Optimisation du schéma de sélection chez le blé tendre : apport des prédictions génomiques et des caractères corrélés

Sarah Ben Sadoun

Université Clermont Auvergne

École Doctorale des Sciences de la vie, Santé, Agronomie & Environnement

Thèse de doctorat

Présentée à l'Université Clermont Auvergne pour l'obtention du grade de

Docteur d'Université

Spécialité : Biologie végétale

Soutenue le 27/05/2020

par

Sarah BEN SADOUN

Optimisation du schéma de sélection chez le blé tendre : apport des prédictions génomiques et des caractères corrélés

Composition du jury

Jacques LE GOUIS Directeur de recherche, INRAE Clermont-Ferrand	Président
Mathilde CAUSSE Directrice de recherche, INRAE Centre PACA	Rapporteur
Laurence MOREAU Directrice de recherche, INRAE Le Moulon	Rapporteur
Leopoldo SANCHEZ-RODRIGUEZ Directeur de recherche, INRAE Orléans	Rapporteur
Julie BOUDET Enseignant-chercheur, Université Clermont Auvergne	Examinatrice
Pascal CROISEAU Chargé de recherche, INRAE Jouy-en-Josas	Examinateur
Bettina LADO Chargée de recherche, Montevideo, Uruguay	Examinatrice
Gilles CHARMET Directeur de recherche, INRAE Clermont-Ferrand	Directeur de Thèse
Sophie BOUCHET Chargée de recherche, INRAE Clermont-Ferrand	Invitée

UMR 1095 INRAE-UCA « Génétique, Diversité et Écophysiologie des Céréales »

LABORATOIRE D'ACCUEIL

GDEC - Génétique Diversité Ecophysiologie des Céréales

UMR 1095 INRAE-UCA

5 chemin de Beaulieu 63000 Clermont-Ferrand

FRANCE

Thèse financée par la région Auvergne-Rhône-Alpes et par le Fonds européen de développement régional (FEDER)

Remerciements

En premier lieu, je tiens à remercier mon directeur de thèse, Gilles Charmet, et mon encadrante de thèse, Sophie Bouchet, pour avoir assuré l'encadrement de mes travaux de thèse et pour avoir contribué à alimenter ma réflexion au cours de ces trois années de thèse. Un très grand merci pour votre gentillesse, votre patience et vos précieux conseils.

Je remercie sincèrement les membres du jury Laurence Moreau, Leopoldo Sanchez-Rodriguez, Mathilde Causse, Jacques Le Gouis, Pascal Croiseau, Bettina Lado et Julie Boudet d'avoir accepté d'évaluer ce travail.

Je souhaite également remercier les membres de mes comités de thèse Brigitte Mangin, Nicolas Heslot, Nourollah Ahmadi et Thierry Tribout qui m'ont aidée à recentrer mes recherches et à en identifier de nouvelles pistes de travail.

Un grand merci également à Renaud Rincent pour son aide très précieuse et pour avoir eu la patience de répondre à mes nombreuses questions. Je remercie aussi Catherine Ravel pour son aide sur l'analyse de marqueurs dérivés des séquences d'ADN des gènes *Glu*.

L'environnement dans lequel j'ai travaillé au GDEC a été essentiel au bon déroulement de ma thèse. J'en remercie chaleureusement Thierry Langin, directeur d'unité, ainsi que l'ensemble des membres l'équipe DGS et plus généralement les membres de l'UMR.

Ces années de thèse auraient eu une saveur différente sans la présence des autres doctorants et post-doctorants du GDEC. Merci à chacun d'entre vous pour avoir su égayer chacune de mes journées au GDEC.

Mes derniers remerciements s'adressent à ma famille et à Rénald. Merci de m'avoir toujours soutenue et encouragée et surtout merci d'avoir embelli ces trois dernières années.

Publications dans des revues scientifiques avec comité de lecture

- **Ben Sadoun S, Rincent R, Auzanneau J, Oury FX, Rolland B, Heumez E, Ravel C, Charmet G, Bouchet S.** « Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: application to bread making quality » Theoretical and Applied Genetics **[En revision, modifications mineures demandées]**

- **Ben Sadoun S, Furgeray-Scarbel A, Oury FX, Heumez E, Rolland B, Auzanneau J, Charmet G, Lemarie S, Bouchet S.** « Integration of genomic selection into bread wheat breeding schemes: a simulation pipeline including economic constraints » **[En préparation]**

- **Ravel C, Faye A, Ben Sadoun S, Ranoux M, et al. (2020). «** SNP markers for early identification of high molecular weight glutenin subunits (HMW-GSs) in bread wheat. » Theoretical and Applied Genetics, 1-20. **[Publié]**

Liste des communications dans des séminaires ou congrès

- **Ben Sadoun S, Rincent R, Auzanneau J, Oury FX, Rolland B, Heumez E, Ravel C, Charmet G, Bouchet S.** « Bread-making quality predictions: optimization of multi-trait assisted genomic selection » EUCARPIA Cereal Section 2018, Clermont-Ferrand, France **[Poster]**

- **Ben Sadoun S, Rincent R, Auzanneau J, Oury FX, Rolland B, Heumez E, Ravel C, Charmet G, Bouchet S.** « Bread-making quality predictions: optimization of multi-trait assisted genomic selection » Gorgon Research Conferences – Quantitative Genetics and Genomics 2019, Lucca, Italie **[Poster]**

- **Ben Sadoun S, Rincent R, Auzanneau J, Oury FX, Rolland B, Heumez E, Ravel C, Charmet G, Bouchet S.** « Economical optimization of Bread-making quality predictions using correlated traits and molecular markers » International Wheat Congress 2019, Saskatoon, Canada **[Communication orale]**

- **Ben Sadoun S, Charmet G, Bouchet S.** « Prédiction génomique de la valeur boulangère », ma thèse en 5 minutes, Phloème 2020, Paris, France **[Communication orale]**

Principales abréviations

ADN : Acide désoxyribonucléique

AFLP : Amplified fragment length polymorphsim

BAF : Blé améliorant ou de force

BAU : Blé autres usages

BAU Imp : Blé autres usages impanifiable

BayesSSVS : Bayesian Stochastic Search Variable Selection

BB : Blé biscuitier

BLUE : Best Linear Unbiased Estimation

BLUP : Best Linear Unbiased Prediction

BMS : Bread-making score

BPS : Blé panifiable supérieur

CDmean : Moyenne du coefficient de détermination généralisé

COV : Certificat d'obtention végétale

CRB : Centre de ressources biologiques

CT : Coût total

CTPS : Comité technique permanent de la sélection des plantes cultivées

DHS : Distinction, homogénéité, stabilité

FAO : Food and Agriculture Organization (ou Organisation des Nations Unies pour l'alimentation et l'agriculture)

Gb : Giga base

GBLUP : Genomic Best Linear Unbiased Prediction

GEBV : Genomic estimated breeding value

GEVES : Groupement d'étude et de contrôle des variétés et des semences

GPD : Grain protein deviation

GPS : Genomic and phenotypic selection

GS : Genomic selection

HD : Haploïdes doublés (ou DH: Doubled haploids)

HMW-GS : High-molecular-weight glutenin subunits

h^2 : Héritabilité

INRAE : Institut National de la Recherche en Agriculture, Alimentation et Environnement

IWGSC : International Wheat Genome Sequencing Consortium

L : Extensibilité de la pâte (paramètre de l'alvéographe de Chopin)

MT : Multi-trait genomic prediction

MVN : Multivariate normal distribution

NBT : New Breeding Techniques

NGS : Next-generation sequencing

NIRS : Near-Infrared Reflectance Spectroscopy

P : Ténacité de la pâte (paramètre de l'alvéographe de Chopin)

PEVmean : Variance d'erreur de prédiction moyenne

Ph1 : Pairing homeologous 1

PS : Phenotypic selection

QTL : Quantitative trait loci

RFLP : Restriction fragment length polymorphism

SAM : Sélection assistée par marqueurs

SNP : Single nucleotide polymorpshim

SSD : Single seed descent

SSR : Single sequence repeat

ST : Single-trait genomic prediction

TA : Trait-assisted genomic prediction

TBV : True breeding value

UC : Usefulness criterion

VATE : Valeur agronomique, technologique et environnementale

W : Force boulangère (paramètre de l'alvéographe de Chopin)

