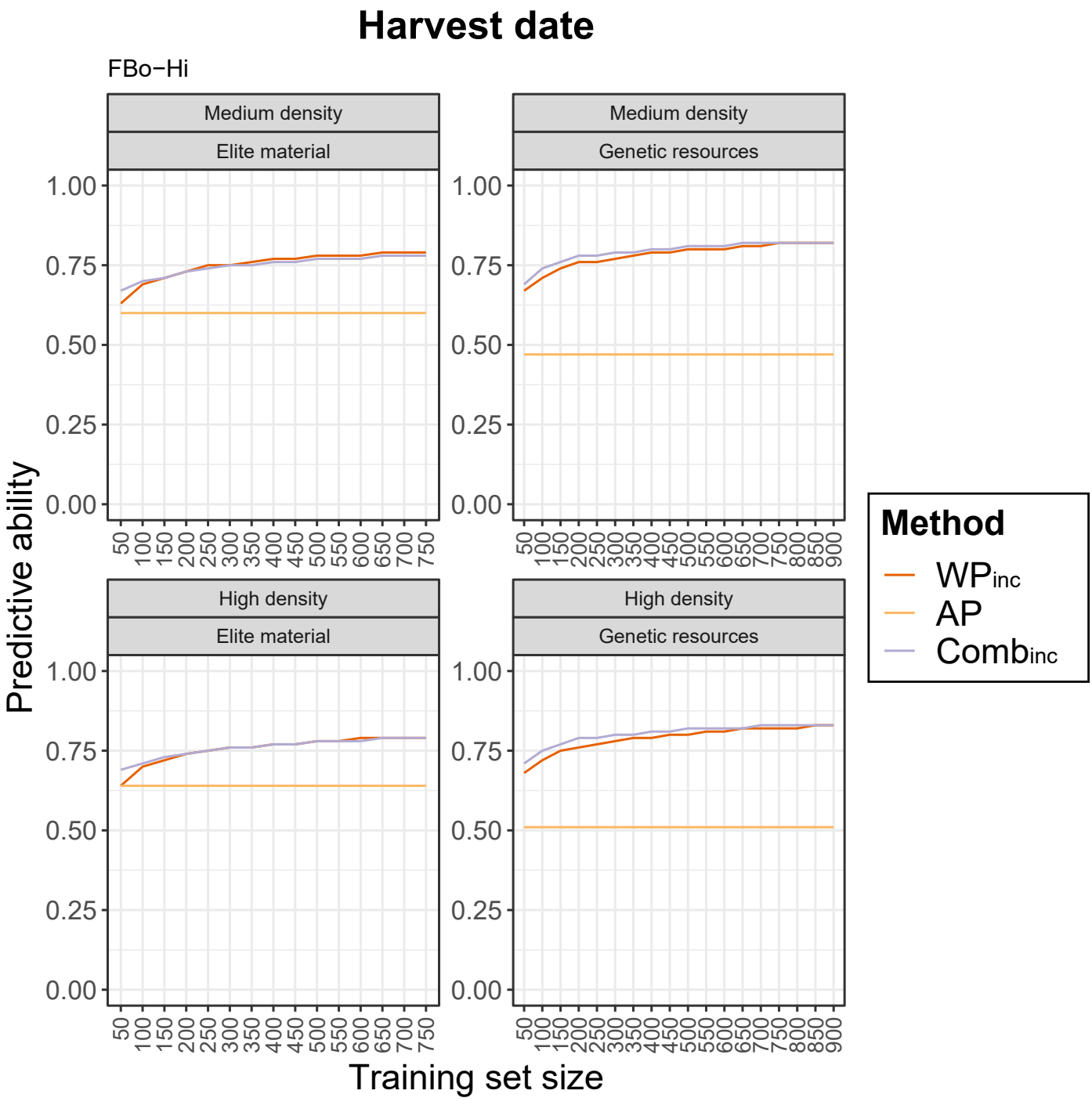


FIGURE S4.16 – Évolution de la précision de prédiction pour la date de récolte lorsque les candidats sont prédits en utilisant un nombre croissant de génotypes provenant de la même population ($\mathbf{WP_{inc}}$), tous les génotypes de la population complémentaire (**AP**) ou une combinaison des deux populations ($\mathbf{Comb_{inc}}$)

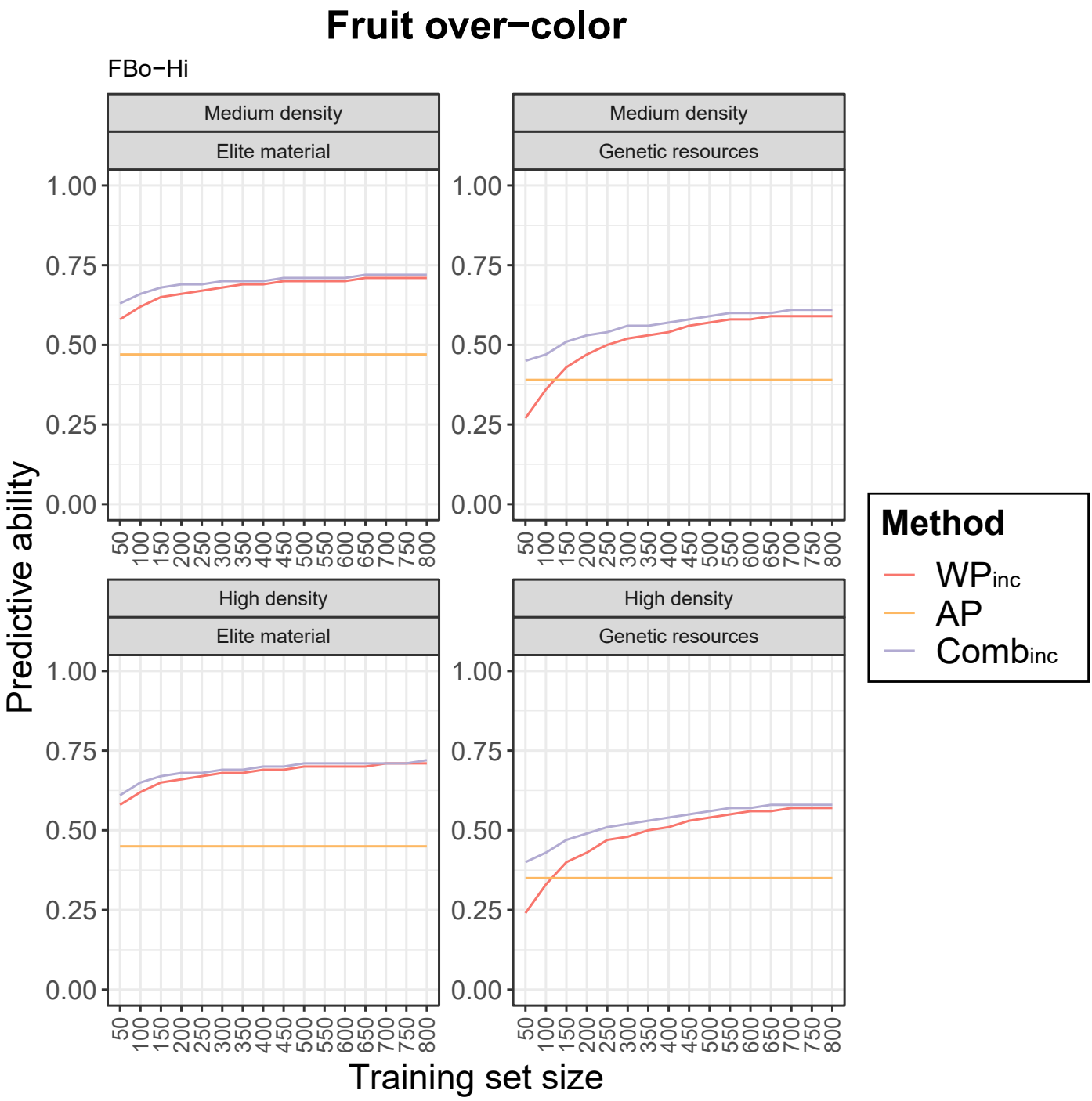


FIGURE S4.17 – Évolution de la précision de prédiction pour la couleur du fruit lorsque les candidats sont prédits en utilisant un nombre croissant de génotypes provenant de la même population ($\mathbf{WP_{inc}}$), tous les génotypes de la population complémentaire (**AP**) ou une combinaison des deux populations ($\mathbf{Comb_{inc}}$)

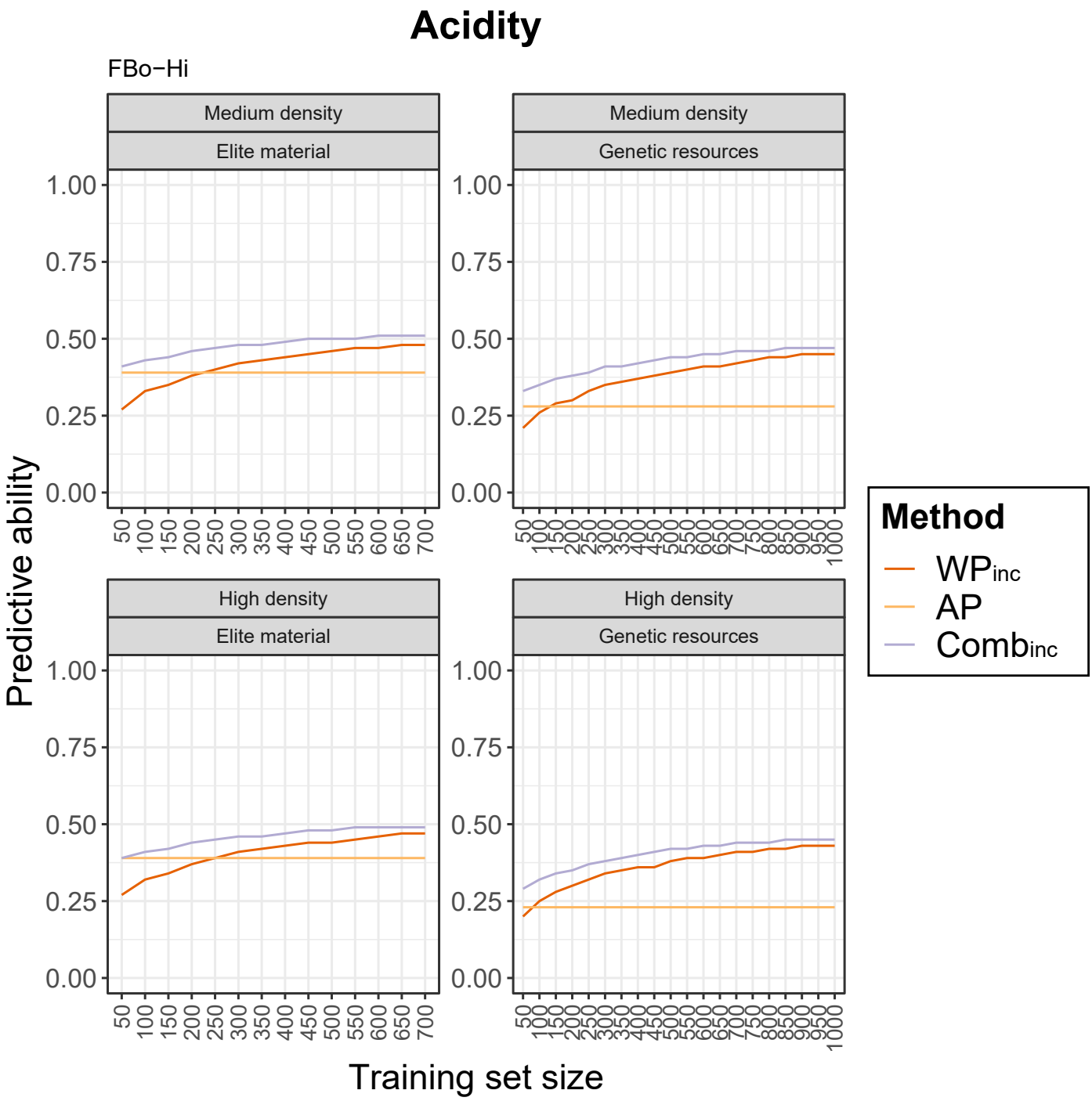


FIGURE S4.18 – Évolution de la précision de prédiction pour l'acidité lorsque les candidats sont prédits en utilisant un nombre croissant de génotypes provenant de la même population ($\mathbf{WP_{inc}}$), tous les génotypes de la population complémentaire (**AP**) ou une combinaison des deux populations ($\mathbf{Comb_{inc}}$)

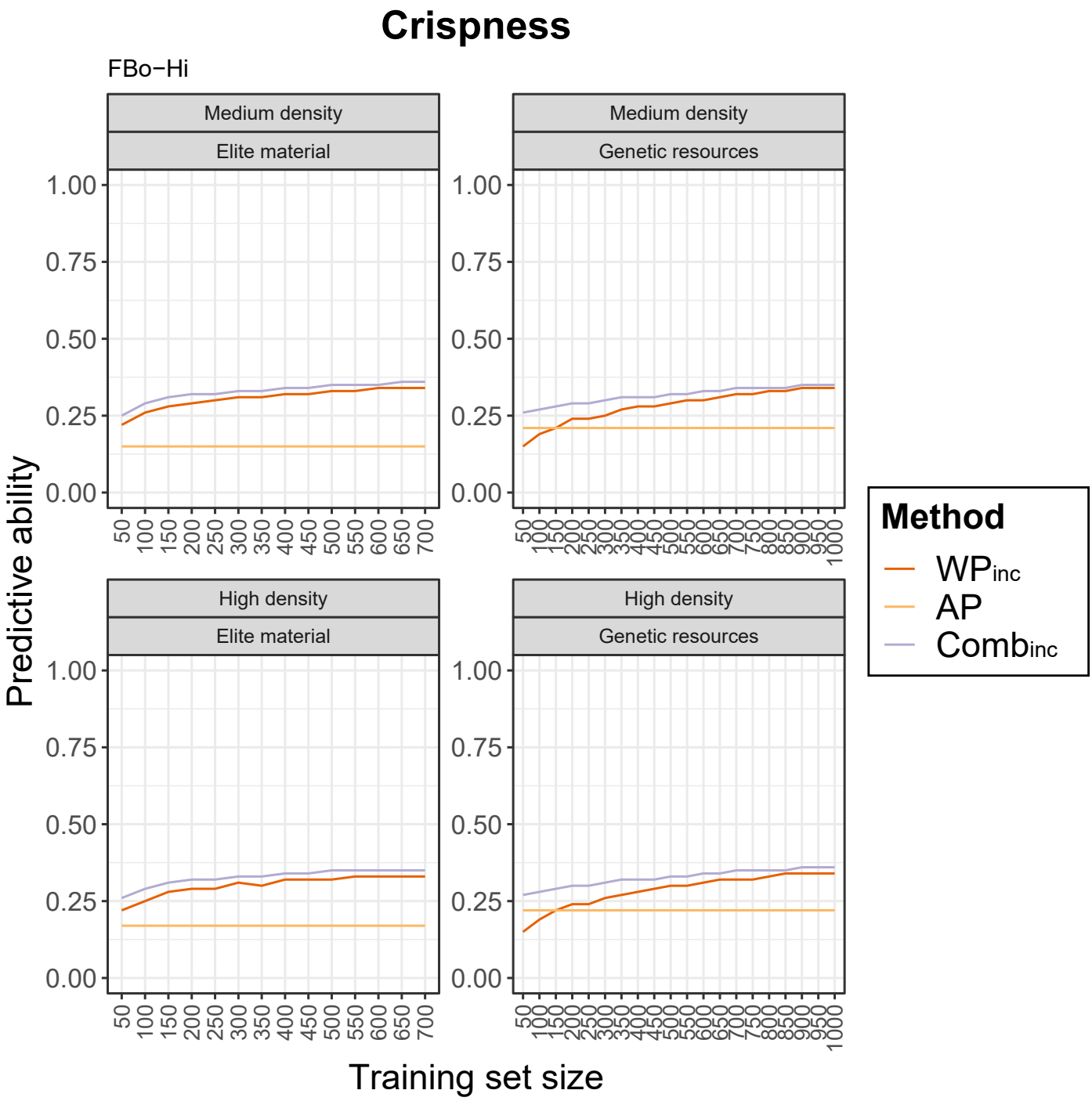


FIGURE S4.19 – Évolution de la précision de prédiction pour le caractère croquant lorsque les candidats sont prédits en utilisant un nombre croissant de génotypes provenant de la même population ($\mathbf{WP_{inc}}$), tous les génotypes de la population complémentaire (**AP**) ou une combinaison des deux populations ($\mathbf{Comb_{inc}}$)

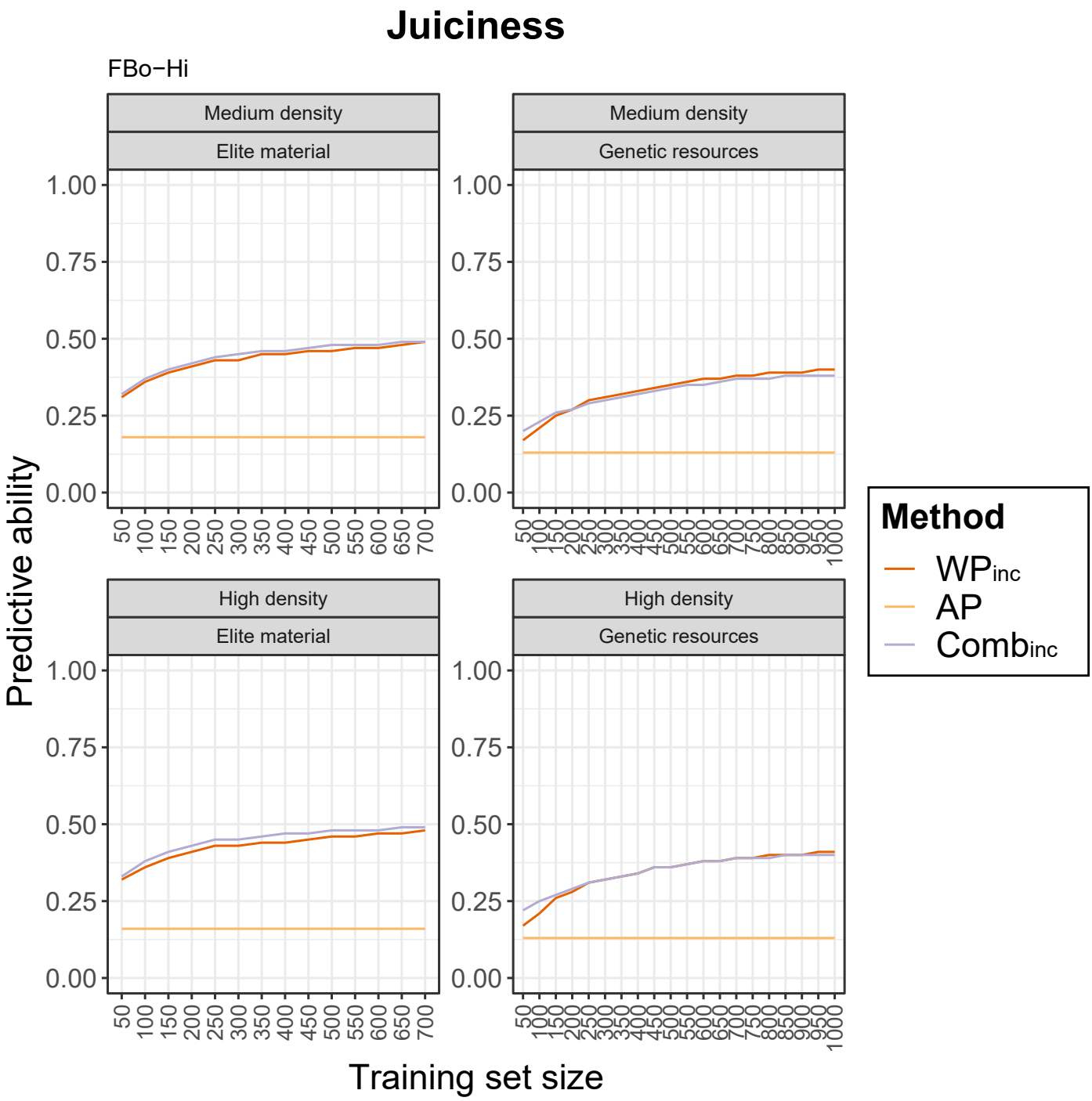


FIGURE S4.20 – Évolution de la précision de prédiction pour la jutosité lorsque les candidats sont prédits en utilisant un nombre croissant de génotypes provenant de la même population ($\mathbf{WP_{inc}}$), tous les génotypes de la population complémentaire (**AP**) ou une combinaison des deux populations ($\mathbf{Comb_{inc}}$)

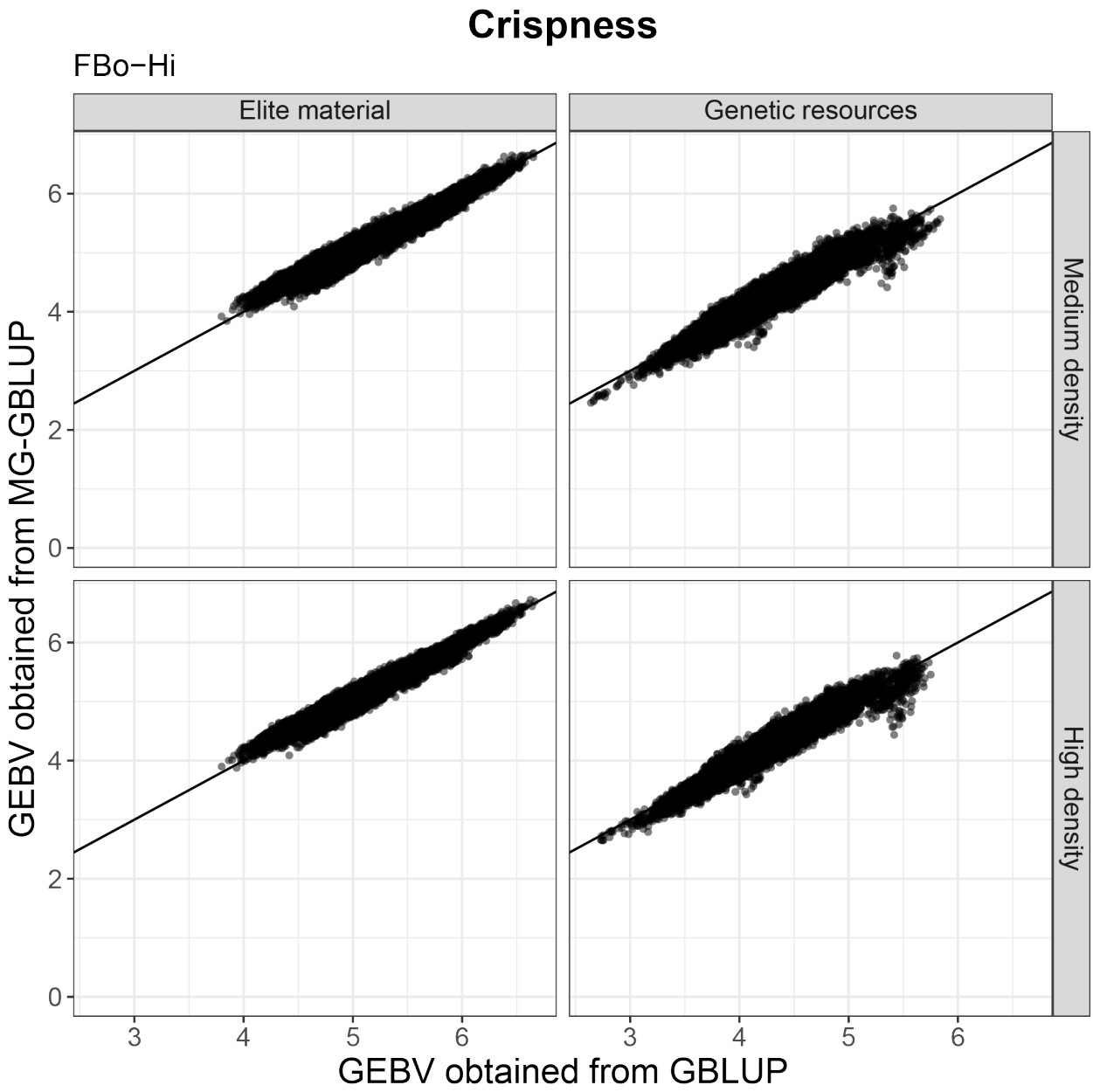


FIGURE S4.21 – Comparaison des GEBV obtenues en utilisant les modèles GBLUP et MG-GBLUP pour le caractère croquant. La corrélation entre les GEBV est indiquée pour chaque population et densité de marqueurs

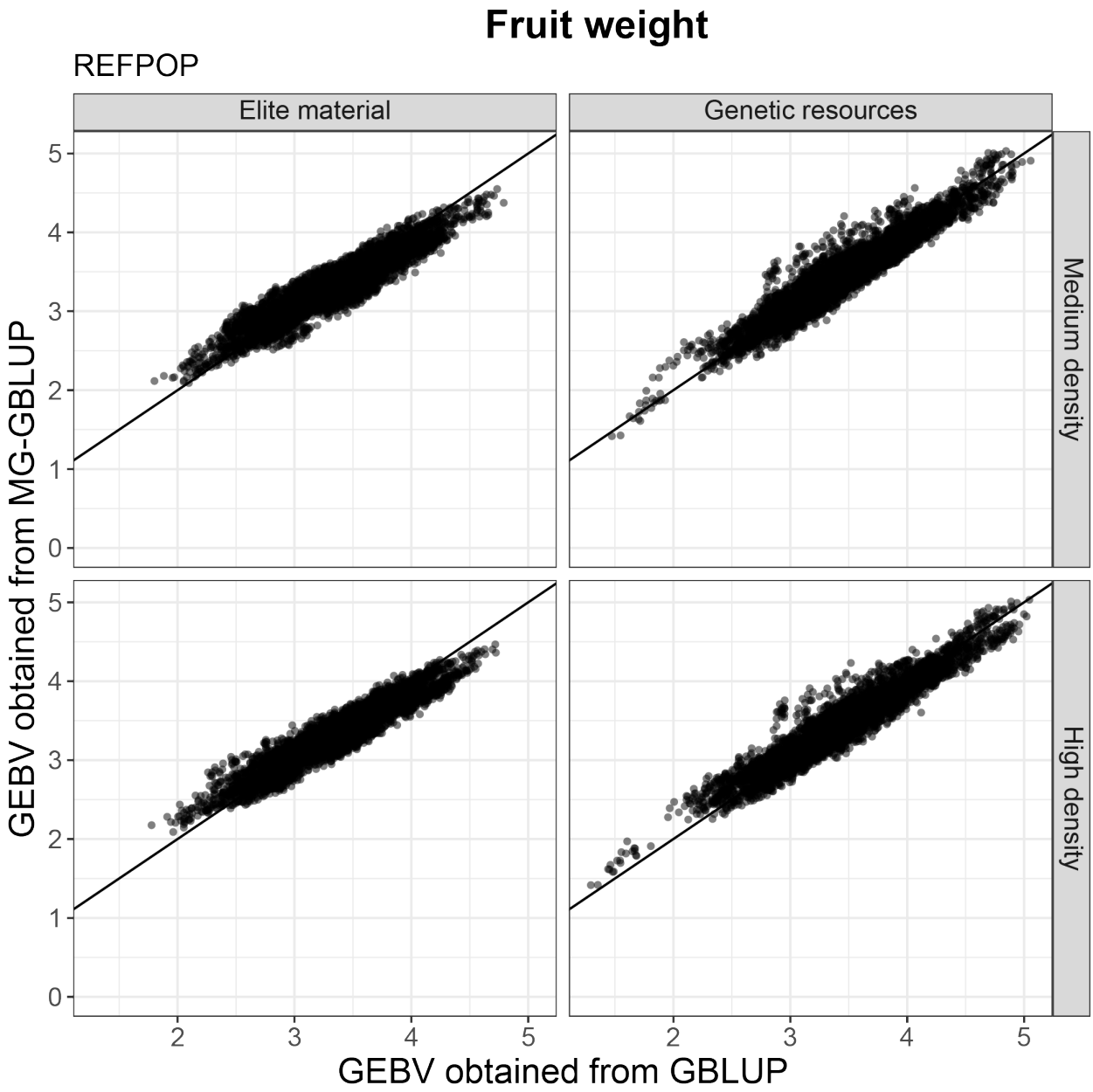


FIGURE S4.22 – Comparaison des GEBV obtenues en utilisant les modèles GBLUP et MG-GBLUP pour le poids de 20 fruits. La corrélation entre les GEBV est indiquée pour chaque population et densité de marqueurs

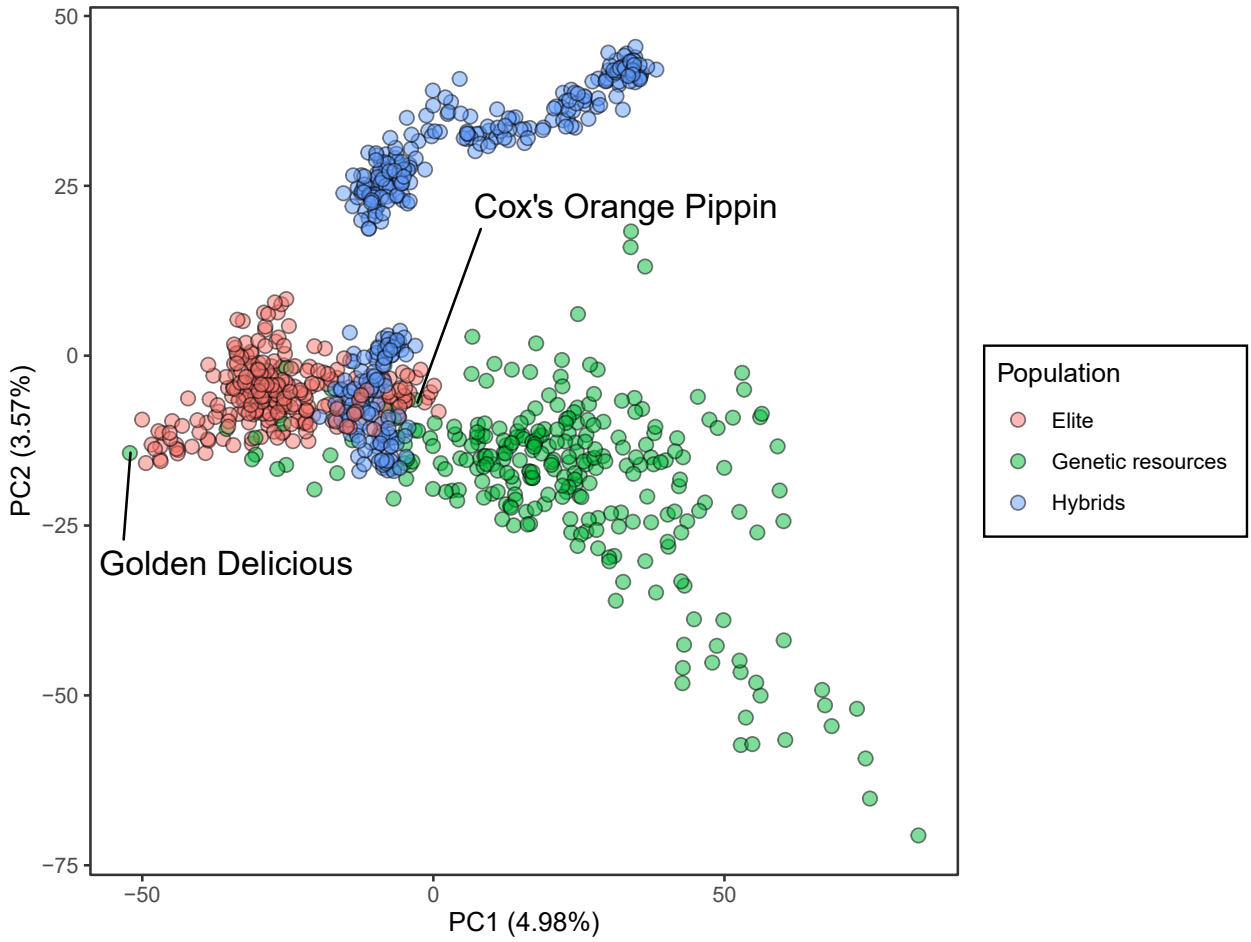


FIGURE S4.23 – Analyse en Composantes Principales (ACP) réalisée en utilisant le jeu de données réduit des panels **REFPOP** et des hybrides RG x E

Tableaux supplémentaires de l'article

TABLEAU 4.1 – Plan de croisement et taille des familles de la population d'hybrides RG x E

		Ressources génétiques				
		X08233	X08483	X02353	X02640	X08488
Elite	X02437	65				
	X03263				39	
	X03318					49
	X06407		38			
	X06683			50		
	X06963	31				
	X07860	51				
	FRAi0085	44	50			
	X08486	56				

TABLEAU 4.2 – Valeurs de F_{ST} entre le matériel élite, les ressources génétiques et les hybrides RG x E obtenues à partie des marqueurs SNP dans les jeux de données FBo-Hi et REFPOP

FBo-Hi		
Population	Matériel élite	Ressources génétiques
Ressources génétiques	0,023	-
Hybrides	0,029	0,017
REFPOP dataset		
Population	Matériel élite	Ressources génétiques
Ressources génétiques	0,021	-
Hybrides	0,029	0,017

4.2 Résultats complémentaires

Dans la discussion de l'article, nous évoquions certaines pistes qui pourraient être explorées pour améliorer la précision des prédictions que nous avions réalisées. Nous avons entre temps testé certaines des approches évoquées, et proposons dans les paragraphes suivants une synthèse des observations qui en découlent.

4.2.1 Prise en compte de l'incertitude liée à l'imputation lors de la prédiction génomique

Pour plusieurs caractères, nous avons obtenu des précisions de prédictions inférieures lorsque les données haute densité ont été utilisées à la place des données moyenne densité. Afin de savoir si cette observation pouvait résulter d'erreurs d'imputation, nous avons de nouveau réalisé les prédictions génomiques pour les cinq caractères du jeu de données FBo-Hi mais en prenant cette fois-ci en compte l'incertitude des données génomiques liée à l'imputation. Pour ce faire, nous avons procédé de trois façons, en nous basant sur les recommandations de Song et al. (Song et al. 2019) :

— Dans le cas de la méthode **Filter**, nous avons supprimé du jeu de données génomiques tous les marqueurs imputés ayant une valeur d'$\mathbf{AR^2}$ inférieure à 0,7 et avons calculé la matrice d'apparentement génomique utilisée pour le GBLUP à partir de ce sous-ensemble de marqueurs.

— Dans le cas de la méthode AR^2, nous avons multiplié chaque élément de la matrice de génotypage (en codant les individus hétérozygotes 1 et les individus homozygotes respectivement 0 et 2 selon qu'ils portaient l'allèle alternatif ou l'allèle de référence) par les valeurs d'$\mathbf{AR^2}$ associées à ces éléments

— Dans le cas de la méthode **GL**, nous avons remplacé le génotype le plus probable après imputation par sa dose estimée, basée sur la vraisemblance de chaque génotype (ou Genotype Likelihood, GL). Ainsi, si pour un marqueur les probabilités associées aux génotypes AA, AB et BB sont respectivement p_{AA}, p_{AB} et p_{BB}, alors la dose estimée sera égale à 0 x p_{AA} + p_{AB} + 2 x p_{BB} si A est l'allèle de référence.

Les précisions de prédiction obtenues sont présentées dans le Tableau 4.3 ci-dessous. La colonne all correspond à la précision de prédiction moyenne obtenue pour ces mêmes caractères en utilisant l'ensemble des marqueurs à haute densité. Pour les méthodes $\mathbf{AR^2}$ et GL, aucune valeur n'est fournie pour le scenario WP lorsque les candidats sont constitués de génotypes provenant des ressources génétiques, car dans ce cas les données utilisées sont des données réelles et non pas des données imputées. Nous avons tout de même calculé les précisions de

prédiction pour les cinq caractères avec la méthode **Filter** dans ce cas de figure, car dans le cas du scenario Comb, les marqueurs ayant une valeur d'AR^2 inférieure à 0,7 dans le jeu de données des individus élite sont également supprimés du jeu de données des ressources génétiques.

Les précisions de prédiction obtenues avec les trois méthodes sont très proches ou moins élevées que les valeurs obtenues en utilisant l'ensemble des marqueurs imputés, à l'exception de la jutosité (lorsque les candidats viennent de la population « matériel élite ») et de la couleur du fruit, dans les deux cas pour le scenario AP. Dans le cas des méthodes $\mathbf{AR^2}$ et GL, ce résultat n'est pas surprenant dans le sens où la majorité des valeurs d'AR^2 obtenues après imputation sont proches de 1 (figure S4.1). Pour de futures prédictions se basant sur des données imputées, nous préconisons donc d'utiliser l'intégralité des données génotypiques obtenues après imputation pour construire les équations de prédiction.

4.2.2 Comparaison des modèles GBLUP et BayesA pour la prédiction des hybrides RG x E

Une autre façon d'améliorer la précision des prédictions génomiques réside dans le choix du modèle utilisé en lien avec l'architecture génétique du caractère d'intérêt. Toutes les prédictions présentées jusqu'à présent ont été obtenues en utilisant un modèle GBLUP (ou MG-GBLUP), qui suppose que les effets des marqueurs proviennent d'une distribution normale. Pour des caractères tels que la couleur du fruit ou l'acidité, des QTL majeurs sont connus et l'utilisation de modèles bayesiens peut alors se révéler avantageuse pour s'affranchir de la contrainte de normalité des effets des marqueurs.
Afin d'étudier l'intérêt de tels modèles, nous avons de nouveau prédit la date de récolte, la couleur du fruit, le poids de 20 fruits et le nombre de fruits des hybrides RG x E en utilisant le modèle BayesA et le package *BGLR* (Pérez et de los Campos 2014) pour chaque proportion et chaque population d'entrainement étudiées précédemment dans ce chapitre. Nous avons utilisé 10 000 itérations, dont 2 000 de burn-in et avons utilisé les données moyenne densité disponibles après imputation afin de raccourcir le temps de calcul.
La figure S4.24 représente la différence entre la précision de prédiction obtenue en utilisant le modèle BayesA et la précision de prédiction obtenue par GBLUP en fonction du caractère et de la composition de la population d'entrainement, que nous avons constituée à partir du jeu de données REFPOP. Les deux modèles donnent des résultats très proches pour la date de récolte, le poids de 20 fruits et le nombre de fruits mais nous avons dans la plupart des cas obtenu des précisions de prédiction plus élevées en utilisant le modèle BayesA pour prédire la couleur du fruit, bien que ces différences soient généralement faibles. La différence entre les deux modèles est la plus marquée pour les proportions 1/0 et 0/1 (différence moyenne de 0,04 dans les deux cas et différence maximum de 0,15 et 0,1 respectivement). Notons que nous avons obtenu des

TABLEAU 4.3 – Précision de prédiction dans les ressources génétiques et le matériel élite du jeu de données FBo-Hi pour l'acidité, le caractère croquant, la date de récolte, la jutosité et la couleur du fruit en fonction de la méthode retenue pour prendre en compte l'incertitude liée à l'imputation des données génotypiques

Candidats	Caractère	Scenario	Filter	$AR2^2$	GL	All
Matériel élite	Acidité	WP	0,47 (0,05)	0,47 (0,05)	0,47 (0,05)	0,47 (0,05)
		AP	0,39 (0,05)	0,39 (0,05)	0,39 (0,05)	0,39 (0,05)
		Comb	0,5 (0,04)	0,5 (0,04)	0,5 (0,04)	0,5 (0,04)
	Croquant	WP	0,33 (0,06)	0,33 (0,06)	0,33 (0,06)	0,33 (0,06)
		AP	0,16 (0,06)	0,16 (0,06)	0,16 (0,06)	0,17 (0,06)
		Comb	0,36 (0,06)	0,36 (0,06)	0,36 (0,06)	0,36 (0,06)
	Date de récolte	WP	0,8 (0,02)	0,8 (0,02)	0,8 (0,02)	0,79 (0,03)
		AP	0,64 (0,04)	0,64 (0,04)	0,64 (0,04)	0,64 (0,04)
		Comb	0,79 (0,03)	0,79 (0,03)	0,8 (0,03)	0,79 (0,03)
	Jutosité	WP	0,48 (0,05)	0,48 (0,05)	0,48 (0,05)	0,48 (0,05)
		AP	0,17 (0,07)	0,16 (0,07)	0,16 (0,07)	0,16 (0,07)
		Comb	0,49 (0,05)	0,49 (0,05)	0,49 (0,05)	0,49 (0,05)
	Couleur	WP	0,71 (0,03)	0,71 (0,03)	0,71 (0,03)	0,71 (0,03)
		AP	0,45 (0,05)	0,44 (0,05)	0,44 (0,05)	0,45 (0,05)
		Comb	0,71 (0,03)	0,71 (0,03)	0,72 (0,03)	0,72 (0,03)
Ressources génétiques	Acidité	WP	0,43 (0,05)	-	-	0,43 (0,05)
		AP	0,22 (0,06)	0,23 (0,06)	0,23 (0,05)	0,23 (0,06)
		Comb	0,45 (0,06)	0,46 (0,06)	0,46 (0,06)	0,46 (0,06)
	Croquant	WP	0,34 (0,05)	-	-	0,34 (0,05)
		AP	0,22 (0,05)	0,21 (0,06)	0,22 (0,05)	0,22 (0,05)
		Comb	0,36 (0,05)	0,36 (0,05)	0,36 (0,05)	0,36 (0,05)
	Date de récolte	WP	0,83 (0,02)	-	-	0,83 (0,02)
		AP	0,52 (0,05)	0,51 (0,05)	0,52 (0,05)	0,51 (0,05)
		Comb	0,83 (0,02)	0,83 (0,02)	0,83 (0,02)	0,83 (0,02)
	Jutosité	WP	0,41 (0,04)	-	-	0,41 (0,05)
		AP	0,13 (0,07)	0,13 (0,07)	0,13 (0,07)	0,14 (0,05)
		Comb	0,4 (0,05)	0,4 (0,05)	0,41 (0,05)	0,4 (0,05)
	Couleur	WP	0,57 (0,05)	-	-	0,57 (0,05)
		AP	0,35 (0,06)	0,35 (0,07)	0,35 (0,07)	0,35 (0,06)
		Comb	0,58 (0,05)	0,58 (0,06)	0,59 (0,06)	0,58 (0,05)

Filter : méthode **AR^2** : méthode AR^2 **GL** : méthode Genotype Likelihood **All** : prédiction initiale en utilisant les marqueurs les plus probables après imputation **WP** : prédiction intra-population **AP** : prédiction inter-population **Comb** : combinaison des populations pour former la population d'entrainement

différences similaires en utilisant le jeu de données FBo-Hi pour la couleur du fruit et la date de récolte (le nombre de fruits et le poids de 20 fruits n'ayant pas été mesurés dans ce jeu de données), mais que la différence de précision de prédiction entre les modèles BayesA et GBLUP était bien moins marquée pour la proportion 1/0. Dans le cadre d'évaluations multi-race, Hayes et al. (2009) et van den Berg et al. (2015) ont également montré que des approches bayesiennes pouvaient donner de meilleurs résultats que le modèle GBLUP, surtout si les races présentes dans la population d'entrainement et dans la population de validation avaient divergé depuis un grand nombre de générations. Dans notre cas, il est possible que les ressources génétiques soient plus éloignées des hybrides RG x E que le matériel élite, expliquant l'avantage du modèle BayesA dans le cas de la proportion 0/1. Nous recommandons donc d'utiliser un modèle bayesien plutôt que le modèle GBLUP pour prédire des caractères pour lesquels des QTL à effet fort sont connus, à condition que le temps de calcul associé à ces prédictions ne soit pas trop important par rapport au temps de calcul nécessaire pour le modèle GBLUP.