Table des matières

Liste des figures

Liste des tables

Introduction générale

Introduction générale

Cultivé dans de nombreuses régions du globe, le blé tendre est la troisième céréale la plus produite au monde. Cette céréale représente environ 20% des apports caloriques chez l'Homme et joue ainsi un rôle essentiel dans l'alimentation humaine (Shewry and Hey 2015). Cependant, la production de blé tendre fait actuellement face à plusieurs enjeux. En effet, l'augmentation de la population mondiale entraine une demande croissante en produits alimentaires, c'est pourquoi le rendement du blé tendre doit être amélioré tout en maintenant une production de grain de qualité satisfaisante. Cette augmentation du rendement doit également aller de pair avec une production plus respectueuse de l'environnement (notamment en réduisant l'utilisation de pesticides et de fertilisants azotés de synthèse). De plus, le changement climatique constitue un défi majeur pour l'agriculture car il s'accompagne d'une augmentation de la variabilité des conditions météorologiques. Afin de faire face à ces enjeux, des innovations agronomiques, telles que le développement d'outils d'aide à la décision, sont mises en œuvres. Le progrès génétique constitue également un levier majeur pour répondre à ces objectifs.

Chez le blé tendre, la sélection variétale consiste à créer des variétés regroupant plusieurs caractères d'intérêt agronomique comme un rendement élevé, une résistance aux maladies ou une bonne qualité boulangère. Certains de ces caractères sont difficiles ou coûteux à phénotyper. C'est notamment le cas de la note de panification qui est utilisée en France lors de l'inscription de nouvelles variétés et permet de classer les variétés en fonction de leur qualité boulangère. Ce caractère complexe ne peut être mesuré qu'à un stade tardif du programme de sélection lorsqu'il y a suffisamment de grains par lignée pour produire la quantité de farine requise pour le test de panification.

Par ailleurs, la baisse du coût de génotypage ainsi que l'amélioration des capacités de calcul et du stockage des données ont rendu possible l'essor d'un nouvel outil pour l'amélioration des plantes: la sélection génomique. Les prédictions génomiques permettent de prédire la performance de candidats à la sélection sans qu'ils ne soient phénotypés pour le caractère d'intérêt. Cette méthode repose sur l'estimation simultanée des effets sur le caractère d'intérêt de tous les marqueurs répartis sur l'ensemble du génome (Whittaker et al. 2000; Meuwissen et al. 2001). La sélection génomique a pour but d'améliorer le progrès génétique par unité de temps et de coût et elle est particulièrement intéressante dans le cas de caractères longs et coûteux à évaluer. La sélection génomique a dans un premier temps fait ses preuves chez les bovins laitiers où le progrès génétique annuel a été doublé (Schaeffer al. 2006; Garcia-Ruiz et al. 2016). Cette méthode est aujourd'hui appliquée à d'autres espèces animales et végétales. Les prédictions génomiques sont également appliquées au domaine de la santé humaine, où l'objectif est alors de prédire le risque de développer certaines maladies.

Actuellement, la majeure partie des travaux sur les prédictions génomiques dans le domaine végétal portent sur un seul caractère. Or la corrélation entre caractères plus ou moins héritables et plus ou moins difficiles ou chers à mesurer peut permettre d'améliorer les prédictions ou de diminuer l'effort de phénotypage à l'échelle du programme de sélection. L'avantage économique de différentes approches utilisant des modèles de prédictions génomiques multi-caractères est rarement testé ou discuté dans la littérature. La plupart des études se concentrent sur la qualité des prédictions obtenues sans s'interroger sur les étapes de sélection au niveau desquelles les prédictions génomiques seraient les plus utiles en termes de gain génétique ou économique.

L'objectif de la thèse est d'étudier l'intérêt des prédictions génomiques à différentes étapes du schéma de sélection du blé tendre, pour optimiser la précision des prédictions, le gain génétique et / ou le coût du programme de sélection. La thèse est découpée en deux axes.

Le premier axe porte sur l'optimisation de l'utilisation du budget alloué au phénotypage et de la composition de la population d'entraînement afin d'améliorer la qualité des prédictions génomiques dans un contexte multi-caractère. La note de panification a été choisi comme caractère d'intérêt à prédire. Nous avons comparé des modèles de prédiction génomique mono-caractère et multi-caractère. Dans les modèles multi-caractères, nous avons pris en compte la force boulangère (paramètre W de l'alvéographe de Chopin) qui est un caractère positivement corrélé à la note de panification et qui est à la fois plus héritable et moins cher à phénotyper. Les modèles multi-caractères ont été évalués en faisant varier le nombre d'individus phénotypés pour chacun des deux caractères avec un budget alloué au phénotypage fixé. De plus, pour optimiser le choix des individus faisant partie de la population d'entraînement, nous avons étendu le critère CDmean (Rincent et al. 2012) au contexte multi-caractère avec un critère « CDmulti ».

Le second axe de la thèse porte sur la comparaison de deux types de schémas de sélection mono-caractères simulés : un schéma de sélection phénotypique (PS) avec deux étapes de sélection basées sur des essais en parcelles, et un schéma de sélection combinée phénotypique et génomique (GPS) avec une première étape de sélection génomique et une seconde étape de sélection basée sur des essais en parcelles. Pour ce deuxième schéma, deux façons de réaliser les croisements ont été comparées : soit les meilleurs individus de la génération précédente ont été croisés de manière aléatoire, soit les croisements ont été choisis en fonction de la complémentarité des parents d'un point de vue allélique. Dans la première partie de la thèse, seuls les coûts de phénotypage et le budget alloué au phénotypage étaient pris en compte pour comparer les différentes approches. En revanche dans ce second volet, les coûts de chacune des étapes du schéma de sélection (notamment les étapes de multiplication ou de génotypage) ont été définis et pris en compte afin de comparer les schémas de sélection pour un budget total fixe. Afin d'évaluer l'impact du budget alloué au programme de sélection, de l'intensité de sélection à chaque

étape, du coût de génotypage et de l'héritabilité du caractère sur les schémas PS et GPS, 36 scénarios faisant varier ces différents paramètres, ont été simulés pendant 3 cycles de sélection successifs.

Le premier chapitre de ce manuscrit correspond à une synthèse bibliographique. Cette synthèse bibliographique présente dans un premier temps le blé tendre et les enjeux liés à la sélection variétale pour cette espèce. Elle montre ensuite l'apport des outils de prédiction génomique pour l'amélioration du blé tendre, avant de décrire le principe des prédictions génomiques multi-caractères et leur application en sélection végétale. Le deuxième chapitre de cette thèse vise à décrire les jeux de données qui ont été utilisés pour réaliser les analyses présentées dans les chapitres 3 et 4. Les deux chapitres suivants sont présentés sous forme d'articles scientifiques écrits en anglais et correspondent aux deux volets de la thèse introduits précédemment. Le premier article a été soumis au journal *Theoretical and Applied Genetics* et est en cours de révision avec des demandes de modifications mineures. Le second article est quant à lui en cours de préparation. Le manuscrit se conclut par une discussion générale qui compare les résultats obtenus au cours de la thèse à ceux disponibles dans la littérature et qui présente les perspectives à court et à plus long terme de cette étude. La liste des références bibliographiques citées dans chacun des chapitres est disponible à la fin du manuscrit.

Chapitre 1 :

Synthèse bibliographique

I. Le blé tendre (*Triticum aestivum* L.)

1. Généralités sur le blé tendre

a. Taxonomie et histoire évolutive du blé tendre

Le blé tendre (*Triticum aestivum* L.), aussi appelé froment, appartient à la tribu des *Triticeae* au sein de la famille des *Poaceae* (anciennement *Gramineae*) et plus largement au groupe des angiospermes monocotylédones. La famille des Poacées regroupe la quasi-totalité du groupe de plantes appelées céréales dont fait partie le blé tendre. Le terme céréale désigne les plantes cultivées principalement pour leurs graines qui servent de base à l'alimentation humaine et animale.

Le genre *Triticum*, auquel appartient le blé tendre, est constitué d'espèces aux niveaux de ploïdie variés. En effet, le genre *Triticum* réunit des espèces diploïdes telles que l'engrain (*Triticum monococcum*), des tétraploïdes comme le blé dur (*Triticum turgidum ssp durum*), ainsi que des espèces hexaploïdes. Le blé tendre est pour sa part une espèce allo-hexaploïde qui résulte de deux hybridations interspécifiques suivies d'un doublement chromosomique (Figure 1.a). La première phase d'allopolyploïdisation correspond à un croisement entre deux espèces sauvages diploïdes : *Triticum urartu* (donneuse du génome A) et une espèce encore inconnue ou disparue proche d'*Aegilops speltoïdes* (donneuse du génome B). Ce croisement s'est produit il y a environ 0,8 million d'années (Marcussen et al. 2014) et a donné après doublement chromosomique un blé tétraploïde (*Triticum turgidum spp dicoccoïdes*) duquel dérivent l'amidonnier *Triticum turgidum ssp dicoccum* à grain vêtu, puis le blé dur *Triticum turgidum ssp durum* à grain nu. La seconde phase d'allopolyploïdisation a eu lieu entre un blé tétraploïde, probablement *Triticum dicoccum* déjà cultivé et *Aegilops tauschii* (espèce diploïde donneuse du génome D) il y a environ 10 000 ans (Gill et al. 2004). C'est de cette seconde hybridation qu'est issu l'ancêtre hexaploïde du blé tendre. Ces évènements d'allopolyploïdisation ont conduit à la création d'espèces avec des génomes complexes. Le génome du blé tendre est donc composé de trois sous-génomes proches dits homéologues A, B et D possédant chacun 7 paires de chromosomes, soit 42 chromosomes au total (Figure 1.b). Malgré la présence de paires de chromosomes présentant des similitudes génétiques, la structure du génome du blé tendre reste stable au fil des générations car seuls les chromosomes homologues peuvent s'apparier lors de la méiose. Ce phénomène est dû à la présence d'un groupe de gènes, dont le plus important est le gène *Pairing homeologous 1* (Ph1) situé sur le chromosome 5B (Riley and Chapman 1958; Sears and Okamoto 1958), qui empêchent l'appariement de chromosomes

homéologues pendant la méiose. Le blé se comporte donc comme une espèce diploïde lors de la méïose, à quelques exceptions près.

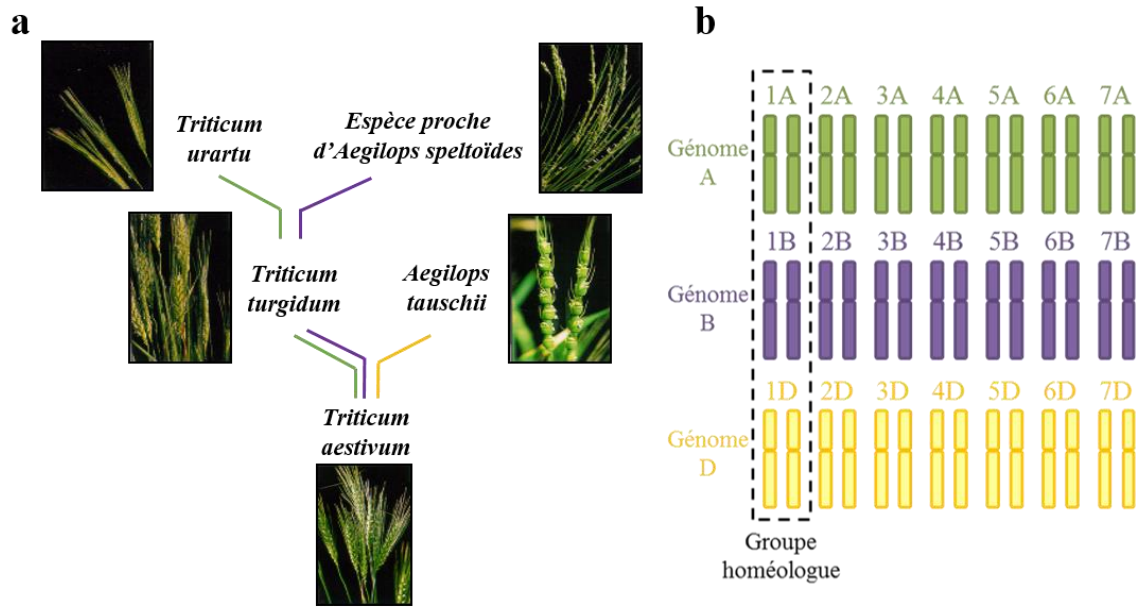

Figure 1 : Représentation schématique de l'histoire phylogénétique du blé tendre et de la structure de son génome.

a. Les événements d'hybridation et les génomes A (vert), B (violet) et D (orange) sont représentés sur la Figure (d'après Gill et al. 2004).
b. Le génome du blé tendre est structuré en trois sous-génomes A (vert), B (violet) et D (orange) diploïdes constitués de 7 paires de chromosomes.

b. Description botanique du blé tendre

Le blé tendre, comme les autres espèces du genre *Triticum*, est une herbacée annuelle à croissance définie. Le blé tendre est constitué d'un appareil racinaire et d'une partie aérienne constituée d'un ensemble de brins appelés talles. Chaque talle est formée d'une tige feuillée ou chaume qui constitue la partie végétative aérienne du blé tendre et porte à son extrémité un épi qui correspond à la partie reproductrice. L'épi de blé est dit composé car il est constitué de plusieurs sous-unités appelées épillets qui sont portés par le rachis. Chaque épillet regroupe généralement trois à quatre fleurs fertiles et chacune des fleurs est entourée de deux glumelles. La fécondation se produit généralement lorsque la fleur est encore fermée par les glumelles, les fleurs du blé tendre sont donc cléistogames, ce qui explique que l'autofécondation est le mode de reproduction le plus fréquent chez le blé tendre. Chez le blé tendre comme chez les autres espèces de la famille des *Poaceae*, le grain est un fruit sec indéhiscent appelé caryopse. Le grain du blé tendre est entouré de glumes et glumelles sur l'épi mais ces dernières

n'adhèrent pas au grain et sont éliminées lors du battage. Les enveloppes du grain entourent l'embryon (également appelé germe) et l'albumen (Figure 2). L'albumen correspond au tissu le plus abondant du grain. En effet il représente entre 80 et 85% de la matière sèche totale du grain (Belderok et al. 2000). Il est composé de deux parties : l'albumen amylacé et la couche à aleurone. L'albumen amylacé contient la totalité de l'amidon présent dans le grain mais également une grande partie des protéines du grain.

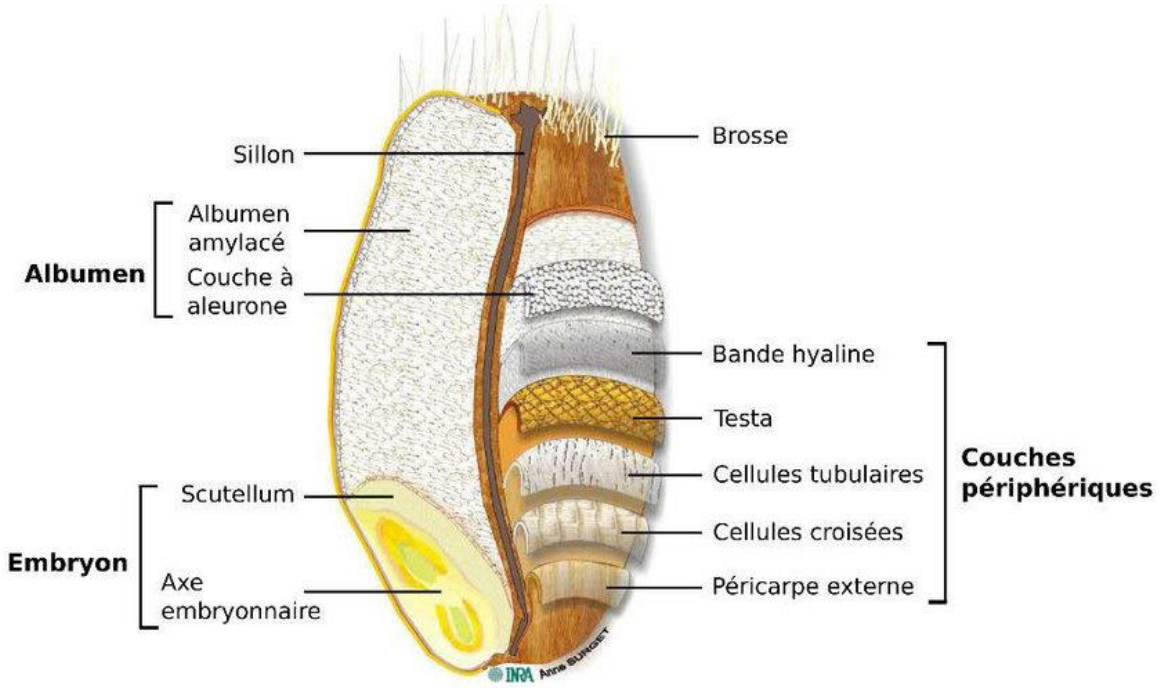

Figure 2 : Anatomie du grain de blé tendre. Le grain de blé est constitué de trois parties : l'embryon, l'albumen et les couches périphériques (d'après Surget and Barron 2005).