4.2.3 Effet des interactions GxE : quelles données utiliser ?

Les prédictions initiales des hybrides RG x E en utilisant le jeu de données REFPOP ont été réalisées en utilisant les données phénotypiques obtenues sur les six sites européens faisant partie du réseau REFPOP. Cependant, Jung et al. (2022) ont montré que les intéractions génotypes x environnement (que nous noterons désormais intéractions GxE) pouvaient être importantes pour un certain nombre de caractères mesurés sur la REFPOP, notamment pour le nombre de fruits, que nous avons prédit pour les hybrides RG x E. Nous avons donc de nouveau prédit les hybrides RG x E à l'aide d'un modèle GBLUP et en utilisant les données de la REFPOP mais en nous limitant aux données phénotypiques obtenues sur le site d'Angers, en considérant que le site de la REFPOP d'Angers et le verger des hybrides RG x E correspondaient au même site. La figure S4.25 représente la différence de précision de prédiction entre le cas où les données phénotypiques de la population d'entrainement ont été ajustées en utilisant les données phénotypiques mesurées sur les six sites du réseau REFPOP ou au contraire les données de la REFPOP uniquement mesurées sur le site d'Angers.
Les précisions de prédiction sont alors dans la majorité des cas supérieures en utilisant les données mesurées sur les six sites, sauf pour le nombre de fruits, qui est un caractère pour lequel les interactions GxE sont fortes. Dans ce dernier cas, les deux approches conduisent à des précisions de prédiction très similaires. Il semble donc préférable d'utiliser les données de tous les partenaires du réseau REFPOP même pour prédire les candidats d'un seul site.

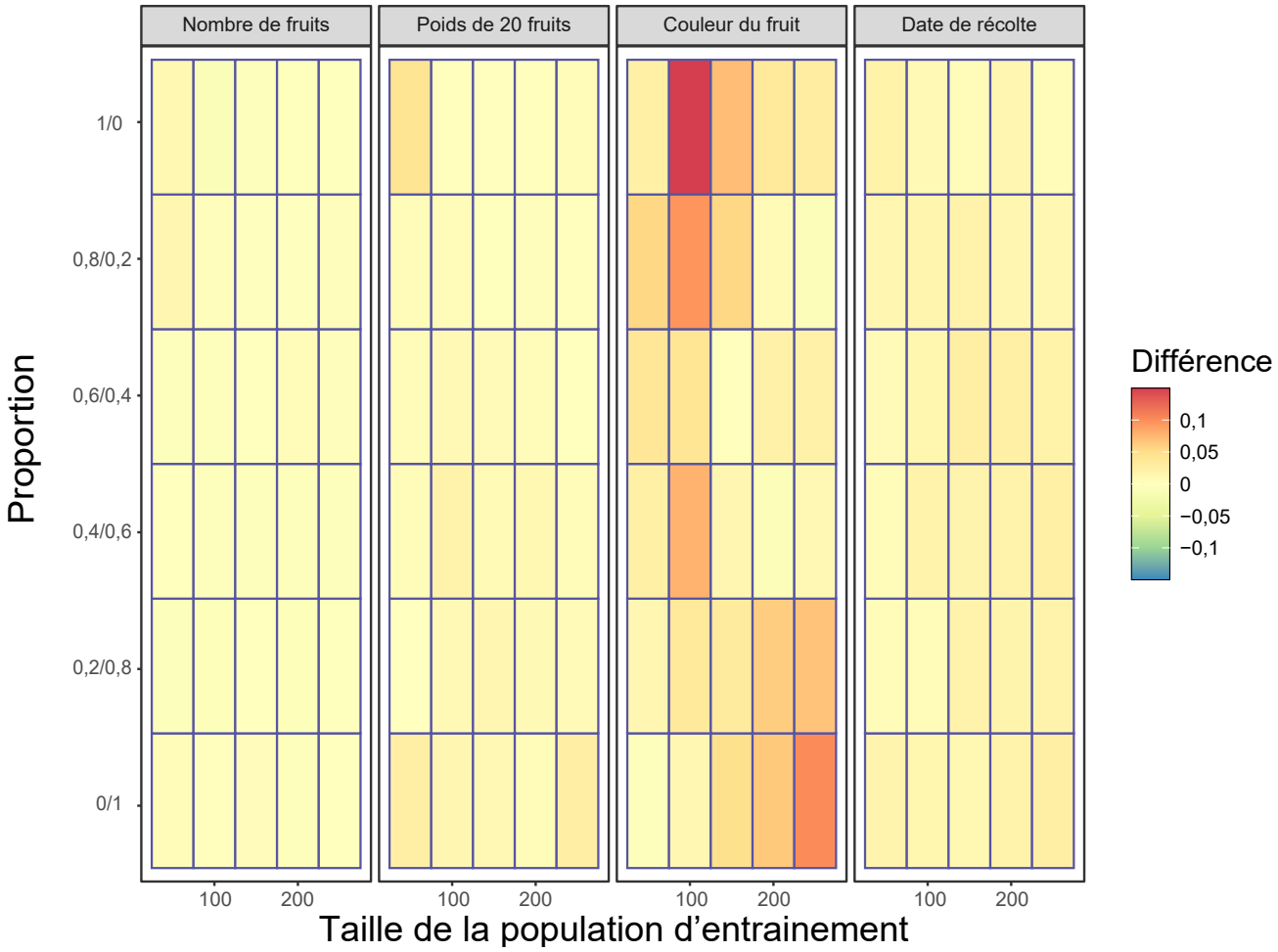


FIGURE S4.24 – Différence de précision de prédiction entre les modèles BayesA et GBLUP pour le nombre de fruits, le poids de 20 fruits, la date des récolte et la couleur du fruit des hybrides RG x E en utilisant une population d'entrainement de taille variable et des proportions variables de ressources génétiques et de matériel élite provenant de la REFPOP dans la population d'entrainement

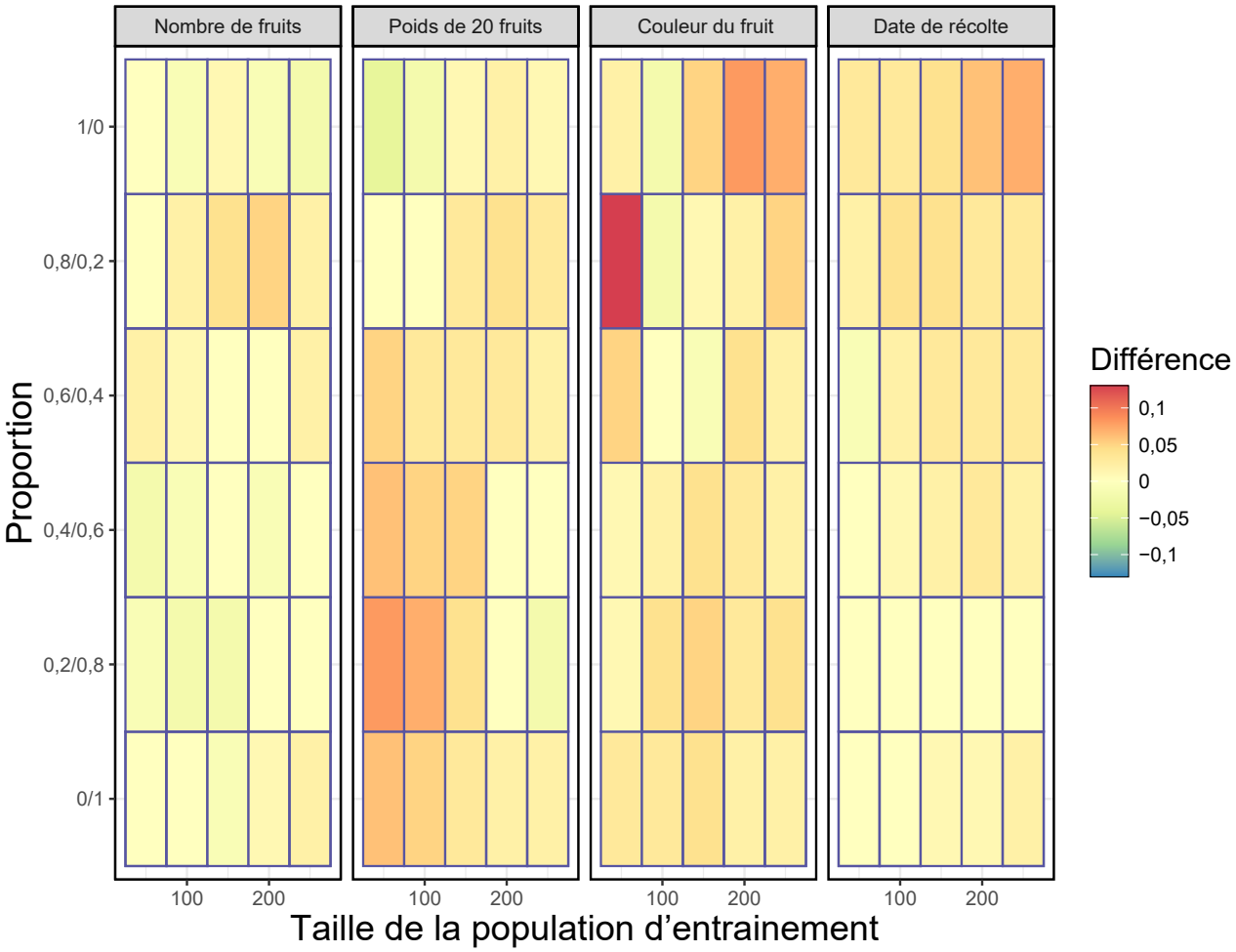


FIGURE S4.25 – Différence de précision de prédiction entre un modèle GBLUP utilisant les données provenant des six sites de la REFPOP et un modèle GBLUP utilisant les données provenant uniquement du site d'Angers pour le nombre de fruits, le poids de 20 fruits, la date des récolte et la couleur du fruit des hybrides RG x E en utilisant une population d'entrainement de taille variable et des proportions variables de ressources génétiques et de matériel élite provenant de la REFPOP dans la population d'entrainement

CHAPITRE 5

Etude par simulations de deux schémas de transfert d'allèles favorables depuis des ressources génétiques vers du matériel élite en utilisant la sélection génomique

5.1 Introduction

Bien que la faible diversité génétique utilisée dans les programmes d'amélioration chez le pommier soit un problème identifié de longue date (Noiton et Shelbourne 1992), peu d'actions de pré-breeding ont été initiées dans le monde comme vu dans l'introduction générale (Kellerhalls et al. 2018 ; Kumar et al. 2010 ; Testolin et al. 2021). Le frein majeur à la mise en place de telles actions est le temps nécessaire avant d'obtenir des géniteurs de bon niveau agronomique provenant de matériel exotique ou de ressources génétiques : plus la divergence avec le matériel élite est importante et plus le nombre de générations d'amélioration de ce matériel sera important afin de retrouver un niveau agronomique acceptable. Par exemple Luo et al. (2019) considèrent que dans le cadre d'un croisement avec une espèce sauvage apparentée, 4 à 5 générations de pseudo-rétrocroisements sont nécessaires afin de ne conserver que 5% du génome du donneur et ainsi éliminer les défauts de qualité du fruit. Compte tenu de la durée de la phase juvénile chez le pommier, les sélectionneurs préfèrent se focaliser sur le progrès génétique à court terme et dans le cas d'utilisation de ressources génétiques, l'accent est mis sur quelques allèles favorables à effet fort introgressés par rétrocroisement dans le matériel élite, principalement pour la résistance aux maladies (cas de gènes majeurs de résistance à la tavelure ou à l'oïdium par exemple). Cette approche est néanmoins limitée pour des caractères dont le déterminisme génétique est plus complexe, d'une part car les variants causaux sont plus difficiles à détecter et

d'autre part parce que les tailles de population nécessaires pour introgresser de nombreux QTL à effets faibles en quelques générations ne sont pas envisageables pour des espèces fruitières compte tenu de la surface limitée des vergers. Pour de tels caractères, la sélection génomique est probablement une approche plus adaptée. Dans le chapitre précédent, nous avons montré qu'il était possible d'obtenir des précisions de prédiction modérées à élevées pour des hybrides « ressources génétiques x élite », ce qui pourrait ouvrir la voie à des actions de pré-breeding basées sur l'utilisation de la sélection génomique et donc une réduction de la durée des cycles de sélection. Sur une durée de 20 ans, l'utilisation de la sélection génomique pourrait ainsi permettre de passer de 3 à 5 cycles de sélection (Kumar et al. 2012b) et donc permettre un gain génétique par unité de temps plus important. Cependant, l'utilisation de la sélection génomique pourrait également conduire à une érosion plus rapide de la diversité et de la variance génétiques que dans le cas de la sélection phénotypique (Doublet et al. 2019), même en quelques générations (Rutkoski et al. 2015). Des stratégies ont été proposées afin de permettre un gain génétique tout en imposant une contrainte sur le maintien de la diversité (Woolliams et al. 2015) mais visent plutôt à optimiser le gain génétique à long terme que dans les premières générations (Allier et al. 2019a ; Santantonio et Robbins 2020). D'autre part, dans le cas de croisements entre ressources génétiques et matériel élite, plusieurs études (Gorjanc et al. 2016 ; Singh et al. 2021) ont montré que l'utilisation de la sélection génomique ou phénotypique tendait à favoriser les allèles provenant du matériel élite au détriment des allèles favorables des ressources génétiques, car ces derniers peuvent se trouver en répulsion avec des allèles favorables du matériel élite. De plus, les effets estimés des allèles rares dans les modèles de sélection génomique sont fortement « rétrécis (shrinkés) » vers 0 (Gianola 2013), minimisant leur poids dans le calcul des GEBV des candidats et donc entrainant un risque accru de perte de ces allèles (Jannink 2010). A terme, les allèles favorables provenant des ressources génétiques et qui sont initialement en fréquence faible dans le matériel élite pourraient donc être facilement perdus du fait de la dérive génétique.

Pour éviter la perte de ces allèles favorables dans les premières générations, il est possible de leur donner un poids supplémentaire afin d'accentuer leur importance dans le calcul des GEBV. Plusieurs formules de pondération basées sur les fréquences alléliques ont été proposées (Goddard 2009 ; Jannink 2010 ; Sun et VanRaden 2014) et plus récemment Liu et al. (2015) ont établi une formule de pondération intégrant également la génération à laquelle la pondération est appliquée afin que le poids donné aux allèles rares diminue au cours du temps. Pour tenir compte de l'incertitude quant à la position et aux effets des QTL, les formules de pondération doivent être basées sur les effets des marqueurs estimés par prédiction génomique.

L'objectif de ce chapitre est de comparer l'intérêt d'utiliser la sélection génomique plutôt que phénotypique dans le cas d'un programme de pré-breeding. Ramasubramanian et Beavis (2021) ont montré que le gain à court et long terme dépendait principalement du choix de la méthode

de sélection des parents, de la métrique utilisée pour réaliser cette sélection et du schéma de sélection. Nous avons dans notre cas simulé deux schémas de sélection utilisant comme matériel de base des ressources génétiques et du matériel élite et avons comparé sur 3 générations (soit environ 20 ans de sélection conventionnelle) différentes façons de choisir les parents de la génération suivante, en utilisant deux approches : la sélection génomique ou phénotypique. L'objectif du chapitre est de (1) comparer les performances des deux approches en termes de gain génétique et d'évolution de la variance génétique, (2) suivre le devenir des allèles favorables rares du matériel élite au cours des générations dans les deux cas. Pour mieux comprendre les résultats de simulation, nous avons par ailleurs mesuré la précision de prédiction qu'il est possible d'atteindre dans le cas de la sélection génomique.

5.2 Matériel et méthodes

L'approche par simulations présentée dans ce chapitre se décompose en trois étapes : dans un premier temps, nous avons simulé des données génomiques et phénotypiques pour 27 familles de plein-frères, considérées par la suite comme matériel élite, en nous appuyant sur le pédigrée connu de ces familles. Nous avons ensuite simulé 20 familles d'hybrides résultant du croisement d'individus provenant du panel « matériel élite » précédemment simulé et des individus provenant d'un panel de ressources génétiques. A partir de ces hybrides, nous avons alors simulé deux schémas de sélection couramment utilisés pour transférer des allèles vers du matériel élite : le premier schéma repose sur des intercroisements successifs entre les hybrides et le deuxième schéma sur l'utilisation de pseudo-rétrocroisements pour lesquels le parent récurrent provient de la population « matériel élite ». Dans ce dernier cas, nous parlerons désormais de pseudo-rétrocroisements par souci de simplification et noterons pBC1, pBC2 et pBC3 les trois générations étudiées, pBC étant l'abréviation de pseudo-backcross. De même, nous noterons IC1, IC2 et IC3 les trois générations étudiées dans le schéma des intercroisements successifs.

5.2.1 Simulation du matériel élite

5.2.1.1 Mise en œuvre des simulations

Nous avons effectué les simulations à l'aide du package R AlphaSimR (Gaynor et al. 2021). Compte tenu du grand nombre de simulations et de la taille des données génomiques, il nous a semblé indispensable d'utiliser un outil pouvant facilement gérer de gros jeux de données, ce qui est une des forces du package : les fonctions majeures sont écrites en C++ et le package est optimisé pour une utilisation minimale de la mémoire vive. De plus, l'utilisation d'un package R a permis de réutiliser des fonctions écrites pour simuler des données dans le chapitre 3. Enfin,

AlphaSimR est l'un des packages les plus utilisés pour simuler des schémas de sélection en amélioration des plantes (Bančič et al. 2021 ; Gorjanc et al. 2018 ; Muleta et al. 2019), ce qui nous a permis de rapidement prendre l'outil en main en nous appuyant sur le code utilisé dans certains de ces articles, en plus de la documentation abondante du package.

5.2.1.2 Utilisation des données des ressources génétiques

Le matériel élite a été simulé à partir des données phasées à haute densité disponibles pour les ressources génétiques (obtenues dans le chapitre 3 lors du phasage du panel de référence) afin de prendre en compte l'apparentement qui peut exister entre les deux populations. Parmi les 1341 individus génotypés à haute densité et provenant du panel des ressources génétiques, nous avons dans un premier temps identifié et retenu les génotypes pour lesquels aucun parent n'était connu, ce qui correspond à 721 génotypes dont les données génomiques pourraient être utilisées lors des simulations. L'utilisation des données choisies de la sorte réduit les chances de choisir aléatoirement des génotypes fortement apparentés lors du processus de simulation.

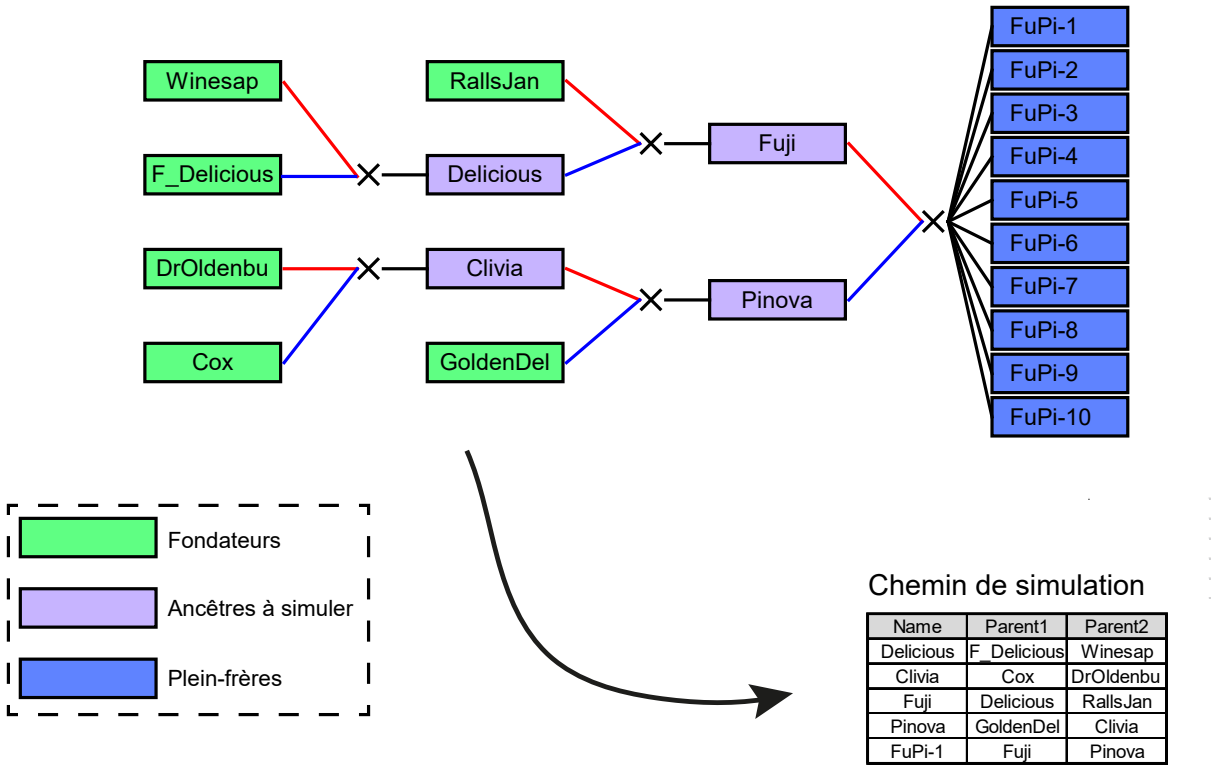
5.2.1.3 Utilisation des données de pédigrée

Les schémas de sélection chez le pommier n'ayant pas de régularité qui permettrait de généraliser les simulations, nous avons utilisé le pedigree connu des 27 familles de la REFPOP comme « squelette » pour les simulations. Pour rappel, le pédigrée des 27 familles est constitué de (généralement) 10 plein-frères qui représentent le matériel élite de la REFPOP et de tous les ancêtres connus de ces plein-frères. Pour chacune des familles, nous avons identifié les individus n'ayant pas de parent connu, que nous qualifierons désormais de « fondateurs » des pédigrées. Pour les 27 familles, nous avons identifié 57 fondateurs de la sorte. Nous avons ensuite attribué des données haute densité à chacun des fondateurs, en tirant aléatoirement et sans remise ces données parmi celles des 721 ressources génétiques sans parent connu. Soulignons que si un individu est identifié comme fondateur dans le pédigrée de deux familles distinctes, les données génomiques qui lui sont attribuées sont les mêmes dans les deux familles.

5.2.1.4 Chemin de simulation du matériel élite

Prenons l'exemple illustratif de la figure 5.1, qui représente une partie du pédigrée de la famille FuPi de la REFPOP. Les fondateurs du pédigrée sont identifiés en vert et les plein-frères de la famille en bleu. Les individus intermédiaires en violet correspondent à des ancêtres des plein-frères qui font le lien entre les fondateurs et les plein-frères et qu'il faut simuler à partir des données génomiques des fondateurs. Le chemin de simulation présenté sur la figure 5.1 correspond ainsi à l'ensemble des croisements permettant d'obtenir les plein-frères à partir des individus fondateurs. Nous avons écrit et utilisé une fonction R permettant d'identifier

les fondateurs et de renvoyer l'ensemble des croisements à simuler afin de pouvoir obtenir les plein-frères au sein de chaque famille.



Name	Parent1	Parent2
Delicious	F_Delicious	Winesap
Clivia	Cox	DrOldenbu
Fuji	Delicious	RallsJan
Pinova	GoldenDel	Clivia
FuPi-1	Fuji	Pinova

FIGURE 5.1 – Représentation d'une partie du pédigrée de la famille FuPi et chemin de simulation déterminé à partir de ce pédigrée

5.2.1.5 Simulation du matériel élite

Tout au long des simulations, nous avons simulé un seul caractère phénotypique purement additif et contrôlé par 50, 200 ou 500 QTL (répartis aléatoirement sur les 17 chromosomes) et avec une héritabilité de 0,2 ; 0,5 ou 0,8 dans la population des ressources génétiques. Il existe donc 9 combinaisons héritabilité/nombre de QTL possibles, que nous nommerons « modalités ». Pour chaque QTL, l'allèle positif a été choisi aléatoirement avec une probabilité de 0,5. Les effets des QTL ont été échantillonnés dans une loi Gamma de paramètre de forme $\alpha = 3$ et d'intensité $\beta = 1$ lorsque le caractère simulé était contrôlé par 50 QTL et de paramètres $\alpha = 1$ et $\beta = 1$ sinon, afin d'obtenir des distributions d'effets différentes selon le contrôle oligogénique ou polygénique du caractère. La figure S5.8 représente la distribution théorique des effets dans les deux cas.

Pour chaque croisement du chemin de simulation, nous avons simulé 1000 descendants, correspondant donc à 1000 plein-frères ne différant que du fait des recombinaisons des gamètes parentaux. Cette valeur de 1000 descendants par croisements a été retenue après comparaison des résultats obtenus en simulant 100 ou 10 000 descendants (voir Résultats). Nous avons attribué une valeur génétique additive à chaque individu simulé, correspondant à la somme des effets des allèles aux QTL portés par cet individu. Nous avons également attribué une valeur phénotypique à chaque individu, cette valeur étant égale à la valeur génétique additive à laquelle est ajoutée un bruit dépendant de l'héritabilité simulée. Une fois les 1000 plein-frères simulés, nous avons conservé celui ayant la valeur phénotypique la plus élevée et ses données génomiques ont été utilisées dans le reste des simulations. Sur le pédigrée de la famille FuPi présenté figure 5.1, nous avons par exemple commencé par simuler 1000 plein-frères résultant du croisement entre Winesap et F_Delicious et avons dénommé « Delicious » le plein-frère ayant la valeur phénotypique la plus élevée parmi ces 1000 plein-frères. Les données génomiques de ce plein-frère sont dès lors utilisées pour chaque croisement faisant intervenir Delicious.
Cette approche nous a permis de simuler des données génomiques pour les parents des 27 familles de la REFPOP. Une fois ces données simulées, nous avons simulé 10 plein-frères par famille (soit le nombre moyen de plein-frères dans la REFPOP), pour lesquels une valeur génétique additive et une valeur phénotypique ont également été calculées comme décrit ci-dessus. Les 270 individus obtenus de cette manière seront désormais dénommés « matériel élite » et les 721 individus de la population des ressources génétiques sans parent connu utilisés pour échantillonner les 57 fondateurs des pédigrées seront dénommés « ressources génétiques ».

5.2.1.6 Validation de la qualité des simulations

Dans le chapitre 4, nous avons montré qu'il existait des différences quantifiables entre les ressources génétiques et le matériel élite de la REFPOP :

- Le déséquilibre de liaison décroit plus rapidement chez les ressources génétiques que chez le matériel élite
- Du fait de la sélection, certains allèles sont présents en fréquence élevée (voire sont fixés) chez le matériel élite mais sont présents en fréquence plus faible chez les ressources génétiques
- La distribution des valeurs phénotypiques diffère pour la plupart des caractères entre les deux groupes

Afin de nous assurer que le matériel élite simulé en utilisant notre approche était effectivement différencié par rapport aux ressources génétiques, nous avons vérifié ces trois points sur les données simulées :

- Nous avons mesuré si la décroissance du déséquilibre de liaison était significativement différente entre ressources génétiques et matériel élite simulés en utilisant un t-test ap-

parié comparant le déséquilibre de liaison moyen au sein de fenêtres définies de la même façon qu'au chapitre 4

— Nous avons comparé les fréquences des allèles favorables et défavorables chez les ressources génétiques et le matériel élite simulés

— Nous avons comparé la moyenne des valeurs génétiques chez les ressources génétiques et le matériel élite simulés

Nous avons en outre calculé la valeur de F_{ST} entre les ressources génétiques et le matériel élite dans nos simulations afin de comparer la valeur obtenue à la valeur de F_{ST} obtenue sur données réelles.

5.2.2 Simulation des hybrides

Nous avons simulé 1000 individus répartis en 20 familles d'hybrides constituées de 50 plein-frères résultant du croisement entre un individu parent provenant de la population « matériel élite » et un individu parent provenant de la population « ressources génétiques ». Nous nommerons ces hybrides « hybrides RG x E ». Nous avons comparé trois approches pour choisir les parents de chaque famille :

— Méthode Pheno : les parents sont choisis sur la base de leur valeur phénotypique. Plus précisément, les 20 individus provenant de la population des ressources génétiques ayant la valeur phénotypique la plus élevée sont choisis d'une part, et les 20 individus provenant de la population du matériel élite ayant la valeur phénotypique la plus élevée sont choisis d'autre part, en choisissant au maximum un individu par famille

— Méthode BV : les parents sont choisis sur la base de leur valeur génétique additive (« Breeding Value »). L'approche est la même que pour la méthode « pheno » mais les valeurs génétiques additives sont utilisées à la place des phénotypes.

— Méthode Random : les parents sont choisis aléatoirement. 20 individus provenant de la population des ressources génétiques et 20 individus provenant de la population du matériel élite sont choisis aléatoirement, en choisissant au maximum un individu par famille. Les 20 familles sont elles aussi choisies aléatoirement parmi les 27.

Nous avons ensuite calculé la matrice d'apparentement pour les 20+20 individus choisis de la sorte à l'aide du package kinship2 (Sinnwell et al. 2014) et établi un plan de croisement reposant sur la minimisation globale de l'apparentement entre les ressources génétiques et le matériel élite utilisés comme géniteurs. Pour ce faire, nous avons utilisé un algorithme du voyageur appliqué à la matrice d'apparentement à l'aide du package TSP (Hahsler et Hornik 2007) afin de trouver les 20 couples de parents ayant un apparentement minimum. Nous avons alors simulé 50 plein-frères par famille, ayant chacun une valeur génétique additive et phénotypique calculée de la même façon que précédemment décrit.

5.2.3 Simulation de générations avancées

Une fois les hybrides simulés, nous avons comparé deux approches couramment utilisées en amélioration variétale : dans le premier cas, nous avons simulé 3 générations d'intercroisements entre les hybrides provenant des 20 familles et dans le second cas, nous avons simulé 3 générations de pseudo-rétrocroisement, en utilisant les hybrides comme parents donneurs et les individus élite comme parents récurrents.
Dans les deux cas, les parents de la génération suivante ont été choisis selon les trois méthodes évoquées précédemment (choix aléatoire, basé sur les valeur génétiques ou phénotypiques), ou sur la base des GEBV des candidats obtenues par prédiction génomique. Nous nommerons cette méthode GS.
A chaque génération, nous avons ainsi retenu comme parents de la génération suivante les candidats pour lesquels l'estimateur de la valeur génétique (le phénotype dans le méthode Pheno, la vraie valeur génétique dans la méthode BV et la GEBV pour la méthode GS) présentait l'écart le plus important à la moyenne de leur famille. Nous avons dans un premier temps choisi 20 hybrides de cette manière, qui ont été intercroisés pour donner 20 familles de 50 individus IC1 dans le cas de la simulation des intercroisements, ou qui ont été croisés aux 20 individus du matériel élite sélectionnés comme parents des hybrides afin de donner 20 familles de 50 pseudo-rétrocroisements dans la simulation des pseudo-rétrocroisements. Tout comme lors de la simulation des hybrides, le choix des croisements des parents a été effectué sur la base de la matrice d'apparentement (ou de la GRM dans le cas de la méthode GS) et de l'algorithme du voyageur afin de minimiser l'apparentement des parents choisis. Dans le cas de la méthode Random, les parents sont croisés de façon totalement aléatoire. Nous avons ensuite répété cette approche pour les générations ultérieures en intercroisant les 20 meilleurs individus IC1 (puis IC2) pour obtenir 20 familles de 50 individus IC2 (puis IC3) dans le cas des intercroisements et en rétrocroisant les 20 meilleurs individus pBC (puis pBC2) pour obtenir 20 familles de 50 individus pBC2 (puis pBC3). L'approche globale employée lors des simulations est récapitulée figure 5.2.

5.2.4 Prédiction génomique pour les générations avancées

5.2.4.1 Constitution de la population d'entrainement et précision de prédiction

A partir de la génération suivant les hybrides RG x E, nous avons également introduit une méthode de sélection des parents basée sur la prédiction génomique. A chaque génération, les valeurs génétiques additives des 50 plein-frères des 20 familles ont été prédites grâce à un modèle RR-BLUP implémenté dans le package AlphaSimR. Nous avons comparé deux façons de construire une population d'entrainement de 500 individus, c'est-à-dire une taille proche de

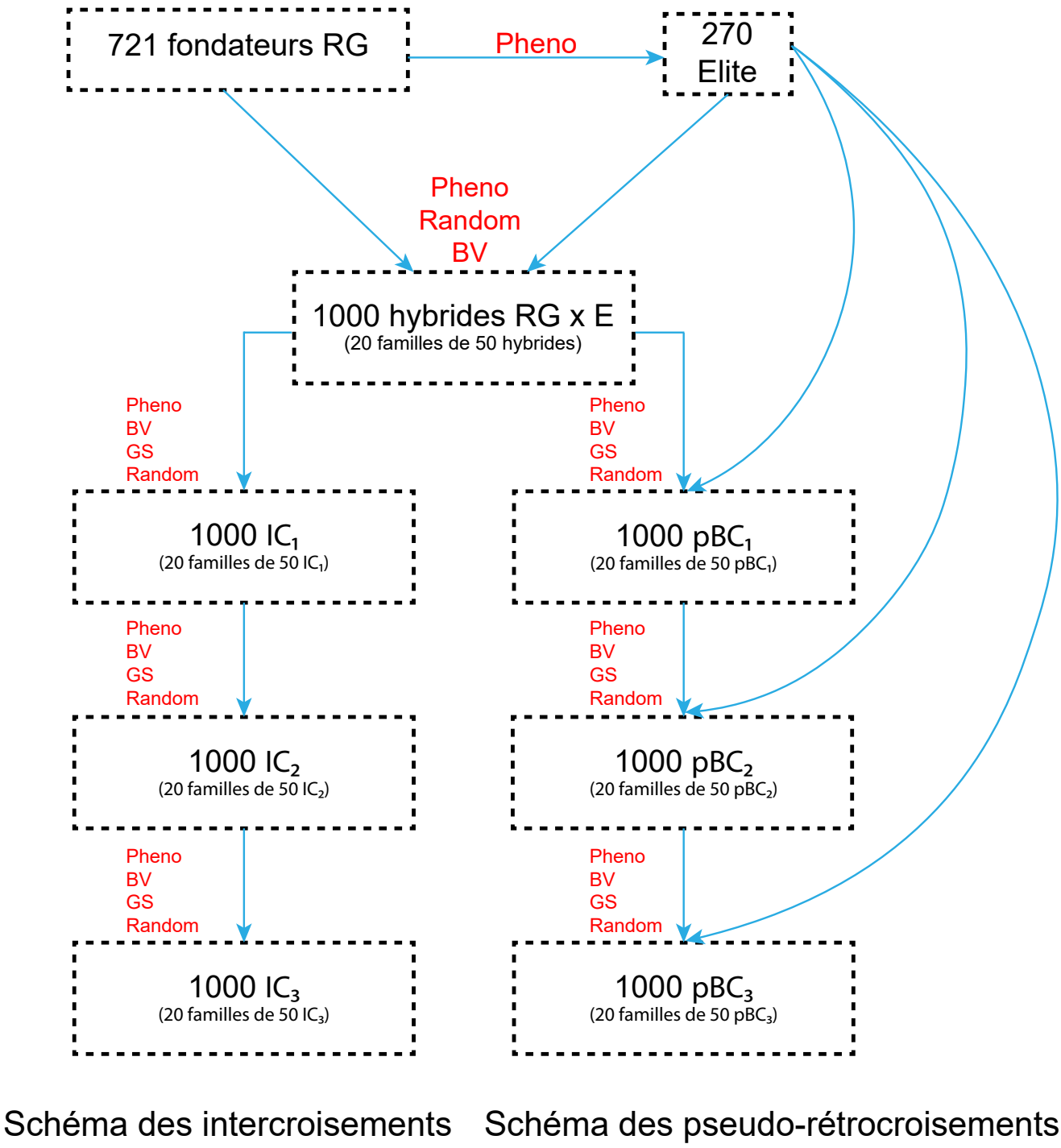


FIGURE 5.2 – Schéma récapitulatif de l'approche utilisée pour les simulations. Les flèches bleues représentent l'origine des parents de la génération suivante. Les méthodes de sélection des parents sont précisées en rouge

celle de la REFPOP. Dans le premier cas, nous avons utilisé une population d'entrainement composée des 270 individus élite simulés à partir des ressources génétiques dans la première

étape et nous avons ensuite ajouté 230 génotypes provenant des ressources génétiques à ce premier ensemble pour constituer une population d'entrainement de 500 individus, soit une population avec un effectif et une répartition proches de la REFPOP. Pour chaque simulation, ces 230 génotypes sont constitués des 57 fondateurs ayant servi à la simulation des individus élite auxquels sont ajoutés 173 génotypes permettant de maximiser la diversité génétique de ce set de 230 individus. Les 173 génotypes restants ont été choisis grâce à la fonction core.set du package R GenoCore (Jeong et al. 2017). Cette population d'entrainement de 500 génotypes a ensuite été utilisée pour prédire les candidats quelle que soit la génération étudiée. Nous appelons cette méthode de constitution de la population d'entrainement TS_no_update. Dans le second cas, nous avons constitué une population d'entrainement dépendant de la génération des candidats. Suivant les recommandations de Neyhart et al. (2017), nous avons constitué la population d'entrainement pour prédire les candidats de la génération n en regroupant les 250 génotypes de la génération précédente ayant la valeur de GEBV la plus élevée et les 250 génotypes de la génération précédente ayant la valeur de GEBV la moins élevée. Afin de maximiser la précision de prédiction intra-famille, ces deux groupes ont été constitués en choisissant les génotypes au sein de chaque famille. Pour ce faire, nous avons créé deux groupes constitués des 13 candidats de chaque famille de la génération n-1 ayant la valeur de GEBV la plus élevée et les 13 candidats de chaque famille de la génération n-1 ayant la valeur de GEBV la moins élevée (soit 13 x 20 familles x 2 groupes = 520 génotypes) et avons ensuite exclu aléatoirement dix génotypes de chaque groupe en nous assurant qu'au maximum un génotype par famille soit exclu afin qu'exactement 25 génotypes par famille constituent la population d'entrainement. Nous appelons cette méthode TS_update. Nous avons enfin étudié un troisième cas pour lequel la taille de la population d'entrainement dépend de la génération des candidats. Dans ce cas, la population d'entrainement a été constituée en regroupant toutes les populations d'entrainement des générations précédentes de la méthode TS_update. Par exemple, lorsque les candidats correspondent à une génération de pBC3, la population d'entrainement est constituée de 2000 génotypes (230 ressources génétiques, 270 génotypes élite, 500 hybrides RG x E, 500 pBC et 500 pBC2). Nous appelons cette méthode TS_update_inc. Pour les trois méthodes étudiées, nous avons mesuré la précision de prédiction intra-famille comme la corrélation entre la vraie valeur génétique des individus d'une famille et leur GEBV et nous définissons pour chaque simulation une précision de prédiction intra-famille moyenne égale à la moyenne des précisions de prédiction intra-famille obtenues sur les 20 familles.

5.2.4.2 Pondération des effets estimés en fonction des fréquences alléliques

Afin de donner davantage de poids aux allèles favorables initalement rares dans le matériel élite, nous avons appliqué les mêmes prédictions que celles décrites au paragraphe précédent mais en pondérant les effets des marqueurs estimés par le modèle RR-BLUP en fonction des fréquences

alléliques dans le matériel élite. Nous avons utilisé la formule proposée par Liu et al. (2015) afin de créer un vecteur des ponderations w, tel que :

$$w = \frac{p^{(\alpha' - 1 + t * \frac{(1-\alpha')}{N})}}{B(\alpha,1)}$$

avec p le vecteur des fréquences alléliques dans le matériel élite, N le nombre de générations du programme d'amélioration pendant lesquelles un poids sera appliqué aux effets estimés des marqueurs (N = 4 dans notre cas), t la génération à laquelle ce poids est appliqué, B la fonction Beta, α' une constante permettant de déterminer les valeurs initiales maximales de w et $\alpha = \alpha' + t * \frac{(1-\alpha')}{N}$. L'utilisation du paramètre t permet de donner un poids fort aux allèles rares dans les premières générations de sélection puis de tendre vers un poids égal pour tous les marqueurs dans les générations ultérieures comme présenté figure S5.9. Dans ce cas, les GEBV obtenues pour chaque candidat à la sélection correspondent à la somme des effets pondérés. Nous avons utilisé une valeur α' égale à 0,01 afin de donner un poids très fort aux allèles rares dans le matériel élite. Ainsi, l'effet estimé d'un allèle favorable avec une fréquence de 0,05 dans le matériel élite sera multiplié par 9,4 à la génération 1 par rapport à un modèle sans pondération, puis par 4,4 à la génération 2 et par 2,1 à la génération 3. Nous nommerons ce type de pondération « pondération dynamique » car le poids donné aux allèles dépend de la génération à laquelle la pondération est appliquée.

5.2.5 Analyse des résultats

Afin d'évaluer l'efficacité du transfert d'allèles rares dans le matériel élite depuis le panel des ressources génétiques, nous avons pour chaque simulation identifié tous les QTL pour lesquels la fréquence de l'allèle favorable dans le matériel élite était inférieure à 10% (que nous nommerons plus simplement allèles rares dans la suite de ce chapitre) et suivi l'évolution de la fréquence de ces allèles au cours des générations. Nous rapportons ainsi ci-après la fréquence moyenne des allèles rares au cours des générations, la proportion moyenne d'allèles favorables perdus au cours des générations ainsi que la proportion moyenne d'allèles favorables pour lesquels la fréquence est supérieure à 10% et est supérieure à la fréquence observée dans les ressources génétiques. Lorsque 50 QTL étaient simulés, nous avons exclu de l'analyse des allèles rares les simulations pour lesquelles moins de 5 allèles rares étaient présents, ce qui a conduit à l'exclusion des résultats de 21 des 50 simulations pour $h^2 = 0{,}2$ et 24 des 50 simulations pour $h^2 = 0{,}5$ et $h^2 = 0{,}8$.