c. Protéines du blé tendre

En 1907, Osborne a proposé une première classification des protéines du grain de blé tendre reposant sur leur solubilité dans différentes solutions. Cette classification a notamment permis de distinguer deux types de protéines : des protéines solubles dans l'eau ou dans des solutions salines qui sont dites fonctionnelles ou métaboliques (les albumines et les globulines) et qui correspondent à 15-20% des protéines des protéines du grain, et des protéines insolubles dans l'eau ou dans des solutions salines qui constituent les protéines de réserve (les gliadines et les gluténines) et qui correspondent à 80-85% des protéines du grain. La classification d'Osborne a été complétée par Shewry et al. (1986) en distinguant les protéines de réserve en fonction de leurs caractéristiques biologiques, chimiques et génétiques. Les gliadines sont des protéines monomériques et sont classées en α-, β-, δ-, γ- et ω-gliadines selon leur séparation par électrophorèse à faible pH (Shewry et al. 2003). Les gènes codant pour les gliadines sont

situés sur le bras court des chromosomes 1 et 6 des 3 sous-génomes du blé tendre (Wrigley and Shepherd 1973). Les gluténines sont quant à elles des protéines polymériques et sont constituées de deux types de sous-unités : des sous-unités de haut poids moléculaire et des sous-unités de faible poids moléculaire. Les sous-unités de haut poids moléculaire sont principalement codées par des gènes situés sur les bras longs des chromosomes homéologues du groupe 1 aux loci *Glu-A1*, *Glu-B1* et *Glu-D1*. En revanche, les sous-unités de faible poids moléculaire sont principalement codées par des gènes situés sur les bras courts des chromosomes homéologues du groupe 1 (Payne et al. 1987).

Les protéines de réserve du blé tendre sont les principaux composants du gluten, un réseau qui se forme lorsque de l'eau est ajoutée à la farine. Le gluten détermine en grande partie les propriétés viscoélastiques des pâtes alimentaires obtenues à partir de farine et d'eau. Les gliadines ont un impact sur la viscosité et l'extensibilité de la pâte tandis que les sous-unités de haut poids moléculaires des gluténines influencent l'élasticité et la ténacité de la pâte (MacRitchie 1999; Shewry et al. 2002). La qualité de la pâte résulte de l'équilibre entre ces différentes propriétés et donc du ratio entre la teneur en gluténines et la teneur en gliadines. En revanche, les albumines et les globulines ont peu d'effet sur les caractéristiques rhéologiques de la pâte. Elles contribuent cependant à la qualité nutritionnelle du blé du fait de leur composition en acides aminés essentiels, et particulier du fait de leur richesse en lysine (Feillet 2000). Ainsi, les protéines contenues dans le grain du blé tendre représentent une importante ressource nutritionnelle et conditionnent les propriétés physico-chimiques de la pâte ce qui a un impact sur les usages industriels du blé tendre.

2. Importance du blé tendre en France et dans le monde

a. Utilisations du blé tendre

Le blé tendre est l'une des céréales les plus consommées par l'Homme et représente 20% des apports caloriques dans l'alimentation humaine (Shewry and Hey 2015). D'un point de vue nutritionnel, le blé tendre est une source de glucides et de protéines. La majeure partie de la production de blé tendre destinée à l'alimentation humaine est transformée en farine qui est ensuite utilisée pour la fabrication de divers produits comme le pain, les biscuits et les gâteaux. La farine est généralement obtenue après avoir retiré l'enveloppe du caryopse mais peut également être obtenue à partir de grains complets, on parle alors de farine de blé complet qui constitue une source de fibres. La qualité du grain de blé tendre est en partie déterminée par sa qualité nutritionnelle et sanitaire. Elle est également déterminée par sa qualité technologique qui garantit la transformation de la farine en pain ou autres produits de boulangerie de

qualité satisfaisante. Etant donné que les utilisations de la farine varient en fonction des pays, les critères de qualité technologique du grain ne sont pas les mêmes dans tous les pays. En France, les variétés de blé tendre proposées à l'inscription au catalogue officiel sont évaluées pour leur aptitude à la panification dite française utilisant seulement de la farine, de l'eau et du sel, sans autre additif (sauf parfois de l'acide ascorbique, un antioxydant). L'évaluation de la qualité boulangère des nouvelles variétés permet de les positionner dans l'une des classes technologiques suivantes : blé améliorant ou de force (BAF), blé panifiable supérieur (BPS), blé panifiable (BP), blé biscuitier (BB), blé autres usages (BAU) et blé autres usages impanifiable (BAU Imp). La note de panification est obtenue grâce à des tests définis par la méthode normalisée NF V03-716, aussi appelée méthode BIPEA. La note de panification est un caractère complexe qui synthétise plusieurs notations traduisant la qualité boulangère à chaque étape de la fabrication du pain. Afin de réaliser ces tests, plusieurs kilogrammes de farine sont nécessaires. Des méthodes indirectes et moins coûteuses sont aussi utilisées en sélection pour caractériser la qualité fermentaire des pâtes et leurs caractéristiques rhéologiques, notamment en utilisant l'alvéographe de Chopin. Cet appareil étudie le comportement d'un disque de pâte formé à partir de farine et d'eau lors de sa déformation sous l'effet d'un déplacement d'air à débit constant (Figure 3.a). La mesure de la pression à l'intérieur de la bulle de pâte et le gonflement de la bulle au cours du temps permettent d'évaluer plusieurs caractéristiques de la pâte (Figure 3.b) : la ténacité de la pâte (notée P), l'extensibilité de la pâte (notée L), la résistance de la pâte aussi appelée force boulangère (notée W), et l'équilibre entre la ténacité et l'extensibilité de la pâte (correspond au rapport P/L).

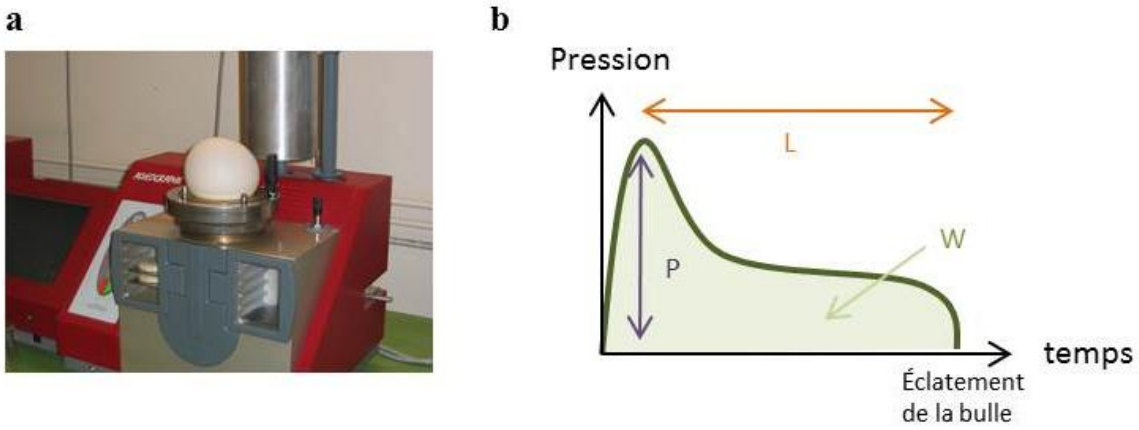

Figure 3 : Alvéographe de Chopin

a. Photo de l'appareil avec une pâte gonflée sous l'effet de la pression de l'air (Moulins Dumée).
b. Courbe représentant l'évolution de la pression à l'intérieur de la bulle de pâte au cours du temps et permettant d'évaluer les caractéristiques rhéologiques de la pâte telles que la ténacité (P), l'extensibilité (L) ou la force boulangère (W)

Par ailleurs, une partie du blé tendre utilisé dans le domaine agro-alimentaire est destiné au secteur de l'amidonnerie afin produire de l'amidon qui sert d'additif alimentaires dans de nombreux produits. L'amidon est aussi utilisé dans la fabrication d'autres produits tels que du papier, des produits cosmétiques, des produits issus de l'industrie pharmaceutique. Il permet également la production de biocarburants. Le co-produit gluten est également valorisé comme additif alimentaire, en alimentation animale ou encore en usage industriel.

L'utilisation du blé tendre ne se limite pas à l'alimentation humaine et au secteur de l'amidonnerie. En effet, le blé tendre est aussi largement utilisé pour l'alimentation animale. Dans ce cas, il peut être distribué en l'état ou bien il peut être incorporé dans des aliments composés. De plus, la paille de blé tendre (c'est-à-dire la tige lignifiée de la plante récoltée à maturité) peut servir de litière pour les élevages. La paille peut également être utilisée pour l'alimentation du bétail bien que cela soit plus rare ou encore être utilisée pour produire du méthane dans des fermenteurs, souvent à la ferme.

b. Production du blé tendre à l'échelle internationale et à l'échelle nationale

La domestication du blé tendre a eu lieu il y a 10 000 ans dans la région dite du Croissant Fertile, dans des territoires actuels de la Syrie et de la Turquie (Lev-Yadun et al. 2000). De nos jours, le blé tendre est cultivé dans de nombreuses régions du monde situées à des latitudes très variées allant du 67ème parallèle nord en Scandinavie au 45ème parallèle sud en Argentine (Feldman 1995; Shewry 2009). Actuellement, environ 95% du blé cultivé sur la planète correspond à du blé tendre, et la majeure partie des 5% restants correspond à du blé dur (Shewry 2009). Bien que le blé soit la céréale cultivée sur la plus grande surface à l'échelle mondiale, il s'agit de la troisième céréale la plus produite au monde après le maïs (*Zea mays*) et le riz (*Oryza sativa*). La production mondiale annuelle de blé s'élève à 742 millions de tonnes en moyenne (valeur calculée sur une période de cinq ans allant de 2013 à 2017) d'après l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO, 2019). Les cinq principaux pays producteurs de blé sont la Chine, l'Inde, la Russie, les Etats-Unis et la France (Figure 4), et ils produisent à eux seuls environ la moitié de la production mondiale de blé chaque année (FAO, 2019).

Etant donné que la France est le 5ème producteur de blé à l'échelle mondiale et le 1er producteur de blé de l'Union européenne (Figure 4), le blé tendre est une espèce d'intérêt économique majeur au niveau national.

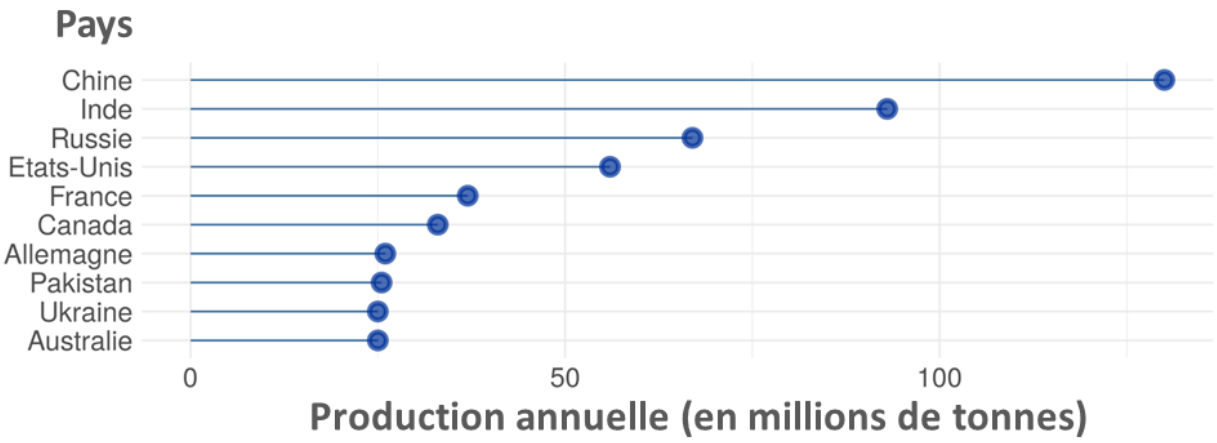

Figure 4 : Production annuelle moyenne de blé dans les dix principaux pays producteurs au niveau mondial (FAO, 2019).

Les moyennes ont été calculées en considérant les données de 2013 à 2017, soit sur une période de cinq ans.

En France métropolitaine, la culture du blé tendre est présente dans la quasi-totalité des départements. Cependant, c'est dans la moitié nord du pays qu'est obtenue la majeure partie de la production française de blé tendre (Agreste). Entre 1950 et 1990, les rendements du blé tendre en France ont connu un essor important. En effet, l'augmentation annuelle moyenne durant cette période était d'environ 1.2 quintaux par hectare et par an. En revanche, depuis les années 90 les rendements de blé tendre en France ainsi que dans d'autres pays d'Europe n'augmentent plus ou peu (FAO, 2019).

c. Enjeux actuels liés à la production du blé tendre

D'après un rapport de l'Organisation des Nations Unies de 2017, la population mondiale pourrait atteindre 9.8 milliards d'habitants en 2050 et 11.2 milliards en 2100. Cette croissance démographique et l'évolution des modes de vie accompagnant l'urbanisation vont de pair avec une augmentation de la demande en denrées alimentaires. Ainsi, la production céréalière, et plus généralement la production agricole mondiale, devra être plus performante durant les prochaines années pour satisfaire la demande croissante. L'augmentation de la production de blé tendre est un défi majeur qui doit être réalisé dans le contexte actuel de changement climatique. Or, le changement climatique est associé à une variabilité accrue du climat et à l'existence de phénomènes météorologiques plus extrêmes et plus fréquents, ce qui affecte l'agriculture. La Figure 5 illustre le réchauffement climatique mesuré à Clermont-Ferrand. Brisson el al. (2010) ont montré que le changement climatique était l'un des principaux facteurs responsables de la stagnation des rendements de blé tendre observée en France depuis les années 90.

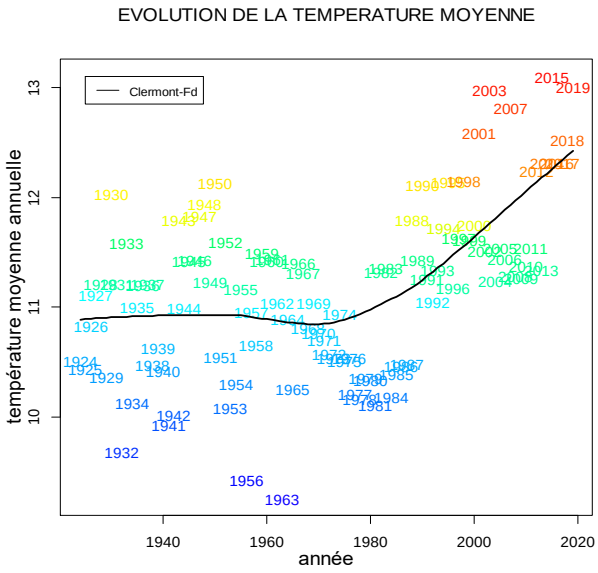

Figure 5 : Evolution de la température moyenne annuelle à Clermont-Ferrand entre 1924 et 2019.