Pour chaque simulation, nous avons également calculé la proportion moyenne d'allèles favorables ainsi que la moyenne des valeurs génétiques additives et la variance génétique au cours des générations. Afin de pouvoir comparer les valeurs génétiques additives d'une simulation à l'autre, la valeur génétique additive de chaque individu d'une simulation a été divisée par la

valeur génétique additive maximale qu'il était possible d'atteindre dans cette même simulation, soit la valeur génétique d'un individu portant l'allèle favorable à l'état homozygote pour chaque QTL simulé. La moyenne des valeurs génétiques additives étant par construction égale à 0 pour les ressources génétiques, les valeurs rapportées pour les générations ultérieures correspondent alors à l'augmentation de la valeur génétique additive standardisée par rapport aux ressources génétiques. Nous nommerons dès lors la moyenne de ces valeurs « valeur génétique standardisée » par souci de concision. De même, la variance génétique étant fixée à 1 pour les ressources génétiques dans toutes les simulations, les valeurs rapportées pour les générations ultérieures peuvent être interprétées comme la proportion de variance disponible par rapport à la variance initiale.

Nous avons enfin mesuré pour chaque génération la proportion du génome des 1000 individus simulés qui provenait du matériel élite, ainsi que la proportion des allèles aux QTL provenant du matériel élite. Afin d'évaluer si les différentes méthodes de sélection étudiées pouvaient conduire à un taux de consanguinité plus élevé, nous avons également mesuré la proportion de marqueurs hétérozygotes à chaque génération. Nous avons effectué 50 simulations pour chaque modalité et chaque valeur rapportée pour un indicateur et une modalité correspond à la moyenne de cet indicateur obtenue sur les 50 simulations. Pour chaque indicateur mesuré de la sorte, nous avons comparé à l'aide d'un test HSD de Tukey si la différence entre les méthodes en troisième génération était siginficative au seuil $\alpha = 5\%$. Pour ce faire, nous avons utilisé la fonction tukey_hsd du package rstatix (Kassambara 2021).

5.3 Résultats

5.3.1 Validation des résultats de simulation du matériel élite

Avant de simuler des générations avancées, nous avons voulu nous assurer que les ressources génétiques et le matériel élite simulés présentaient des caractéristiques comparables à ce qui est observé sur données réelles. Les différences majeures entre ces deux groupes concernent le déséquilibre de liaison, la fréquence des allèles favorables et défavorables fixés, ainsi que les valeurs génétiques et phénotypiques observées pour différents caractères. Nous avons donc évalué ces différences sur 100 simulations. Le déséquilibre de liaison était significativement différent entre le matériel élite et les ressources génétiques dans les 100 simulations, bien que les valeurs moyennes de r^2 à différentes distances soient systématiquement inférieures aux valeurs observées sur données réelles (figure S5.10).

La moyenne des valeurs génétiques ainsi que la fréquence moyenne des allèles favorables et défavorables fixés sont présentées dans le tableau 5.1 pour les ressources génétiques et le matériel élite. Nous rapportons également le FST moyen estimé entre ressources génétiques et matériel

élite à partir des 100 simulations. Nous présentons les résultats obtenus en simulant 100, 1000 ou 10000 descendants pour chaque croisement du chemin de simulation. Une simulation prend dans ce cas respectivement 20 secondes, une minute ou 10 minutes en moyenne, indépendamment du nombre de QTL ou de l'héritabilité simulés. La valeur génétique moyenne est supérieure chez le matériel élite par rapport aux ressources génétiques, et ce d'autant plus que l'héritabilité du caractère simulé est élevée. De même, la fréquence des allèles favorables fixés est plus élevée pour le matériel élite que pour les ressources génétiques et la fréquence des allèles défavorables est en moyenne moins élevée pour toutes les modalités. Comme observé sur données réelles, la différenciation entre ressources génétiques et matériel élite mesurée par le FST est faible dans les simulations, bien que les valeurs obtenues (allant de 0,012 à 0,015) soient inférieures aux valeurs obtenues sur données réelles à partir des jeux de données FBo-Hi (0,023) et REFPOP (0,021).
Au regard des résultats obtenus, nous avons validé l'approche présentée pour simuler du matériel élite et avons choisi une valeur de 1000 descendants par croisement du chemin de simulation, ce qui constitue un compromis entre la différenciation ressources génétiques/matériel élite et le temps de calcul des simulations.

5.3.2 Valeur génétique moyenne et variance génétique dans les générations avancées

Dans toutes les simulations, la valeur génétique standardisée la plus élevée à chaque génération a été obtenu lorsque les parents ont été choisis sur la base de leur vraie valeur génétique additive, alors que la valeur génétique standardisée la plus faible et la variance génétique la plus élevée ont été obtenues lorsque les parents ont été choisis aléatoirement. Nous ne mentionnerons donc par la suite que les cas de figure où les parents ont été choisis sur la base de leur valeur phénotypique ou sur la base des prédictions génomiques. Dans les deux schémas simulés, nous avons observé une augmentation de la valeur génétique standardisée lorsque l'héritabilité du caractère simulé augmentait et une diminution lorsque le nombre de QTL contrôlant le caractère simulé augmentait (tableau 5.2). La valeur génétique standardisée à chaque génération était plus élevée dans le schéma des intercroisements que dans celui des pseudo-rétrocroisements, quels que soient la méthode de sélection des parents utilisée, l'héritabilité du caractère simulé et le nombre de QTL contrôlant ce caractère (figures 5.3 et 5.4). De plus, la différence de valeur génétique standardisée en troisième génération entre deux méthodes de sélection des parents n'était jamais significative dans le schéma des pseudo-rétrocroisements. Dans la suite du chapitre, nous mettrons donc l'accent sur les résultats relatifs au schéma des intercroisements et illustrons les résultats associés pour le schéma des pseudo-rétrocroisements en annexe. Dans le schéma des intercroisements, la méthode TS_update_inc a permis d'obtenir une valeur gé-

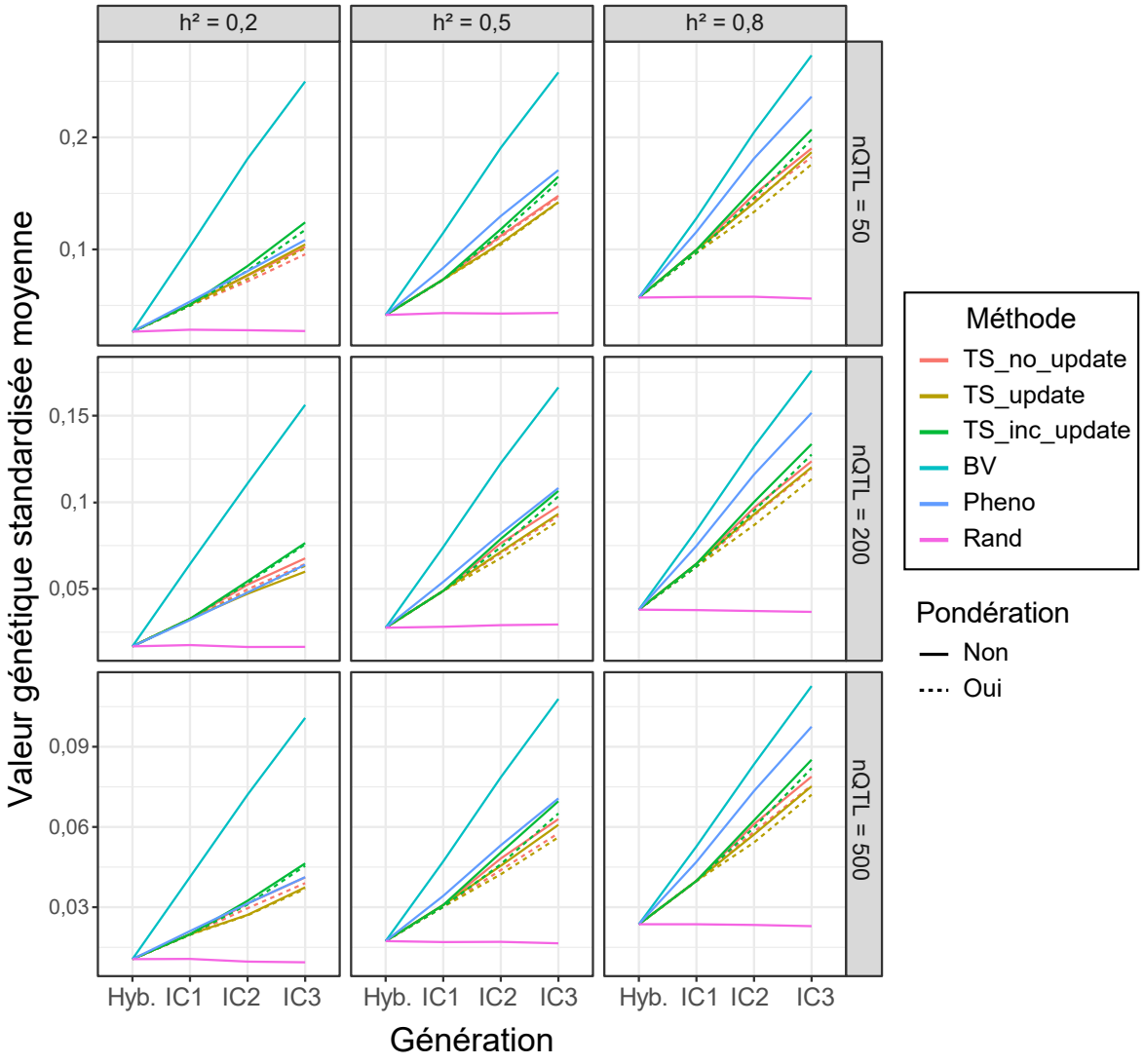


FIGURE 5.3 – Valeur génétique standardisée au cours des générations dans le schéma des intercroisements en fonction de la méthode de sélection des parents. Les prédictions ont été réalisées en utilisant les données haute densité. Notez que l'échelle de l'axe des y varie en fonction du nombre de QTL

nétique standardisée en troisième génération plus élevée qu'avec les méthodes TS_no_update et TS_update pour toutes les modalités, et ce d'autant plus que l'héritabilité était faible : pour $h^2 = 0{,}2$ la valeur génétique standardisée obtenue avec la méthode TS_update_inc était en moyenne supérieure de 24% à la valeur génétique standardisée obtenue avec la méthode TS_no_update (17% pour $h^2 = 0{,}5$ et 12% pour $h^2 = 0{,}8$). Les valeurs génétiques standardisées obtenues avec les méthodes TS_no_update et TS_update n'étaient par ailleurs jamais significativement différentes. A faible héritabilité ($h^2 = 0{,}2$), la valeur génétique standardisée était également plus élevée en troisième génération avec la méthode TS_update_inc qu'avec la méthode Pheno (valeur génétique standardisée supérieure de 15% en moyenne). Pour $h^2 = 0{,}5$, aucune différence significative n'a pu être observée entre les deux méthodes, alors que la

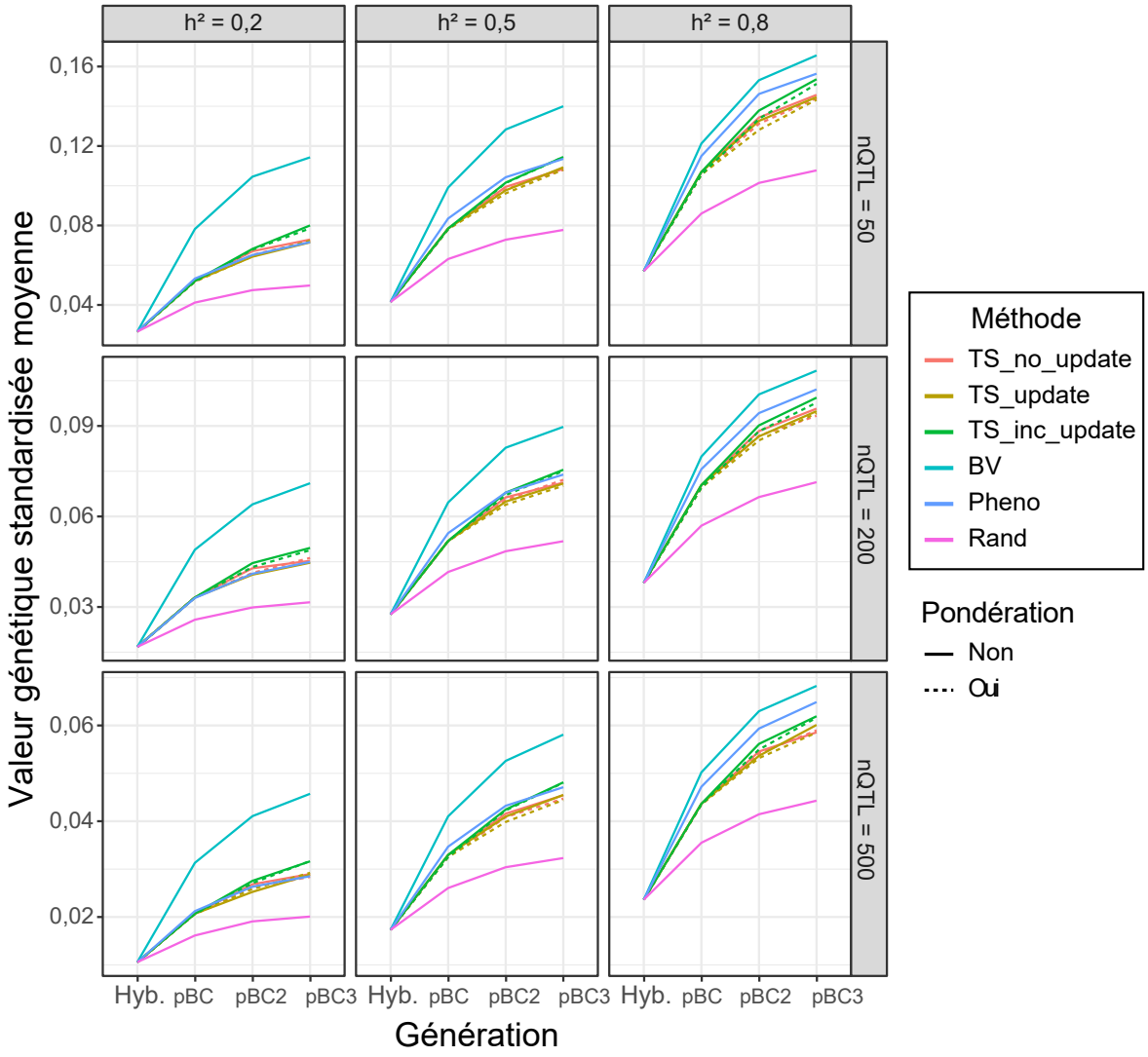


FIGURE 5.4 – Valeur génétique standardisée au cours des générations dans le schéma des pseudo-rétrocroisements en fonction de la méthode de sélection des parents (prédictions haute densité). Notez que l'échelle de l'axe des y varie en fonction du nombre de QTL

méthode Pheno a permis d'obtenir une valeur génétique standardisée plus élevée pour $h^2 = 0{,}8$ (valeur génétique standardisée supérieure de 20% en moyenne). Les valeurs génétiques standardisées moyennes obtenues avec les méthodes TS_no_update, TS_update ou Pheno n'étaient pas significativement différentes pour $h^2 = 0{,}2$. En revanche, lorsque $h^2 = 0{,}5$ ou $h^2 = 0{,}8$, la méthode Pheno a conduit à une valeur génétique standardisée en troisième génération plus élevée qu'avec les méthode TS_no_update et TS_update. Notons enfin qu'aucune différence n'a pu être observée entre les méthodes basées sur la pondération dynamique et la méthode équivalente sans pondération pour toutes les modalités. La proportion moyenne d'allèles favorables en troisième génération suit exactement les mêmes tendances que celles décrites ci-dessus pour la valeur génétique standardisée (figure S5.11) Dans le schéma des intercroisements, nous avons observé une diminution de la variance génétique au cours des générations (figure S5.12). Cette

diminution de la variance était indépendante de l'héritabilité mais était d'autant plus marquée que le nombre de QTL contrôlant le caractère était faible. La variance génétique moyenne en IC3 était la même entre la méthode GS et Pheno (tableau 5.3), sauf pour $h^2 = 0{,}2$ et 50 ou 200 QTL, cas pour lesquels la variance génétique moyenne après utilisation de la méthode TS_update_inc était significativement moins élevée que la variance génétique moyenne après utilisation de la méthode Pheno, et inversement pour $h^2 = 0{,}8$ et 500 QTL, cas pour lequel la variance génétique moyenne après utilisation de la méthode Pheno était significativement moins élevée que la variance génétique moyenne après utilisation des méthodes basées sur la prédiction génomique. Là encore, aucune différence significative n'a pu être observée pour une méthode donnée entre l'utilisation ou non de la pondération dynamique.

5.3.3 Précision de prédiction dans les générations avancées

Dans les deux schémas de sélection simulés, la précision de prédiction pour les différentes générations de candidats augmente lorsque l'héritabilité du caractère simulé augmente (par exemple dans le schéma des intercroisements, augmentation moyenne de $0{,}12 \pm 0{,}02$ entre $h^2 = 0{,}2$ et $h^2 = 0{,}5$ et de $0{,}15 \pm 0{,}01$ entre $h^2 = 0{,}5$ et $h^2 = 0{,}8$ à 50 QTL et de $0{,}08 \pm 0{,}02$ entre $h^2 = 0{,}2$ et $h^2 = 0{,}5$ à 500 QTL et de $0{,}1 \pm 0{,}01$ entre $h^2 = 0{,}5$ et $h^2 = 0{,}8$ à 200 ou 500 QTL). Pour une héritabilité donnée, les précisions de prédiction moyennes n'étaient pas significativement différentes lorsque le caractère était contrôlé par 200 ou 500 QTL mais étaient légèrement supérieures aux valeurs obtenues pour 50 QTL (en passant de 50 à 200 ou 500 QTL, augmentation moyenne de $0{,}03 \pm 0{,}02$ à $h^2 = 0{,}2$, de $0{,}05 \pm 0{,}02$ à $h^2 = 0{,}5$ et de $0{,}07 \pm 0{,}02$ à $h^2 = 0{,}8$ dans les deux schémas simulés).

Dans les deux schémas de sélection simulés, les précisions de prédiction moyennes les plus élevées ont toujours été obtenues avec la méthode TS_update_inc (figures S5.13 et 5.5) et étaient modérées à élevées selon la génération étudiée (0,3 et 0,37 en première génération à $h^2 = 0{,}2$ et 0,8 et 0,77 en troisième génération à $h^2 = 0{,}8$ respectivement pour le schéma des pseudo-rétrocroisements et des intercroisements). Dans le schéma des pseudo-rétrocroisements (figure S5.13), la précision de prédiction moyenne reste constante au cours des générations avec la méthode TS_no_update, et augmente légèrement au cours des générations avec la méthode TS_update. Dans le schéma des intercroisements (figure 5.5), la précision de prédiction moyenne a diminué au cours des générations avec la méthode TS_no_update, et ce d'autant plus que l'héritabilité du caractère simulé est élevée. La précision de prédiction moyenne augmente légèrement au cours des générations avec la méthode TS_update et augmente de façon plus marquée avec la méthode TS_update_inc. Les précisions de prédiction moyennes obtenues avec les deux méthodes ne sont pas significativement différentes sauf lorsque l'héritabilité du caractère simulé est de 0,8. Comme dans le schéma des intercroisements, la précision de pré-

diction moyenne augmente au cours des générations avec la méthode TS_update_inc, surtout entre la génération des hybrides et la première génération de pseudo-rétrocroisements.

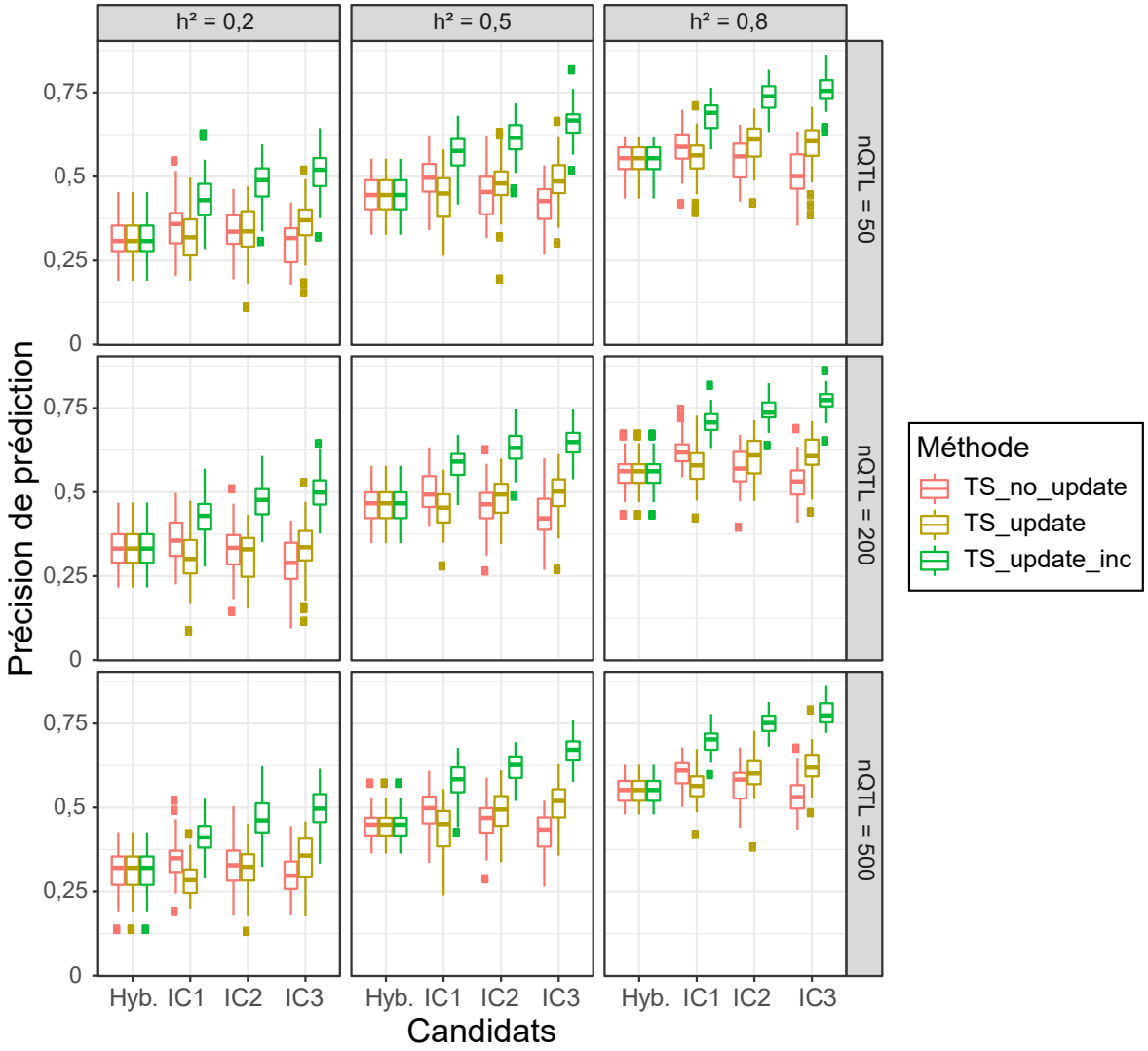


FIGURE 5.5 – Précision de prédiction moyenne intra-famille dans le schéma des intercroisements en fonction de la méthode de constitution de la population d'entrainement

5.3.4 Evolution de la fréquence des allèles rares au cours des générations

Sur les 50 simulations, les allèles rares représentaient en moyenne près de 10% des QTL (5 allèles rares en moyenne pour 50 QTL simulés, 18 allèles rares pour 200 QTL simulés et 49 allèles rares pour 500 QTL simulés). Les résultats de simulation montrent que la fréquence moyenne des allèles rares a suivi une trajectoire opposée dans les deux schémas de sélection simulés. Dans le schéma des pseudo-rétrocroisements, nous avons observé une diminution de la fréquence des allèles rares à chaque génération dans toutes les simulations. La fréquence

moyenne des allèles rares en troisième génération avoisinait 0,05 pour toutes les modalités, et l'utilisation de la pondération dynamique n'a pas eu d'effet sur la fréquence moyenne des allèles rares. Nous avons en revanche observé une augmentation de la fréquence moyenne des allèles rares au cours des générations dans le schéma des intercroisements, la sélection phénotypique et les méthodes basées sur la pondération dynamique donnant toujours des résultats supérieurs à ceux observés lorsque les méthodes basées sur la prédiction génomique seule étaient utilisées (figure 5.6 et tableau 5.4). Pour toutes les méthodes étudiées, nous n'avons observé d'effet ni de l'héritabilité ni du nombre de QTL sur la fréquence moyenne des allèles rares en troisième génération. De même, aucune différence n'a été observée entre les méthodes étudiées quant au nombre d'allèles rares favorables perdus au cours des générations (figure S5.14).

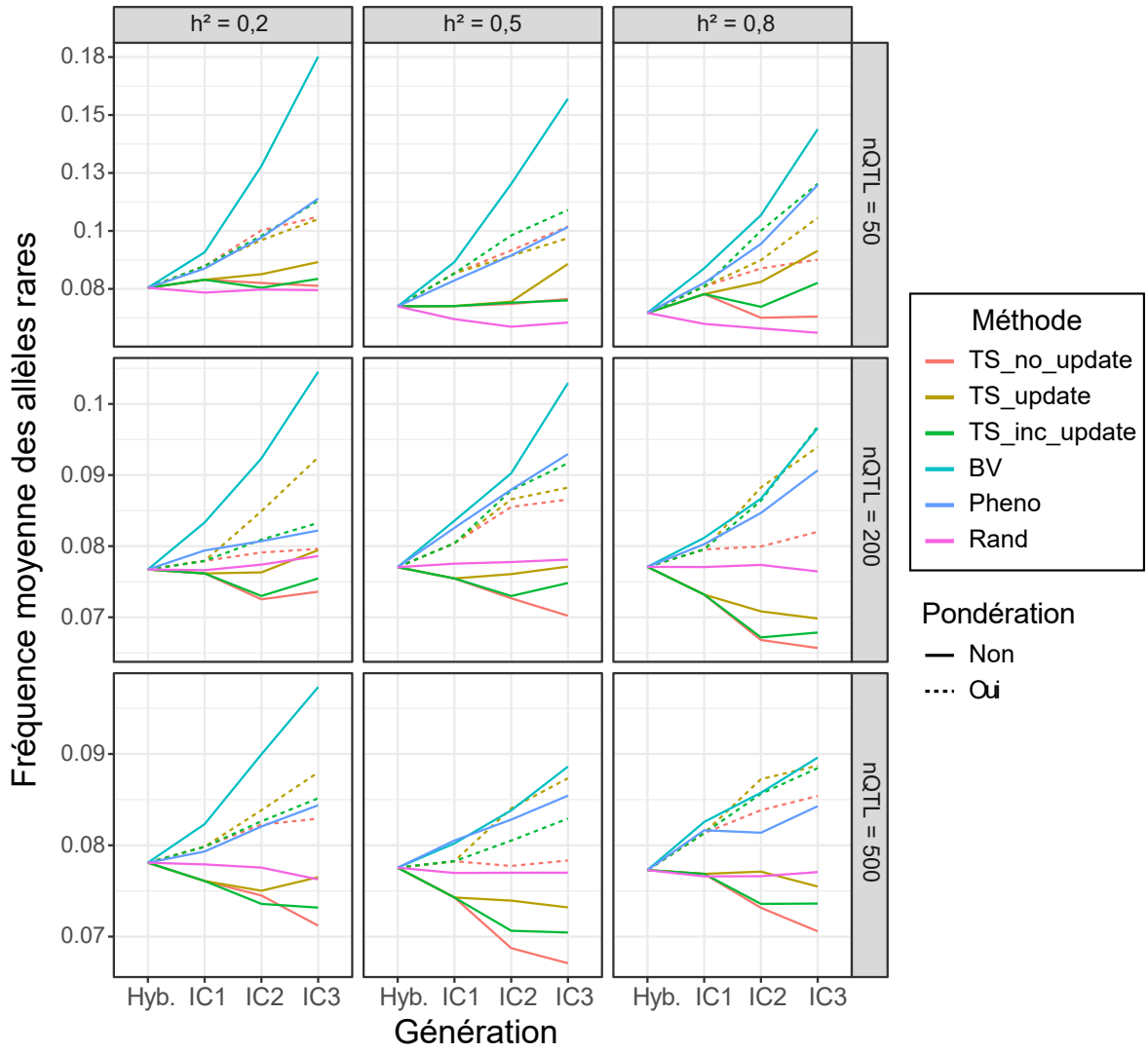


FIGURE 5.6 – Fréquence moyenne des allèles favorables rares au cours des générations dans le schéma des intercroisements en fonction de la méthode de sélection des parents

5.4 Discussion

Dans ce chapitre, nous nous sommes intéressés à la mise en place de schémas de sélection initiés à partir d'un panel de ressources génétiques avec l'objectif d'améliorer la valeur agronomique des génotypes les plus prometteurs en les croisant avec un panel constitué de matériel élite. Dans ce cas, nous avons cherché à savoir si l'utilisation de la sélection génomique pouvait présenter un avantage par rapport à une sélection phénotypique classique. Nous avons également évalué l'évolution de la diversité génétique et en particulier les changements de fréquence des allèles favorables rares du matériel élite, avec l'objectif d'augmenter la fréquence de tels allèles au cours du temps. Le travail présenté est basé sur l'analyse de résultats de simulations pour différentes architectures génétiques.

5.4.1 Intérêt de la sélection génomique pour des actions de pré-breeding chez le pommier

Dans nos simulations, la valeur génétique moyenne des différentes générations était toujours supérieure en appliquant la sélection phénotypique plutôt que la sélection génomique, sauf dans le cas où l'héritabilité du caractère simulé était de 0,2. Dans ce cas, l'utilisation d'une population d'entrainement comprenant des données phénotypiques de toutes les générations précédentes (méthode TS_update_inc) a permis d'obtenir des valeurs génétiques moyennes plus élevées en troisième génération que si les parents avaient été choisis sur la base de leur valeur phénotypique. Même si le gain génétique par génération est donc comparable entre sélection phénotypique et génomique, voire en faveur de la sélection phénotypique, le gain génétique par unité de temps est plus important lorsque la sélection génomique est utilisée du fait de la réduction de l'intervalle de temps entre deux générations. Nous illustrons cette idée figure 5.7 en considérant que l'intervalle entre deux générations est de quatre ans dans le cas de l'utilisation de la sélection génomique et de sept ans dans le cas de l'utilisation de la sélection phénotypique (Kumar et al. 2012b). Les simulations montrent également que plus la précision de prédiction est élevée et plus la valeur génétique standardisée moyenne au bout de trois générations est importante. De ce fait, la méthode TS_update_inc permet d'obtenir un gain génétique plus élevé que les méthodes TS_no_update et TS_update en troisième génération, ces deux dernières méthodes produisant des résultats similaires en termes de gain génétique et de précision de prédiction. La précision de prédiction augmente généralement lorsque la taille de la population d'entrainement augmente (Edwards et al. 2019 ; Norman et al. 2018), ce qui justifie ici l'intérêt de la méthode TS_update_inc quels que soient l'héritabilité du caractère simulé et le nombre de QTL contrôlant ce caractère. Ce résultat peut sembler en contradiction avec les observations présentées au chapitre 4, qui mettaient en avant le gain limité de précision de

prédiction obtenu en augmentant la taille de la population d'entrainement sur données réelles. Il faut toutefois noter que les précisions mesurées dans ce chapitre correspondent à la corrélation entre les GEBV et les TBV, qui ne sont connues que dans le cadre des simulations alors qu'elles sont uniquement approchées par les valeurs phénotypiques dans le chapitre 4. Nous pouvons ainsi observer sur la figure S5.15 que les gains de précision de prédiction sont modestes dans nos simulations si nous remplaçons les TBV par les valeurs phénotypiques des candidats.

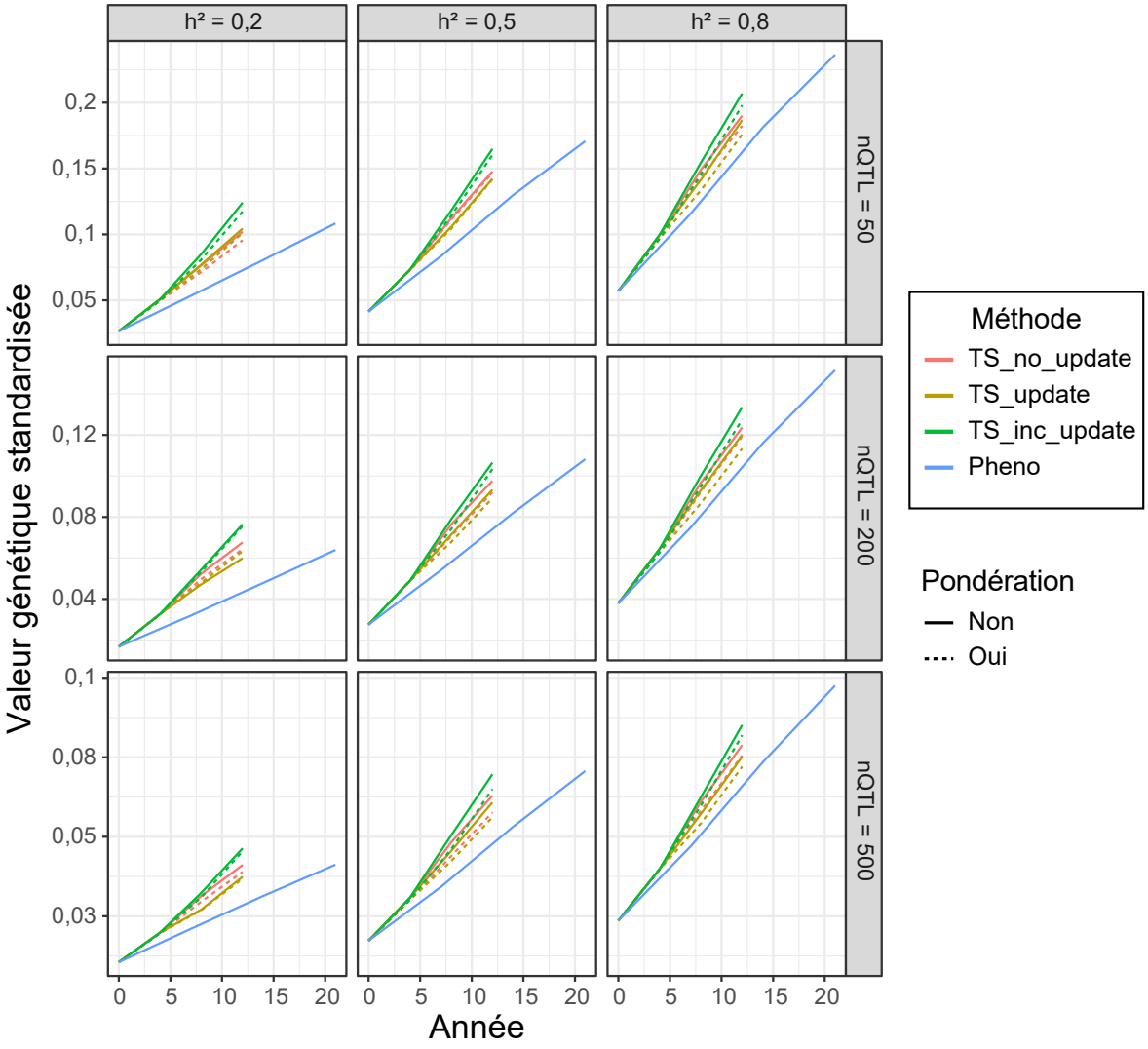


FIGURE 5.7 – Valeur génétique standardisée par unité de temps dans le schéma des intercroisements en fonction de la méthode de sélection des parents

Dans le cas où la taille de la population d'entrainement restait constante, les méthodes TS_update et TS_no_update ont conduit aux mêmes valeurs génétiques standardisées en troisième génération dans les deux schémas simulés, alors que les précisions de prédiction obtenues avec les deux méthodes différaient généralement. S'il n'est pas possible de mettre à jour la population d'entrainement à chaque génération du fait de contraintes économiques ou logistiques, il est donc envisageable d'estimer les effets des marqueurs à partir de la même population d'entraine-

ment pendant les premières générations d'un programme de pré-breeding tel qu'envisagé dans ce chapitre. Ici, la population d'entrainement initiale, regroupant des génotypes provenant du panel des ressources génétiques et du panel du matériel élite, a été constituée de façon à ressembler dans sa composition à la REFPOP présentée au chapitre 2. Nous en concluons donc que la REFPOP peut servir de population d'entrainement pour initier des programmes de pré-breeding chez le pommier. L'utilisation de la pondération dynamique a en outre permis d'augmenter la fréquence moyenne des allèles rares en troisième génération sans compromettre le gain génétique à court terme. Toutefois, la perte d'allèles rares du fait de la dérive génétique était aussi marquée avec ou sans pondération (figure S5.14), suggérant que la perte initiale d'allèles rares est probablement inévitable lorsque la sélection génomique est appliquée (Jannink 2010). L'utilisation de la pondération dynamique peut donc présenter des avantages par rapport aux prédictions génomiques classiques, mais un certain nombre d'éléments sont à prendre en compte afin de juger de la pertinence de cette approche. Plusieurs difficultés peuvent en effet être rencontrées lors de l'utilisation de la pondération des effets des marqueurs telle que présentée dans ce chapitre. D'une part, une sélection basée sur les GEBV pondérées permettra de donner un poids plus fort aux allèles rares seulement si leur effet est correctement estimé. Or dans les modèles de prédiction, les effets de ces allèles peuvent être mal estimés, surtout s'ils sont également rares dans la population d'entrainement. Dans le cas où le signe de l'effet estimé d'un allèle favorable est erroné, la pondération aura même pour conséquence de contre-sélectionner cet allèle. Le choix de la formule de pondération employée peut également influencer le gain génétique à court et long terme (Sun et VanRaden 2014). Dans le cas de la pondération dynamique utilisée dans ce chapitre, l'utilisateur peut faire varier les pondérations assignées aux marqueurs à travers les paramètres α et N (l'horizon de sélection). Nous avons obtenu des résultats très proches de ceux présentés pour des valeurs de α comprises entre 0,05 et 0,5 (avec un pas de 0,05) ou en fixant l'horizon de sélection à 10 générations (résultats non présentés). A court terme, comme dans le cas de ce chapitre, nous pensons donc que l'influence de ces paramètres est moindre du fait du nombre de générations réduit, mais des simulations plus poussées devraient être menées afin d'étudier l'impact de ces paramètres sur le gain génétique à long terme (au-delà d'une dizaine de générations). Enfin, les pondérations appliquées aux effets des marqueurs sont fonction des fréquences alléliques pour une population donnée et le choix de cette population dépend des objectifs à atteindre. Dans ce chapitre, nous avons calculé les pondérations à chaque génération en utilisant les fréquences alléliques initiales du matériel élite, ce qui permet de donner un poids plus fort aux mêmes régions génomiques à chaque génération et donc d'augmenter la fréquence des allèles favorables à quelques endroits du génome. Cette approche est utile dans une perspective de gain génétique à court terme, comme étudié dans ce chapitre, mais n'est pas compatible avec une amélioration à long terme du matériel élite, car pour des générations plus avancées que celles étudiées ici, le risque existe d'appliquer un

poids plus fort à des régions génomiques pour lesquelles la fréquence de l'allèle favorable aurait déjà fortement augmenté. Dans ce cas, il est préférable de calculer les pondérations en fonction des fréquences alléliques de la génération des candidats, permettant de donner un poids plus important aux régions pour lesquelles les allèles favorables sont encore rares dans cette génération plutôt que dans le matériel élite. Par ailleurs, nous avons considéré que le matériel élite n'évoluait pas en parallèle des générations d'intercroisements ou de pseudo-rétrocroisements, alors qu'en réalité il pourrait évoluer, entrainant donc un changement de la fréquence des allèles rares par la même occasion.

5.4.2 Mise en place de la sélection génomique pour des actions de pré-breeding

Compte tenu de l'importance de la précision de prédiction, la constitution de la population d'entrainement est un aspect crucial à considérer dans la mise en place de la sélection génomique chez le pommier. Dans nos simulations, nous avons montré qu'il était préférable de mettre à jour la population d'entrainement à chaque génération en utilisant et cumulant les données phénotypiques des générations précédentes (méthode TS_update_inc). Il est important de souligner que cette observation est valable dans notre cas car nous avons simulé un nombre restreint de générations. Etant donné que la précision de prédiction décroît au fur et à mesure que la génération des candidats s'éloigne de la population d'entrainement (Hidalgo et al. 2021 ; Müller et al. 2017), inclure les données phénotypiques et génotypiques de toutes les générations dans la population d'entrainement est probablement moins utile dans le cas de générations plus avancées (Wolc et al. 2011), auquel cas il est préférable d'inclure uniquement les données des dernières générations. La densité de marqueurs utilisée est également un facteur important de la précision de prédiction. Les résultats présentés dans ce chapitre ont été obtenus en utilisant une haute densité de marqueurs, mais génotyper tous les candidats à haute densité semble irréalisable pour une espèce comme le pommier compte tenu des coûts actuels de génotypage. Une solution alternative consisterait alors à génotyper les candidats avec une puce moyenne voire basse densité, avant d'imputer les données génomiques de ces candidats pour obtenir des données génomique haute densité. Lorsque la précision d'imputation est élevée, les GEBV calculées à partir des données imputées sont proches des GEBV calculées avec les données haute densité (Berry et Kearney 2011 ; Dassonneville et al. 2011), ce qui devrait permettre la mise en place des schémas proposés dans ce chapitre compte tenu de la précision d'imputation élevée chez le pommier rapportée au chapitre 3. Cependant, il faut souligner que même si la précision d'imputation peut être élevée au niveau du génome, la précision d'imputation des allèles rares est souvent problématique (Si et al. 2021), ce qui pourrait remettre en cause l'intérêt de la sélection génomique pondérée présentée précédemment dans ce chapitre. En particulier, l'imputation

d'allèles rares basée sur l'utilisation d'un panel de référence peut se révéler problématique si ces allèles sont présents en fréquence faible dans ce panel. Dans ce cas, nous préconisons donc l'utilisation d'un panel représentatif de la diversité génétique du pommier, tel que celui présenté au chapitre 3. S'il est possible d'imputer les données génomiques en utilisant les informations du pédigrée, nous recommandons alors de génotyper à haute densité les parents des croisements à chaque génération. Soulignons tout de même que lorsque nous avons effectué les mêmes simulations que celles présentées dans ce chapitre en utilisant des données moyenne densité, nous avons dans certains cas obtenu une valeur génétique standardisée similaire à celle observée lorsque les données haute densité étaient utilisées (figure S5.16), alors que la précision de prédiction en utilisant ces données était toujours inférieure à la précision de prédiction obtenue en utilisant les données haute densité (figure S5.17). Ainsi, nous n'avons observé aucune différence pour la valeur génétique standard en troisième génération de pseudo-rétrocroisements, alors que la différence était presque toujours significative dans le schéma des intercroisements. En revanche, nous n'avons observé aucune différence de fréquence des allèles rares en troisième génération dans les deux schémas simulés lorsque nous avons comparé l'utilisation de données moyenne et haute densité (figure S5.18). Compte tenu de la précision d'imputation élevée qu'il est possible d'obtenir chez le pommier et du coût peu élevé que représente le génotypage à haute densité des parents des candidats à chaque génération, nous recommandons dans tous les cas d'utiliser des données haute densité dans les modèles de prédiction génomique. Cependant, si l'utilisation de telles données se révélait difficile, nous soulignons que l'utilisation de données moyenne densité peut être une alternative valable dans un programme de pré-breeding.