De plus, il est nécessaire de réduire l'impact environnemental de la production des céréales. Afin de répondre à cet enjeu, il faut diminuer l'utilisation des pesticides et des fertilisants azotés de synthèse qui peuvent avoir des effets néfastes sur les écosystèmes tout en maintenant des rendements élevés et une production de bonne qualité.

Le développement d'une production de blé tendre économe en azote et avec un rendement élevé malgré des conditions climatiques variables passe par des innovations agronomiques (notamment en améliorant les systèmes de culture et en utilisant des outils d'aide à la décision) et par un progrès génétique, notamment pour une meilleure efficience de l'azote.

3. Amélioration génétique du blé tendre

a. Ressources génétiques du blé tendre

L'amélioration génétique du blé tendre a pour objectif de créer de nouvelles variétés plus performantes ou plus adaptées à certains types environnements que les variétés déjà existantes. La création variétale traditionnelle du blé tendre consiste à croiser deux plantes présentant des caractères d'intérêts agronomiques complémentaires. Cette méthode de sélection utilise ainsi la diversité génétique existante. Le progrès génétique est fonction de l'intensité et de la précision de la sélection, de l'intervalle de temps

entre les générations ainsi que de la variabilité génétique (De Rochambeau 1992). Afin d'assurer un progrès génétique sur le long terme, il est donc important de ne pas épuiser la variabilité génétique de l'espèce.

L'enjeu de la gestion des ressources génétiques est de garantir la disponibilité de la diversité génétique d'une espèce cultivée pour maintenir le progrès génétique et pour produire des variétés adaptées à des conditions climatiques variables. De plus, les ressources génétiques peuvent être analysées dans des programmes de recherche afin d'améliorer les connaissances sur l'espèce étudiée. En France, la gestion des ressources génétiques du blé tendre est principalement assurée par le centre de ressources biologiques (CRB) des céréales à pailles de Clermont-Ferrand (https://www6.clermont.inrae.fr/umr1095/Organisation/Infrastructures-experimentales/Centre-de-Ressources-Biologiques). Ce CRB regroupe les céréales à pailles d'intérêt agronomique majeur, dont fait partie le blé tendre, et leurs espèces apparentées sauvages. En ce qui concerne le blé tendre, environ 12 000 accessions sont conservées au CRB de Clermont-Ferrand. Cette collection rassemble des blés français (des variétés populations, des lignées de sélections et des lignées élites fixées) ainsi que des variétés et des lignées issues d'une soixantaine de pays étrangers. Certaines de ces accessions sont des lignées porteuses de caractères particuliers tels que des résistances aux maladies. La collection du CRB de Clermont-Ferrand a entre autre permis de constituer une core collection de 372 accessions de blé tendre représentatives de la diversité existante (Balfourier et al. 2007). Cette core collection, nommée 372CC, a fait l'objet de divers projets de recherche (dont Bordes et al. 2011; Rousset et al. 2011; Le Gouis et al. 2012; Bordes et al. 2013; Wilhelm et al. 2013; Bogard et al. 2014).

Par ailleurs, en France et dans l'Union européenne, une variété végétale ne peut pas être protégée par un brevet. Il existe cependant un titre de protection appelé certificat d'obtention végétale (COV) qui est délivré pour toute nouvelle variété inscrite au catalogue. Le COV permet de rétribuer le travail de l'obtenteur tout en laissant la ressource génétique libre d'accès aux autres personnes à des fins de recherche ou de sélection. Les variétés inscrites au catalogue constituent donc une réserve de ressources génétiques utilisables par les sélectionneurs pour produire de nouvelles variétés. Ce système de protection avec le COV assure la continuité de l'amélioration génétique chez les espèces végétales.

b. Objectifs de sélection chez le blé tendre

Pour être commercialisées en France, les nouvelles variétés de blé tendre doivent être inscrites sur le catalogue officiel français. Avant leur inscription, les variétés doivent passer différents tests qui sont réalisés par le secteur d'étude des variétés du groupement d'étude et de contrôle des variétés et des semences (GEVES) dans plusieurs essais et sur une période durant en moyenne 2 ans (Abecassis et Bergez 2009). Le premier test, appelé test DHS (Distinction, Homogénéité, Stabilité), évalue si la variété est distincte des variétés déjà inscrites dans l'Union européenne, si elle est homogène entre les individus de la variété et si elle est stable dans le temps. Chez le blé tendre, la majorité des variétés sont des lignées « pures » (homozygotes), obtenues par autofécondations ou haplo-diploïdisation. Les critères d'homogénéité pour les variétés lignées sont assez rigoureux (la variété est dite homogène si moins de 0.4% des plantes correspondent à des « hors type »). Il existe également des variétés hybrides pour lesquelles les critères d'homogénéité sont moins sévères. La seconde catégorie d'épreuves évalue la valeur agronomique, technologique et environnementale (VATE) des variétés. Seules les variétés apportant une amélioration agronomique et/ou technologique et qui sont adaptées à des conditions de culture utilisant peu de fongicides peuvent être inscrites au catalogue. L'inscription d'une variété est décidée par le Ministère de l'Agriculture après avis du Comité Technique Permanent de la Sélection des Plantes Cultivées (CTPS) sur la base des synthèses des études DHS et VATE présentées par le GEVES. Les critères évalués lors des études VATE sont définis par le CTPS de façon à obtenir la meilleure adéquation entre les objectifs des utilisateurs des variétés et les capacités scientifiques et techniques des sélectionneurs. Ces critères orientent les objectifs de sélection des variétés françaises de blé tendre. Les quatre catégories de critères sont les suivantes : le rendement, la valeur technologique, les caractéristiques physiologiques et les résistances aux bio-agresseurs (Table 1). Selon la classe technologique à laquelle appartient la variété évaluée, le seuil de rendement exprimé en pourcentage par rapport aux variétés témoins n'est pas le même. En effet, une variété de blé BPS devra avoir un rendement supérieur à 102% du rendement moyen des témoins alors qu'une variété de blé BAU Imp devra quant à elle avoir un rendement supérieur à 109% du rendement moyen des témoins. D'autres critères tels que les résistances aux maladies, au froid ou à la verse entraînement l'obtention de bonus ou de malus sur la note finale de la variété. Une fois inscrites au catalogue, les variétés sont encore évaluées par ARVALIS, Institut du végétal. L'objectif de l'évaluation en post-inscription est de caractériser ces variétés dans un plus grand nombre d'environnements afin de définir leur adaptation régionale et leur robustesse vis-à-vis des aléas climatiques.

Rendement	Valeur technologique	Caractéristiques physiologiques	Résistances aux bio-agresseurs
- Rendement dans les essais traités et non traités (sans fongicides)	- Teneur en protéines - Poids Spécifique - Paramètres de l'alvéographe de Chopin - Dureté - Indice de Hagberg - Valeur en panification française - Viscosité pour alimentation animale (volailles) - Test biscuitier pour les blés biscuitiers	- *Grain Protein Deviation* (GPD) : écart à la régression négative existant entre le rendement et la teneur en protéines - Alternativité - Précocité d'épiaison et à montaison - Hauteur de plante à maturité - Résistance à la verse, au froid et à la germination sur pied	- Rouille jaune - Rouille Brune - Piétin-verse - Oïdium - Septorioses - Fusarioses - Tolérance globale aux maladies (écart de rendement en conditions traitées et non-traitées) - Mosaïques - Cécidomyie orange

Table 1 : Principaux caractères évalués en vue de l'inscription d'une variété au catalogue officiel français (GEVES, 2017)

Des corrélations positives ou négatives existent entre certains caractères à sélectionner. C'est notamment le cas de la teneur en protéines qui est négativement corrélée au rendement (Simmonds 1995). Les corrélations négatives existant entre les différents objectifs de sélection constituent un enjeu pour les sélectionneurs. Des indices de sélection prenant en compte différents caractères ont été développés dans des études théoriques. Par exemple l'indice de sélection proposé par Smith (1936) et Hazel (1943) peut être utilisé quand l'objectif du sélectionneur consiste à améliorer plusieurs caractères simultanément de façon optimale. Par la suite, Kempthorne et Nordskog (1959) ont développé un autre indice de sélection pour les situations où le sélectionneur cherche à améliorer un caractère sans dégrader les performances des autres caractères. Dans les indices de sélection, un poids peut être attribué à chacun des caractères en fonction de leur importance économique.