Enfin, il convient de noter que nous avons simulé un seul caractère dans nos simulations et basé la sélection des individus sur ce seul caractère au cours des générations. Or la création variétale chez le pommier prend en compte un grand nombre de caractères, et il est certain qu'un programme de pré-breeding chercherait au moins à évaluer les candidats pour la qualité du fruit, la résistance aux maladies et le rendement. Un programme de pré-breeding basé sur la sélection génomique devrait donc idéalement prédire plusieurs caractères et utiliser l'information de ces prédictions pour choisir les candidats servant de parents à la prochaine génération. Cette prise en compte de plusieurs caractères peut se faire sous la forme d'un index pondérant les GEBV des différents caractères en fonction de pondérations déterminés par le sélectionneur (voir par exemple (Li et al. 2022) pour un exemple récent de simulation de programme de pré-breeding utilisant un index de sélection), avec la difficulté que comporte le choix des coefficients de pondération à appliquer. D'autre part, nous n'avons pas non plus pris en compte les défauts qui surviennent souvent dans la descendance de croisements « ressources génétiques x élite » chez le pommier et qui sont rédhibitoires pour un sélectionneur. Le déterminisme génétique de ces défauts est généralement mal connu, ce qui pose des difficultés pour les intégréer dans des simulations. Il est néanmoins nécessaire de se poser la question de la prise en compte de ces

caractères dans les modèles de prédiction : si les défauts sont prédits de manière suffisamment précise, il est possible de les prendre en compte dans un index de sélection au même titre que les caractères désirables, ou d'exclure d'emblée les génotypes présentant des défauts des parents potentiels dans le programme de pré-breeding.

5.4.3 Pistes à explorer dans la mise en place des simulations

Les simulations sont une approche intéressante pour étudier l'intérêt de l'utilisation de la sélection génomique et sont de plus en plus utilisées dans la recherche actuelle (Daetwyler et al. 2013). Il n'en reste pas moins que certaines hypothèses utilisées pour simuler les données peuvent s'éloigner de la réalité et il convient donc d'analyser les résultats de simulation à la lumière de ces possibles limitations. Nous avons par exemple supposé dans nos simulations que le caractère simulé dépendait uniquement d'effets additifs, alors que les effets de dominance et d'épistasie peuvent expliquer une part importante de la variance phénotypique chez le pommier, comme cela a pu être montré dans le cas de caractères contrôlés par quelques gènes (par exemple pour l'acidité (Jia et al. 2018), le port colonnaire (Otto et al. 2014) ou la résistance à la tavelure (Gessler et al. 2006) ou dans le cas de caractères liés à la qualité du fruit (Kumar et al. 2015) pour lesquels de nombreux QTL à effet mineur sont impliqués. Dans le cas où le degré de dominance est important, Werner et al. (Werner et al. 2020) préconisent même d'éviter de choisir les parents à intercroiser sur la base de leur GEBV (approche que nous avons utilisée dans ce chapitre) et de choisir les parents maximisant la valeur génétique moyenne du croisement estimée à partir des effets prédits des marqueurs. Il est donc important d'adapter les modèles de prédiction et de choix des parents à l'architecture génétique des caractères à prédire lorsque cela est possible, voire d'adapter les schémas de sélection pour tirer profit des effets de dominance.

Soulignons également qu'il aurait été possible de choisir d'autres critères que ceux étudiés ici pour le choix des parents à chaque génération. Plusieurs critères basés sur la complémentarité entre les parents à croiser ont récemment été proposés dans la littérature, tels que la prédiction de la variance résultant du croisement (Lado et al. 2017 ; Wolfe et al. 2021), la complémentarité de régions génomiques (Allier et al. 2020a ; Vanavermaete et al. 2021) ou encore la complémentarité des allèles rares à transférer vers le matériel élite (De Beukelaer et al. 2017 ; Moeinizade et al. 2021). Nous n'avons pas exploré ces pistes afin de ne pas multiplier les scénarios à étudier mais il est clair que de telles approches doivent être considérés dans la mise en place d'actions de pré-breeding.

Enfin, nous n'avons pas non plus pris en compte la dimension économique des différents programmes étudiés dans nos simulations. Puisque nous avons utilisé le même nombre d'individus et de familles quelle que soit la méthode adoptée, le coût économique associé à chaque méthode

est différent, du fait du phénotypage plus ou moins important (par exemple phénotypage de tous les candidats dans le cas de la méthode Pheno mais aucun phénotypage dans le cas de la méthode TS_no_update, phénotypage). Il aurait été intéressant de raisonner à budget constant pour chaque méthode et de chercher à optimiser le programme de pré-breeding dans ce cas (Ben Sadoun 2021). Le retour sur investissement de la mise en place de la sélection génomique est assurément l'élément qui déterminera si cette approche est adoptée chez le pommier et il nous semble donc que les travaux présentés dans ce chapitre mériteraient d'être approfondis afin d'optimiser l'allocation des ressources au sein du programme de pré-breeding.

Figures supplémentaires du chapitre

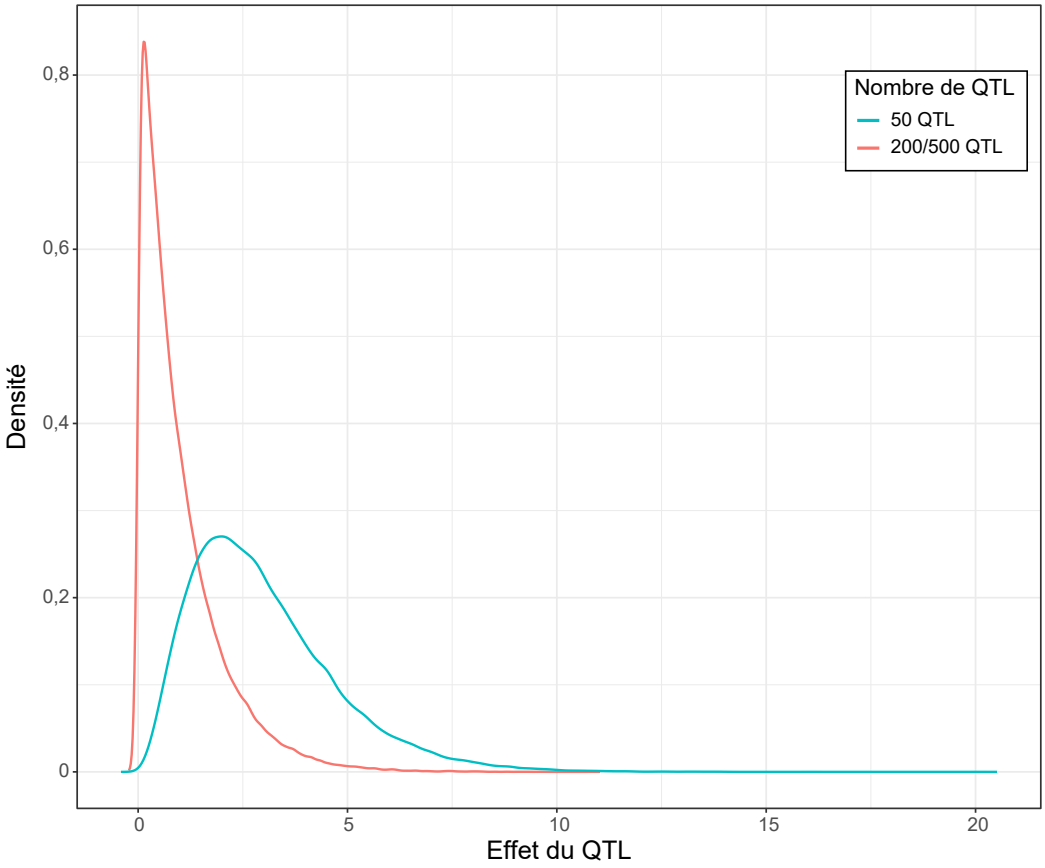


FIGURE S5.8 – Distribution des effets des QTL en fonction de l'architecture génétique simulée

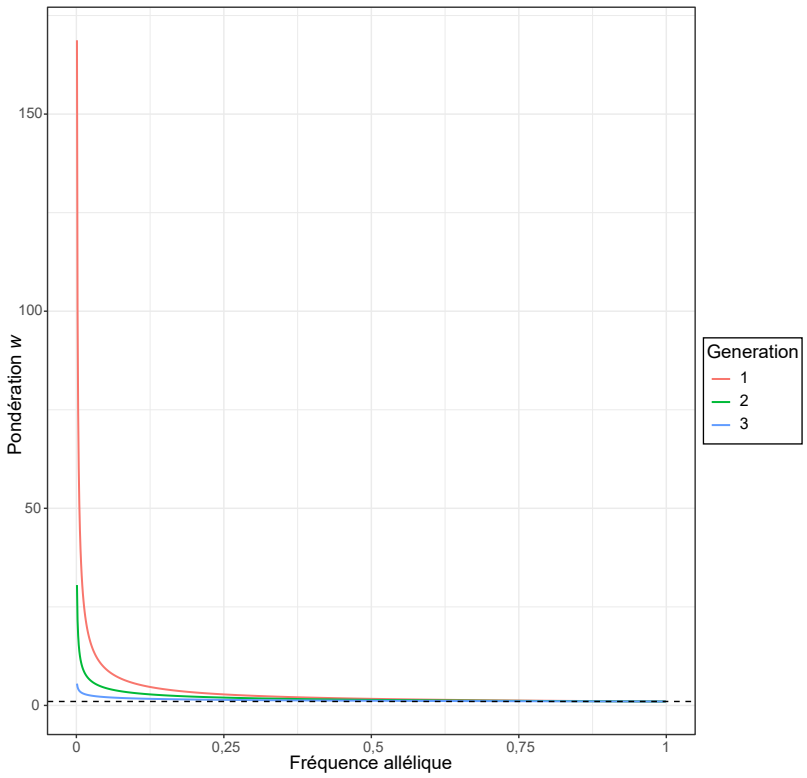


FIGURE S5.9 – Pondération des effets des marqueurs en fonction de la fréquence allélique dans le matériel élite et de la génération considérée. La ligne en pointillés représente une pondération égale à 1

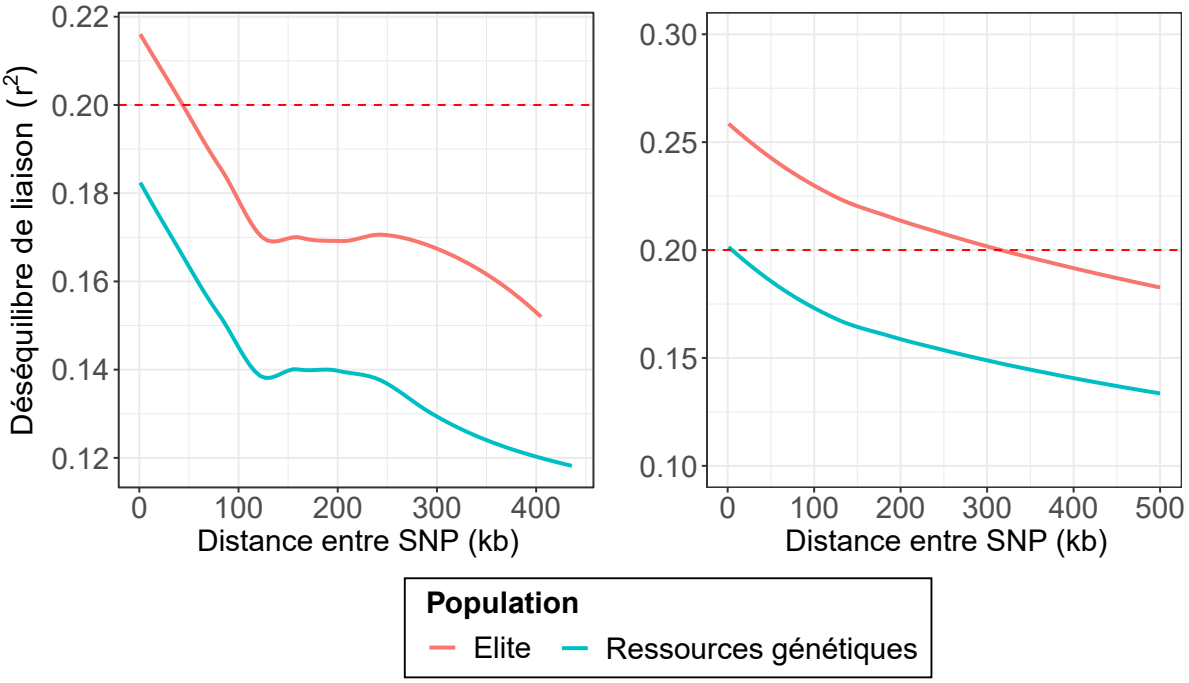


FIGURE S5.10 – Comparaison du déséquilibre de liaison entre ressources génétiques et matériel élite. A gauche : données issues d'une simulation, à droite : données réelles issues du panel REFPOP. **Rouge** : matériel élite ; **Bleu** : ressources génétiques

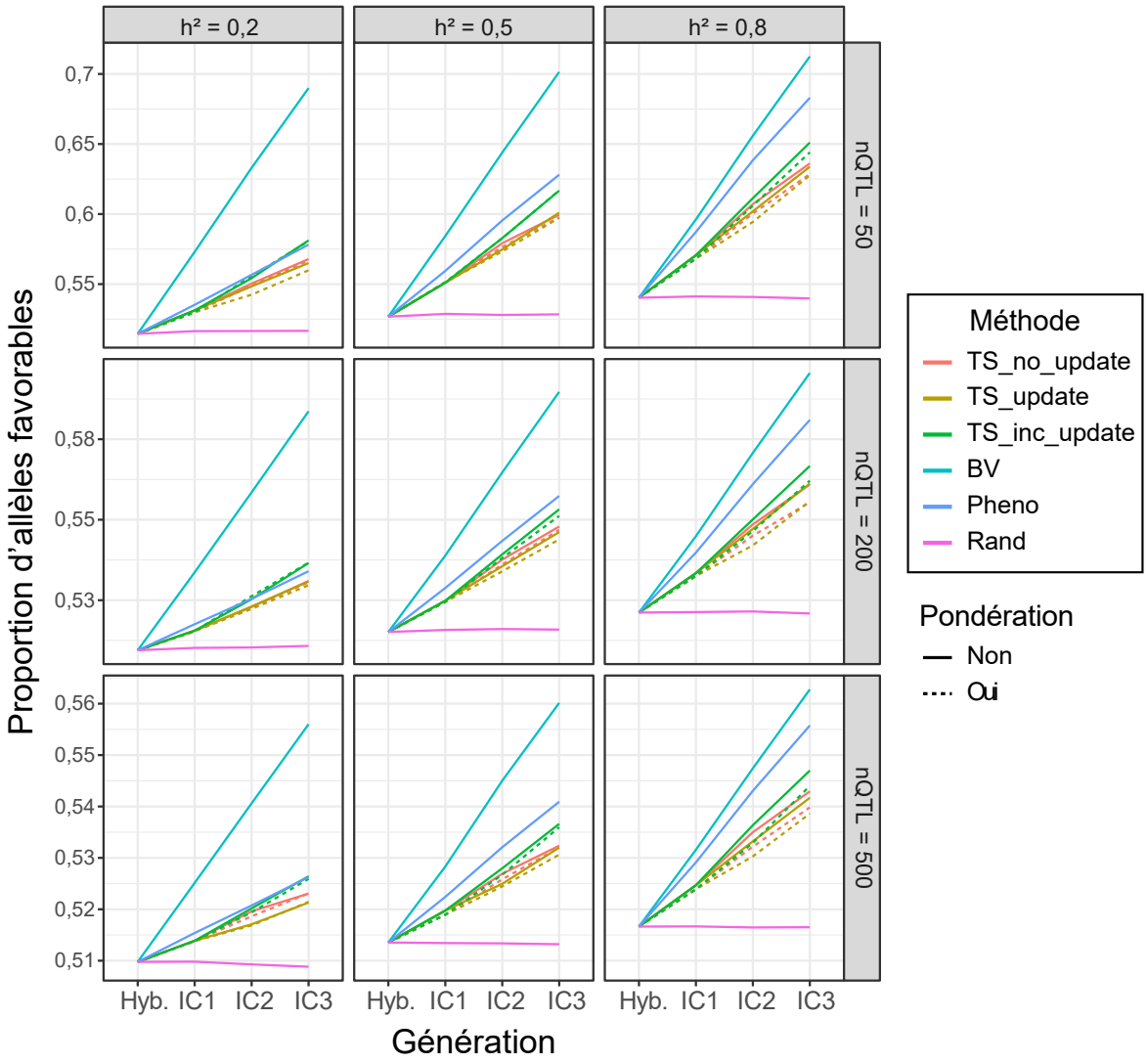


FIGURE S5.11 – Proportion moyenne d'allèles favorables au cours des générations dans le schéma des intercroisements en fonction de la méthode de sélection des parents

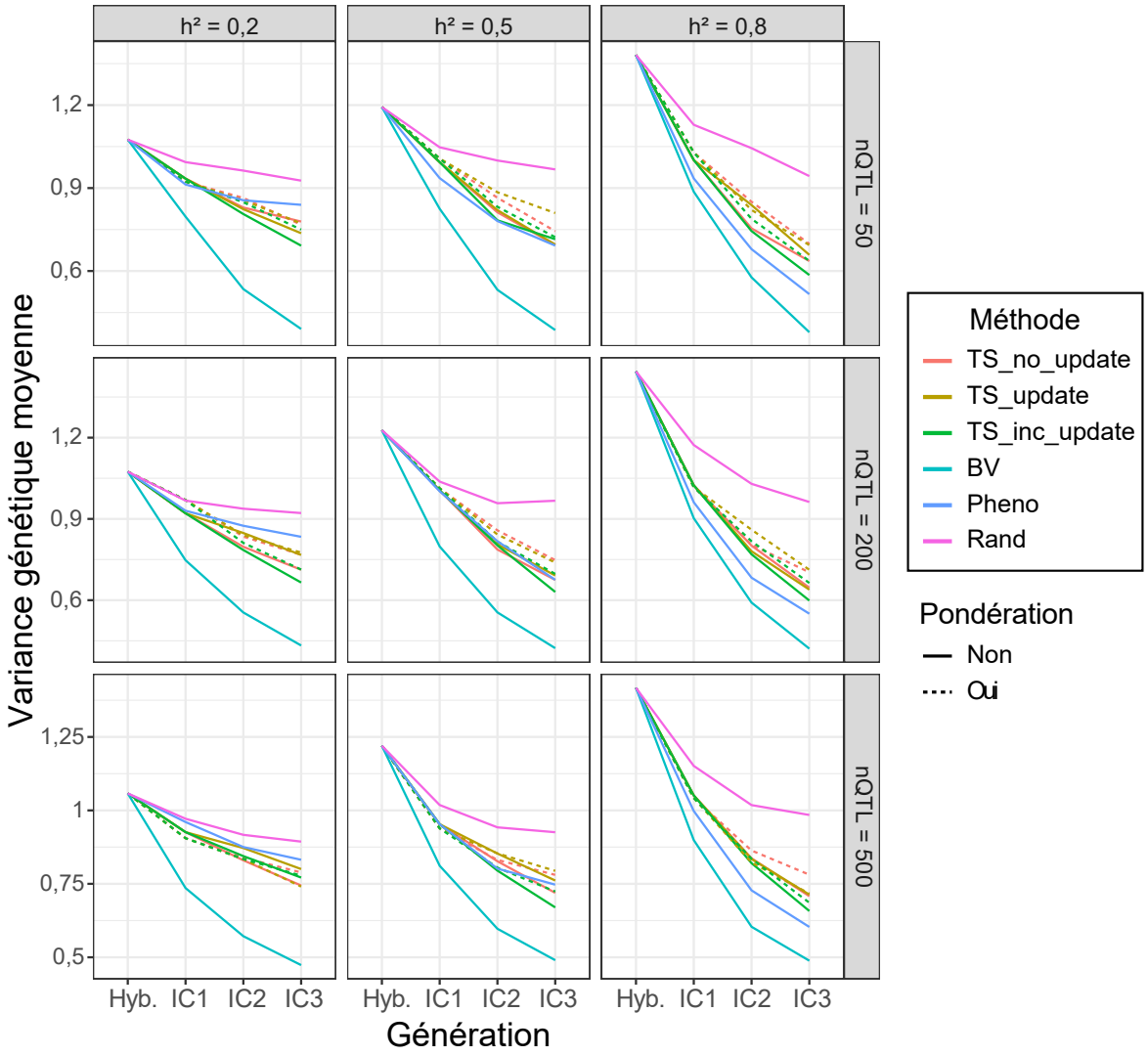


FIGURE S5.12 – Variance génétique moyenne au cours des générations dans le schéma des intercroisements en fonction de la méthode de sélection des parents

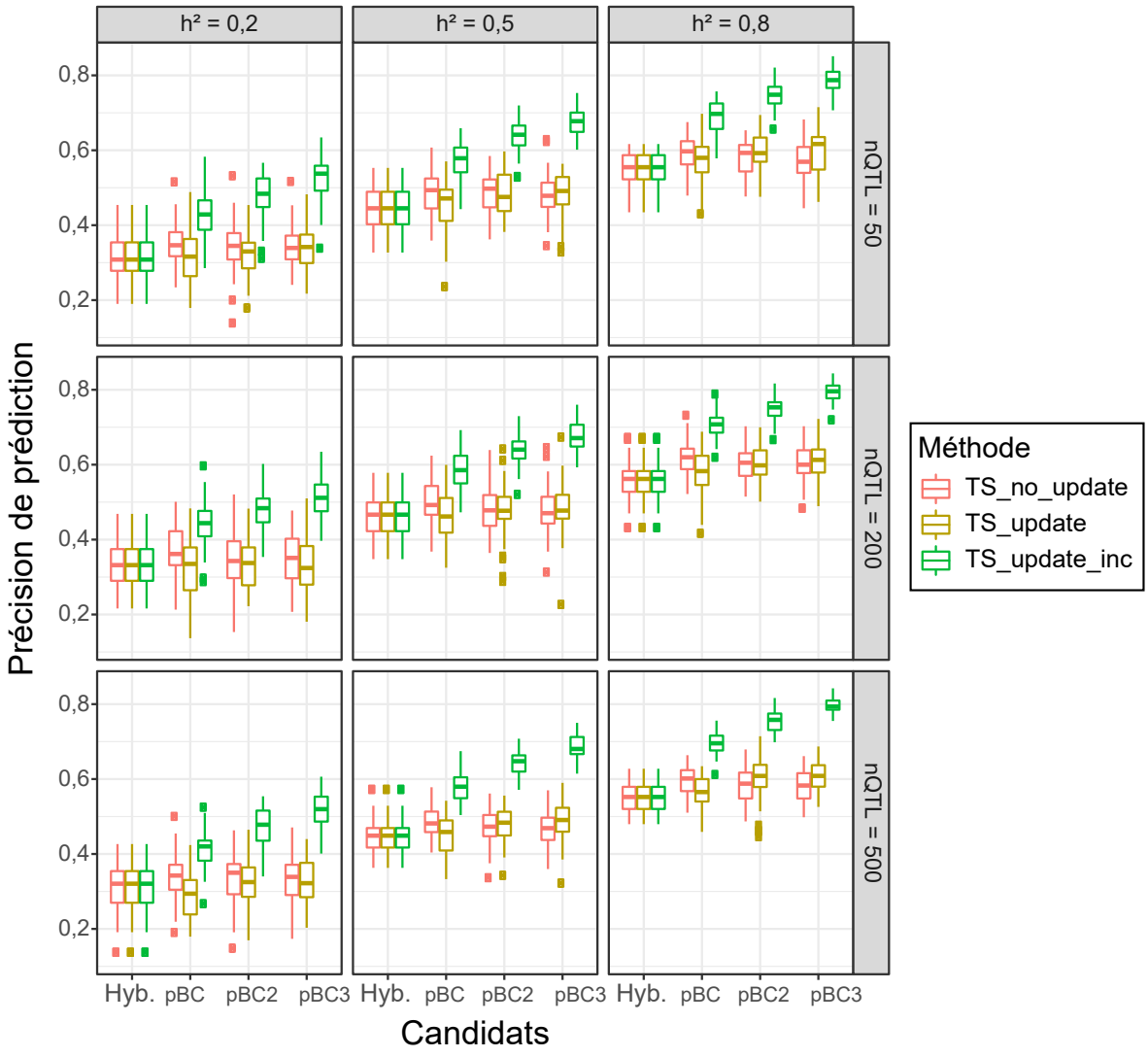


FIGURE S5.13 – Précision de prédiction moyenne intra-famille dans le schéma des rétrocroisements en fonction de la méthode de constitution de la population d'entrainement

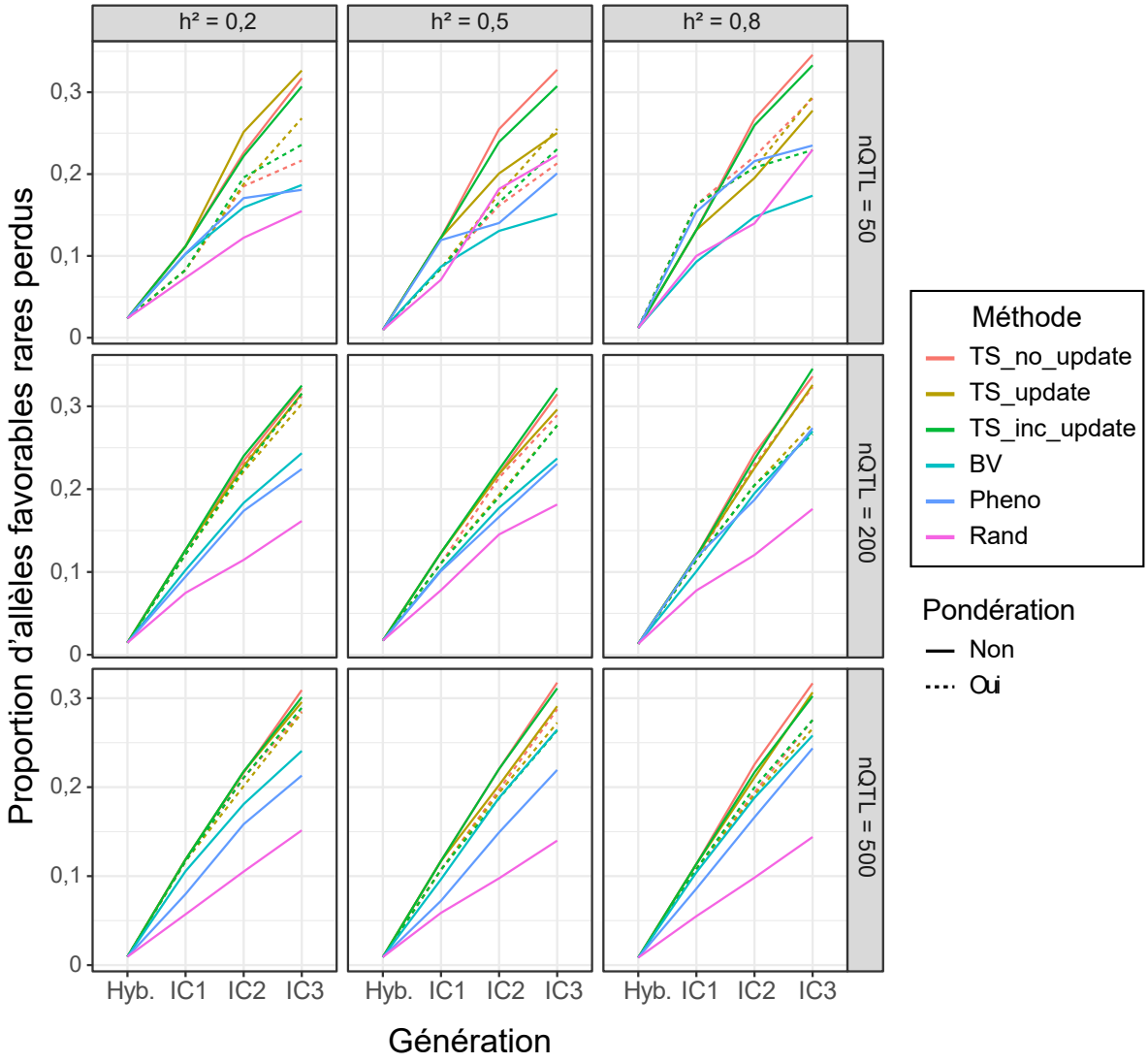


FIGURE S5.14 – Proportion moyenne d'allèles favorables rares perdus au cours des générations dans le schéma des intercroisements en fonction de la méthode de sélection des parents

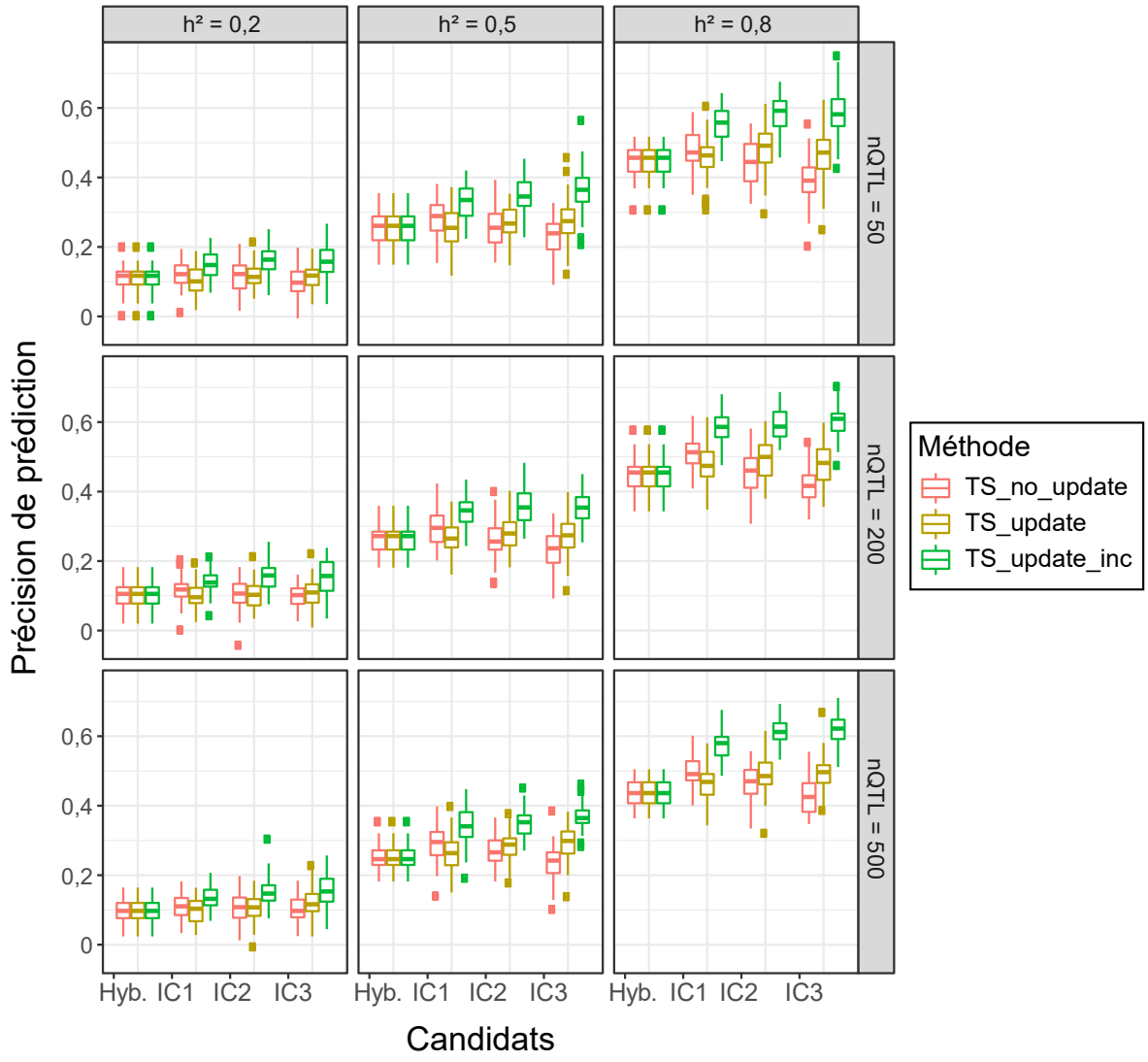


FIGURE S5.15 – Précision de prédiction moyenne intra-famille correspondant à la corrélation entre les GEBV et les valeurs phénotypiques dans le schéma des intercroisements en fonction de la méthode de constitution de la population d'entrainement

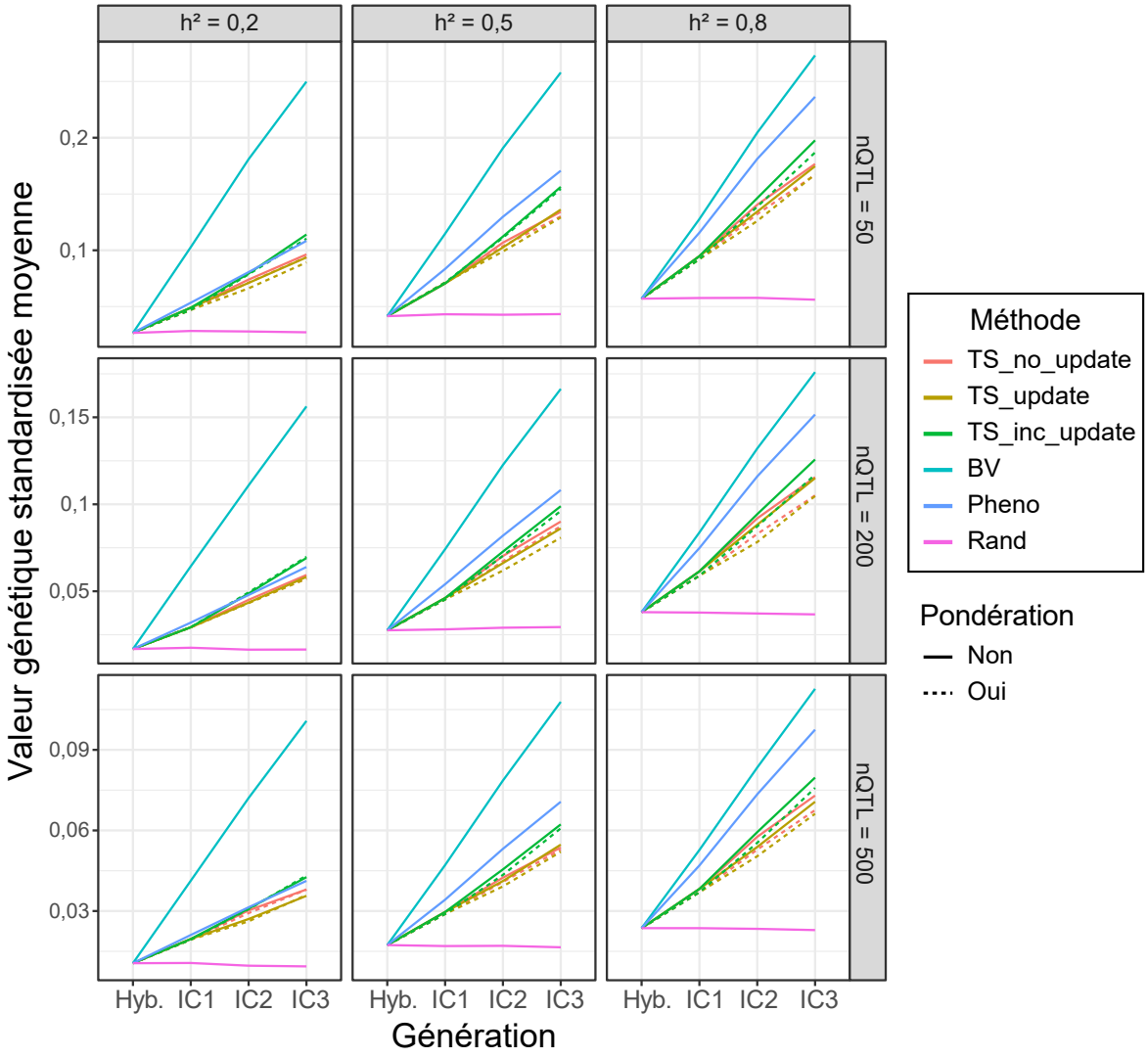


FIGURE S5.16 – Valeur génétique standardisée au cours des générations dans le schéma des intercroisements en fonction de la méthode de sélection des parents. Les prédictions ont été réalisées en utilisant les données moyenne densité. L'échelle de l'axe des y varie en fonction du nombre de QTL

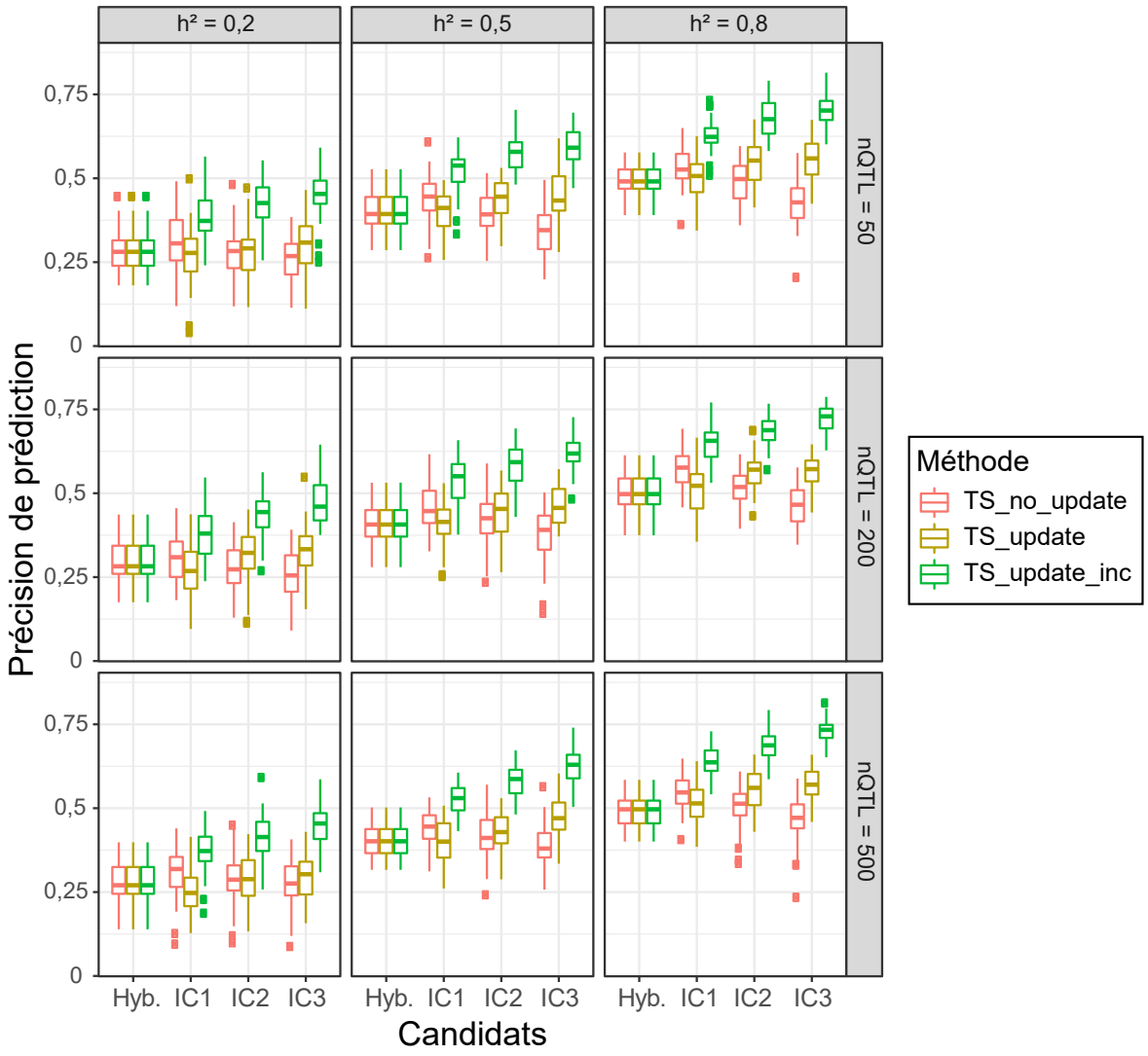


FIGURE S5.17 – Précision de prédiction moyenne intra-famille dans le schéma des intercroisements en fonction de la méthode de constitution de la population d'entrainement (prédictions moyenne densité)