c. Schéma de sélection du blé tendre

L'introduction de la sélection généalogique par Vilmorin en 1856 et la découverte des lois de l'hérédité par Mendel en 1866 ont permis d'améliorer les méthodes de sélection des plantes. Ces nouvelles méthodes ont conduit à une séparation du métier de sélectionneur de celui d'agriculteur (Gallais 2018). Le blé tendre est une espèce autogame, c'est pourquoi son amélioration repose historiquement sur la création de variétés lignées pures. L'objectif est d'obtenir à la fin du schéma de sélection des lignées homozygotes combinant un maximum d'allèles favorables au niveau des gènes contrôlant l'expression des caractères d'intérêt agronomique. Le schéma de sélection classique du blé tendre débute par des croisements de lignées présentant des caractères complémentaires. Il est par exemple possible de croiser une lignée ayant un rendement élevé mais une mauvaise qualité boulangère avec une autre lignée avec un rendement plus faible mais une meilleure qualité boulangère. La génération F1 est alors constituée d'hybrides, et les hybrides issus d'un même croisement sont tous identiques d'un point de vue génétique. Or la variabilité génétique est nécessaire pour qu'il y ait une réponse à la sélection. Puis, la génération F2 est obtenue après autofécondation des hybrides F1, ce qui permet de générer une variabilité génétique (Gallais 2015). La sélection est réalisée parmi les descendances au cours des générations successives d'autofécondation. Ainsi, le nombre de génotypes à évaluer à chaque étape décroit. Lors des premières générations, la sélection ne se fait que sur les caractères les plus héritables. Le rendement n'est évalué qu'à un stade plus tardif du schéma pour lequel le niveau de fixation est suffisant et le nombre de génotypes est restreint. L'évaluation du rendement est réalisée dans plusieurs essais ce qui permet d'évaluer la stabilité des performances de chaque génotype. Les caractères coûteux à phénotyper tels que la note de panification ne sont évalués que pour les individus présents dans les dernières générations du schéma de sélection. Les lignées obtenues à la fin du schéma de sélection peuvent être utilisées pour des croisements lors des cycles de sélection suivants. Le schéma de sélection du blé tendre est un processus long qui s'étend sur huit à dix ans. Des méthodes ont été développées pour réduire la durée du schéma de sélection. La production d'haploïdes doublés permet par exemple d'obtenir des lignées fixées rapidement en provoquant le doublement du stock chromosomique de plantes haploïdes issues de gamétogénèse in vivo ou in vitro (Snape 1989; Raina 1997; Tadesse et al. 2012).

Afin de répondre aux enjeux actuels auxquels est confrontée la production de blé tendre, il est nécessaire de développer des variétés plus performantes, plus adaptées à des systèmes économes en intrants, et pouvant être cultivée dans des environnements aux conditions climatiques variables. De plus, le progrès génétique doit être obtenu le plus rapidement possible. L'accélération du progrès génétique peut être réalisée en améliorant les schémas de sélection grâce à l'optimisation des designs expérimentaux et de l'allocation des ressources, ainsi qu'à l'utilisation de méthodes issues du développement des biotechnologies et de la génomique.

II. L'apport des outils de prédiction génomique pour l'amélioration du blé tendre

1. Avancées génomiques récentes

a. Développement des technologies de séquençage

Le séquençage de l'ADN consiste à déterminer la succession des nucléotides pour un fragment d'ADN donné. Les deux premières méthodes de séquençage ont été développées indépendamment à la fin des années 1970, environ 20 ans après que Watson et Crick (1953) aient mis en évidence la structure en double hélice de l'ADN. L'approche de Maxam et Gilbert (1977) est une méthode par dégradation chimique de l'ADN. En revanche, l'approche de Sanger (1977), plus largement utilisée, repose sur la synthèse enzymatique sélective.

De nouvelles technologies sont progressivement apparues par la suite pour répondre à la demande croissante en séquençage de l'ADN de différentes espèces. Le séquençage nouvelle génération (*next-generation sequencing* ; NGS), également appelé séquençage à haut débit, permet d'obtenir des séquences d'ADN plus rapidement que les premières méthodes de séquençage.

Le développement de technologies de séquençage a permis de détecter différents types de polymorphismes au niveau des séquences d'ADN et ainsi de développer des marqueurs moléculaires. Ces polymorphismes peuvent correspondre à des polymorphismes au niveau des nucléotides (*Single Nucleotide Polymorpshim* ; SNP) ou à des zones d'insertions/délétions plus ou moins longues. Le génotypage utilisant de nombreux marqueurs SNP de façon simultanée a été rendu possible par le développement de puces à ADN. L'analyse des séquences d'ADN a également conduit à la détection de variants structuraux. Ces NGS ont remplacé d'anciennes méthodes d'analyse du polymorphisme basées sur la présence ou l'absence de sites de restrictions (*Restriction Fragment Length Polymorphism*; RFLP), sur différentes tailles des fragments amplifiés (*Amplified Fragment Length Polymorphsim* ; AFLP), ou sur différents nombres de motifs répétés appelés micro-satellites (*Single Sequence Repeat* ; SSR).

Le développement de ces méthodes de séquençage s'est accompagné d'une forte diminution des coûts de séquençage (Poland and Rife 2012). L'exemple de la réduction du coût de séquençage du génome humain est illustré sur la Figure 6.

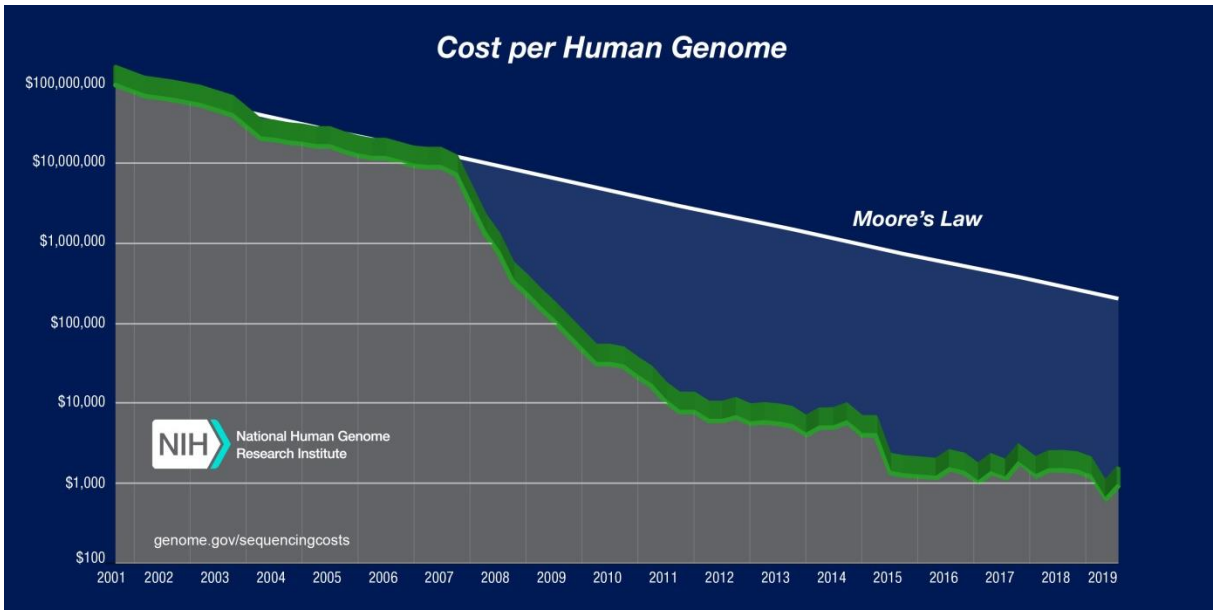

Figure 6 : Evolution du coût de séquençage du génome humain depuis 2001 (National Human Genome Research Institute, 2019)

La courbe verte correspond à l'évolution du coût de génotypage au cours du temps. La droite blanche correspond au coût de génotypage qui serait attendu si son évolution suivait la loi de Moore.