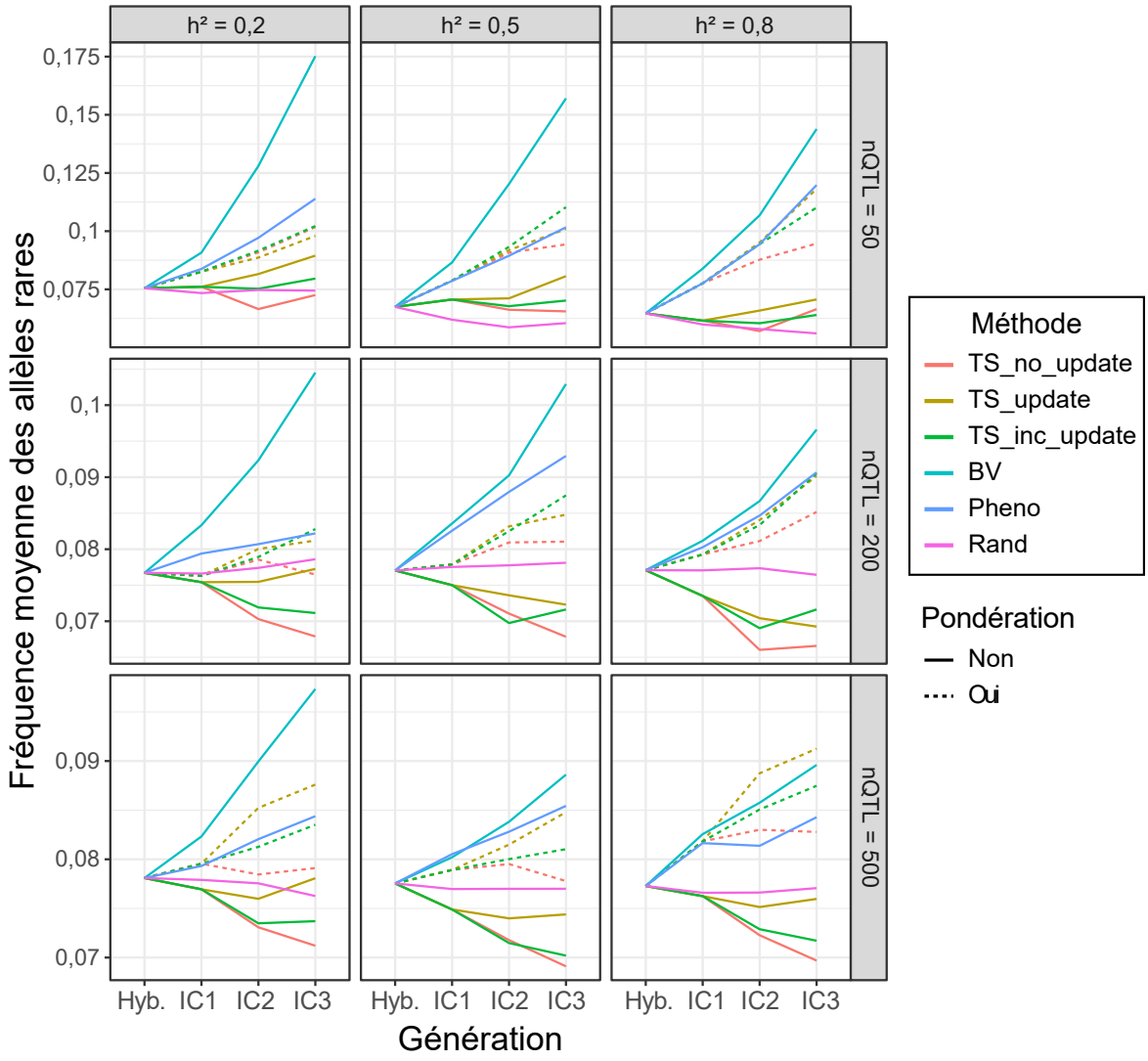


FIGURE S5.18 – Fréquence moyenne des allèles favorables rares au cours des générations dans le schéma des intercroisements en fonction de la méthode de sélection des parents (prédictions moyenne densité)

Tableaux supplémentaires du chapitre

TABLEAU 5.1 – Valeur génétique additive moyenne dans les ressources génétiques et le matériel élite, nombre moyen d'allèles favorables et défavorables fixés dans les ressources génétiques et le matériel élite et F_{ST} moyen entre ressources génétiques et matériel élite obtenus sur 100 simulations. L'écart-type de chaque valeur est indiqué entre parenthèses

h^2	nQTL	n	Valeur génétique		Allèles favorables fixés		Allèles défavorables fixés		F_{ST}
			RG	Elite	RG	Elite	RG	Elite	
0,2	50	100	0	0,04 (0,01)	0,33 (0,04)	0,36 (0,05)	0,33 (0,04)	0,32 (0,04)	0,012 (0,001)
		1000	0	0,049 (0,02)	0,33 (0,04)	0,38 (0,04)	0,33 (0,04)	0,32 (0,04)	0,012 (0,001)
		10000	0	0,054 (0,01)	0,33 (0,04)	0,39 (0,05)	0,32 (0,04)	0,31 (0,04)	0,012 (0,001)
	200	100	0	0,04 (0,01)	0,33 (0,02)	0,36 (0,02)	0,33 (0,02)	0,33 (0,02)	0,012 (0,001)
		1000	0	0,051 (0,02)	0,32 (0,02)	0,36 (0,02)	0,33 (0,02)	0,32 (0,02)	0,012 (0,001)
		10000	0	0,059 (0,01)	0,33 (0,02)	0,36 (0,02)	0,33 (0,02)	0,33 (0,02)	0,012 (0,001)
	500	100	0	0,04 (0,02)	0,33 (0,01)	0,36 (0,01)	0,33 (0,01)	0,34 (0,01)	0,012 (0,001)
		1000	0	0,051 (0,02)	0,33 (0,01)	0,36 (0,02)	0,33 (0,01)	0,33 (0,01)	0,012 (0,001)
		10000	0	0,06 (0,02)	0,33 (0,01)	0,36 (0,01)	0,33 (0,01)	0,33 (0,01)	0,012 (0,001)
0,5	50	100	0	0,062 (0,01)	0,33 (0,04)	0,39 (0,05)	0,33 (0,04)	0,3 (0,04)	0,013 (0,001)
		1000	0	0,073 (0,01)	0,33 (0,04)	0,4 (0,05)	0,33 (0,04)	0,3 (0,04)	0,013 (0,001)
		10000	0	0,088 (0,02)	0,32 (0,05)	0,39 (0,05)	0,34 (0,04)	0,3 (0,04)	0,013 (0,001)
	200	100	0	0,066 (0,01)	0,33 (0,02)	0,37 (0,02)	0,33 (0,02)	0,32 (0,02)	0,013 (0,001)
		1000	0	0,083 (0,01)	0,33 (0,02)	0,38 (0,02)	0,33 (0,02)	0,31 (0,02)	0,013 (0,001)
		10000	0	0,093 (0,01)	0,33 (0,02)	0,38 (0,02)	0,33 (0,02)	0,31 (0,02)	0,013 (0,001)
	500	100	0	0,068 (0,01)	0,33 (0,01)	0,36 (0,01)	0,33 (0,01)	0,33 (0,01)	0,013 (0,001)
		1000	0	0,083 (0,01)	0,33 (0,01)	0,37 (0,01)	0,33 (0,01)	0,32 (0,01)	0,013 (0,001)
		10000	0	0,097 (0,02)	0,33 (0,01)	0,37 (0,01)	0,33 (0,01)	0,32 (0,01)	0,014 (0,001)
0,8	50	100	0	0,087 (0,02)	0,33 (0,04)	0,41 (0,05)	0,33 (0,04)	0,29 (0,02)	0,013 (0,001)
		1000	0	2,08 (0,02)	0,34 (0,04)	0,43 (0,05)	0,32 (0,04)	0,27 (0,04)	0,013 (0,001)
		10000	0	2,41 (0,02)	0,33 (0,04)	0,44 (0,05)	0,32 (0,05)	0,28 (0,05)	0,013 (0,001)
	200	100	0	0,093 (0,02)	0,33 (0,02)	0,38 (0,02)	0,33 (0,02)	0,31 (0,02)	0,014 (0,001)
		1000	0	0,116 (0,02)	0,33 (0,02)	0,39 (0,02)	0,33 (0,02)	0,31 (0,02)	0,015 (0,001)
		10000	0	0,132 (0,01)	0,33 (0,02)	0,4 (0,02)	0,33 (0,02)	0,3 (0,02)	0,015 (0,001)
	500	100	0	0,094 (0,02)	0,33 (0,01)	0,37 (0,01)	0,33 (0,01)	0,33 (0,01)	0,013 (0,001)
		1000	0	0,117 (0,02)	0,33 (0,01)	0,37 (0,02)	0,33 (0,01)	0,32 (0,01)	0,015 (0,001)
		10000	0	0,139 (0,01)	0,33 (0,01)	0,38 (0,01)	0,33 (0,01)	0,31 (0,01)	0,015 (0,001)

TABLEAU 5.2 – Valeur génétique additive en troisième génération en fonction de la méthode de sélection des parents, de la densité de marquage et de la modalité dans les deux schémas de sélection simulés. L'écart-type de chaque valeur est indiqué entre parenthèses

Génération	Méthode	Pondération	Densité	$h^2 = 0,2$			$h^2 = 0,5$			$h^2 = 0,8$		
				50	200	500	50	200	500	50	200	500
BC3	TS_no_update	Non	High	0,073 (0,02)	0,045 (0,01)	0,029 (0,01)	0,108 (0,02)	0,071 (0,01)	0,045 (0,01)	0,146 (0,02)	0,096 (0,01)	0,059 (0,01)
			Medium	0,071 (0,02)	0,045 (0,01)	0,027 (0,01)	0,107 (0,02)	0,069 (0,01)	0,045 (0,01)	0,141 (0,02)	0,092 (0,01)	0,058 (0,01)
		Oui	High	0,072 (0,02)	0,046 (0,01)	0,028 (0,01)	0,108 (0,02)	0,072 (0,01)	0,045 (0,01)	0,144 (0,02)	0,093 (0,01)	0,059 (0,01)
			Medium	0,069 (0,02)	0,045 (0,01)	0,028 (0,01)	0,104 (0,02)	0,068 (0,01)	0,043 (0,01)	0,141 (0,02)	0,093 (0,01)	0,057 (0,01)
	TS_update	Non	High	0,072 (0,02)	0,045 (0,01)	0,029 (0,01)	0,109 (0,02)	0,071 (0,01)	0,045 (0,01)	0,145 (0,02)	0,095 (0,01)	0,06 (0,01)
			Medium	0,069 (0,02)	0,044 (0,01)	0,028 (0,01)	0,107 (0,02)	0,07 (0,01)	0,043 (0,01)	0,141 (0,02)	0,093 (0,01)	0,058 (0,01)
		Oui	High	0,072 (0,02)	0,045 (0,01)	0,029 (0,01)	0,108 (0,02)	0,07 (0,01)	0,044 (0,01)	0,143 (0,02)	0,094 (0,01)	0,059 (0,01)
			Medium	0,071 (0,02)	0,044 (0,01)	0,028 (0,01)	0,104 (0,02)	0,069 (0,01)	0,044 (0,01)	0,139 (0,02)	0,093 (0,01)	0,057 (0,01)
	TS_update_inc	Non	High	0,08 (0,02)	0,05 (0,01)	0,032 (0,01)	0,114 (0,02)	0,075 (0,01)	0,048 (0,01)	0,154 (0,02)	0,099 (0,01)	0,062 (0,01)
			Medium	0,078 (0,02)	0,048 (0,01)	0,03 (0,01)	0,112 (0,02)	0,073 (0,01)	0,047 (0,01)	0,147 (0,02)	0,096 (0,01)	0,061 (0,01)
		Oui	High	0,079 (0,02)	0,049 (0,01)	0,032 (0,01)	0,114 (0,02)	0,075 (0,01)	0,048 (0,01)	0,151 (0,02)	0,098 (0,01)	0,062 (0,01)
			Medium	0,078 (0,02)	0,049 (0,01)	0,03 (0,01)	0,112 (0,02)	0,073 (0,01)	0,047 (0,01)	0,145 (0,02)	0,095 (0,01)	0,06 (0,01)
	BV	-	-	0,114 (0,02)	0,071 (0,01)	0,046 (0,01)	0,14 (0,02)	0,09 (0,01)	0,058 (0,01)	0,166 (0,02)	0,108 (0,01)	0,068 (0,01)
	Pheno	-	-	0,072 (0,02)	0,045 (0,01)	0,029 (0,01)	0,114 (0,02)	0,074 (0,01)	0,047 (0,01)	0,156 (0,02)	0,102 (0,01)	0,065 (0,01)
	Rand	-	-	0,05 (0,01)	0,032 (0,01)	0,02 (0,01)	0,078 (0,02)	0,052 (0,01)	0,032 (0,01)	0,108 (0,02)	0,071 (0,01)	0,044 (0,01)
IC3	TS_no_update	Non	High	0,102 (0,02)	0,068 (0,02)	0,041 (0,01)	0,148 (0,03)	0,098 (0,02)	0,063 (0,01)	0,19 (0,02)	0,124 (0,02)	0,079 (0,01)
			Medium	0,096 (0,02)	0,059 (0,02)	0,038 (0,01)	0,134 (0,03)	0,09 (0,02)	0,054 (0,01)	0,177 (0,03)	0,116 (0,02)	0,073 (0,01)
		Oui	High	0,096 (0,02)	0,064 (0,02)	0,039 (0,01)	0,147 (0,02)	0,092 (0,02)	0,058 (0,01)	0,182 (0,02)	0,119 (0,02)	0,076 (0,01)
			Medium	0,094 (0,03)	0,057 (0,02)	0,038 (0,01)	0,13 (0,02)	0,087 (0,02)	0,053 (0,01)	0,167 (0,03)	0,105 (0,02)	0,068 (0,01)
	TS_update	Non	High	0,104 (0,03)	0,06 (0,02)	0,037 (0,01)	0,142 (0,03)	0,093 (0,02)	0,061 (0,01)	0,187 (0,02)	0,12 (0,02)	0,075 (0,01)
			Medium	0,094 (0,03)	0,058 (0,02)	0,036 (0,01)	0,136 (0,03)	0,086 (0,02)	0,055 (0,01)	0,175 (0,03)	0,115 (0,02)	0,071 (0,01)
		Oui	High	0,101 (0,03)	0,063 (0,02)	0,037 (0,01)	0,142 (0,03)	0,089 (0,02)	0,056 (0,01)	0,176 (0,03)	0,113 (0,02)	0,072 (0,01)
			Medium	0,089 (0,02)	0,057 (0,02)	0,036 (0,01)	0,129 (0,02)	0,081 (0,02)	0,052 (0,01)	0,167 (0,03)	0,105 (0,02)	0,066 (0,01)
	TS_update_inc	Non	High	0,124 (0,02)	0,076 (0,02)	0,046 (0,01)	0,165 (0,03)	0,107 (0,02)	0,07 (0,01)	0,207 (0,02)	0,134 (0,02)	0,085 (0,01)
			Medium	0,114 (0,02)	0,069 (0,02)	0,043 (0,01)	0,156 (0,03)	0,099 (0,02)	0,062 (0,01)	0,198 (0,03)	0,126 (0,02)	0,08 (0,01)
		Oui	High	0,117 (0,03)	0,076 (0,02)	0,045 (0,01)	0,16 (0,02)	0,103 (0,02)	0,065 (0,01)	0,198 (0,03)	0,128 (0,02)	0,082 (0,01)
			Medium	0,11 (0,02)	0,07 (0,02)	0,043 (0,01)	0,155 (0,03)	0,096 (0,02)	0,061 (0,01)	0,187 (0,03)	0,117 (0,02)	0,076 (0,01)
	BV	-	-	0,25 (0,03)	0,156 (0,02)	0,101 (0,01)	0,258 (0,02)	0,166 (0,02)	0,108 (0,01)	0,273 (0,03)	0,176 (0,02)	0,113 (0,01)
	Pheno	-	-	0,108 (0,02)	0,064 (0,01)	0,041 (0,01)	0,171 (0,02)	0,108 (0,02)	0,071 (0,01)	0,236 (0,02)	0,152 (0,02)	0,098 (0,01)
	Rand	-	-	0,027 (0,02)	0,016 (0,01)	0,009 (0,01)	0,043 (0,02)	0,029 (0,01)	0,016 (0,01)	0,056 (0,02)	0,037 (0,01)	0,023 (0,01)

TABLEAU 5.3 – Variance génétique en troisième génération en fonction de la méthode de sélection des parents, de la densité de marquage et de la modalité dans les deux schémas de sélection simulés. L'écart-type de chaque valeur est indiqué entre parenthèses

Génération	Méthode	Pondération	Densité	$h^2 = 0,2$			$h^2 = 0,5$			$h^2 = 0,8$		
				50	200	500	50	200	500	50	200	500
BC3	TS_no_update	Non	High	1,046 (0,28)	1,017 (0,23)	1,044 (0,28)	1,175 (0,32)	1,165 (0,25)	1,234 (0,30)	1,415 (0,36)	1,544 (0,38)	1,407 (0,41)
			Medium	1,046 (0,25)	0,985 (0,22)	1,044 (0,26)	1,178 (0,29)	1,255 (0,32)	1,175 (0,30)	1,463 (0,48)	1,439 (0,34)	1,47 (0,45)
		Oui	High	1,108 (0,34)	1,043 (0,25)	1,059 (0,23)	1,227 (0,36)	1,222 (0,35)	1,225 (0,34)	1,424 (0,49)	1,558 (0,49)	1,465 (0,37)
			Medium	1,061 (0,31)	1,041 (0,24)	1,078 (0,3)	1,228 (0,34)	1,228 (0,33)	1,215 (0,35)	1,395 (0,37)	1,473 (0,42)	1,495 (0,47)
	TS_update	Non	High	1,071 (0,26)	0,969 (0,24)	1,038 (0,30)	1,19 (0,33)	1,257 (0,33)	1,225 (0,32)	1,373 (0,42)	1,393 (0,38)	1,516 (0,46)
			Medium	1,084 (0,35)	1,047 (0,23)	1,075 (0,27)	1,163 (0,30)	1,183 (0,31)	1,261 (0,37)	1,457 (0,4)	1,488 (0,44)	1,541 (0,49)
		Oui	High	1,094 (0,33)	1,005 (0,26)	1,057 (0,26)	1,235 (0,35)	1,214 (0,31)	1,235 (0,34)	1,437 (0,45)	1,448 (0,34)	1,474 (0,44)
			Medium	1,068 (0,29)	1,012 (0,20)	1,041 (0,23)	1,22 (0,31)	1,187 (0,38)	1,266 (0,42)	1,481 (0,51)	1,51 (0,42)	1,455 (0,37)
	TS_update_inc	Non	High	1,066 (0,29)	1,001 (0,25)	1,052 (0,24)	1,136 (0,26)	1,173 (0,32)	1,208 (0,38)	1,34 (0,46)	1,446 (0,42)	1,449 (0,41)
			Medium	1,078 (0,28)	0,994 (0,20)	1,034 (0,26)	1,156 (0,28)	1,176 (0,32)	1,2 (0,37)	1,373 (0,43)	1,52 (0,37)	1,4 (0,40)
		Oui	High	1,09 (0,27)	1,053 (0,25)	0,984 (0,23)	1,153 (0,34)	1,205 (0,32)	1,163 (0,32)	1,368 (0,37)	1,527 (0,42)	1,38 (0,31)
			Medium	1,081 (0,25)	1,009 (0,24)	0,986 (0,26)	1,181 (0,32)	1,201 (0,33)	1,211 (0,35)	1,448 (0,37)	1,439 (0,35)	1,524 (0,50)
	BV	-	-	0,956 (0,25)	0,919 (0,23)	0,932 (0,25)	1,074 (0,34)	1,144 (0,33)	1,148 (0,33)	1,333 (0,40)	1,341 (0,39)	1,446 (0,39)
	Pheno	-	-	1,107 (0,27)	1,025 (0,24)	1,055 (0,27)	1,2 (0,36)	1,18 (0,34)	1,197 (0,32)	1,397 (0,42)	1,42 (0,35)	1,375 (0,38)
	Rand	-	-	1,099 (0,23)	1,065 (0,29)	1,068 (0,26)	1,307 (0,32)	1,299 (0,36)	1,284 (0,28)	1,505 (0,54)	1,608 (0,46)	1,613 (0,47)
IC3	TS_no_update	Non	High	0,779 (0,20)	0,712 (0,18)	0,744 (0,19)	0,695 (0,15)	0,675 (0,15)	0,719 (0,16)	0,636 (0,19)	0,646 (0,13)	0,708 (0,14)
			Medium	0,786 (0,19)	0,815 (0,20)	0,748 (0,16)	0,722 (0,18)	0,727 (0,22)	0,744 (0,15)	0,717 (0,22)	0,684 (0,15)	0,746 (0,17)
		Oui	High	0,767 (0,20)	0,768 (0,20)	0,789 (0,18)	0,744 (0,18)	0,748 (0,17)	0,78 (0,15)	0,698 (0,21)	0,703 (0,18)	0,781 (0,16)
			Medium	0,777 (0,17)	0,793 (0,21)	0,795 (0,15)	0,767 (0,18)	0,788 (0,18)	0,809 (0,20)	0,745 (0,20)	0,739 (0,18)	0,797 (0,20)
	TS_update	Non	High	0,737 (0,18)	0,767 (0,19)	0,8 (0,21)	0,696 (0,18)	0,691 (0,20)	0,761 (0,18)	0,658 (0,19)	0,639 (0,15)	0,713 (0,20)
			Medium	0,779 (0,22)	0,762 (0,18)	0,807 (0,19)	0,734 (0,19)	0,751 (0,17)	0,804 (0,17)	0,698 (0,18)	0,66 (0,16)	0,749 (0,18)
		Oui	High	0,772 (0,19)	0,776 (0,20)	0,741 (0,15)	0,81 (0,21)	0,739 (0,16)	0,794 (0,18)	0,694 (0,18)	0,712 (0,18)	0,717 (0,11)
			Medium	0,797 (0,17)	0,811 (0,23)	0,799 (0,20)	0,762 (0,18)	0,754 (0,16)	0,763 (0,17)	0,712 (0,21)	0,755 (0,18)	0,786 (0,17)
	TS_update_inc	Non	High	0,692 (0,19)	0,665 (0,13)	0,771 (0,18)	0,715 (0,18)	0,63 (0,12)	0,67 (0,12)	0,585 (0,15)	0,599 (0,13)	0,658 (0,13)
			Medium	0,747 (0,17)	0,762 (0,20)	0,738 (0,17)	0,663 (0,16)	0,716 (0,18)	0,737 (0,17)	0,575 (0,18)	0,641 (0,16)	0,695 (0,17)
		Oui	High	0,75 (0,17)	0,713 (0,14)	0,778 (0,17)	0,722 (0,16)	0,697 (0,19)	0,724 (0,16)	0,636 (0,17)	0,663 (0,17)	0,686 (0,13)
			Medium	0,74 (0,14)	0,753 (0,19)	0,773 (0,15)	0,673 (0,18)	0,727 (0,19)	0,732 (0,16)	0,647 (0,14)	0,687 (0,16)	0,731 (0,13)
	BV	-	-	0,39 (0,12)	0,433 (0,09)	0,473 (0,08)	0,386 (0,11)	0,423 (0,11)	0,49 (0,10)	0,379 (0,10)	0,421 (0,09)	0,488 (0,08)
	Pheno	-	-	0,839 (0,22)	0,834 (0,22)	0,832 (0,18)	0,692 (0,19)	0,675 (0,15)	0,747 (0,14)	0,516 (0,14)	0,55 (0,12)	0,603 (0,13)
	Rand	-	-	0,927 (0,16)	0,922 (0,20)	0,894 (0,17)	0,967 (0,19)	0,967 (0,25)	0,926 (0,21)	0,943 (0,18)	0,963 (0,21)	0,984 (0,25)

TABLEAU 5.4 – Fréquence des allèles rares en troisième génération en fonction de la méthode de sélection des parents, de la densité de marquage et de la modalité dans les deux schémas de sélection simulés. L'écart-type de chaque valeur est indiqué entre parenthèses

Génération	Méthode	Pondération	Densité	h^2 = 0,2			h^2= 0,5			h^2 = 0,8		
				50	200	500	50	200	500	50	200	500
BC3	TS_no_update	Non	High	0,054 (0,05)	0,051 (0,04)	0,053 (0,05)	0,055 (0,05)	0,053 (0,04)	0,052 (0,04)	0,055 (0,05)	0,052 (0,04)	0,052 (0,05)
			Medium	0,057 (0,05)	0,049 (0,04)	0,054 (0,05)	0,049 (0,04)	0,052 (0,04)	0,053 (0,05)	0,057 (0,05)	0,051 (0,04)	0,053 (0,04)
		Oui	High	0,06 (0,05)	0,053 (0,05)	0,055 (0,05)	0,059 (0,05)	0,055 (0,05)	0,055 (0,05)	0,054 (0,05)	0,055 (0,05)	0,054 (0,05)
			Medium	0,062 (0,06)	0,053 (0,05)	0,055 (0,05)	0,059 (0,06)	0,054 (0,04)	0,055 (0,05)	0,058 (0,06)	0,055 (0,05)	0,055 (0,05)
	TS_update	Non	High	0,058 (0,05)	0,052 (0,04)	0,053 (0,05)	0,052 (0,05)	0,055 (0,05)	0,054 (0,05)	0,062 (0,06)	0,051 (0,05)	0,054 (0,05)
			Medium	0,057 (0,05)	0,053 (0,04)	0,055 (0,05)	0,054 (0,05)	0,055 (0,05)	0,053 (0,05)	0,057 (0,06)	0,052 (0,04)	0,054 (0,05)
		Oui	High	0,058 (0,05)	0,051 (0,04)	0,053 (0,05)	0,056 (0,05)	0,054 (0,05)	0,054 (0,05)	0,059 (0,06)	0,053 (0,05)	0,054 (0,05)
			Medium	0,06 (0,05)	0,057 (0,05)	0,057 (0,05)	0,061 (0,06)	0,056 (0,05)	0,057 (0,05)	0,063 (0,06)	0,057 (0,05)	0,058 (0,05)
	TS_update_inc	Non	High	0,059 (0,06)	0,055 (0,05)	0,057 (0,05)	0,061 (0,06)	0,056 (0,05)	0,056 (0,05)	0,06 (0,06)	0,057 (0,05)	0,056 (0,05)
			Medium	0,056 (0,05)	0,051 (0,04)	0,055 (0,05)	0,058 (0,06)	0,053 (0,05)	0,053 (0,05)	0,057 (0,05)	0,053 (0,04)	0,053 (0,05)
		Oui	High	0,062 (0,06)	0,056 (0,05)	0,057 (0,05)	0,062 (0,05)	0,06 (0,05)	0,058 (0,05)	0,065 (0,06)	0,059 (0,05)	0,058 (0,05)
			Medium	0,062 (0,06)	0,055 (0,05)	0,058 (0,05)	0,06 (0,05)	0,057 (0,05)	0,056 (0,05)	0,061 (0,06)	0,057 (0,05)	0,056 (0,05)
	BV	-	-	0,074 (0,07)	0,059 (0,05)	0,06 (0,05)	0,071 (0,07)	0,061 (0,05)	0,059 (0,05)	0,069 (0,07)	0,059 (0,05)	0,058 (0,05)
	Pheno	-	-	0,06 (0,05)	0,054 (0,04)	0,056 (0,04)	0,062 (0,05)	0,058 (0,05)	0,056 (0,05)	0,069 (0,06)	0,058 (0,05)	0,057 (0,05)
	Rand	-	-	0,054 (0,04)	0,052 (0,04)	0,055 (0,04)	0,054 (0,04)	0,054 (0,04)	0,055 (0,04)	0,052 (0,05)	0,053 (0,04)	0,054 (0,04)
IC3	TS_no_update	Non	High	0,076 (0,11)	0,074 (0,09)	0,071 (0,09)	0,071 (0,1)	0,07 (0,09)	0,067 (0,09)	0,063 (0,09)	0,066 (0,08)	0,071 (0,09)
			Medium	0,073 (0,11)	0,068 (0,08)	0,071 (0,09)	0,066 (0,09)	0,068 (0,09)	0,069 (0,09)	0,066 (0,09)	0,067 (0,09)	0,07 (0,09)
		Oui	High	0,106 (0,13)	0,08 (0,11)	0,083 (0,1)	0,102 (0,12)	0,087 (0,1)	0,078 (0,1)	0,088 (0,11)	0,082 (0,11)	0,085 (0,11)
			Medium	0,102 (0,13)	0,076 (0,1)	0,079 (0,1)	0,094 (0,12)	0,081 (0,1)	0,078 (0,1)	0,095 (0,1)	0,085 (0,11)	0,083 (0,11)
	TS_update	Non	High	0,087 (0,11)	0,079 (0,1)	0,077 (0,1)	0,086 (0,11)	0,077 (0,1)	0,073 (0,09)	0,091 (0,12)	0,07 (0,09)	0,075 (0,1)
			Medium	0,089 (0,12)	0,077 (0,09)	0,078 (0,1)	0,081 (0,11)	0,072 (0,09)	0,074 (0,09)	0,071 (0,1)	0,069 (0,09)	0,076 (0,09)
		Oui	High	0,105 (0,14)	0,092 (0,12)	0,088 (0,11)	0,097 (0,12)	0,088 (0,11)	0,087 (0,11)	0,106 (0,14)	0,094 (0,12)	0,089 (0,11)
			Medium	0,098 (0,12)	0,081 (0,1)	0,088 (0,11)	0,101 (0,15)	0,085 (0,11)	0,085 (0,1)	0,118 (0,15)	0,09 (0,11)	0,091 (0,11)
	TS_update_inc	Non	High	0,079 (0,1)	0,075 (0,1)	0,073 (0,09)	0,07 (0,1)	0,075 (0,09)	0,07 (0,09)	0,078 (0,11)	0,068 (0,09)	0,074 (0,09)
			Medium	0,08 (0,1)	0,071 (0,09)	0,074 (0,09)	0,07 (0,1)	0,072 (0,09)	0,07 (0,09)	0,064 (0,09)	0,072 (0,09)	0,072 (0,09)
		Oui	High	0,113 (0,14)	0,083 (0,11)	0,085 (0,11)	0,109 (0,15)	0,092 (0,11)	0,083 (0,1)	0,12 (0,14)	0,097 (0,12)	0,088 (0,11)
			Medium	0,102 (0,13)	0,083 (0,11)	0,084 (0,11)	0,11 (0,14)	0,087 (0,11)	0,081 (0,1)	0,11 (0,14)	0,091 (0,11)	0,087 (0,11)
	BV	-	-	0,175 (0,19)	0,105 (0,13)	0,097 (0,11)	0,157 (0,15)	0,103 (0,12)	0,089 (0,11)	0,144 (0,17)	0,097 (0,12)	0,09 (0,1)
	Pheno	-	-	0,114 (0,11)	0,082 (0,09)	0,084 (0,09)	0,102 (0,12)	0,093 (0,1)	0,085 (0,09)	0,12 (0,13)	0,091 (0,1)	0,084 (0,09)
	Rand	-	-	0,074 (0,07)	0,079 (0,07)	0,076 (0,07)	0,06 (0,06)	0,078 (0,07)	0,077 (0,07)	0,056 (0,06)	0,076 (0,07)	0,077 (0,07)

CHAPITRE 6

Discussion générale

L'évolution des méthodes de sélection depuis le début du vingtième siècle a indéniablement permis un progrès génétique important chez de nombreuses espèces, mais souvent au détriment de la diversité génétique utilisée, les programmes d'amélioration s'appuyant souvent sur un nombre restreint de géniteurs plus ou moins fortement apparentés. Cette base génétique étroite peut conduire à long terme à un progrès génétique limité, ce qui amène de nombreux acteurs du monde de la sélection à se questionner sur la meilleure façon d'utiliser la diversité génétique non exploitée à ce jour. Chez le pommier, la faible diversité génétique utilisée dans les schémas de sélection est un problème identifié de longue date (Noiton et Shelbourne 1992), mais peu de programmes de création variétale se sont intéressés à l'exploitation de la diversité existante dans les variétés anciennes et les espèces sauvages apparentées, probablement du fait du temps nécessaire pour mener à bien des actions de pré-breeding. Récemment, l'avènement de la génomique a permis le développement de méthodes de prédiction basées sur les données génétiques d'individus non phénotypés, permettant d'envisager une utilisation plus efficace des sources de diversité génétique. Dans ce contexte, l'objectif de mon travail de thèse était d'étudier l'intérêt de la sélection génomique dans le cadre d'actions de pré-breeding chez le pommier. La précision des prédictions génomiques repose sur plusieurs facteurs, parmi lesquels se trouvent la densité de marqueurs utilisée et la constitution de la population d'entrainement. Dans le chapitre 3 de la thèse, nous avons étudié la précision d'imputation qu'il était possible d'atteindre chez le pommier lorsque les données génomiques étaient imputées depuis une moyenne densité vers une haute densité de marqueurs. Dans le chapitre 4, nous avons utilisé ces données imputées afin de mesurer la précision de prédiction pour différents caractères lorsque les individus à prédire étaient représentés par un panel de ressources génétiques, de matériel élite ou d'hybrides

obtenus en croisant des individus de chacun des deux panels. Pour ce faire, nous avons étudié différentes façons de constituer la population d'entrainement nécessaire à la calibration des modèles de prédiction. Dans le chapitre 5 de la thèse, nous avons étudié par simulation deux schémas de sélection permettant d'augmenter la fréquence d'allèles favorables rares dans du matériel élite mais présents dans les ressources génétiques en s'appuyant sur la sélection génomique. Dans cette discussion générale, je reviens sur les résultats principaux obtenus pendant la thèse et dresse le bilan de pistes restant à explorer pour la mise en place de la sélection génomique dans des actions de pré-breeding chez le pommier.

5.1 Intérêt de la sélection génomique chez le pommier

La sélection génomique a très tôt été envisagée comme une approche prometteuse chez les espèces pérennes car elle pouvait en théorie permettre une réduction importante des cycles de sélection (Grattapaglia et al. 2018 ; Kumar et al. 2012a) grâce à la mise en place d'une sélection précoce. Cependant, avant le début de cette thèse peu d'études avaient été menées quant à la précision de prédiction qu'il était possible d'atteindre pour différents caractères chez le pommier, et les candidats dans ces études étaient dans la plupart des cas des génotypes issus de familles de plein-frères (Kumar et al. 2012b ; 2015). L'identification de ressources génétiques de bonne valeur agronomique par prédiction génomique était alors une piste peu explorée (Migicovsky et al. 2016). Nous avons montré dans le chapitre 4 qu'il était possible d'obtenir des précisions de prédiction modérées (0,34 pour le croquant du fruit) à élevées (0,82 pour la date de récolte) en utilisant une population d'entrainement constituée de ressources génétiques pour prédire d'autres ressources génétiques, et que la précision de prédiction pouvait également être élevée lorsque ces mêmes candidats étaient prédits en utilisant des données provenant de matériel élite (0,13 pour la jutosité du fruit à 0,69 pour la couleur du fruit). Dans le cas de génotypes résultant d'un croisement entre un parent élite et un parent provenant d'un panel de ressources génétiques, nous avons montré à partir de données réelles (chapitre 4) et simulées (chapitre 5) que des précisions de prédiction moyennes à élevées pouvaient être obtenues si la population d'entrainement était constituée de génotypes provenant à la fois des panels de matériel élite et des ressources génétiques ou si les candidats des générations précédentes étaient utilisés comme population d'entrainement dans le cas de générations avancées de croisements. L'ensemble de ces résultats montrent que l'identification, via la prédiction génomique, de génotypes intéressants provenant des ressources génétiques est tout à fait envisageable dans le cadre d'actions de pré-breeding. Nous avons aussi montré que l'amélioration de ces génotypes par croisements avec du matériel élite est possible grâce à la sélection génomique, auquel cas, la durée entre deux générations pourrait être réduite. En considérant une durée de 4 ans entre deux générations grâce à la sélection génomique contre 7 ans au minimum pour la sélection phénotypique dans

nos simulations, nous avons ainsi observé un gain génétique par unité de temps supérieur lorsque la sélection génomique était utilisée plutôt que la sélection phénotypique. Nous avons également montré par simulation que l'utilisation de la sélection génomique dans des schémas de pré-breeding pouvait être un atout pour transférer des allèles rares favorables vers du élite depuis des variétés issues de ressources génétiques lorsqu'un poids plus important était donné à de tels allèles dans les modèles de prédiction. La proportion moyenne du génome provenant des ressources génétiques chez les descendants de la 3$^{\text{ème}}$ génération était en outre plus élevée que lorsque la fréquence des allèles rares n'était pas prise en compte dans le modèle. Si le choix des géniteurs d'un programme de pré-breeding est basé sur des prédictions génomiques, comme dans le cas du chapitre 5, il est possible que les individus issus des ressources génétiques ayant les GEBV les plus élevées soient proches du matériel élite d'un point de vue génétique du fait du relativement faible nombre de générations entre le matériel élite et les variétés provenant des ressources génétiques. Dans ce cas, ces individus transmettront peu de nouveaux allèles rares dans les générations successives. Au contraire, si le choix des géniteurs du programme de pré-breeding est basé sur la sélection phénotypique, il est possible que ce choix initial des géniteurs soit pertinent du point de vue de l'élargissement de la base génétique du matériel élite, mais il existe un risque de sélectionner des hybrides dont une partie importante du génome provient du matériel élite dans les générations ultérieures. L'utilisation de marqueurs moléculaires dans les programmes de pré-breeding et la pondération des allèles rares dans les modèles de prédiction permettent de limiter les risques rencontrés dans les deux situations évoquées. L'utilisation de la pondération des allèles rares est d'autant plus intéressante que, contrairement à d'autres études basées sur de telles méthodes de pondération (Liu et al. 2015 ; Sun et VanRaden 2014), nous n'avons pas observé de pénalisation du gain génétique à court terme lorsque nous avons appliqué des pondérations aux prédictions génomiques. En revanche, nous avons observé une perte de variance génétique plus importante dans le cas de l'utilisation de la sélection génomique que dans le cas de la sélection phénotypique, comme cela a pu être observé dans d'autres études menées par simulation (Breider et al. 2022 ; Wientjes et al. 2022). Dans notre cas, l'étude a porté sur un faible nombre de générations et il serait intéressant de savoir si la perte de variance à plus long terme peut être un frein à l'utilisation de la sélection génomique chez le pommier. Cependant, il est peu probable que la variance génétique disponible soit épuisée au bout de quelques générations (par exemple Silva et al. (2021) montrent par simulation que pour une intensité de sélection de 2,5% il faut au moins 50 générations avant que la variance génétique ne soit épuisée), ce qui nous amène à penser que la sélection génomique est une approche de choix pour du pré-breeding chez le pommier. Les potentiels freins à sa mise en place sont donc plutôt d'ordre logistique et économique, comme précisé ci-après.

5.2 Mise en place de la sélection génomique dans les programmes de pré-breeding : limites et opportunités

Bien que la sélection génomique présente un réel intérêt chez le pommier, nous avons identifié un certain nombre de contraintes qui pourraient retarder sa mise en place dans les programmes de pré-breeding, voire de création variétale en général. Je discute ici certaines de ces limites et des pistes potentielles permettant de les dépasser.

5.2.1 Constitution de la population d'entrainement

Chez les espèces arboricoles, générer et maintenir des populations de grande taille est un investissement onéreux et pouvant conduire à un travail de phénotypage fastidieux. La taille de la population d'entrainement peut donc être un premier frein à la mise en place de la sélection génomique chez le pommier. Dans la majorité des programmes actuels d'amélioration du pommier à l'échelle internationale, les actions de pré-breeding et de développement de variétés sont distincts, ce qui pourrait avoir pour conséquence de devoir constituer une population d'entrainement destinée à la prédiction du matériel développé dans le cadre du pré-breeding et une autre population d'entrainement dédiée au matériel élite en cours d'évaluation et donc être une contrainte pour les raisons évoquées en début de paragraphe. Nous avons montré dans le chapitre 4 qu'une population représentative de la diversité du pommier (dans notre cas en comprenant à la fois du matériel élite et des variétés anciennes) pouvait être utilisée pour prédire différentes populations de candidats, ce qui signifie qu'une seule et même population pourrait servir de population d'entrainement chez le pommier. De ce point de vue, la population REFPOP, constituée à la fois de ressources génétiques et de descendances élite, pourrait être utilisée comme population d'entrainement, d'autant plus qu'elle présente l'avantage d'être déjà en place sur 6 sites européens, ce qui représente un réseau expérimental extrêmement utile pour la promotion de la sélection génomique chez le pommier. Dans le cadre d'un programme de pré-breeding, il est de plus important d'utiliser une population d'entrainement suffisamment diverse afin de représenter les différents allèles aux QTL. D'une part, nous avons vu dans le chapitre 4 que la précision de prédiction pouvait drastiquement chuter lorsqu'un allèle était fixé dans la population d'entrainement et ségrégeait chez les candidats. D'autre part, nous avons rappelé au chapitre 5 l'importance d'estimer avec précision l'effet des allèles rares du matériel élite en vue de les valoriser par pondération dynamique. Or estimer les effets de tels allèles avec précision implique d'utiliser une population d'entrainement dans laquelle ces allèles sont présents en fréquence suffisamment élevée. Enfin, nous pensons que certains défauts pouvant apparaître lorsque des ressources génétiques sont utilisées dans des programmes de pré-breeding doivent être pris en compte sur la base de prédictions génomiques s'ils sont contrôlés par un

grand nombre de loci (comme cela est le cas pour la mauvaise conservation du fruit ou l'alternance, qui sont des défauts considérés comme rédhibitoires par les sélectionneurs actuellement). Cela implique d'inclure un certain nombre de génotypes présentant ces défauts et de mettre en place un phénotypage fiable de ces défauts dans la population d'entrainement afin de les prédire correctement, ce qui renforce la nécessité d'utiliser une population d'entrainement diverse.

Nous avons également montré que des données historiques pouvaient être utilisées pour augmenter la taille de la population d'entrainement, et qu'inversement il était possible d'optimiser la constitution de la population d'entrainement, par exemple en utilisant l'approche proposée par Rincent et al. (2017), afin de réduire sa taille et donc les coûts associés. Là encore, le réseau REFPOP peut être utilisé pour phénotyper des caractères difficiles à évaluer ou peu étudiés jusqu'à présent. Cependant, pour certains caractères tels que la résistance à des maladies émergentes, de nouveaux vergers devront être mis en place pour ne pas compromettre les évaluations de caractères agronomiques de production, de phénologie et de qualité du fruit. Dans le cas de caractères liés au changement climatique, il sera potentiellement nécessaire d'évaluer les génotypes d'intérêt dans de nouveaux environnements dont les caractéristiques se rapprochent des conditions climatiques à venir en France. Pour ce faire, la collaboration entre instituts de recherche est une voie à privilégier et le réseau REFPOP y contribue largement. Une telle collaboration pourrait par la même occasion permettre de mutualiser les coûts liés au génotypage et au phénotypage en exploitant conjointement des données acquises par les différents instituts (Jung et al. 2022).

Enfin, nous avons vu dans le chapitre 5 qu'une même population d'entrainement pouvait être utilisée pendant plusieurs générations dans le cadre d'actions de pré-breeding, même s'il reste préférable d'accumuler et exploiter les phénotypes des générations précédentes. Compte tenu des contraintes associées à une mise à jour de la population d'entrainement à chaque génération, l'utilisation au cours des générations d'une unique population d'entrainement à large diversité représente là encore une possibilité. Nous nous sommes limités à des schémas de pré-breeding s'étalant sur trois générations dans nos simulations, ce qui ne nous permet pas de donner des préconisations quant au renouvellement de la population d'entrainement dans le cas où elle n'est pas mise à jour à chaque génération. Ce renouvellement dépendra vraisemblablement du schéma de sélection adopté d'une part, puisque nous avons vu que la diminution de la précision de prédiction était moins importante dans le schéma des rétrocroisements que dans celui des intercroisements, et de la taille de la population d'entrainement d'autre part, étant donné qu'une population d'entrainement de grande taille permet généralement d'obtenir de meilleures précisions de prédiction, auquel cas la diminution de la précision de prédiction est sans doute moins préjudiciable. Enfin, il est évident que la population d'entrainement devra être mise à jour si de nouveaux caractères sont pris en compte dans le programme de pré-breeding ou si de nouvelles sources de diversité qui ne sont pas représentées dans la population d'entrainement

sont exploitées (par exemple du matériel exotique ou des espèces sauvages apparentées, voir plus bas).

5.2.2 Phénotypage de la population d'entrainement

Le développement rapide de nouveaux outils de phénotypage chez le pommier pourrait permettre de phénotyper un plus grand nombre d'individus qu'actuellement, tout en assurant un phénotypage plus précis que dans le cas de notations manuelles pour certains caractères. Par exemple, des travaux menés par les équipes ImHorPhen et VadiPom de l'IRHS d'Angers ont permis de mettre en œuvre des algorithmes d'analyse d'images pour progressivement améliorer l'évaluation de la couleur du fruit (Couasnet et al. 2021). De même, des caractères liés à la forme du fruit sont généralement notés par les sélectionneurs en utilisant une description du fruit ou dans certains cas une échelle ordinale pouvant varier d'un institut à l'autre, ce qui complique l'utilisation de données acquises sur différents sites ou par différents sélectionneurs. Différents projets utilisant des algorithmes de machine learning et de l'analyse d'images sont actuellement en cours (Z.Migicovsky, communication personnelle ; C.Dujak, communication personnelle) et pourraient permettre une meilleure étude des caractères liés à la forme du fruit. Afin de minimiser le travail de récolte, ces nouveaux outils de phénotypage devraient dans l'idéal être utilisés en verger, par exemple via l'utilisation de drones (Coupel-Ledru et al. 2019) ou d'appareils d'acquisition d'images peu encombrants et facilement transportables. Des essais de ce type ont lieu sur la REFPOP dans le cadre du projet INVITE et ont déjà révélé le potentiel de ces méthodes. Par exemple, Zine-El-Abidine et al. (2021) ont développé un algorithme capable de détecter les pommes d'un arbre à partir de photographies avec un taux de réussite de 95%.

5.2.3 Génotypage des candidats à la sélection

En parallèle du phénotypage, le génotypage des candidats à la sélection peut également représenter un autre frein à l'utilisation de la sélection génomique. Le nombre de marqueurs utilisés dans les prédictions génomiques peut avoir une influence sur la précision de prédiction obtenue, surtout si les candidats sont issus de ressources génétiques. En effet dans ce cas le déséquilibre de liaison décroît plus vite que pour un panel constitué de matériel élite, ce qui peut nécessiter d'utiliser des données de marquage haute densité pour atteindre des précisions de prédiction correctes. Or nous avons d'une part montré au chapitre 3 qu'il était possible d'imputer précisément des données moyenne densité vers de la haute densité chez le pommier, et d'autre part nous avons observé aux chapitres 4 et 5 des différences minimes de précision de prédiction lorsque les données moyenne densité étaient utilisées à la place des données haute densité. L'utilisation de données haute densité a tout de même permis une meilleure gestion des allèles rares dans les programmes simulés de pré-breeding, et de manière générale une haute densité

de marqueurs est préconisée pour la gestion de la diversité chez différentes espèces (Eynard et al. 2016). Le coût de génotypage représente actuellement environ 35€/échantillon pour du génotypage moyenne densité et 85 à 90€/échantillon pour du génotypage haute densité, ce qui nous a amené à conclure que la stratégie la plus prometteuse consiste à génotyper les candidats à la sélection à moyenne densité puis à les imputer, tout en génotypant les parents des candidats ainsi que leurs ancêtres à haute densité afin de régulièrement mettre à jour le panel de référence utilisé lors de l'étape d'imputation.

5.3 Suite à donner aux actions de pré-breeding

Enfin, une dernière limitation potentielle à la mise en place de la sélection génomique dans le cadre d'actions de pré-breeding est liée à la vision à long terme dans les programmes de sélection chez le pommier. En effet, compte tenu de la durée des programmes d'amélioration, les acteurs de la filière se projettent actuellement à court terme pour définir les objectifs de sélection et les croisements à effectuer. Dans le chapitre 5, c'est la raison pour laquelle nous nous sommes projetés sur une durée de trois générations et les actions à mener au-delà de ces trois générations ne sont pas claires, ce qui rend une planification à long terme basée sur la sélection génomique incertaine. D'une part, il est possible que les personnes impliquées au début des actions de pré-breeding ne soient plus en activité au bout de la troisième génération, ce qui pourrait entrainer une déviation par rapport à la vision initialement envisagée pour les actions de pré-breeding. D'autre part, les objectifs de sélection ont largement le temps de changer en 20 ans chez le pommier, en particulier en lien avec le changement climatique et les attentes de la société. Enfin, du fait du grand nombre de critères impliqués dans la création variétale chez le pommier, il est probable que les acteurs de la sélection désirent phénotyper de manière poussée le matériel issu des actions de pré-breeding à la fin de la troisième génération de croisements (voire avant). La stratégie adoptée dans les programmes de pré-breeding a alors de grandes chances d'être modifiée en fonction des observations. Au-delà de 3 générations, plusieurs scénarios peuvent être envisagés :

— Compte tenu de l'évolution du matériel élite en parallèle du programme de pré-breeding, un premier scénario basique consisterait à reprendre la même stratégie que dans les 3 premières générations mais en utilisant le matériel élite amélioré à la place du matériel élite initial et en renouvelant les variétés anciennes provenant des ressources génétiques utilisées dans les croisements. Il serait d'autant plus nécessaire de renouveler le choix des variétés provenant des ressources génétiques dans le cas où de nouveaux objectifs de sélection apparaitraient. Si une même population d'entrainement a été utilisée pour construire les équations de prédiction des trois premières générations, nous recommandons de mettre à jour cette population d'entrainement afin de prendre en compte

l'évolution du matériel élite. Dans ce cas, les données génotypiques et phénotypiques des individus évalués dans le cadre du développement du matériel élite pourraient être ajoutées à la population d'entrainement précédemment utilisée.

— Le scénario précédent se base sur l'identification de génotypes provenant de ressources génétiques sur la base de leur valeur agronomique estimée au travers d'observations phénotypiques ou éventuellement via leur GEBV. Comme évoqué au chapitre 5, une autre approche consisterait à sélectionner des génotypes pour leur complémentarité avec le matériel élite, en identifiant par prédiction génomique des régions favorables chez certains individus des ressources génétiques et pour lesquelles les allèles favorables sont absents ou rares chez le matériel élite. Cette approche nécessite d'estimer précisément les effets des marqueurs, ce qui n'est pas forcément possible au début d'un programme de pré-breeding compte tenu de la taille restreinte de la population d'entrainement et donc de la représentation limitée de certains allèles. Après 3 générations, la taille de la population d'entrainement pourra être supérieure à la taille initiale grâce à l'utilisation des données phénotypiques et génotypiques des générations précédentes si la population d'entrainement est mise à jour à chaque génération, permettant alors une meilleure estimation de ces effets. De plus, si la fréquence des allèles rares favorables augmente au cours des générations, la mise à jour de la population devrait là encore permettre une meilleure estimation des effets de ces marqueurs. Dans le cas d'une population d'entrainement à large diversité, il nous semblerait également intéressant de valoriser les pépins obtenus par pollinisation libre en verger en en plantant et génotypant un certain nombre. Prenons l'exemple de la REFPOP, qui correspond à une telle population. Les fruits de la REFPOP résultent de pollinisations libres (aussi dites "open") impliquant les génotypes du matériel élite et des ressources génétiques avoisinants et certains allèles rares dans la REFPOP pourraient alors être présents en plus grande quantité dans les descendances "open" de la REFPOP. Etant donné que la REFPOP donne déjà des fruits, leur valorisation permettrait en outre de gagner une génération de croisement et d'exploiter davantage de données

— Dans le cas mentionné plus haut où les objectifs de sélection au-delà des 3 générations ont évolué, il pourrait être nécessaire d'utiliser des génotypes provenant de matériel exotique ou d'espèces sauvages apparentées, notamment afin de faire face aux contraintes liées au changement climatique. L'utilisation de tels génotypes dans un programme de pré-breeding peut se révéler complexe, notamment du fait de leur grande distance génétique avec le matériel élite et de leurs différences en termes de performances agronomiques. Nous discutons certains aspects de cette problématique plus en détail dans la section suivante.

5.4 Quels caractères sélectionner dans les programmes de pré-breeding et quel matériel végétal utiliser pour y parvenir ?

A ce jour, l'utilisation de ressources génétiques ou d'espèces sauvages apparentées s'est largement focalisée sur la résistance aux stress biotiques chez le pommier. Les demandes sociétales actuelles vont dans le sens d'une agriculture dépourvue de pesticides, ce qui devrait continuer à placer de telles résistances au cœur des programmes de sélection. La majorité des variétés inscrites sont résistantes à la tavelure et dans une moindre mesure au feu bactérien, mais des contournements de gènes majeurs ont été observés depuis plusieurs années dans les deux cas (Emeriewen et al. 2019 ; Papp et al. 2020b ; Parisi et al. 1993), rendant l'introgression de nouveaux QTL de résistance désirable. Dans le cas de la tavelure, il existe plusieurs sources de résistance au sein du pool domestique chez le pommier (Gessler et al. 2006), mais pour d'autres caractères comme la résistance à l'oïdium ou au feu bactérien, les allèles d'intérêt se trouvent plutôt dans le pool sauvage. Les caractères liés à la qualité du fruit sont également systématiquement considérés lors du processus de création variétale chez le pommier, et il est tout à fait envisageable d'utiliser des ressources génétiques pour introduire de nouveaux arômes dans les variétés élite (Larsen et al. 2019), ou pour développer de nouvelles caractéristiques originales pour le consommateur et la filière, tel qu'illustré récemment par la création de variétés à chair rouge (Wang et al. 2018). L'utilisation de ressources génétiques pour améliorer des caractères liés à la qualité est cependant plus incertaine que dans le cas de la résistance aux stress biotiques. D'une part, certains caractères n'ont pas besoin d'être améliorés dans le matériel élite, et l'utilisation de ressources génétiques pourrait conduire à casser des associations favorables pour ces caractères. Par exemple, la conservation du fruit est un élément central des programmes d'amélioration chez le pommier et les variétés élite actuelles ont généralement une longue durée de conservation (Migicovsky et al. 2021), tandis que de nombreuses variétés anciennes se conservent mal. D'autre part, les sélectionneurs recherchent des optimums plutôt que des maximums pour de nombreux caractères liés à la qualité organoleptique chez la pomme, tels que la perception du sucre et de l'acidité du fruit. Dans ce cas, le transfert de nouveaux allèles favorables depuis les ressources génétiques n'est pas une priorité puisque la gestion de ces optimums peut être envisagée au sein du matériel élite au travers de nouvelles associations d'allèles déjà présents.

Enfin, l'intérêt des ressources génétiques et des espèces sauvages apparentées pour de nouveaux objectifs de sélection liés au changement climatique est indéniable. Jusqu'à présent, les programmes d'amélioration chez le pommier visaient à développer des variétés adaptées à une large gamme de conditions pédoclimatiques, rendant potentiellement ces variétés mal adaptées

TABLEAU 6.1 – Caractères d'intérêt pour lesquels les espèces sauvages apparentées du pommier domestique pourraient être utilisés

Caractère	*M. sieversii*	*M. sylvestris*	*M. orientalis*	*M. prunifolia*
Tavelure	X		X	X
Feu bactérien	X			X
Chancre à Nectria		X		
Oïdium	X		X	X
Puceron lanigère	X			
Floraison tardive	X		X	
WUE*	X			
Tolérance à la sècheresse	X			

Adapté de Bramel et Volk (2019)
* WUE : Water-Use Efficiency

au changement climatique dans certaines zones du fait d'événements extrêmes à venir (par exemple des périodes prolongées de sècheresse) ou même déjà observés telles que les gelées d'avril 2021 ayant fortement touché les arboriculteurs français. Anticiper de tels événements pourrait conduire à la création de variétés à floraison plus tardive qu'actuellement ou à l'utilisation de génotypes issus des ressources génétiques et nécessitant une utilisation plus modérée d'eau (Coupel-Ledru et al. 2022). Certaines espèces sauvages apparentées présentent également une tolérance à la sècheresse qui pourrait être exploitée dans les programmes de pré-breeding (Bramel et Volk 2019). Pour de tels caractères, la prédiction génomique peut être plus complexe du fait des interactions génotype x environnement et de l'architecture génétique, ce qui a conduit plusieurs auteurs à préconiser l'utilisation de modèles écophysiologiques dans lesquels les paramètres génétiques d'un modèle écophysiologique sont déterminés par prédiction génomique (Cooper et al. 2016 ; Heslot et al. 2014). La prédiction génomique de l'adaptation différentielle de génotypes en fonction des environnements peut aussi être mise en œuvre par le biais de modèles combinant des données acquises sur plateforme de phénotypage haut-débit et des données de caractérisation fine de l'environnement (Millet et al. 2019). Comme évoqué dans les paragraphes précédents, certaines espèces sauvages apparentées présentent donc un intérêt indéniable pour l'amélioration du pommier, en particulier vis-à-vis de la réponse aux stress biotiques et abiotiques (tableau 6.1). Dans le cadre de ma thèse, nous nous sommes uniquement intéressés à des donneurs potentiels provenant du pool primaire du pommier, mais il me semble pertinent de se demander si les mêmes schémas de sélection pourraient être utilisés pour des espèces sauvages apparentées. Bien que les croisements entre *M. domestica* et les espèces apparentées les plus proches soient relativement faciles, il nous semble encore difficile d'envisager l'utilisation directe de ce type de matériel dans les programmes de pré-breeding en mobilisant la prédiction génomique. D'une part, les puces de génotypage disponibles actuellement chez le pommier ne sont pas adaptées au génotypage d'espèces sauvages, rendant les stratégies basées

sur l'utilisation de prédictions génomiques hasardeuses. Notons qu'à ce titre, une puce de génotypage comprenant un peu moins de 50 000 marqueurs SNP est en cours de développement au Julius Kühn Institut de Dresden en Allemagne (A.Peil, communication personnelle) et qu'environ 4000 de ces marqueurs ciblent 6 espèces apparentées du pommier. Une approche alternative pourrait consister à utiliser des modèles de prédiction dont la matrice d'apparentement a été estimée en utilisant des données phénotypiques acquises de manière non-destructive (Van Tassel et al. 2022), comme dans le cas de données NIRS (Rincent et al. 2018) ou hyperspectrales (Krause et al. 2019). D'autre part, les fruits des espèces sauvages apparentées sont souvent de petite taille et riches en composés phénoliques, rendant alors l'évaluation de la qualité du fruit impossible dans les premières générations d'intercroisement. Des défauts additionnels tels que le port de l'arbre ou la position de l'inflorescence peuvent compliquer le phénotypage. Dans une optique d'évaluation des génotypes basée sur la prédiction génomique, il faudrait alors évaluer le potentiel des espèces sauvages dans un fond génétique élite. Serra et al. (2016) ont à ce titre proposé une approche s'appuyant sur le principe des lignées d'introgressions afin d'introduire des fragments d'un parent exotique dans un fond génétique élite chez les espèces arboricoles de façon à ce que la majorité du génome du parent exotique soit représenté dans la descendance du croisement tout en limitant la taille des fragments exotiques par descendant. Compte tenu du coût de la mise en place d'une telle population et du temps nécessaire pour générer de tels génotypes, l'utilisation de cette approche devrait être envisagée dès la mise en place d'un programme de pré-breeding.

5.5 Conclusion générale

Les résultats acquis au cours de cette thèse montrent que la sélection génomique est une approche pertinente pour assurer l'efficacité d'actions de pré-breeding chez le pommier : l'utilisation de prédictions génomiques peut permettre d'une part d'identifier des variétés issues des ressources génétiques à valoriser et d'autre part guider le choix des candidats à la sélection dans des schémas plus avancés de pré-breeding, en permettant dans ce cas un gain génétique par unité de temps plus important que si la sélection phénotypique avait été utilisée. Pour définitivement valider l'intérêt de la sélection génomique chez le pommier, les travaux menés par simulations dans cette thèse gagneraient maintenant à être confirmés par des résultats acquis sur des populations en verger.

Si l'utilisation de la sélection génomique dans les programmes de pré-breeding du pommier est alors envisagée, il me semble nécessaire de prendre en compte trois aspects que nous avons peu explorés dans le cadre de la thèse. Dans un premier temps, la dimension économique de la mise en place de la sélection génomique devra être envisagée et les coûts comparés à un programme reposant sur la sélection phénotypique afin de définir entre autres le nombre de croisements à

effectuer, la taille des familles et l'intensité de sélection. L'équipe Vadipom de l'IRHS d'Angers est actuellement en train de chiffrer le coût que représente chaque étape d'un programme d'amélioration du pommier, ce qui devrait permettre de s'appuyer sur des éléments concrets pour optimiser la mise en place d'un programme de pré-breeding. Le choix des géniteurs à utiliser dans un programme de pré-breeding est également un élément prépondérant. Les parents à la base du programme de pré-breeding peuvent être déterminés sur la base des connaissances empiriques des sélectionneurs, ce qui présente l'avantage d'éviter le choix d'individus présentant des défauts rédhibitoires, mais présente l'inconvénient d'exclure tout matériel qui n'aurait pas été phénotypé pendant plusieurs saisons. L'utilisation de prédictions génomiques peut alors guider le choix d'individus intéressants à utiliser comme géniteurs, soit en se basant sur leurs valeurs de GEBV, soit en cherchant à utiliser des parents complémentaires du matériel élite. Enfin, les caractères à prendre en compte dans le programme de pré-breeding devront être clairement définis en amont afin de déterminer sur quels critères les géniteurs seront choisis au cours des générations du programme. Chez le pommier, les critères de sélection sont sans doute trop nombreux pour être tous pris en compte dans un programme de pré-breeding, mais la qualité du fruit, le rendement et la qualité d'adaptation à différents environnements doivent au minimum être pris en compte, en plus d'éventuels autres caractères liés à de nouveaux objectifs de sélection. Dans cette thèse, nous avons suggéré d'utiliser un index de sélection afin de simultanément prendre en compte ces caractères dans un programme de pré-breeding. Il serait aussi envisageable de mettre en place plusieurs sous-programmes portant sur des sous-populations d'amélioration différenciées, chacune étant consacrée à l'amélioration d'un ou quelques caractères en particulier, et d'intercroiser par la suite les individus sélectionnés dans ces programmes.

Références citées

Adam, D. (2021). How Far Will Global Population Rise ? Researchers Can't Agree. *Nature* 597 : 462-465 (cf. p. 1).

Akdemir, D., J. I. Sanchez et J.-L. Jannink (2015). Optimization of Genomic Selection Training Populations with a Genetic Algorithm. *Genet Sel Evol* 47 : 38 (cf. p. 12).

Allier, A., C. Lehermeier, A. Charcosset, L. Moreau et S. Teyssèdre (2019a). Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection. *Front. Genet.* 10 : 1006 (cf. p. 16, 17, 128).

Allier, A., L. Moreau, A. Charcosset, S. Teyssèdre et C. Lehermeier (2019b). Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses : Application to Polygenic Trait Introgression. *G3* 9 : 1469-1479 (cf. p. 17).

Allier, A., S. Teyssèdre, C. Lehermeier, A. Charcosset et L. Moreau (2020a). Genomic Prediction with a Maize Collaborative Panel : Identification of Genetic Resources to Enrich Elite Breeding Programs. *Theor Appl Genet* 133 : 201-215 (cf. p. 150).

Allier, A., S. Teyssèdre, C. Lehermeier, L. Moreau et A. Charcosset (2020b). Optimized Breeding Strategies to Harness Genetic Resources with Different Performance Levels. *BMC Genomics* 21 : 349 (cf. p. 3, 17, 18).

Antolín, R., C. Nettelblad, G. Gorjanc, D. Money et J. M. Hickey (2017). A Hybrid Method for the Imputation of Genomic Data in Livestock Populations. *Genet Sel Evol* 49 : 30 (cf. p. 59).

Asoro, F. G., M. A. Newell, W. D. Beavis, M. P. Scott et J.-L. Jannink (2011). Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. *The Plant Genome* 4 : 132-144 (cf. p. 10).

Azodi, C. B., E. Bolger, A. McCarren, M. Roantree, G. de los Campos et S.-H. Shiu (2019). Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3* 9 : 3691-3702 (cf. p. 9).

Bančič, J., C. R. Werner, R. C. Gaynor, G. Gorjanc, D. A. Odeny, H. F. Ojulong, I. K. Dawson, S. P. Hoad et J. M. Hickey (2021). Modeling Illustrates That Genomic Selection Provides New Opportunities for Intercrop Breeding. *Front. Plant Sci.* 12 : 605172 (cf. p. 130).

Baumgartner, I. O., M. Kellerhals, F. Costa, L. Dondini, G. Pagliarani, R. Gregori, S. Tartarini, L. Leumann, F. Laurens et A. Patocchi (2016). Development of SNP-based Assays for Disease Resistance and Fruit Quality Traits in Apple (Malus × Domestica Borkh.) and Validation in Breeding Pilot Studies. *Tree Genetics & Genomes* 12 : 35 (cf. p. 26).

Bayer, P. E., B. Valliyodan, H. Hu, J. I. Marsh, Y. Yuan, T. D. Vuong, G. Patil, Q. Song, J. Batley, R. K. Varshney, H.-M. Lam, D. Edwards et H. T. Nguyen (2022). Sequencing the USDA Core Soybean Collection Reveals Gene Loss during Domestication and Breeding. *The Plant Genome* 15 (cf. p. 2).

Ben Sadok, I., A. Tiecher, D. Galvez-Lopez, M. Lahaye, P. Lasserre-Zuber, M. Bruneau, S. Hanteville, R. Robic, R. Cournol et F. Laurens (2015). Apple Fruit Texture QTLs : Year and Cold Storage Effects on Sensory and Instrumental Traits. *Tree Genetics & Genomes* 11 : 119 (cf. p. 23).

Ben Sadoun, S. (2021). Integration of Genomic Selection into Winter-Type Bread Wheat Breeding Schemes : A Simulation Pipeline Including Economic Constraints. *Crop Breed Genet Genom* 3 : 4 (cf. p. 151).

Bernardo, R. (1994). Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. *Crop Sci.* 34 : 20-25 (cf. p. 6).

Bernardo, R. (2008). Molecular Markers and Selection for Complex Traits in Plants : Learning from the Last 20 Years. *Crop Sci.* 48 : 1649-1664 (cf. p. 26).

Bernardo, R. (2009). Genomewide Selection for Rapid Introgression of Exotic Germplasm in Maize. *Crop Sci.* 49 : 419-425 (cf. p. 17).

Bernardo, R. (2014). Genomewide Selection When Major Genes Are Known. *Crop Sci.* 54 : 68-75.

Bernardo, R. (2016). Genomewide Predictions for Backcrossing a Quantitative Trait from an Exotic to an Adapted Line. *Crop Sci.* 56 : 1067-1075 (cf. p. 17).

Bernardo, R. et J. Yu (2007). Prospects for Genomewide Selection for Quantitative Traits in Maize. *Crop Sci.* 47 : 1082-1090 (cf. p. 7).

Berry, D. et J. Kearney (2011). Imputation of Genotypes from Low- to High-Density Genotyping Platforms and Implications for Genomic Selection. *Animal* 5 : 1162-1169 (cf. p. 148).

Bianco, L., A. Cestaro, G. Linsmith, H. Muranty, C. Denancé, A. Théron, C. Poncet, D. Micheletti, E. Kerschbamer, E. A. Di Pierro, S. Larger, M. Pindo, E. Van de Weg, A. Davassi, F. Laurens, R. Velasco, C.-E. Durel et M. Troggio (2016). Development and Validation of the Axiom ® Apple480K SNP Genotyping Array. *Plant J* 86 : 62-74 (cf. p. 24, 37, 43).

Bianco, L., A. Cestaro, D. J. Sargent, E. Banchi, S. Derdak, M. Di Guardo, S. Salvi, J. Jansen, R. Viola, I. Gut, F. Laurens, D. Chagné, R. Velasco, E. van de Weg et M. Troggio (2014). Development and Validation of a 20K Single Nucleotide Polymorphism (SNP) Whole Genome Genotyping Array for Apple (Malus × Domestica Borkh). *PLoS ONE* 9 : e110377 (cf. p. 24, 37).

Bink, M. C. A. M., J. Jansen, M. Madduri, R. E. Voorrips, C.-E. Durel, A. B. Kouassi, F. Laurens, F. Mathis, C. Gessler, D. Gobbin, F. Rezzonico, A. Patocchi, M. Kellerhals, A. Boudichevskaia, F. Dunemann, A. Peil, A. Nowicka, B. Lata, M. Stankiewicz-Kosyl, K. Jeziorek, E. Pitera, A. Soska, K. Tomala, K. M. Evans, F. Fernández-Fernández, W. Guerra, M. Korbin, S. Keller, M. Lewandowski, W. Plocharski, K. Rutkowski, E. Zurawicz, F. Costa, S. Sansavini, S. Tartarini, M. Komjanc, D. Mott, A. Antofie, M. Lateur, A. Rondia, L. Gianfranceschi et W. E. van de Weg (2014). Bayesian QTL Analyses Using Pedigreed Families of an Outcrossing Species, with Application to Fruit Firmness in Apple. *Theor Appl Genet* 127 : 1073-1090 (cf. p. 24).

Biscarini, F., N. Nazzicari, M. Bink, P. Arús, M. J. Aranzana, I. Verde, S. Micali, T. Pascal, B. Quilot-Turion, P. Lambert, C. da Silva Linge, I. Pacheco, D. Bassi, A. Stella et L. Rossini (2017). Genome-Enabled Predictions for Fruit Weight and Quality from Repeated Records in European Peach Progenies. *BMC Genomics* 18 : 432 (cf. p. 6).

Blissett, E. (2019). « Genomewide Selection in Apple : Prediction and Postdiction in the University of Minnesota Apple Breeding Program ». Univeristy of Minnesota. 102 p. (cf. p. 27).

Boichard, D., V. Ducrocq, P. Croiseau et S. Fritz (2016). Genomic Selection in Domestic Animals : Principles, Applications and Perspectives. *Comptes Rendus Biologies* 339 : 274-277 (cf. p. 6).

Bramel, P. et G. Volk (2019). *A Global Strategy for the Conservation and Use of Apple Genetic Resources* (cf. p. 30, 178).

Brandariz, S. P. et R. Bernardo (2019). Small Ad Hoc versus Large General Training Populations for Genomewide Selection in Maize Biparental Crosses. *Theor Appl Genet* 132 : 347-353.

Brard, S. et A. Ricard (2015). Is the Use of Formulae a Reliable Way to Predict the Accuracy of Genomic Selection ? *J. Anim. Breed. Genet.* 132 : 207-217 (cf. p. 9).

Breider, I. S., R. C. Gaynor, G. Gorjanc, S. Thorn, M. K. Pandey, R. K. Varshney et J. M. Hickey (2022). *A Multi-Part Strategy for Introgression of Exotic Germplasm into Elite Plant Breeding Programs Using Genomic Selection.* preprint. bioRXiv (cf. p. 171).

Browning, B. L. et S. R. Browning (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics* 84 : 210-223 (cf. p. 49).

Browning, S. R. et B. L. Browning (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics* 81 : 1084-1097 (cf. p. 44, 55).

Burdick, J. T., W.-M. Chen, G. R. Abecasis et V. G. Cheung (2006). In Silico Method for Inferring Genotypes in Pedigrees. *Nat Genet* 38 : 1002-1004 (cf. p. 14).

Bus, V. G. M., H. C. M. Bassett, D. Bowatte, D. Chagné, C. A. Ranatunga, D. Ulluwishewa, C. Wiedow et S. E. Gardiner (2010). Genome Mapping of an Apple Scab, a Powdery Mildew and a Woolly Apple Aphid Resistance Gene from Open-Pollinated Mildew Immune Selection. *Tree Genetics & Genomes* 6 : 477-487 (cf. p. 32).

Bus, V. G. M., F. N. D. Laurens, W. E. Van De Weg, R. L. Rusholme, E. H. A. Rikkerink, S. E. Gardiner, H. C. M. Bassett, L. P. Kodde et K. M. Plummer (2005). The *Vh8* Locus of a New Gene-for-gene Interaction between *Venturia Inaequalis* and the Wild Apple *Malus Sieversii* Is Closely Linked to the *Vh2* Locus in *Malus Pumila* R12740-7A. *New Phytologist* 166 : 1035-1049 (cf. p. 58).

Bustos-Korts, D., M. Malosetti, S. Chapman, B. Biddulph et F. van Eeuwijk (2016). Improvement of Predictive Ability by Uniform Coverage of the Target Genetic Space. *G3* 6 : 3733-3747 (cf. p. 12).

Calenge, F., D. Drouet, C. Denancé, W. E. Van de Weg, M.-N. Brisset, J.-P. Paulin et C.-E. Durel (2005). Identification of a Major QTL Together with Several Minor Additive or Epistatic QTLs for Resistance to Fire Blight in Apple in Two Related Progenies. *Theor Appl Genet* 111 : 128-135 (cf. p. 23).

Calenge, F. et C.-E. Durel (2006). Both Stable and Unstable QTLs for Resistance to Powdery Mildew Are Detected in Apple after Four Years of Field Assessments. *Mol Breeding* 17 : 329-339 (cf. p. 23).

Calenge, F., A. Faure, M. Goerre, C. Gebhardt, W. E. Van de Weg, L. Parisi et C.-E. Durel (2004). Quantitative Trait Loci (QTL) Analysis Reveals Both Broad-Spectrum and Isolate-Specific QTL for Scab Resistance in an Apple Progeny Challenged with Eight Isolates of *Venturia Inaequalis*. *Phytopathology* 94 : 370-379 (cf. p. 23).

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos et R. F. Veerkamp (2008). Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. *Genetics* 178 : 553-561 (cf. p. 13).

Calus, M., A. Bouwman, J. Hickey, R. Veerkamp et H. Mulder (2014). Evaluation of Measures of Correctness of Genotype Imputation in the Context of Genomic Prediction : A Review of Livestock Applications. *Animal* 8 : 1743-1753 (cf. p. 49).

Ceci, A. T., M. Bassi, W. Guerra, M. Oberhuber, P. Robatscher, F. Mattivi et P. Franceschi (2021). Metabolomic Characterization of Commercial, Old, and Red-Fleshed Apple Varieties. *Metabolites* 11 : 378.

Chagné, D., C. M. Carlisle, C. Blond, R. K. Volz, C. J. Whitworth, N. C. Oraguzie, R. N. Crowhurst, A. C. Allan, R. V. Espley, R. P. Hellens et S. E. Gardiner (2007). Mapping a Candidate Gene (MdMYB10) for Red Flesh and Foliage Colour in Apple. *BMC Genomics* 8 : 212 (cf. p. 23).

Chagné, D., R. N. Crowhurst, M. Troggio, M. W. Davey, B. Gilmore, C. Lawley, S. Vanderzande, R. P. Hellens, S. Kumar, A. Cestaro, R. Velasco, D. Main, J. D. Rees, A. Iezzoni, T. Mockler, L. Wilhelm, E. Van de Weg, S. E. Gardiner, N. Bassil et C. Peace (2012). Genome-Wide SNP Detection, Validation, and Development of an 8K SNP Array for Apple. *PLoS ONE* 7 : e31745 (cf. p. 24).

Chagné, D., C. Kirk, N. How, C. Whitworth, C. Fontic, G. Reig, G. Sawyer, S. Rouse, L. Poles, S. E. Gardiner, S. Kumar, R. Espley, R. K. Volz, M. Troggio et I. Iglesias (2016). A Functional Genetic Marker for Apple Red Skin Coloration across Different Environments. *Tree Genetics & Genomes* 12 : 67.

Chen, P., Z. Li, D. Zhang, W. Shen, Y. Xie, J. Zhang, L. Jiang, X. Li, X. Shen, D. Geng, L. Wang, C. Niu, C. Bao, M. Yan, H. Li, C. Li, Y. Yan, Y. Zou, D. Micheletti, E. Koot, F. Ma et Q. Guan (2021). Insights into the Effect of Human Civilization on *Malus* Evolution and Domestication. *Plant Biotechnol J* 19 : 2206-2220 (cf. p. 28).

Civan, P., R. Rincent, A. Danguy-Des-Deserts, J.-M. Elsen et S. Bouchet (2021). « Population Genomics Along With Quantitative Genetics Provides a More Efficient Valorization of Crop Plant Genetic Diversity in Breeding and Pre-breeding Programs » (cf. p. 17).

Clark, S. A., J. M. Hickey, H. D. Daetwyler et J. H. van der Werf (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 (cf. p. 12).

Conner, P. J., S. K. Brown et N. F. Weeden (1997). Randomly Amplified Polymorphic DNA-based Genetic Linkage Maps of Three Apple Cultivars. *JASHS* 122 : 350-359 (cf. p. 23).

Cooper, M., F. Technow, C. Messina, C. Gho et L. R. Totir (2016). Use of Crop Growth Models with Whole-Genome Prediction : Application to a Maize Multienvironment Trial. *Crop Sci.* 56 : 2141-2156 (cf. p. 178).

Cornille, A., F. Antolín, E. Garcia, C. Vernesi, A. Fietta, O. Brinkkemper, W. Kirleis, A. Schlumbaum et I. Roldán-Ruiz (2019). A Multifaceted Overview of Apple Tree Domestication. *Trends in Plant Science* 24 : 770-782 (cf. p. 20).

Cornille, A., T. Giraud, M. J. Smulders, I. Roldán-Ruiz et P. Gladieux (2014). The Domestication and Evolutionary Ecology of Apples. *Trends in Genetics* 30 : 57-65 (cf. p. 19, 20, 28).

Cornille, A., P. Gladieux, M. J. M. Smulders, I. Roldán-Ruiz, F. Laurens, B. Le Cam, A. Nersesyan, J. Clavel, M. Olonova, L. Feugey, I. Gabrielyan, X.-G. Zhang, M. I. Tenaillon

et T. Giraud (2012). New Insight into the History of Domesticated Apple : Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties. *PLoS Genet* 8 : e1002703 (cf. p. 20, 28).

Costa, F. (2015). MetaQTL Analysis Provides a Compendium of Genomic Loci Controlling Fruit Quality Traits in Apple. *Tree Genetics & Genomes* 11 : 819 (cf. p. 23).

Couasnet, G., M. Z. El Abidine, F. Laurens, H. Dutagaci et D. Rousseau (2021). « Machine Learning Meets Distinctness in Variety Testing ». *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW),1303-1311 (cf. p. 174).

Coupel-Ledru, A., B. Pallas, M. Delalande, F. Boudon, E. Carrié, S. Martinez, J.-L. Regnard et E. Costes (2019). Multi-Scale High-Throughput Phenotyping of Apple Architectural and Functional Traits in Orchard Reveals Genotypic Variability under Contrasted Watering Regimes. *Hortic Res* 6 : 52 (cf. p. 174).

Coupel-Ledru, A., B. Pallas, M. Delalande, V. Segura, B. Guitton, H. Muranty, C.-E. Durel, J.-L. Regnard et E. Costes (2022). Tree Architecture, Light Interception and Water Use Related Traits Are Controlled by Different Genomic Regions in an Apple Tree Core Collection. *New Phytologist* (cf. p. 23, 31, 178).

Crosby, J., J. Janick, P. Pecknold, S. Korban, P. O'Connor, S. Ries, J. Goffreda et A. Voordeckers (1992). Breeding Apples for Scab Resistance 1945 – 1990. *Acta Hortic.* : 43-70 (cf. p. 50).

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, J. L. Araus, D. Makumbi, R. P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger et H.-J. Braun (2010). Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. *Genetics* 186 : 713-724 (cf. p. 6).

Crossa, J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, C. Saint-Pierre, P. Vikram, C. Sansaloni, C. Petroli, D. Akdemir, C. Sneller, M. Reynolds, M. Tattaris, T. Payne, C. Guzman, R. J. Peña, P. Wenzl et S. Singh (2016). Genomic Prediction of Gene Bank Wheat Landraces. *G3* 6 : 1819-1834 (cf. p. 17).

Daccord, N., J.-M. Celton, G. Linsmith, C. Becker, N. Choisne, E. Schijlen, H. van de Geest, L. Bianco, D. Micheletti, R. Velasco, E. A. Di Pierro, J. Gouzy, D. J. G. Rees, P. Guérif, H. Muranty, C.-E. Durel, F. Laurens, Y. Lespinasse, S. Gaillard, S. Aubourg, H. Quesneville, D. Weigel, E. van de Weg, M. Troggio et E. Bucher (2017). High-Quality de Novo Assembly of the Apple Genome and Methylome Dynamics of Early Fruit Development. *Nat Genet* 49 : 1099-1106 (cf. p. 21, 44, 61).

Daetwyler, H., B. Villanueva, P. Bijma et J. Woolliams (2007). Inbreeding in Genome-Wide Selection. *J. Anim. Breed. Genet.* 124 : 369-376 (cf. p. 16).

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos et J. M. Hickey (2013). Genomic Prediction in Animals and Plants : Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics* 193 : 347-365 (cf. p. 150).

Daetwyler, H. D., B. Villanueva et J. A. Woolliams (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PLoS ONE* 3 : e3395 (cf. p. 9-11).

Dassonneville, R., R. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M. Lund, V. Ducrocq et G. Su (2011). Effect of Imputing Markers from a Low-Density Chip on the Reliability of Genomic Breeding Values in Holstein Populations. *Journal of Dairy Science* 94 : 3679-3686 (cf. p. 15, 148).

De Beukelaer, H., Y. Badke, V. Fack et G. De Meyer (2017). Moving Beyond Managing Realized Genomic Relationship in Long-Term Genomic Selection. *Genetics* 206 : 1127-1138 (cf. p. 150).

De los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel et J. M. Cotes (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics* 182 : 375-385 (cf. p. 7).

De Koning, D.-J. (2016). Meuwissen *et al.* on Genomic Selection. *Genetics* 203 : 5-7 (cf. p. 6).

De Roos, A. P. W., B. J. Hayes, R. J. Spelman et M. E. Goddard (2008). Linkage Disequilibrium and Persistence of Phase in Holstein–Friesian, Jersey and Angus Cattle. *Genetics* 179 : 1503-1512.

De Roos, A. P. W., B. J. Hayes et M. E. Goddard (2009). Reliability of Genomic Predictions Across Multiple Populations. *Genetics* 183 : 1545-1553 (cf. p. 11, 12).

Desta, Z. A. et R. Ortiz (2014). Genomic Selection : Genome-Wide Prediction in Plant Improvement. *Trends in Plant Science* 19 : 592-601 (cf. p. 7).

Doublet, A.-C., P. Croiseau, S. Fritz, A. Michenet, C. Hozé, C. Danchin-Burge, D. Laloë et G. Restoux (2019). The Impact of Genomic Selection on Genetic Diversity and Genetic Gain in Three French Dairy Cattle Breeds. *Genet Sel Evol* 51 : 52 (cf. p. 16, 128).

DoVale, J. C., H. F. Carvalho, F. Sabadin et R. Fritsche-Neto (2021). *Reduction of Genotyping Marker Density for Genomic Selection Is Not an Affordable Approach to Long-Term Breeding in Cross-Pollinated Crops*. preprint. bioRXiv.

Druet, T., I. M. Macleod et B. J. Hayes (2014). Toward Genomic Prediction from Whole-Genome Sequence Data : Impact of Sequencing Design on Genotype Imputation and Accuracy of Predictions. *Heredity* 112 : 39-47 (cf. p. 13).

Druet, T. et M. Georges (2010). A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. *Genetics* 184 : 789-798 (cf. p. 15).

Duan, N., Y. Bai, H. Sun, N. Wang, Y. Ma, M. Li, X. Wang, C. Jiao, N. Legall, L. Mao, S. Wan, K. Wang, T. He, S. Feng, Z. Zhang, Z. Mao, X. Shen, X. Chen, Y. Jiang, S. Wu, C. Yin, S.

Ge, L. Yang, S. Jiang, H. Xu, J. Liu, D. Wang, C. Qu, Y. Wang, W. Zuo, L. Xiang, C. Liu, D. Zhang, Y. Gao, Y. Xu, K. Xu, T. Chao, G. Fazio, H. Shu, G.-Y. Zhong, L. Cheng, Z. Fei et X. Chen (2017). Genome Re-Sequencing Reveals the History of Apple and Supports a Two-Stage Model for Fruit Enlargement. *Nat Commun* 8 : 249 (cf. p. 28).

Durel, C.-E., C. Denancé et M.-N. Brisset (2009). Two Distinct Major QTL for Resistance to Fire Blight Co-Localize on Linkage Group 12 in Apple Genotypes 'Evereste' and *Malus Floribunda* Clone 821. *Genome* 52 : 139-147 (cf. p. 32).

Durel, C.-E. (2016). Caractériser La Diversité Des Variétés Anciennes de Pommier Conservées En France. *Jardins de France* : 13-15 (cf. p. 30).

Edwards, S. M., J. B. Buntjer, R. Jackson, A. R. Bentley, J. Lage, E. Byrne, C. Burt, P. Jack, S. Berry, E. Flatman, B. Poupard, S. Smith, C. Hayes, R. C. Gaynor, G. Gorjanc, P. Howell, E. Ober, I. J. Mackay et J. M. Hickey (2019). The Effects of Training Population Design on Genomic Prediction Accuracy in Wheat. *Theor Appl Genet* 132 : 1943-1952 (cf. p. 10, 145).

Emeriewen, O. F., T. Wöhner, H. Flachowsky et A. Peil (2019). Malus Hosts–Erwinia Amylovora Interactions : Strain Pathogenicity and Resistance Mechanisms. *Front. Plant Sci.* 10 : 551 (cf. p. 177).

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* 4 : 250-255.

Evans, K. (2013). Apple Breeding in the Pacific Northwest. *Acta Hortic.* : 75-78 (cf. p. 24).

Evans, K. M., A. Patocchi, F. Rezzonico, F. Mathis, C. E. Durel, F. Fernández-Fernández, A. Boudichevskaia, F. Dunemann, M. Stankiewicz-Kosyl, L. Gianfranceschi, M. Komjanc, M. Lateur, M. Madduri, Y. Noordijk et W. E. van de Weg (2011). Genotyping of Pedigreed Apple Breeding Material with a Genome-Covering Set of SSRs : Trueness-to-Type of Cultivars and Their Parentages. *Mol Breeding* 28 : 535-547 (cf. p. 61).

Evenson, R. E. et D. Gollin (2003). Assessing the Impact of the Green Revolution, 1960 to 2000. *Science* 300 : 758-762 (cf. p. 1).

Eynard, S. E., P. Croiseau, D. Laloë, S. Fritz, M. P. L. Calus et G. Restoux (2018). Which Individuals To Choose To Update the Reference Population ? Minimizing the Loss of Genetic Diversity in Animal Genomic Selection Programs. *G3* 8 : 113-121 (cf. p. 11).

Eynard, S. E., J. J. Windig, S. J. Hiemstra et M. P. L. Calus (2016). Whole-Genome Sequence Data Uncover Loss of Genetic Diversity Due to Selection. *Genet Sel Evol* 48 : 33 (cf. p. 175).

FAO (2021). *The State of Food Security and Nutrition in the World*. FAO (cf. p. 1).

Fazio, G., P. Forsline, H. Aldwinckle et L. Pons (2008). The Apple Collection in Geneva, NY : A Resource for the Apple Industry Today and for Generations to Come. *New York Fruit Quarterly* 16 : 5-8 (cf. p. 30).

Fernández, J., B. Villanueva et M. A. Toro (2021). Optimum Mating Designs for Exploiting Dominance in Genomic Selection Schemes for Aquaculture Species. *Genet Sel Evol* 53 : 14 (cf. p. 17).

Feuillet, C., P. Langridge et R. Waugh (2008). Cereal Breeding Takes a Walk on the Wild Side. *Trends in Genetics* 24 : 24-32 (cf. p. 4).

Feurtey, A., E. Guitton, M. De Gracia Coquerel, L. Duvaux, J. Shiller, M.-N. Bellanger, P. Expert, M. Sannier, V. Caffier, T. Giraud, B. Le Cam et C. Lemaire (2020). Threat to Asian Wild Apple Trees Posed by Gene Flow from Domesticated Apple Trees and Their "Pestified" Pathogens. *Mol Ecol* 29 : 4925-4941 (cf. p. 30).

Fischer, C. (1994). « Shortening of the Juvenile Period in Apple Breeding ». *Progress in Temperate Fruit Breeding*. T. 1,161-164 (cf. p. 25).

Flachowsky, H., M.-V. Hanke, A. Peil, S. H. Strauss et M. Fladung (2009). A Review on Transgenic Approaches to Accelerate Breeding of Woody Plants. *Plant Breeding* 128 : 217-226 (cf. p. 25).

Flachowsky, H., P.-M. Le Roux, A. Peil, A. Patocchi, K. Richter et M.-V. Hanke (2011). Application of a High-Speed Breeding Technology to Apple (Malus × Domestica) Based on Transgenic Early Flowering Plants and Marker-Assisted Selection. *New Phytologist* 192 : 364-377 (cf. p. 32).

Frey, J. E., B. Frey, C. Sauer et M. Kellerhals (2004). Efficient Low-Cost DNA Extraction and Multiplex Fluorescent PCR Method for Marker-Assisted Selection in Breeding. *Plant Breeding* 123 : 554-557 (cf. p. 26).

Frischknecht, M., T. H. Meuwissen, B. Bapst, F. R. Seefried, C. Flury, D. Garrick, H. Signer-Hasler, C. Stricker, A. Bieber, R. Fries, I. Russ, J. Sölkner, A. Bagnato et B. Gredler-Grandl (2018). Short Communication : Genomic Prediction Using Imputed Whole-Genome Sequence Variants in Brown Swiss Cattle. *Journal of Dairy Science* 101 : 1292-1296 (cf. p. 13).

Fritsche-Neto, R., G. Galli, K. L. R. Borges, G. Costa-Neto, F. C. Alves, F. Sabadin, D. H. Lyra, P. P. P. Morais, L. R. Braatz de Andrade, I. Granato et J. Crossa (2021). Optimizing Genomic-Enabled Prediction in Small-Scale Maize Hybrid Breeding Programs : A Roadmap Review. *Front. Plant Sci.* 12 : 658267.

García-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-López et C. P. Van Tassell (2016). Changes in Genetic Selection Differentials and Generation Intervals in US Holstein Dairy Cattle as a Result of Genomic Selection. *Proc. Natl. Acad. Sci. U.S.A.* 113 : 3995-4004 (cf. p. 6).

Gardner, K. M., P. Brown, T. F. Cooke, S. Cann, F. Costa, C. Bustamante, R. Velasco, M. Troggio et S. Myles (2014). Fast and Cost-Effective Genetic Mapping in Apple Using Next-Generation Sequencing. *G3* 4 : 1681-1687 (cf. p. 24).

Gaynor, R. C., G. Gorjanc et J. M. Hickey (2021). AlphaSimR : An R Package for Breeding Program Simulations. *G3* 11 (cf. p. 129).

Genetic Resources for Food and Agriculture, C. on (2010). *The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture* (cf. p. 4).

Gessler, C., A. Patocchi, S. Sansavini, S. Tartarini et L. Gianfranceschi (2006). *Venturia Inaequalis* Resistance in Apple. *Critical Reviews in Plant Sciences* 25 : 473-503 (cf. p. 32, 58, 150, 177).

Gessler, C. et I. Pertot (2012). Vf Scab Resistance of Malus. *Trees* 26 : 95-108.

Gianfranceschi, L. et V. Soglio (2004). The European Project HiDRAS : Innovative Multidisciplinary Approaches to Breeding High Quality Disease Resistant Apples. *Acta Hortic.* : 327-330 (cf. p. 35).

Gianola, D. (2013). Priors in Whole-Genome Regression : The Bayesian Alphabet Returns. *Genetics* 194 : 573-596 (cf. p. 7, 9, 128).

GIEC (2022). *Climate Change 2022 : Impacts, Adaptation and Vulnerability*. Sixth Assessment Report of the Intergovernmental Panel on Climate Change (cf. p. 1-3).

Goddard, M. E. et B. J. Hayes (2009). Mapping Genes for Complex Traits in Domestic Animals and Their Use in Breeding Programmes. *Nat Rev Genet* 10 : 381-391 (cf. p. 11).

Goddard, M. (2009). Genomic Selection : Prediction of Accuracy and Maximisation of Long Term Response. *Genetica* 136 : 245-257 (cf. p. 17, 128).

Godfray, H. C. J., J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas et C. Toulmin (2010). Food Security : The Challenge of Feeding 9 Billion People. *Science* 327 : 812-818 (cf. p. 2).

Gonen, S., V. Wimmer, R. C. Gaynor, E. Byrne, G. Gorjanc et J. M. Hickey (2018). A Heuristic Method for Fast and Accurate Phasing and Imputation of Single-Nucleotide Polymorphism Data in Bi-Parental Plant Populations. *Theor Appl Genet* 131 : 2345-2357 (cf. p. 60).

Gonzalez, M. Y., Y. Zhao, Y. Jiang, N. Stein, A. Habekuss, J. C. Reif et A. W. Schulthess (2021). Genomic Prediction Models Trained with Historical Records Enable Populating the German Ex Situ Genebank Bio-Digital Resource Center of Barley (Hordeum Sp.) with Information on Resistances to Soilborne Barley Mosaic Viruses. *Theor Appl Genet* 134 : 2181-2196 (cf. p. 10).

Gorjanc, G., M. Battagin, J.-F. Dumasy, R. Antolin, R. C. Gaynor et J. M. Hickey (2017). Prospects for Cost-Effective Genomic Selection via Accurate Within-Family Imputation. *Crop Sci.* 57 : 216-228 (cf. p. 43).

Gorjanc, G., R. C. Gaynor et J. M. Hickey (2018). Optimal Cross Selection for Long-Term Genetic Gain in Two-Part Programs with Rapid Recurrent Genomic Selection. *Theor Appl Genet* 131 : 1953-1966 (cf. p. 130).

Gorjanc, G., J. Jenko, S. J. Hearne et J. M. Hickey (2016). Initiating Maize Pre-Breeding Programs Using Genomic Selection to Harness Polygenic Variation from Landrace Populations. *BMC Genomics* 17 : 30 (cf. p. 17, 128).

Goudet, J. (2005). Hierfstat, a Package for r to Compute and Test Hierarchical F-statistics. *Mol Ecol Notes* 5 : 184-186.

Grattapaglia, D. et M. D. V. Resende (2011). Genomic Selection in Forest Tree Breeding. *Tree Genetics & Genomes* 7 : 241-255 (cf. p. 9).

Grattapaglia, D., O. B. Silva-Junior, R. T. Resende, E. P. Cappa, B. S. F. Müller, B. Tan, F. Isik, B. Ratcliffe et Y. A. El-Kassaby (2018). Quantitative Genetics and Genomics Converge to Accelerate Forest Tree Breeding. *Front. Plant Sci.* 9 : 1693 (cf. p. 170).

Gross, B. L., A. D. Henk, C. M. Richards, G. Fazio et G. M. Volk (2014). Genetic Diversity in *Malus* × *Domestica* (Rosaceae) through Time in Response to Domestication. *Am. J. Bot.* 101 : 1770-1779 (cf. p. 28).

Gross, B. L. et K. M. Olsen (2010). Genetic Perspectives on Crop Domestication. *Trends in Plant Science* 15 : 529-537 (cf. p. 2).

Guitton, B., J.-J. Kelner, R. Velasco, S. E. Gardiner, D. Chagné et E. Costes (2012). Genetic Control of Biennial Bearing in Apple. *Journal of Experimental Botany* 63 : 131-149 (cf. p. 23).

Habier, D., R. L. Fernando et J. C. M. Dekkers (2009). Genomic Selection Using Low-Density Marker Panels. *Genetics* 182 : 343-353 (cf. p. 11, 14, 43).

Habier, D., R. L. Fernando et D. J. Garrick (2013). Genomic BLUP Decoded : A Look into the Black Box of Genomic Prediction. *Genetics* 194 : 597-607 (cf. p. 13).

Hahsler, M. et K. Hornik (2007). **TSP** - Infrastructure for the Traveling Salesperson Problem. *J. Stat. Soft.* 23 (cf. p. 133).

Han, Y., D. Zheng, S. Vimolmangkang, M. A. Khan, J. E. Beever et S. S. Korban (2011). Integration of Physical and Genetic Maps in Apple Confirms Whole-Genome and Segmental Duplications in the Apple Genome. *Journal of Experimental Botany* 62 : 5117-5130 (cf. p. 21).

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla et M. E. Goddard (2009). Accuracy of Genomic Breeding Values in Multi-Breed Dairy Cattle Populations. *Genet Sel Evol* 41 : 51 (cf. p. 11, 13, 123).

Heffner, E. L., M. E. Sorrells et J.-L. Jannink (2009). Genomic Selection for Crop Improvement. *Crop Sci.* 49 : 1-12 (cf. p. 7).

Herry, F., F. Hérault, D. Picard Druet, A. Varenne, T. Burlot, P. Le Roy et S. Allais (2018). Design of Low Density SNP Chips for Genotype Imputation in Layer Chicken. *BMC Genetics* 19 : 108 (cf. p. 43).

Heslot, N., D. Akdemir, M. E. Sorrells et J.-L. Jannink (2014). Integrating Environmental Covariates and Crop Modeling into the Genomic Selection Framework to Predict Genotype by Environment Interactions. *Theor Appl Genet* 127 : 463-480 (cf. p. 178).

Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. van der Werf et M. A. Cleveland (2012a). A Phasing and Imputation Method for Pedigreed Populations That Results in a Single-Stage Genomic Evaluation. *Genet Sel Evol* 44 : 9 (cf. p. 59).

Hickey, J. M., J. Crossa, R. Babu et G. de los Campos (2012b). Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. *Crop Sci.* 52 : 654-663 (cf. p. 15, 49).

Hickey, J. M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, B. M. Prasanna, M. Grondona, A. Zambelli, V. S. Windhausen, K. Mathews et G. Gorjanc (2014). Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. *Crop Sci.* 54 : 1476-1488.

Hidalgo, J., D. Lourenco, S. Tsuruta, Y. Masuda, V. Breen, R. Hawken, M. Bermann et I. Misztal (2021). Investigating the Persistence of Accuracy of Genomic Predictions over Time in Broilers. *J. Anim. Breed. Genet.* 99 (cf. p. 148).

Howard, N. P., D. C. Albach et J. J. Lubby (2018a). « The Identification of Apple Pedigree Information on a Large Diverse Set of Apple Germplasm and Its Application in Apple Breeding Using New Genetic Tools ». 18th International Conference on Organic Fruit-Growing : Proceedings of the Conference (cf. p. 39).

Howard, N. P., D. C. Albach, J. J. Luby, W. E. Van De Weg, M. Troggio, C.-E. Durel, C. Peace, S. Vanderzande et G. Volk (2018b). « Collaborative Project to Identify Direct and Distant Pedigree Relationships in Apple ». Poster (cf. p. 29).

Howard, N. P., C. Peace, K. A. T. Silverstein, A. Poets, J. J. Luby, S. Vanderzande, C.-E. Durel, H. Muranty, C. Denancé et E. van de Weg (2021). The Use of Shared Haplotype Length Information for Pedigree Reconstruction in Asexually Propagated Outbreeding Crops, Demonstrated for Apple and Sweet Cherry. *Hortic Res* 8 : 202 (cf. p. 30, 37, 44, 51, 61).

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini et G. R. Abecasis (2012). Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-Phasing. *Nat Genet* 44 : 955-959 (cf. p. 60).

Howie, B., J. Marchini et M. Stephens (2011). Genotype Imputation with Thousands of Genomes. *G3* 1 : 457-470 (cf. p. 60).

Hozé, C., M.-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard et P. Croiseau (2013). High-Density Marker Imputation Accuracy in Sixteen French Cattle Breeds. *Genet Sel Evol* 45 : 33 (cf. p. 15).

Ibáñez-Escriche, N., R. L. Fernando, A. Toosi et J. C. Dekkers (2009). Genomic Selection of Purebreds for Crossbred Performance. *Genet Sel Evol* 41 : 12.

Iezzoni, A. F., J. McFerson, J. Luby, K. Gasic, V. Whitaker, N. Bassil, C. Yue, K. Gallardo, V. McCracken, M. Coe, C. Hardner, J. D. Zurn, S. Hokanson, E. van de Weg, S. Jung, D. Main, C. da Silva Linge, S. Vanderzande, T. M. Davis, L. L. Mahoney, C. Finn et C. Peace (2020). RosBREED : Bridging the Chasm between Discovery and Application to Enable DNA-informed Breeding in Rosaceous Crops. *Hortic Res* 7 : 177 (cf. p. 26).

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot et M. E. Sorrells (2015). Training Set Optimization under Population Structure in Genomic Selection. *Theor Appl Genet* 128 : 145-158.

Isidro y Sánchez, J. et D. Akdemir (2021). Training Set Optimization for Sparse Phenotyping in Genomic Selection : A Conceptual Overview. *Front. Plant Sci.* 12 : 715910 (cf. p. 12).

Isik, F., J. Holland et C. Maltecca (2017). *Genetic Data Analysis for Plant and Animal Breeding*. 1st ed. 2017 (cf. p. 9, 10).

Janick, J. et J. N. Moore (1975). *Advances in Fruit Breeding*. 623 p.

Jannink, J.-L. (2010). Dynamics of Long-Term Genomic Selection. *Genet Sel Evol* 42 : 35 (cf. p. 6, 16, 17, 128, 147).

Jannink, J.-L., H. Iwata, P. R. Bhat, S. Chao, P. Wenzl et G. J. Muehlbauer (2009). Marker Imputation in Barley Association Studies. *The Plant Genome* 2 : 11-22 (cf. p. 14).

Jarquin, D., J. Specht et A. Lorenz (2016). Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection : Historical Data Creates Robust Models for Enhancing Selection of Accessions. *G3* 6 : 2329-2341 (cf. p. 17).

Jenks, M. A. et P. J. Bebeli (2011). *Breeding for Fruit Quality*. 384 p. (cf. p. 23).

Jeong, S., J.-Y. Kim, S.-C. Jeong, S.-T. Kang, J.-K. Moon et N. Kim (2017). GenoCore : A Simple and Fast Algorithm for Core Subset Selection from Large Genotype Datasets. *PLoS ONE* 12 : e0181420 (cf. p. 136).

Jia, D., F. Shen, Y. Wang, T. Wu, X. Xu, X. Zhang et Z. Han (2018). Apple Fruit Acidity Is Genetically Diversified by Natural Variations in Three Hierarchical Epistatic Genes : *MdSAUR37* , *MdPP2CH* and *MdALMTII*. *Plant J* 95 : 427-443 (cf. p. 150).

Jung, M., B. Keller, M. Roth, M. J. Aranzana, A. Auwerkerken, W. Guerra, M. Al-Rifaï, M. Lewandowski, N. Sanin, M. Rymenants, F. Didelot, C. Dujak, C. F. Forcada, A. Knauf, F. Laurens, B. Studer, H. Muranty et A. Patocch (2022). Genetic Architecture and Genomic Predictive Ability of Apple Quantitative Traits across Environments. *Hortic Res* : uhac028 (cf. p. 38, 42, 123, 173).

Jung, M., M. Roth, M. J. Aranzana, A. Auwerkerken, M. Bink, C. Denancé, C. Dujak, C.-E. Durel, C. Font i Forcada, C. M. Cantin, W. Guerra, N. P. Howard, B. Keller, M. Lewandowski, M. Ordidge, M. Rymenants, N. Sanin, B. Studer, E. Zurawicz, F. Laurens, A. Patocchi et H. Muranty (2020). The Apple REFPOP—a Reference Population for Genomics-Assisted Breeding in Apple. *Hortic Res* 7 : 189 (cf. p. 13, 28, 36, 37, 45).

Kahiluoto, H., J. Kaseva, J. Balek, J. E. Olesen, M. Ruiz-Ramos, A. Gobin, K. C. Kersebaum, J. Takáč, F. Ruget, R. Ferrise, P. Bezak, G. Capellades, C. Dibari, H. Mäkinen, C. Nendel, D. Ventrella, A. Rodríguez, M. Bindi et M. Trnka (2019). Decline in Climate Resilience of European Wheat. *Proc. Natl. Acad. Sci. U.S.A.* 116 : 123-128 (cf. p. 2).

Karoui, S., M. J. Carabaño, C. Díaz et A. Legarra (2012). Joint Genomic Evaluation of French Dairy Cattle Breeds Using Multiple-Trait Models. *Genet Sel Evol* 44 : 39 (cf. p. 11).

Kassambara, A. (2021). *Rstatix : Pipe-Friendly Framework for Basic Statistical Tests*. Version 0.7.0 (cf. p. 138).

Kellerhalls, M., S. S., Baumgartner I.O., Andreoli R., Gassmann J., Bolliger N., Schärer H.-J., Ludwig M. et Steinemann B. (2018). « Broaden the Genetic Basis in Apple Breeding by Using Genetic Resources ». Proceedings of the 18th International Conference on Organic Fruit-Growing. 19-21. Februar, Ed. Foerdergemeinschaft Ökologischer Obstbau e.V. (Cf. p. 32, 127).

Kenis, K. et J. Keulemans (2005). Genetic Linkage Maps of Two Apple Cultivars (Malus x Domestica Borkh.) Based on AFLP and Microsatellite Markers. *Mol Breeding* 15 : 205-219 (cf. p. 23).

Khan, M. A., K. M. Olsen, V. Sovero, M. M. Kushad et S. S. Korban (2014a). Fruit Quality Traits Have Played Critical Roles in Domestication of the Apple. *The Plant Genome* 7 (cf. p. 19).

Khan, M. A., Y. Han, Y. F. Zhao, M. Troggio et S. S. Korban (2012). A Multi-Population Consensus Genetic Map Reveals Inconsistent Marker Order among Maps Likely Attributed to Structural Variations in the Apple Genome. *PLoS ONE* 7 : e47864 (cf. p. 23).

Khan, S. A., Y. Tikunov, P.-Y. Chibon, C. Maliepaard, J. Beekwilder, E. Jacobsen et H. J. Schouten (2014b). Metabolic Diversity in Apple Germplasm. *Plant Breeding* 133 : 281-290.

Khoury, C. K., S. Brush, D. E. Costich, H. A. Curry, S. Haan, J. M. M. Engels, L. Guarino, S. Hoban, K. L. Mercer, A. J. Miller, G. P. Nabhan, H. R. Perales, C. Richards, C. Riggins et I. Thormann (2022). Crop Genetic Erosion : Understanding and Responding to Loss of Crop Diversity. *New Phytologist* 233 : 84-118 (cf. p. 3).

King, G. J., S. Tartarini, L. Brown, F. Gennari et S. Sansavini (1999). Introgression of the Vf Source of Scab Resistance and Distribution of Linked Marker Alleles within the Malus Gene Pool : *Theor Appl Genet* 99 : 1039-1046 (cf. p. 59).

Kono, T. J. Y., C. Liu, E. E. Vonderharr, D. Koenig, J. C. Fay, K. P. Smith et P. L. Morrell (2019). The Fate of Deleterious Variants in a Barley Genomic Prediction Population. *Genetics* 213 : 1531-1544 (cf. p. 16).

Korban, S. S. (2021). *The Apple Genome*. Corrected publication 2021. 412 p. (cf. p. 19).

Kouassi, A. B., C.-E. Durel, F. Costa, S. Tartarini, E. van de Weg, K. Evans, F. Fernandez-Fernandez, C. Govan, A. Boudichevskaja, F. Dunemann, A. Antofie, M. Lateur, M. Stankiewicz-

Kosyl, A. Soska, K. Tomala, M. Lewandowski, K. Rutkovski, E. Zurawicz, W. Guerra et F. Laurens (2009). Estimation of Genetic Parameters and Prediction of Breeding Values for Apple Fruit-Quality Traits Using Pedigreed Plant Material in Europe. *Tree Genetics & Genomes* 5 : 659-672 (cf. p. 38).

Krause, M. R., L. González-Pérez, J. Crossa, P. Pérez-Rodríguez, O. Montesinos-López, R. P. Singh, S. Dreisigacker, J. Poland, J. Rutkoski, M. Sorrells, M. A. Gore et S. Mondal (2019). Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat. *G3* 9 : 1231-1247 (cf. p. 179).

Kumar, S., M. C. A. M. Bink, R. K. Volz, V. G. M. Bus et D. Chagné (2012a). Towards Genomic Selection in Apple (Malus × Domestica Borkh.) Breeding Programmes : Prospects, Challenges and Strategies. *Tree Genetics & Genomes* 8 : 1-14 (cf. p. 6, 24, 26, 27, 170).

Kumar, S., D. Chagné, M. C. A. M. Bink, R. K. Volz, C. Whitworth et C. Carlisle (2012b). Genomic Selection for Fruit Quality Traits in Apple (Malus×domestica Borkh.) *PLoS ONE* 7 : e36674 (cf. p. 27, 28, 128, 145, 170).

Kumar, S., E. Hilario, C. H. Deng et C. Molloy (2020). Turbocharging Introgression Breeding of Perennial Fruit Crops : A Case Study on Apple. *Hortic Res* 7 : 47 (cf. p. 27, 28).

Kumar, S., C. Molloy, P. Muñoz, H. Daetwyler, D. Chagné et R. Volz (2015). Genome-Enabled Estimates of Additive and Nonadditive Genetic Variances and Prediction of Apple Phenotypes Across Environments. *G3* 5 : 2711-2718 (cf. p. 27, 28, 150, 170).

Kumar, S., P. Raulier, D. Chagné et C. Whitworth (2014). Molecular-Level and Trait-Level Differentiation between the Cultivated Apple (*Malus* × *Domestica* Borkh.) and Its Main Progenitor *Malus Sieversii*. *Plant Genet. Res.* 12 : 330-340 (cf. p. 24).

Kumar, S., R. K. Volz, P. A. Alspach et V. G. M. Bus (2010). Development of a Recurrent Apple Breeding Programme in New Zealand : A Synthesis of Results, and a Proposed Revised Breeding Strategy. *Euphytica* 173 : 207-222 (cf. p. 32, 33, 127).

Labroo, M. R., A. J. Studer et J. E. Rutkoski (2021). Heterosis and Hybrid Crop Breeding : A Multidisciplinary Review. *Front. Genet.* 12 : 643761 (cf. p. 17).

Lado, B., S. Battenfield, C. Guzmán, M. Quincke, R. P. Singh, S. Dreisigacker, R. J. Peña, A. Fritz, P. Silva, J. Poland et L. Gutiérrez (2017). Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs. *The Plant Genome* 10 (cf. p. 17, 150).

Larsen, B., Z. Migicovsky, A. A. Jeppesen, K. M. Gardner, T. B. Toldam-Andersen, S. Myles, M. Ørgaard, M. A. Petersen et C. Pedersen (2019). Genome-Wide Association Studies in Apple Reveal Loci for Aroma Volatiles, Sugar Composition, and Harvest Date. *The Plant Genome* 12 : 180104 (cf. p. 177).

Lasserre-Zuber, P., V. Caffier, R. Stievenard, A. Lemarquand, B. Le Cam et C.-E. Durel (2018). Pyramiding Quantitative Resistance with a Major Resistance Gene in Apple : From Ephe-

meral to Enduring Effectiveness in Controlling Scab. *Plant Disease* 102 : 2220-2223 (cf. p. 26).

Lassois, L., C. Denancé, E. Ravon, A. Guyader, R. Guisnel, L. Hibrand-Saint-Oyant, C. Poncet, P. Lasserre-Zuber, L. Feugey et C.-E. Durel (2016). Genetic Diversity, Population Structure, Parentage Analysis, and Construction of Core Collections in the French Apple Germplasm Based on SSR Markers. *Plant Mol Biol Rep* 34 : 827-844 (cf. p. 21, 30, 31).

Laurens, F. (1998). Review of the Current Apple Breedin Programmes in the World : Objectives for Scion Cultivar Improvement. *Acta Hortic.* : 163-170 (cf. p. 22).

Laurens, F., M. J. Aranzana, P. Arus, D. Bassi, M. Bink, J. Bonany, A. Caprera, L. Corelli-Grappadelli, E. Costes, C.-E. Durel, J.-B. Mauroux, H. Muranty, N. Nazzicari, T. Pascal, A. Patocchi, A. Peil, B. Quilot-Turion, L. Rossini, A. Stella, M. Troggio, R. Velasco et E. van de Weg (2018). An Integrated Approach for Increasing Breeding Efficiency in Apple and Peach in Europe. *Hortic Res* 5 : 11 (cf. p. 26, 35).

Le Mézec, P., C. Danchin-Burge et S. Moureaux (2018). *Les Programmes de Sélection et de Diffusion de Taureaux d'IA à l'ère de La Génomique et Leurs Effets Sur La Diversité Génétique*. IDELE (cf. p. 16).

Leforestier, D., E. Ravon, H. Muranty, A. Cornille, C. Lemaire, T. Giraud, C.-E. Durel et A. Branca (2015). Genomic Basis of the Differences between Cider and Dessert Apple Varieties. *Evol Appl* 8 : 650-661 (cf. p. 31).

Legave, J.-M. (2022). *Les Productions Fruitières à l'heure Du Changement Climatique*. Quae (cf. p. 23).

Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan, L. Campo, P. Flament, A. E. Melchinger, M. Menz, N. Meyer, L. Moreau, J. Moreno-González, M. Ouzunova, H. Pausch, N. Ranc, W. Schipprack, M. Schönleben, H. Walter, A. Charcosset et C.-C. Schön (2014). Usefulness of Multiparental Populations of Maize (*Zea Mays* L.) for Genome-Based Prediction. *Genetics* 198 : 3-16 (cf. p. 12).

Lehermeier, C., C.-C. Schön et G. de los Campos (2015). Assessment of Genetic Heterogeneity in Structured Plant Populations Using Multivariate Whole-Genome Regression Models. *Genetics* 201 : 323-337.

Lenth, V. R. (2021). *Emmeans : Estimated Marginal Means, Aka Least-Squares Means. R Package Version 1.6.0.*

Li, L., Y. Li, S. R. Browning, B. L. Browning, A. J. Slater, X. Kong, J. L. Aponte, V. E. Mooser, S. L. Chissoe, J. C. Whittaker, M. R. Nelson et M. G. Ehm (2011). Performance of Genotype Imputation for Rare Variants Identified in Exons and Flanking Regions of Genes. *PLoS ONE* 6 : e24945 (cf. p. 15).

Li, Y., S. Kaur, L. W. Pembleton, H. Valipour-Kahrood, G. M. Rosewarne et H. D. Daetwyler (2022). Strategies of Preserving Genetic Diversity While Maximizing Genetic Response from Implementing Genomic Selection in Pulse Breeding Programs. *Theor Appl Genet* (cf. p. 149).

Liao, L., W. Zhang, B. Zhang, T. Fang, X.-F. Wang, Y. Cai, C. Ogutu, L. Gao, G. Chen, X. Nie, J. Xu, Q. Zhang, Y. Ren, J. Yu, C. Wang, C. H. Deng, B. Ma, B. Zheng, C.-X. You, D.-G. Hu, R. Espley, K. Lin-Wang, J.-L. Yao, A. C. Allan, A. Khan, S. S. Korban, Z. Fei, R. Ming, Y.-J. Hao, L. Li et Y. Han (2021). Unraveling a Genetic Roadmap for Improved Taste in the Domesticated Apple. *Molecular Plant* (cf. p. 28).

Liebhard, R., L. Gianfranceschi, B. Koller, C. Ryder, R. Tarchini, E. Van De Weg et C. Gessler (2002). Development and Characterisation of 140 New Microsatellites in Apple (Malus x Domestica Borkh.) *Mol Breeding* 10 : 217-241 (cf. p. 23).

Liebhard, R., M. Kellerhals, W. Pfammatter, M. Jertmini et C. Gessler (2003). Mapping Quantitative Physiological Traits in Apple (Malus x Domestica Borkh.) *Plant Molecular Biology* 52 : 511-526 (cf. p. 23).

Lin, P., S. M. Hartz, Z. Zhang, S. F. Saccone, J. Wang, J. A. Tischfield, H. J. Edenberg, J. R. Kramer, A. M.Goate, L. J. Bierut, J. P. Rice et for the COGA Collaborators COGEND Collaborators, GENEVA (2010). A New Statistic to Evaluate Imputation Reliability. *PLoS ONE* 5 : e9697 (cf. p. 49).

Liu, H., T. Meuwissen, A. C. Sørensen et P. Berg (2015). Upweighting Rare Favourable Alleles Increases Long-Term Genetic Gain in Genomic Selection Programs. *Genet Sel Evol* 47 : 19 (cf. p. 17, 128, 137, 171).

Liu, X., H. Wang, H. Wang, Z. Guo, X. Xu, J. Liu, S. Wang, W.-X. Li, C. Zou, B. M. Prasanna, M. S. Olsen, C. Huang et Y. Xu (2018). Factors Affecting Genomic Selection Revealed by Empirical Evidence in Maize. *The Crop Journal* 6 : 341-352 (cf. p. 10).

Longhi, S., L. Cappellin, W. Guerra et F. Costa (2013). Validation of a Functional Molecular Marker Suitable for Marker-Assisted Breeding for Fruit Texture in Apple (Malus × Domestica Borkh.) *Mol Breeding* 32 : 841-852 (cf. p. 26).

Longin, C. F. H. et J. C. Reif (2014). Redesigning the Exploitation of Wheat Genetic Resources. *Trends in Plant Science* 19 : 631-636 (cf. p. 4).

Lorenz, A. J. et K. P. Smith (2015). Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. *Crop Sci.* 55 : 2657-2667 (cf. p. 11).

Lozada-Soto, E. A., C. Maltecca, D. Lu, S. Miller, J. B. Cole et F. Tiezzi (2021). Trends in Genetic Diversity and the Effect of Inbreeding in American Angus Cattle under Genomic Selection. *Genet Sel Evol* 53 : 50 (cf. p. 16).

Luby, J. J., N. P. Howard, J. R. Tillman et D. S. Bedford (2022). Extended Pedigrees of Apple Cultivars from the University of Minnesota Breeding Program Elucidated Using SNP Array Markers. *HortScience* 57 : 472-477 (cf. p. 30).

Lund, M., I. van den Berg, P. Ma, R. Brøndum et G. Su (2016). Review : How to Improve Genomic Predictions in Small Dairy Cattle Populations. *Animal* 10 : 1042-1049.

Luo, F., K. Evans, J. L. Norelli, Z. Zhang et C. Peace (2020a). Prospects for Achieving Durable Disease Resistance with Elite Fruit Quality in Apple Breeding. *Tree Genetics & Genomes* 16 : 21 (cf. p. 32).

Luo, F., J. L. Norelli, N. P. Howard, M. Wisniewski, H. Flachowsky, M.-V. Hanke et C. Peace (2020b). Introgressing Blue Mold Resistance into Elite Apple Germplasm by Rapid Cycle Breeding and Foreground and Background DNA-informed Selection. *Tree Genetics & Genomes* 16 : 28 (cf. p. 32).

Luo, F., P. Sandefur, K. Evans et C. Peace (2019). A DNA Test for Routinely Predicting Mildew Resistance in Descendants of Crabapple 'White Angel'. *Mol Breeding* 39 : 33 (cf. p. 127).

Lush, J. L. (1937). *Animal Breeding Plans*. Iowa State College Press, Iowa (cf. p. 3).

Lyra, D. H., Í. S. C. Granato, P. P. P. Morais, F. C. Alves, A. R. M. dos Santos, X. Yu, T. Guo, J. Yu et R. Fritsche-Neto (2018). Controlling Population Structure in the Genomic Prediction of Tropical Maize Hybrids. *Mol Breeding* 38 : 126.

Makanjuola, B. O., F. Miglior, E. A. Abdalla, C. Maltecca, F. S. Schenkel et C. F. Baes (2020). Effect of Genomic Selection on Rate of Inbreeding and Coancestry and Effective Population Size of Holstein and Jersey Cattle Populations. *Journal of Dairy Science* 103 : 5183-5199 (cf. p. 16).

Maliepaard, C., F. H. Alston, G. van Arkel, L. M. Brown, E. Chevreau, F. Dunemann, K. M. Evans, S. Gardiner, P. Guilford, A. W. van Heusden, J. Janse, F. Laurens, J. R. Lynn, A. G. Manganaris, A. P. M. den Nijs, N. Periam, E. Rikkerink, P. Roche, C. Ryder, S. Sansavini, H. Schmidt, S. Tartarini, J. J. Verhaegh, M. Vrielink-van Ginkel et G. J. King (1998). Aligning Male and Female Linkage Maps of Apple (Malus Pumila Mill.) Using Multi-Allelic Markers : *Theor Appl Genet* 97 : 60-73 (cf. p. 23).

Mangin, B., R. Rincent, C.-E. Rabier, L. Moreau et E. Goudemand-Dugue (2019). Training Set Optimization of Genomic Prediction by Means of EthAcc. *PLoS ONE* 14 : e0205629.

Marchini, J. et B. Howie (2010). Genotype Imputation for Genome-Wide Association Studies. *Nat Rev Genet* 11 : 499-511 (cf. p. 14).

Martini, J. W. R., T. L. Molnar, J. Crossa, S. J. Hearne et K. V. Pixley (2021). Opportunities and Challenges of Predictive Approaches for Harnessing the Potential of Genetic Resources. *Front. Plant Sci.* 12 : 674036 (cf. p. 17).

McClure, K. A., K. M. Gardner, G. M. Douglas, J. Song, C. F. Forney, J. DeLong, L. Fan, L. Du, P. M. Toivonen, D. J. Somers, I. Rajcan et S. Myles (2018). A Genome-Wide Association Study of Apple Quality and Scab Resistance. *The Plant Genome* 11 : 170075 (cf. p. 27, 28).

McClure, K. A., Y. Gong, J. Song, M. Vinqvist-Tymchuk, L. Campbell Palmer, L. Fan, K. Burgher-MacLellan, Z. Zhang, J.-M. Celton, C. F. Forney, Z. Migicovsky et S. Myles (2019). Genome-Wide Association Studies in Apple Reveal Loci of Large Effect Controlling Apple Polyphenols. *Hortic Res* 6 : 107 (cf. p. 27, 28).

McCouch, S. R., K. L. McNally, W. Wang et R. Sackville Hamilton (2012). Genomics of Gene Banks : A Case Study in Rice. *Am. J. Bot.* 99 : 407-423 (cf. p. 4).

Meuwissen, T. H. E., B. J. Hayes et M. E. Goddard (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 11 : 1819-1829 (cf. p. 6, 13).

Meuwissen, T. H. E., A. K. Sonesson, G. Gebregiwergis et J. A. Woolliams (2020). Management of Genetic Diversity in the Era of Genomics. *Front. Genet.* 11 : 880 (cf. p. 18).

Miar, Y., M. Sargolzaei et F. S. Schenkel (2017). A Comparison of Different Algorithms for Phasing Haplotypes Using Holstein Cattle Genotypes and Pedigree Data. *Journal of Dairy Science* 100 : 2837-2849 (cf. p. 59).

Migicovsky, Z., K. M. Gardner, D. Money, J. Sawler, J. S. Bloom, P. Moffett, C. T. Chao, H. Schwaninger, G. Fazio, G.-Y. Zhong et S. Myles (2016). Genome to Phenome Mapping in Apple Using Historical Data. *The Plant Genome* 9 : 1 (cf. p. 27, 28, 170).

Migicovsky, Z., T. H. Yeats, S. Watts, J. Song, C. F. Forney, K. Burgher-MacLellan, D. J. Somers, Y. Gong, Z. Zhang, J. Vrebalov, R. van Velzen, J. G. Giovannoni, J. K. C. Rose et S. Myles (2021). Apple Ripening Is Controlled by a NAC Transcription Factor. *Front. Genet.* 12 : 671300 (cf. p. 26, 177).

Miglior, F., A. Fleming, F. Malchiodi, L. F. Brito, P. Martin et C. F. Baes (2017). A 100-Year Review : Identification and Genetic Selection of Economically Important Traits in Dairy Cattle. *Journal of Dairy Science* 100 : 10251-10271 (cf. p. 1).

Miller, A. J. et B. L. Gross (2011). From Forest to Field : Perennial Fruit Crop Domestication. *Am. J. Bot.* 98 : 1389-1414 (cf. p. 28).

Millet, E. J., W. Kruijer, A. Coupel-Ledru, S. Alvarez Prado, L. Cabrera-Bosquet, S. Lacube, A. Charcosset, C. Welcker, F. van Eeuwijk et F. Tardieu (2019). Genomic Prediction of Maize Yield across European Environmental Conditions. *Nat Genet* 51 : 952-956 (cf. p. 178).

Minamikawa, M. F., M. Kunihisa, K. Noshita, S. Moriya, K. Abe, T. Hayashi, Y. Katayose, T. Matsumoto, C. Nishitani, S. Terakami, T. Yamamoto et H. Iwata (2021). Tracing Founder Haplotypes of Japanese Apple Varieties : Application in Genomic Prediction and Genome-Wide Association Study. *Hortic Res* 8 : 49 (cf. p. 27, 28).

Moeinizade, S., Y. Han, H. Pham, G. Hu et L. Wang (2021). A Look-Ahead Monte Carlo Simulation Method for Improving Parental Selection in Trait Introgression. *Sci Rep* 11 : 3918 (cf. p. 150).

Moeinizade, S., M. Wellner, G. Hu et L. Wang (2020). Complementarity-based Selection Strategy for Genomic Selection. *Crop Sci.* 60 : 149-156 (cf. p. 17).

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, G.-Y. Zhong et S. Myles (2015). LinkImpute : Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3* 5 : 2383-2390 (cf. p. 44).

Moriya, S., M. Kunihisa, K. Okada, T. Shimizu, C. Honda, T. Yamamoto, H. Muranty, C. Denancé, Y. Katayose, H. Iwata et K. Abe (2017). Allelic Composition of MdMYB1 Drives Red Skin Color Intensity in Apple (Malus × Domestica Borkh.) and Its Application to Breeding. *Euphytica* 213 : 78 (cf. p. 23).

Muir, W. (2007). Comparison of Genomic and Traditional BLUP-estimated Breeding Value Accuracy and Selection Response under Alternative Trait and Genomic Parameters : Comparison of BLUP and GEBV Selection. *J. Anim. Breed. Genet.* 124 : 342-355 (cf. p. 7, 10).

Muleta, K. T., G. Pressoir et G. P. Morris (2019). Optimizing Genomic Selection for a Sorghum Breeding Program in Haiti : A Simulation Study. *G3* 9 : 391-401 (cf. p. 130).

Muleta, K. T., P. Bulli, Z. Zhang, X. Chen et M. Pumphrey (2017). Unlocking Diversity in Germplasm Collections via Genomic Selection : A Case Study Based on Quantitative Adult Plant Resistance to Stripe Rust in Spring Wheat. *The Plant Genome* 10 (cf. p. 17).

Müller, D., P. Schopp et A. E. Melchinger (2017). Persistency of Prediction Accuracy and Genetic Gain in Synthetic Populations Under Recurrent Genomic Selection. *G3* 7 : 801-811 (cf. p. 148).

Muranty, H., C. Denancé, L. Feugey, J.-L. Crépin, Y. Barbier, S. Tartarini, M. Ordidge, M. Troggio, M. Lateur, H. Nybom, F. Paprstein, F. Laurens et C.-E. Durel (2020). Using Whole-Genome SNP Data to Reconstruct a Large Multi-Generation Pedigree in Apple Germplasm. *BMC Plant Biol* 20 : 2 (cf. p. 29-31, 36, 38, 44).

Muranty, H., M. Troggio, I. B. Sadok, M. A. Rifaï, A. Auwerkerken, E. Banchi, R. Velasco, P. Stevanato, W. E. van de Weg, M. Di Guardo, S. Kumar, F. Laurens et M. C. A. M. Bink (2015). Accuracy and Responses of Genomic Selection on Key Traits in Apple Breeding. *Hortic Res* 2 : 15060 (cf. p. 27, 28, 38).

Myles, S. (2013). Improving Fruit and Wine : What Does Genomics Have to Offer ? *Trends in Genetics* 29 : 190-196.

Naz, A. A., S. Dadshani, A. Ballvora, K. Pillen et J. Léon (2019). Genetic Analysis and Transfer of Favorable Exotic QTL Alleles for Grain Yield Across D Genome Using Two Advanced Backcross Wheat Populations. *Front. Plant Sci.* 10 : 711 (cf. p. 4).

Nejati-Javaremi, A., C. Smith et J. P. Gibson (1997). Effect of Total Allelic Relationship on Accuracy of Evaluation and Response to Selection. *J. Anim. Breed. Genet.* 75 : 1738 (cf. p. 6).

Neyhart, J. L., T. Tiede, A. J. Lorenz et K. P. Smith (2017). Evaluating Methods of Updating Training Data in Long-Term Genomewide Selection. *G3* : 12 (cf. p. 136).

Noiton, D. et C. J. A. Shelbourne (1992). Quantitative Genetics in an Apple Breeding Strategy. *Euphytica* 60 : 213-219 (cf. p. 32, 127, 169).

Noiton, D. A. et P. A. Alspach (1996). Founding Clones, Inbreeding, Coancestry, and Status Number of Modern Apple Cultivars. *JASHS* 121 : 773-782 (cf. p. 29).

Norman, A., J. Taylor, J. Edwards et H. Kuchel (2018). Optimising Genomic Selection in Wheat : Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3* 8 : 2889-2899 (cf. p. 10, 145).

Nsibi, M., B. Gouble, S. Bureau, T. Flutre, C. Sauvage, J.-M. Audergon et J.-L. Regnard (2020). Adoption and Optimization of Genomic Selection To Sustain Breeding for Apricot Fruit Quality. *G3* 10 : 4513-4529.

Nybom, H., J. Sehic et L. Garkava-Gustavsson (2008). Modern Apple Breeding Is Associated with a Significant Change in the Allelic Ratio of the Ethylene Production Gene *Md-ACS1*. *The Journal of Horticultural Science and Biotechnology* 83 : 673-677.

Nyine, M., S. Wang, K. Kiani, K. Jordan, S. Liu, P. Byrne, S. Haley, S. Baenziger, S. Chao, R. Bowden et E. Akhunov (2019). Genotype Imputation in Winter Wheat Using First-Generation Haplotype Map SNPs Improves Genome-Wide Association Mapping and Genomic Prediction of Traits. *G3* 9 : 125-133 (cf. p. 43).

Olatoye, M. O., L. V. Clark, N. R. Labonte, H. Dong, M. S. Dwiyanti, K. G. Anzoua, J. E. Brummer, B. K. Ghimire, E. Dzyubenko, N. Dzyubenko, L. Bagmet, A. Sabitov, P. Chebukin, K. Głowacka, K. Heo, X. Jin, H. Nagano, J. Peng, C. Y. Yu, J. H. Yoo, H. Zhao, S. P. Long, T. Yamada, E. J. Sacks et A. E. Lipka (2020). Training Population Optimization for Genomic Selection in *Miscanthus*. *G3* 10 : 2465-2476 (cf. p. 11).

Omasheva, M. Y., H. Flachowsky, N. A. Ryabushkina, A. S. Pozharskiy, N. N. Galiakparov et M.-V. Hanke (2017). To What Extent Do Wild Apples in Kazakhstan Retain Their Genetic Integrity ? *Tree Genetics & Genomes* 13 : 52 (cf. p. 30).

Otto, D., R. Petersen, B. Brauksiepe, P. Braun et E. R. Schmidt (2014). The Columnar Mutation (“Co Gene”) of Apple (Malus × Domestica) Is Associated with an Integration of a Gypsy-like Retrotransposon. *Mol Breeding* 33 : 863-880 (cf. p. 150).

Ou, J.-H. et C.-T. Liao (2019). Training Set Determination for Genomic Selection. *Theor Appl Genet* 132 : 2781-2792.

Papp, D., L. Gao, R. Thapa, D. Olmstead et A. Khan (2020a). Field Apple Scab Susceptibility of a Diverse Malus Germplasm Collection Identifies Potential Sources of Resistance for Apple Breeding. *CABI Agric Biosci* 1 : 16 (cf. p. 58).

Papp, D., J. Singh, D. Gadoury et A. Khan (2020b). New North American Isolates of *Venturia Inaequalis* Can Overcome Apple Scab Resistance of *Malus Floribunda* 821. *Plant Disease* 104 : 649-655 (cf. p. 177).

Parisi, L., Y. Lespinasse, J. Guillaumes et J. Krüger (1993). A New Race of *Venturia Inaequalis* Virulent to Apples with Resistance Due to the Vf Gene. *Phytopathology* 83 : 533 (cf. p. 177).

Pasquariello, M., S. Berry, C. Burt, C. Uauy et P. Nicholson (2020). Yield Reduction Historically Associated with the Aegilops Ventricosa 7DV Introgression Is Genetically and Physically Distinct from the Eyespot Resistance Gene Pch1. *Theor Appl Genet* 133 : 707-717 (cf. p. 4).

Patocchi, A., A. Wehrli, P.-H. Dubuis, A. Auwerkerken, C. Leida, G. Cipriani, T. Passey, M. Staples, F. Didelot, V. Philion, A. Peil, H. Laszakovits, T. Rühmer, K. Boeck, D. Baniulis, K. Strasser, R. Vávra, W. Guerra, S. Masny, F. Ruess, F. Le Berre, H. Nybom, S. Tartarini, A. Spornberger, A. Pikunova et V. G. M. Bus (2020). Ten Years of VINQUEST : First Insight for Breeding New Apple Cultivars With Durable Apple Scab Resistance. *Plant Disease* 104 : 2074-2081 (cf. p. 58).

Peace, C. P., J. J. Luby, W. E. van de Weg, M. C. A. M. Bink et A. F. Iezzoni (2014). A Strategy for Developing Representative Germplasm Sets for Systematic QTL Validation, Demonstrated for Apple, Peach, and Sweet Cherry. *Tree Genetics & Genomes* 10 : 1679-1694 (cf. p. 24).

Peil, A., H. Flachowsky, F. Dunemann, K. Richter, M. Hoefer, I. Király et M.-V. Hanke (2008). « Resistance Breeding in Apple at Dresden-Pillnitz ». Ecofruit - 13th International Conference on Cultivation Technique and Phytopathological Problems in Organic Fruit-Growing,220-225 (cf. p. 26).

Peng, T., X. Sun et R. H. Mumm (2014). Optimized Breeding Strategies for Multiple Trait Integration : I. Minimizing Linkage Drag in Single Event Introgression. *Mol Breeding* 33 : 89-104 (cf. p. 4).

Pérez, P. et G. de los Campos (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* 198 : 483-495 (cf. p. 121).

Pérez-Enciso et Zingaretti (2019). A Guide for Using Deep Learning for Complex Trait Genomic Prediction. *Genes* 10 : 553 (cf. p. 9).

Pimentel, E., C. Edel, R. Emmerling et K.-U. Götz (2015). How Imputation Errors Bias Genomic Predictions. *Journal of Dairy Science* 98 : 4131-4138 (cf. p. 15).

Pook, T., M. Mayer, J. Geibel, S. Weigend, D. Cavero, C. C. Schoen et H. Simianer (2020). Improving Imputation Quality in BEAGLE for Crop and Livestock Data. *G3* 10 : 177-188 (cf. p. 44, 60).

Prada, D. (2009). Molecular Population Genetics and Agronomic Alleles in Seed Banks : Searching for a Needle in a Haystack ? *Journal of Experimental Botany* 60 : 2541-2552 (cf. p. 4).

Pszczola, M. et M. P. L. Calus (2016). Updating the Reference Population to Achieve Constant Genomic Prediction Reliability across Generations. *Animal* 10 : 1018-1024 (cf. p. 11).

Pszczola, M., T. Strabel, H. Mulder et M. Calus (2012). Reliability of Direct Genomic Values for Animals with Different Relationships within and to the Reference Population. *Journal of Dairy Science* 95 : 389-400 (cf. p. 11).

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly et P. C. Sham (2007). PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81 : 559-575.

Ramasubramanian, V. et W. D. Beavis (2021). Strategies to Assure Optimal Trade-Offs Among Competing Objectives for the Genetic Improvement of Soybean. *Front. Genet.* 12 : 675500 (cf. p. 128).

Ramstein, G. P. et M. D. Casler (2019). Extensions of BLUP Models for Genomic Prediction in Heterogeneous Populations : Application in a Diverse Switchgrass Sample. *G3* 9 : 789-805.

Reif, J. C., P. Zhang, S. Dreisigacker, M. L. Warburton, M. van Ginkel, D. Hoisington, M. Bohn et A. E. Melchinger (2005). Wheat Genetic Diversity Trends during Domestication and Breeding. *Theor Appl Genet* 110 : 859-864 (cf. p. 3).

Resende, M. F. R., P. Muñoz, J. J. Acosta, G. F. Peter, J. M. Davis, D. Grattapaglia, M. D. V. Resende et M. Kirst (2012). Accelerating the Domestication of Trees Using Genomic Selection : Accuracy of Prediction Models across Ages and Environments. *New Phytologist* 193 : 617-624 (cf. p. 6).

Rincent, R., A. Charcosset et L. Moreau (2017). Predicting Genomic Selection Efficiency to Optimize Calibration Set and to Assess Prediction Accuracy in Highly Structured Populations. *Theor Appl Genet* 130 : 2231-2247 (cf. p. 12, 173).

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V. Rodríguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C.-C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset et L. Moreau (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 (cf. p. 12).

Rincent, R., J.-P. Charpentier, P. Faivre-Rampant, E. Paux, J. Le Gouis, C. Bastien et V. Segura (2018). Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions : Proof of Concept on Wheat and Poplar. *G3* 8 : 3961-3972 (cf. p. 179).

Rio, S., L. Moreau, A. Charcosset et T. Mary-Huard (2020). Accounting for Group-Specific Allele Effects and Admixture in Genomic Predictions : Theory and Experimental Evaluation in Maize. *Genetics* 216 : 27-41.

Rizzo, G., J. P. Monzon, F. A. Tenorio, R. Howard, K. G. Cassman et P. Grassini (2022). Climate and Agronomy, Not Genetics, Underpin Recent Maize Yield Gains in Favorable Environments. *Proc. Natl. Acad. Sci. U.S.A.* 119 : e2113629119 (cf. p. 2).

Rodríguez-Álvarez, M. X., M. P. Boer, F. A. van Eeuwijk et P. H. Eilers (2018). Correcting for Spatial Heterogeneity in Plant Breeding Experiments with P-splines. *Spatial Statistics* 23 : 52-71.

Ros-Freixedes, R., A. Whalen, C.-Y. Chen, G. Gorjanc, W. O. Herring, A. J. Mileham et J. M. Hickey (2020). Accuracy of Whole-Genome Sequence Imputation Using Hybrid Peeling in Large Pedigreed Livestock Populations. *Genet Sel Evol* 52 : 17 (cf. p. 61).

Roth, M., H. Muranty, M. Di Guardo, W. Guerra, A. Patocchi et F. Costa (2020). Genomic Prediction of Fruit Texture and Training Population Optimization towards the Application of Genomic Selection in Apple. *Hortic Res* 7 : 148 (cf. p. 27, 28).

Roux-Cuvelier, M., M. Grisoni, A. Bellec, E. Bloquel, C. Charron, M. Delalande, M. Delmas, A. Didier, C.-E. Durel, C.-H. Duval, F. Esnault, L. Feugey, E. Geoffriau, B. Khadari, S. Lepers-Andrzejewski, F. Luro, C. Marchal, A. Pernet, J. Salinier, M. Seguin, R. Stevens, B. van Issum-Groyer et R. Kahane (2021). Conservation of Horticultural Genetic Resources in France. *Chronica Horticulturae* 61 : 21-36 (cf. p. 31).

Rowan, T. N., J. L. Hoff, T. E. Crum, J. F. Taylor, R. D. Schnabel et J. E. Decker (2019). A Multi-Breed Reference Panel and Additional Rare Variants Maximize Imputation Accuracy in Cattle. *Genet Sel Evol* 51 : 77 (cf. p. 53, 60).

Ru, S., D. Main, K. Evans et C. Peace (2015). Current Applications, Challenges, and Perspectives of Marker-Assisted Seedling Selection in Rosaceae Tree Fruit Breeding. *Tree Genetics & Genomes* 11 : 8 (cf. p. 26).

Rutkoski, J., R. Singh, J. Huerta-Espino, S. Bhavani, J. Poland, J. Jannink et M. Sorrells (2015). Genetic Gain from Phenotypic and Genomic Selection for Quantitative Resistance to Stem Rust of Wheat. *The Plant Genome* 8 (cf. p. 16, 128).

Santantonio, N. et K. Robbins (2020). *A Hybrid Optimal Contribution Approach to Drive Short-Term Gains While Maintaining Long-Term Sustainability in a Modern Plant Breeding Program*. preprint. bioRxiv (cf. p. 128).

Sargolzaei, M., J. P. Chesnais et F. S. Schenkel (2014). A New Approach for Efficient Genotype Imputation Using Information from Relatives. *BMC Genomics* 15 : 478 (cf. p. 15, 59).

Sarinelli, J. M., J. P. Murphy, P. Tyagi, J. B. Holland, J. W. Johnson, M. Mergoum, R. E. Mason, A. Babar, S. Harrison, R. Sutton, C. A. Griffey et G. Brown-Guedira (2019). Trai-

ning Population Selection and Use of Fixed Effects to Optimize Genomic Predictions in a Historical USA Winter Wheat Panel. *Theor Appl Genet* 132 : 1247-1261.

Schaeffer, L. (2006). Strategy for Applying Genome-Wide Selection in Dairy Cattle. *J. Anim. Breed. Genet.* 123 : 218-223 (cf. p. 16).

Schauberger, B., T. Ben-Ari, D. Makowski, T. Kato, H. Kato et P. Ciais (2018). Yield Trends, Variability and Stagnation Analysis of Major Crops in France over More than a Century. *Sci Rep* 8 : 16865 (cf. p. 1).

Scheet, P. et M. Stephens (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data : Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics* 78 : 629-644 (cf. p. 14).

Schleif, N., S. M. Kaeppler et H. F. Kaeppler (2021). Generating Novel Plant Genetic Variation via Genome Editing to Escape the Breeding Lottery. *In Vitro Cell.Dev.Biol.-Plant* 57 : 627-644 (cf. p. 5).

Schouten, H. J., Y. Tikunov, W. Verkerke, R. Finkers, A. Bovy, Y. Bai et R. G. Visser (2019). Breeding Has Increased the Diversity of Cultivated Tomato in The Netherlands. *Front. Plant Sci.* 10 : 1606 (cf. p. 3).

Scott, B., M. Haile-Mariam, B. Cocks et J. Pryce (2021a). How Genomic Selection Has Increased Rates of Genetic Gain and Inbreeding in the Australian National Herd, Genomic Information Nucleus, and Bulls. *Journal of Dairy Science* 104 : 11832-11849 (cf. p. 16).

Scott, M. F., N. Fradgley, A. R. Bentley, T. Brabbs, F. Corke, K. A. Gardner, R. Horsnell, P. Howell, O. Ladejobi, I. J. Mackay, R. Mott et J. Cockram (2021b). Limited Haplotype Diversity Underlies Polygenic Trait Architecture across 70 Years of Wheat Breeding. *Genome Biol* 22 : 137 (cf. p. 2).

Segura, V., C.-E. Durel et E. Costes (2009). Dissecting Apple Tree Architecture into Genetic, Ontogenetic and Environmental Effects : QTL Mapping. *Tree Genetics & Genomes* 5 : 165-179 (cf. p. 23).

Serra, O., J. M. Donoso, R. Picañol, I. Batlle, W. Howad, I. Eduardo et P. Arús (2016). Marker-Assisted Introgression (MAI) of Almond Genes into the Peach Background : A Fast Method to Mine and Integrate Novel Variation from Exotic Sources in Long Intergeneration Species. *Tree Genetics & Genomes* 12 : 96 (cf. p. 179).

Shan, Q., Z. Wang, H. Ling, G. Zhang, J. Yan et F. Han (2021). Unreasonable Human Disturbance Shifts the Positive Effect of Climate Change on Tree-Ring Growth of Malus Sieversii in the Origin Area of World Cultivated Apples. *Journal of Cleaner Production* 287 : 125008 (cf. p. 30).

Si, Y., B. Vanderwerff et S. Zöllner (2021). Why Are Rare Variants Hard to Impute ? Coalescent Models Reveal Theoretical Limits in Existing Algorithms. *Genetics* 217 : iyab011 (cf. p. 148).

Silfverberg-Dilworth, E., C. L. Matasci, W. E. Van de Weg, M. P. W. Van Kaauwen, M. Walser, L. P. Kodde, V. Soglio, L. Gianfranceschi, C. E. Durel, F. Costa, T. Yamamoto, B. Koller, C. Gessler et A. Patocchi (2006). Microsatellite Markers Spanning the Apple (Malus x Domestica Borkh.) Genome. *Tree Genetics & Genomes* 2 : 202-224 (cf. p. 23).

Silva, É. D. B. da, A. Xavier et M. V. Faria (2021). Impact of Genomic Prediction Model, Selection Intensity, and Breeding Strategy on the Long-Term Genetic Gain and Genetic Erosion in Soybean Breeding. *Front. Genet.* 12 : 637133 (cf. p. 16, 171).

Singh, J., M. Sun, S. B. Cannon, J. Wu et A. Khan (2021). An Accumulation of Genetic Variation and Selection across the Disease-Related Genes during Apple Domestication. *Tree Genetics & Genomes* 17 : 29 (cf. p. 128).

Singh, S., P. Vikram, D. Sehgal, J. Burgueño, A. Sharma, S. K. Singh, C. P. Sansaloni, R. Joynson, T. Brabbs, C. Ortiz, E. Solis-Moya, V. Govindan, N. Gupta, H. S. Sidhu, A. K. Basandrai, D. Basandrai, L. Ledesma-Ramires, M. P. Suaste-Franco, G. Fuentes-Dávila, J. I. Moreno, K. Sonder, V. K. Singh, S. Singh, S. Shokat, M. A. R. Arif, K. A. Laghari, P. Srivastava, S. Bhavani, S. Kumar, D. Pal, J. P. Jaiswal, U. Kumar, H. K. Chaudhary, J. Crossa, T. S. Payne, M. Imtiaz, V. S. Sohu, G. P. Singh, N. S. Bains, A. Hall et K. V. Pixley (2018). Harnessing Genetic Potential of Wheat Germplasm Banks through Impact-Oriented-Prebreeding for Future Food and Nutritional Security. *Sci Rep* 8 : 12527 (cf. p. 4).

Sinnwell, J. P., T. M. Therneau et D. J. Schaid (2014). The Kinship2 R Package for Pedigree Data. *Hum Hered* 78 : 91-93 (cf. p. 133).

Sommer, L., M. Spiller, G. Stiewe, K. Pillen, J. C. Reif et A. W. Schulthess (2020). Proof of Concept to Unmask the Breeding Value of Genetic Resources of Barley (*Hordeum Vulgare*) with a Hybrid Strategy. *Plant Breeding* 139 : 536-549 (cf. p. 4).

Song, H., S. Ye, Y. Jiang, Z. Zhang, Q. Zhang et X. Ding (2019). Using Imputation-Based Whole-Genome Sequencing Data to Improve the Accuracy of Genomic Prediction for Combined Populations in Pigs. *Genet Sel Evol* 51 : 58 (cf. p. 120).

Sood, S., Z. Lin, B. Caruana, A. T. Slater et H. D. Daetwyler (2020). Making the Most of All Data : Combining Non-genotyped and Genotyped Potato Individuals with HBLUP. *The Plant Genome* 13.

Spengler, R. N. (2019). Origins of the Apple : The Role of Megafaunal Mutualism in the Domestication of Malus and Rosaceous Trees. *Front. Plant Sci.* 10 : 617 (cf. p. 19, 28).

Su, W., Y. Jing, S. Lin, Z. Yue, X. Yang, J. Xu, J. Wu, Z. Zhang, R. Xia, J. Zhu, N. An, H. Chen, Y. Hong, Y. Yuan, T. Long, L. Zhang, Y. Jiang, Z. Liu, H. Zhang, Y. Gao, Y. Liu, H. Lin, H. Wang, L. Yant, S. Lin et Z. Liu (2021). Polyploidy Underlies Co-Option and Diversification of Biosynthetic Triterpene Pathways in the Apple Tribe. *Proc. Natl. Acad. Sci. U.S.A.* 118 : e2101767118 (cf. p. 21).

Sun, C. et P. M. VanRaden (2014). Increasing Long-Term Response by Selecting for Favorable Minor Alleles. *PLoS ONE* 9 : e88510 (cf. p. 17, 128, 147, 171).

Sun, X., C. Jiao, H. Schwaninger, C. T. Chao, Y. Ma, N. Duan, A. Khan, S. Ban, K. Xu, L. Cheng, G.-Y. Zhong et Z. Fei (2020). Phased Diploid Genome Assemblies and Pan-Genomes Provide Insights into the Genetic History of Apple Domestication. *Nat Genet* 52 : 1423-1432 (cf. p. 20, 21).

Sverrisdóttir, E., E. H. R. Sundmark, H. Ø. Johnsen, H. G. Kirk, T. Asp, L. Janss, G. Bryan et K. L. Nielsen (2018). The Value of Expanding the Training Population to Improve Genomic Selection Models in Tetraploid Potato. *Front. Plant Sci.* 9 : 1118 (cf. p. 11).

Swarts, K., H. Li, J. A. Romero Navarro, D. An, M. C. Romay, S. Hearne, C. Acharya, J. C. Glaubitz, S. Mitchell, R. J. Elshire, E. S. Buckler et P. J. Bradbury (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The Plant Genome* 7 (cf. p. 44).

Swarup, S., E. J. Cargill, K. Crosby, L. Flagel, J. Kniskern et K. C. Glenn (2021). Genetic Diversity Is Indispensable for Plant Breeding to Improve Crops. *Crop Sci.* 61 : 839-852 (cf. p. 2).

Takeda, M., K. Inoue, H. Oyama, K. Uchiyama, K. Yoshinari, N. Sasago, T. Kojima, M. Kashima, H. Suzuki, T. Kamata, M. Kumagai, W. Takasugi, T. Aonuma, Y. Soma, S. Konno, T. Saito, M. Ishida, E. Muraki, Y. Inoue, M. Takayama, S. Nariai, R. Hideshima, R. Nakamura, S. Nishikawa, H. Kobayashi, E. Shibata, K. Yamamoto, K. Yoshimura, H. Matsuda, T. Inoue, A. Fujita, S. Terayama, K. Inoue, S. Morita, R. Nakashima, R. Suezawa, T. Hanamure, A. Zoda et Y. Uemoto (2021). Exploring the Size of Reference Population for Expected Accuracy of Genomic Prediction Using Simulated and Real Data in Japanese Black Cattle. *BMC Genomics* 22 : 799 (cf. p. 10).

Tanaka, R., S. T. Mandaharisoa, M. Rakotondramanana, H. N. Ranaivo, J. Pariasca-Tanaka, H. Kajiya-Kanegae, H. Iwata et M. Wissuwa (2021). From Gene Banks to Farmer's Fields : Using Genomic Selection to Identify Donors for a Breeding Program in Rice to Close the Yield Gap on Smallholder Farms. *Theor Appl Genet* 134 : 3397-3410 (cf. p. 17).

Tanksley, S. D. et S. R. McCouch (1997). Seed Banks and Molecular Maps : Unlocking Genetic Potential from the Wild. *Science* 277 : 1063-1066 (cf. p. 4).

Technow, F. et L. R. Totir (2015). Using Bayesian Multilevel Whole Genome Regression Models for Partial Pooling of Training Sets in Genomic Prediction. *G3* 5 : 1603-1612.

Testolin, R., L. Falginella, A. De Carli, G. De Mori et G. Cipriani (2021). Pyramiding Resistance Genes and Widening the Genetic Base of the Apple (Malus × Domestica Borkh.) Crop. *Italus Hortus* 28 : 32 (cf. p. 32, 127).

Thudi, M., R. Palakurthi, J. C. Schnable, A. Chitikineni, S. Dreisigacker, E. Mace, R. K. Srivastava, C. T. Satyavathi, D. Odeny, V. K. Tiwari, H.-M. Lam, Y. B. Hong, V. K. Singh,

G. Li, Y. Xu, X. Chen, S. Kaila, H. Nguyen, S. Sivasankar, S. A. Jackson, T. J. Close, W. Shubo et R. K. Varshney (2021). Genomic Resources in Plant Breeding for Sustainable Agriculture. *Journal of Plant Physiology* 257 : 153351 (cf. p. 2).

Tiezzi, F. et C. Maltecca (2015). Accounting for Trait Architecture in Genomic Predictions of US Holstein Cattle Using a Weighted Realized Relationship Matrix. *Genet Sel Evol* 47 : 24.

Tiplady, K. et S. Ric (2015). « Imputation Accuracy Measurement and Post-Imputation Quality in Imputed SNP Genotypes for Dairy Cattle ». 21st Conference of the Association for the Advancement of Animal Breeding and Genetics (cf. p. 61).

Tiret, M., M. Pégard et L. Sánchez (2021). How to Achieve a Higher Selection Plateau in Forest Tree Breeding ? Fostering Heterozygote × Homozygote Relationships in Optimal Contribution Selection in the Case Study of *Populus Nigra*. *Evol Appl* 14 : 2635-2646 (cf. p. 18).

Trucchi, E., A. Benazzo, M. Lari, A. Iob, S. Vai, L. Nanni, E. Bellucci, E. Bitocchi, F. Raffini, C. Xu, S. A. Jackson, V. Lema, P. Babot, N. Oliszewski, A. Gil, G. Neme, C. T. Michieli, M. De Lorenzi, L. Calcagnile, D. Caramelli, B. Star, H. de Boer, S. Boessenkool, R. Papa et G. Bertorelle (2021). Ancient Genomes Reveal Early Andean Farmers Selected Common Beans While Preserving Diversity. *Nature Plants* 7 : 123-128 (cf. p. 2).

Tsairidou, S., A. Hamilton, D. Robledo, J. E. Bron et R. D. Houston (2020). Optimizing Low-Cost Genotyping and Imputation Strategies for Genomic Selection in Atlantic Salmon. *G3* 10 : 581-590 (cf. p. 43).

Urrestarazu, J., C. Denancé, E. Ravon, A. Guyader, R. Guisnel, L. Feugey, C. Poncet, M. Lateur, P. Houben, M. Ordidge, F. Fernandez-Fernandez, K. M. Evans, F. Paprstein, J. Sedlak, H. Nybom, L. Garkava-Gustavsson, C. Miranda, J. Gassmann, M. Kellerhals, I. Suprun, A. V. Pikunova, N. G. Krasova, E. Torutaeva, L. Dondini, S. Tartarini, F. Laurens et C.-E. Durel (2016). Analysis of the Genetic Diversity and Structure across a Wide Range of Germplasm Reveals Prominent Gene Flow in Apple at the European Level. *BMC Plant Biol* 16 : 130 (cf. p. 29, 30).

Urrestarazu, J., H. Muranty, C. Denancé, D. Leforestier, E. Ravon, A. Guyader, R. Guisnel, L. Feugey, S. Aubourg, J.-M. Celton, N. Daccord, L. Dondini, R. Gregori, M. Lateur, P. Houben, M. Ordidge, F. Paprstein, J. Sedlak, H. Nybom, L. Garkava-Gustavsson, M. Troggio, L. Bianco, R. Velasco, C. Poncet, A. Théron, S. Moriya, M. C. A. M. Bink, F. Laurens, S. Tartarini et C.-E. Durel (2017). Genome-Wide Association Mapping of Flowering and Ripening Periods in Apple. *Front. Plant Sci.* 8 : 1923 (cf. p. 24, 29, 31, 43).

Van den Berg, I., P. J. Bowman, I. M. MacLeod, B. J. Hayes, T. Wang, S. Bolormaa et M. E. Goddard (2017). Multi-Breed Genomic Prediction Using Bayes R with Sequence Data and Dropping Variants with a Small Effect. *Genet Sel Evol* 49 : 70 (cf. p. 15).

Van den Berg, S., M. P. L. Calus, T. H. E. Meuwissen et Y. C. J. Wientjes (2015). Across Population Genomic Prediction Scenarios in Which Bayesian Variable Selection Outperforms GBLUP. *BMC Genetics* 16 : 146 (cf. p. 123).

Van de Wouw, M., C. Kik, T. van Hintum, R. van Treuren et B. Visser (2010). Genetic Erosion in Crops : Concept, Research Results and Challenges. *Plant Genet. Res.* 8 : 1-15 (cf. p. 2).

Van Tassel, D. L., L. R. DeHaan, L. Diaz-Garcia, J. Hershberger, M. J. Rubin, B. Schlautman, K. Turner et A. J. Miller (2022). Re-Imagining Crop Domestication in the Era of High Throughput Phenomics. *Current Opinion in Plant Biology* 65 : 102150 (cf. p. 179).

Vanavermaete, D., J. Fostier, S. Maenhout et B. De Baets (2021). Deep Scoping : A Breeding Strategy to Preserve, Reintroduce and Exploit Genetic Variation. *Theor Appl Genet* 134 : 3845-3861 (cf. p. 150).

Van Nocker, S. et S. E. Gardiner (2014). Breeding Better Cultivars, Faster : Applications of New Technologies for the Rapid Deployment of Superior Horticultural Tree Crops. *Hortic Res* 1 : 14022 (cf. p. 27).

VanRaden, P. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91 : 4414-4423 (cf. p. 8).

Varshney, R. K. (2021). *The Plant Genome* Special Issue : Advances in Genomic Selection and Application of Machine Learning in Genomic Prediction for Crop Improvement. *The Plant Genome* 14 (cf. p. 9).

Velasco, R., A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana, S. K. Bhatnagar, M. Troggio, D. Pruss, S. Salvi, M. Pindo, P. Baldi, S. Castelletti, M. Cavaiuolo, G. Coppola, F. Costa, V. Cova, A. Dal Ri, V. Goremykin, M. Komjanc, S. Longhi, P. Magnago, G. Malacarne, M. Malnoy, D. Micheletti, M. Moretto, M. Perazzolli, A. Si-Ammour, S. Vezzulli, E. Zini, G. Eldredge, L. M. Fitzgerald, N. Gutin, J. Lanchbury, T. Macalma, J. T. Mitchell, J. Reid, B. Wardell, C. Kodira, Z. Chen, B. Desany, F. Niazi, M. Palmer, T. Koepke, D. Jiwan, S. Schaeffer, V. Krishnan, C. Wu, V. T. Chu, S. T. King, J. Vick, Q. Tao, A. Mraz, A. Stormo, K. Stormo, R. Bogden, D. Ederle, A. Stella, A. Vecchietti, M. M. Kater, S. Masiero, P. Lasserre, Y. Lespinasse, A. C. Allan, V. Bus, D. Chagné, R. N. Crowhurst, A. P. Gleave, E. Lavezzo, J. A. Fawcett, S. Proost, P. Rouzé, L. Sterck, S. Toppo, B. Lazzari, R. P. Hellens, C.-E. Durel, A. Gutin, R. E. Bumgarner, S. E. Gardiner, M. Skolnick, M. Egholm, Y. Van de Peer, F. Salamini et R. Viola (2010). The Genome of the Domesticated Apple (Malus × Domestica Borkh.) *Nat Genet* 42 : 833-839 (cf. p. 21, 61).

Verma, S., K. Evans, Y. Guan, J. J. Luby, U. R. Rosyara, N. P. Howard, N. Bassil, M. C. A. M. Bink, W. E. van de Weg et C. P. Peace (2019). Two Large-Effect QTLs, Ma and Ma3, Determine Genetic Potential for Acidity in Apple Fruit : Breeding Insights from a Multi-Family Study. *Tree Genetics & Genomes* 15 : 18.

Voss-Fels, K. P., M. Cooper et B. J. Hayes (2019). Accelerating Crop Genetic Gains with Genomic Selection. *Theor Appl Genet* 132 : 669-686 (cf. p. 1, 7).

Wang, C., S. Hu, C. Gardner et T. Lübberstedt (2017). Emerging Avenues for Utilization of Exotic Germplasm. *Trends in Plant Science* 22 : 624-637 (cf. p. 4).

Wang, N., S. Jiang, Z. Zhang, H. Fang, H. Xu, Y. Wang et X. Chen (2018). Malus Sieversii : The Origin, Flavonoid Synthesis Mechanism, and Breeding of Red-Skinned and Red-Fleshed Apples. *Hortic Res* 5 : 70 (cf. p. 177).

Washburn, J. D., E. Cimen, G. Ramstein, T. Reeves, P. O'Briant, G. McLean, M. Cooper, G. Hammer et E. S. Buckler (2021). Predicting Phenotypes from Genetic, Environment, Management, and Historical Data Using CNNs. *Theor Appl Genet* 134 : 3997-4011 (cf. p. 10).

Watkins, R., R. A. Smith, International Board for Plant Genetic Resources, Commission of the European Communities et Committee on Disease Resistance Breeding and Use of Genebanks (1982). *Descriptor List for Apple (Malus)*.

Watts, S., Z. Migicovsky, K. A. McClure, C. H. J. Yu, B. Amyotte, T. Baker, D. Bowlby, K. Burgher-MacLellan, L. Butler, R. Donald, L. Fan, S. Fillmore, J. Flewelling, K. Gardner, M. Hodges, T. Hughes, V. Jagadeesan, N. Lewis, E. MacDonell, L. MacVicar, M. McElroy, D. Money, M. O'Hara, Q. Ong, L. Campbell Palmer, J. Sawler, M. Vinqvist-Tymchuk, H. V. Rupasinghe, J. M. DeLong, C. F. Forney, J. Song et S. Myles (2021). Quantifying Apple Diversity : A Phenomic Characterization of Canada's Apple Biodiversity Collection. *Plants People Planet* 1 : 1-14.

Wedger, M. J., A. C. Schumann et B. L. Gross (2021). Candidate Genes and Signatures of Directional Selection on Fruit Quality Traits during Apple Domestication. *Am. J. Bot.* 108 : 616-627.

Weeden, N. F., M. Hemmatt, D. M. Lawson, M. Lodhi, R. L. Bell, A. G. Manganaris, B. I. Reischs, S. K. Brown et G. .-. Ye (1994). Development and Application of Molecular Marker Linkage Maps in Woody Fruit Crops. *Euphytica* 77 : 71-75 (cf. p. 23).

Weng, Z., A. Wolc, X. Shen, R. L. Fernando, J. C. M. Dekkers, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan et D. J. Garrick (2016). Effects of Number of Training Generations on Genomic Prediction for Various Traits in a Layer Chicken Population. *Genet Sel Evol* 48 : 22 (cf. p. 11).

Werner, C. R., R. C. Gaynor, D. J. Sargent, A. Lillo, G. Gorjanc et J. M. Hickey (2020). *Genomic Selection Strategies for Clonally Propagated Crops*. preprint. bioRXiv (cf. p. 150).

Whalen, A., G. Gorjanc, R. Ros-Freixedes et J. M. Hickey (2018a). Assessment of the Performance of Hidden Markov Models for Imputation in Animal Breeding. *Genet Sel Evol* 50 : 44 (cf. p. 44, 60).

Whalen, A., R. Ros-Freixedes, D. L. Wilson, G. Gorjanc et J. M. Hickey (2018b). Hybrid Peeling for Fast and Accurate Calling, Phasing, and Imputation with Sequence Data of Any Coverage in Pedigrees. *Genet Sel Evol* 50 : 67 (cf. p. 60).

Whittaker, J. C., R. Thompson et M. C. Denham (2000). Marker-Assisted Selection Using Ridge Regression. *Genet. Res.* 75 : 249-252 (cf. p. 6).

Wientjes, Y., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten et M. Calus (2015). Empirical and Deterministic Accuracies of Across-Population Genomic Prediction. *Genet Sel Evol* 47 : 5.

Wientjes, Y. C. J., R. F. Veerkamp et M. P. L. Calus (2013). The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics* 193 : 621-631 (cf. p. 12, 13).

Wientjes, Y. C. J., P. Bijma, M. P. L. Calus, B. J. Zwaan, Z. G. Vitezica et J. van den Heuvel (2022). The Long-Term Effects of Genomic Selection : 1. Response to Selection, Additive Genetic Variance, and Genetic Architecture. *Genet Sel Evol* 54 : 19 (cf. p. 16, 171).

Wientjes, Y. C. J., P. Bijma, R. F. Veerkamp et M. P. L. Calus (2016). An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments. *Genetics* 202 : 799-823.

Wimmer, V., C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang et C.-C. Schön (2013). Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection. *Genetics* 195 : 573-587.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick et J. C. Dekkers (2011). Persistence of Accuracy of Genomic Estimated Breeding Values over Generations in Layer Chickens. *Genet Sel Evol* 43 : 23 (cf. p. 11, 148).

Wolfe, M. D., A. W. Chan, P. Kulakow, I. Rabbi et J.-L. Jannink (2021). Genomic Mating in Outbred Species : Predicting Cross Usefulness with Additive and Total Genetic Covariance Matrices. *Genetics* 219 : iyab122 (cf. p. 17, 150).

Woolliams, J., P. Berg, B. Dagnachew et T. Meuwissen (2015). Genetic Contributions and Their Optimization. *J. Anim. Breed. Genet.* 132 : 89-99 (cf. p. 128).

Xu, Y., X. Liu, J. Fu, H. Wang, J. Wang, C. Huang, B. M. Prasanna, M. S. Olsen, G. Wang et A. Zhang (2020). Enhancing Genetic Gain through Genomic Selection : From Livestock to Plants. *Plant Communications* 1 : 100005.

Yamagishi, N., R. Kishigami et N. Yoshikawa (2014). Reduced Generation Time of Apple Seedlings to within a Year by Means of a Plant Virus Vector : A New Plant-Breeding Technique with No Transmission of Genetic Modification to the next Generation. *Plant Biotechnol J* 12 : 60-68 (cf. p. 25).

Yang, C. J., R. Sharma, G. Gorjanc, S. Hearne, W. Powell et I. Mackay (2020). Origin Specific Genomic Selection : A Simple Process To Optimize the Favorable Contribution of Parents to Progeny. *G3* 10 : 2445-2455 (cf. p. 5, 17).

Yao, J.-L., J. Xu, A. Cornille, S. Tomes, S. Karunairetnam, Z. Luo, H. Bassett, C. Whitworth, J. Rees-George, C. Ranatunga, A. Snirc, R. Crowhurst, N. de Silva, B. Warren, C. Deng, S. Kumar, D. Chagné, V. G. M. Bus, R. K. Volz, E. H. A. Rikkerink, S. E. Gardiner, T. Giraud, R. MacDiarmid et A. P. Gleave (2015). A *microRNA* Allele That Emerged Prior to Apple Domestication May Underlie Fruit Size Evolution. *Plant J* 84 : 417-427 (cf. p. 19).

Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu, S. E. Mitchell, K. L. Roozeboom, D. Wang, M. L. Wang, G. A. Pederson, T. T. Tesso, P. S. Schnable, R. Bernardo et J. Yu (2016). Genomic Prediction Contributing to a Promising Global Strategy to Turbocharge Gene Banks. *Nature Plants* 2 : 16150 (cf. p. 5, 17).

Yue, X.-P., C. Dechow et W.-S. Liu (2015). A Limited Number of Y Chromosome Lineages Is Present in North American Holsteins. *Journal of Dairy Science* 98 : 2738-2745 (cf. p. 2).

Zapata-Valenzuela, J., R. W. Whetten, D. Neale, S. McKeand et F. Isik (2013). Genomic Estimated Breeding Values Using Genomic Relationship Matrices in a Cloned Population of Loblolly Pine. *G3* 3 : 909-916 (cf. p. 6).

Zeitler, L., J. Ross-Ibarra et M. G. Stetter (2020). Selective Loss of Diversity in Doubled-Haploid Lines from European Maize Landraces. *G3* 10 : 2497-2506 (cf. p. 2).

Zhang, A., H. Wang, Y. Beyene, K. Semagn, Y. Liu, S. Cao, Z. Cui, Y. Ruan, J. Burgueño, F. San Vicente, M. Olsen, B. M. Prasanna, J. Crossa, H. Yu et X. Zhang (2017). Effect of Trait Heritability, Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 Bi-Parental Tropical Maize Populations. *Front. Plant Sci.* 8 : 1916.

Zhang, L., J. Hu, X. Han, J. Li, Y. Gao, C. M. Richards, C. Zhang, Y. Tian, G. Liu, H. Gul, D. Wang, Y. Tian, C. Yang, M. Meng, G. Yuan, G. Kang, Y. Wu, K. Wang, H. Zhang, D. Wang et P. Cong (2019). A High-Quality Apple Genome Assembly Reveals the Association of a Retrotransposon and Red Fruit Colour. *Nat Commun* 10 : 1494 (cf. p. 21, 44, 61).

Zheng, C., M. P. Boer et F. A. van Eeuwijk (2018). Accurate Genotype Imputation in Multi-parental Populations from Low-Coverage Sequence. *Genetics* 210 : 71-82 (cf. p. 44).

Zheng, H.-F., J.-J. Rong, M. Liu, F. Han, X.-W. Zhang, J. B. Richards et L. Wang (2015). Performance of Genotype Imputation for Low Frequency and Rare Variants from the 1000 Genomes. *PLoS ONE* 10 : e0116487 (cf. p. 15).

Zhu, G., S. Wang, Z. Huang, S. Zhang, Q. Liao, C. Zhang, T. Lin, M. Qin, M. Peng, C. Yang, X. Cao, X. Han, X. Wang, E. van der Knaap, Z. Zhang, X. Cui, H. Klee, A. R. Fernie, J. Luo et S. Huang (2018). Rewiring of the Fruit Metabolome in Tomato Breeding. *Cell* 172 : 249-261.e12 (cf. p. 4).

Zhu, Y. et B. H. Barritt (2008). Md-ACS1 and Md-ACO1 Genotyping of Apple (Malus x Domestica Borkh.) Breeding Parents and Suitability for Marker-Assisted Selection. *Tree Genetics & Genomes* 4 : 555-562 (cf. p. 26).

Zine-El-Abidine, M., H. Dutagaci, G. Galopin et D. Rousseau (2021). Assigning Apples to Individual Trees in Dense Orchards Using 3D Colour Point Clouds. *Biosystems Engineering* 209 : 30-52 (cf. p. 174).

DOCTORAT ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

université angers

Titre : Valorisation des ressources génétiques du pommier dans une population d'amélioration élite grâce à la sélection génomique

Mots clés : *Malus domestica*, prédiction génomique, diversité génétique, pré-breeding, simulations

Résumé : Les programmes d'amélioration à l'échelle mondiale chez le pommier utilisent de façon récurrente un petit nombre de variétés comme géniteurs. Cette base génétique étroite des populations d'amélioration élite est une préoccupation pour les sélectionneurs. Dans ce contexte, l'utilisation de ressources génétiques présentant des allèles favorables rares pourrait permettre d''enrichir cette base génétique. La sélection génomique pourrait alors représenter une approche intéressante pour valoriser de tels génotypes dans un programme de pré-breeding. L'objectif de cette thèse est d'étudier l'intérêt de la sélection génomique dans de tels programmes chez le pommier. Afin de construire des modèles de prédiction basés sur un grand nombre de marqueurs, nous avons dans un premier temps étudié par simulations la précision d'imputation qu'il était possible d'atteindre dans des familles biparentales et avons montré qu'il était possible d'obtenir des données imputées de haute qualité.
Nous avons par la suite évalué l'intérêt de combiner des ressources génétiques et du matériel élite afin de constituer une population d'entrainement à large diversité utilisable dans différents contextes et avons obtenu des précisions de prédiction modérées à élevées selon le trait étudié. Nous avons enfin simulé deux schémas de pré-breeding et avons montré que la sélection génomique pouvait permettre un gain génétique par unité de temps et une augmentation de la fréquence des allèles favorables rares plus importants que la sélection phénotypique. Les résultats de la thèse montrent que la sélection génomique peut permettre d'améliorer l'efficacité des programmes de pré-breeding chez le pommier.

Title: Harnessing apple genetic resources in an elite breeding population using genomic selection

Keywords: *Malus domestica*, genomic prediction, genetic diversity, pre-breeding, simulations

Abstract: Apple breeding programs worldwide have usually been using a few recurrent parents for decades, which has resulted in a narrow genetic base in elite populations that is worrisome for many breeders. In this context, genotypes from genetic resources that contain rare favourable alleles could be used to broaden the genetic base of elite material. Genomic selection could then be an appealing approach in order to efficiently harness this diversity in a pre-breeding program. The aim of this thesis is to explore the value of genomic selection in such programs in apple. In order to use high density genotypic data to build prediction models, we first used simulations to quantify the imputation accuracy that could be obtained in several biparental families and showed that high quality imputed genotypic data could be obtained.
We then explored the benefits of combining genetic resources and elite material into a single, diverse training set that could be used in multiple contexts and obtained moderate to high predictive abilities in doing so. We finally simulated two pre-breeding schemes and showed that genomic selection could lead to a higher genetic gain per unit of time and increases in rare allele frequencies than using phenotypic selection. The results from this work suggest that genomic selection could help improve the efficiency of pre-breeding programs in apple.