b. Avancées génomiques chez le blé tendre

Le génome du blé tendre est d'une complexité considérable. En effet, ce génome hexaploïde fait environ 5 fois celui de l'homme (génome haploïde du blé tendre : environ 16Gb; Arumuganathan and Earle 1991) et est composé à plus de 85% de séquences d'ADN répétées. Le consortium international de séquençage du génome du blé (*International Wheat Genome Sequencing Consortium*, IWGSC) a été créé en 2005 avec pour objectif de séquencer le génome du blé tendre. L'IWGSC regroupe des producteurs de blé, des sélectionneurs et des scientifiques issus des secteurs publics et privés provenant de 68 pays. En 2018, l'IWGSC a fourni une première version de la séquence de référence annotée du génome du blé (https://plants.ensembl.org/Triticum_aestivum/Info/Index). Actuellement, de nombreux SNP sont disponibles pour les génomes A, B et D du blé tendre (Rimbert et al. 2018).

c. Avancées génomiques au service de la sélection

Les marqueurs moléculaires peuvent être utilisés pour établir des cartes génétiques et détecter des régions du génome dont le polymorphisme est impliqué dans la variabilité du caractère étudié. C'est régions, appelées QTL (*Quantitative Trait Loci*), sont des zones du génome où sont localisés un ou plusieurs gènes à l'origine de la variabilité du caractère. Une fois les QTL détectés, il est possible d'utiliser des marqueurs moléculaires dans les QTL ou en déséquilibre de liaison avec les QTL pour sélectionner les plantes présentant des allèles favorables même si la mutation causale n'a pas été identifiée (Lande and Thompson 1990; Paterson et al. 1991). Cette méthode, appelée sélection assistée par marqueurs (SAM), vise à cumuler au sein d'un même individu les allèles favorables aux différents loci contrôlant la variation des caractères d'intérêt. La SAM est particulièrement intéressante lorsque le phénotypage du caractère d'intérêt est coûteux ou difficile à réaliser, lorsque celui-ci ne peut être obtenu qu'à un stade tardif du développement de la plante, ou encore lorsqu'il ne s'exprime qu'à l'état homozygote. L'apport de la SAM par rapport à la sélection conventionnelle d'un point de vue économique dépend également du coût de génotypage de l'héritabilité du caractère d'intérêt (Moreau et al. 2000). En effet, plus le coût de génotypage et l'héritabilité du caractère sont faibles, plus l'intérêt économique de la SAM est important. Cependant, une héritabilité très faible limite la capacité de détection des QTL.

Chez le blé tendre, la SAM a été largement étudiée et utilisée dans les programmes de sélection. Par exemple Eagles et al. (2001) ont étudié l'utilisation de la SAM dans l'amélioration du blé en Australie et ont montré que la SAM était utile pour cumuler des allèles favorables liés à la résistance à différentes maladies et pouvait également être utilisée pour sélectionner les allèles favorables aux loci *Glu* qui sont liés aux caractéristiques rhéologique de la pâte. D'autres études ont montré l'intérêt d'utiliser la SAM pour sélectionner des variétés de blé ayant des caractéristiques rhéologiques intéressantes (Ahmad 2000; Charmet et al. 2001; de Bustos et al. 2001) ou des résistantes à différentes maladies dont la fusariose (revue scientifique de Buerstmayr et al. 2009), la rouille (Bariana et al. 2007; Murphy et al. 2009) et l'oïdium (Liu et al. 2000).

Bien que l'intérêt de la SAM en amélioration des plantes ait été démontré pour plusieurs caractères, cette méthode rencontre des limites pour l'étude des caractères agronomiques hautement polygéniques tels que le rendement. De nombreux marqueurs moléculaires sont à présent disponibles et il est possible d'estimer de façon simultanée les effets de tous les marqueurs sur un caractère d'intérêt : on parle alors de prédictions génomiques.

2. Présentation de la sélection génomique

a. Principe général

L'objectif de la sélection génomique est d'augmenter le gain génétique par unité de temps et de coût en prédisant la valeur génétique (*Genomic Estimated Breeding Value* ; GEBV) de chacun des candidats à la sélection à partir d'un marquage moléculaire dense et sans que les candidats ne soient phénotypés, puis en sélectionnant les candidats ayant la meilleure valeur génétique. Par rapport aux méthodes utilisant les QTL pour calculer un score moléculaire (Lande et Thompson 1990), la principale nouveauté apportée par les prédictions génomiques est que cette méthode estime de façon simultanée les effets sur le caractère d'intérêt de tous les marqueurs répartis sur l'ensemble du génome (Whittaker et al. 2000; Meuwissen et al. 2001). Les prédictions génomiques reposent sur l'hypothèse que les caractères quantitatifs sont contrôlés par un grand nombre de QTL à faible effet (modèle infinitésimal de Fisher). Grâce au marquage haute-densité (densité supérieure à l'étendue du déséquilibre de liaison, marqueurs indépendants), tous les QTL sont en déséquilibre de liaison avec au moins un marqueur. La somme des effets de tous les marqueurs devrait alors expliquer toute la variance génétique additive du caractère étudié.

La Figure 7 illustre le principe de la sélection génomique (Heffner et al. 2009).

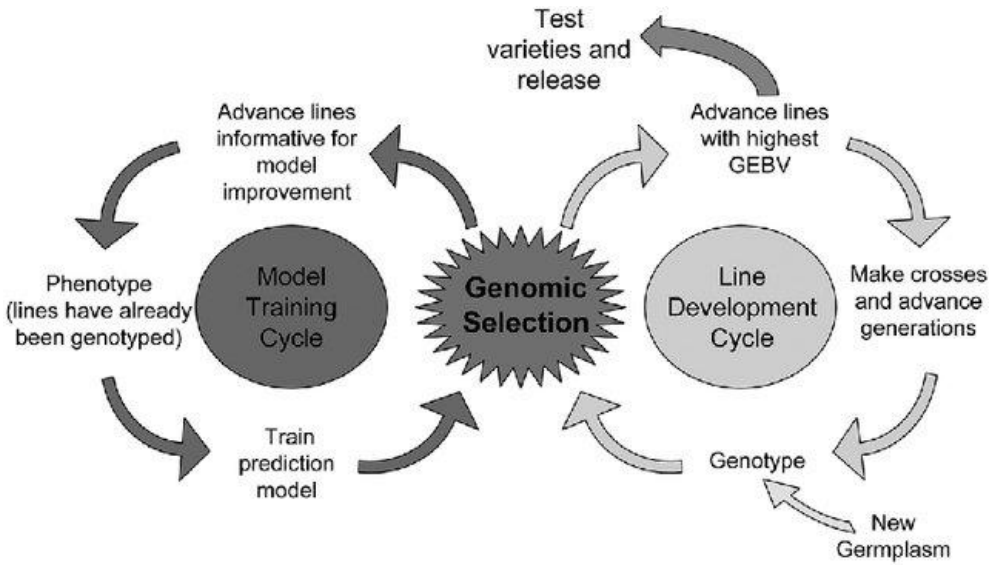

Figure 7 : Principe de base de la sélection génomique (Heffner et al. 2009)

GEBV : *Genomic estimated breeding value*

La calibration du modèle de prédiction génomique est réalisée à partir d'une population dite d'entraînement qui est à la fois génotypée et phénotypée pour le caractère d'intérêt. C'est donc grâce aux données de génotypage et de phénotypage de cette population que sont estimés simultanément les effets de tous les marqueurs sur la variabilité du caractère. Pour estimer la qualité prédictive du modèle, des méthodes de validation croisée sont généralement utilisées. La validation croisée consiste à diviser la population initiale en k groupes de taille égale. L'un des k échantillons sert alors de population de validation tandis que les k-1 autres échantillons permettent de calibrer le modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé une fois comme population de validation. La corrélation de Pearson entre les valeurs prédites et les valeurs phénotypiques des individus de la population de validation permet d'estimer la capacité prédictive du modèle (Figure 8). Une fois les effets de chacun des marqueurs déterminés, le modèle est appliqué à des individus qui ne sont pas phénotypés pour le caractère d'intérêt mais qui sont génotypés, ce qui permet de prédire leur performance. La sélection des meilleurs candidats est ensuite basée sur ces valeurs génétiques prédites (GEBV).

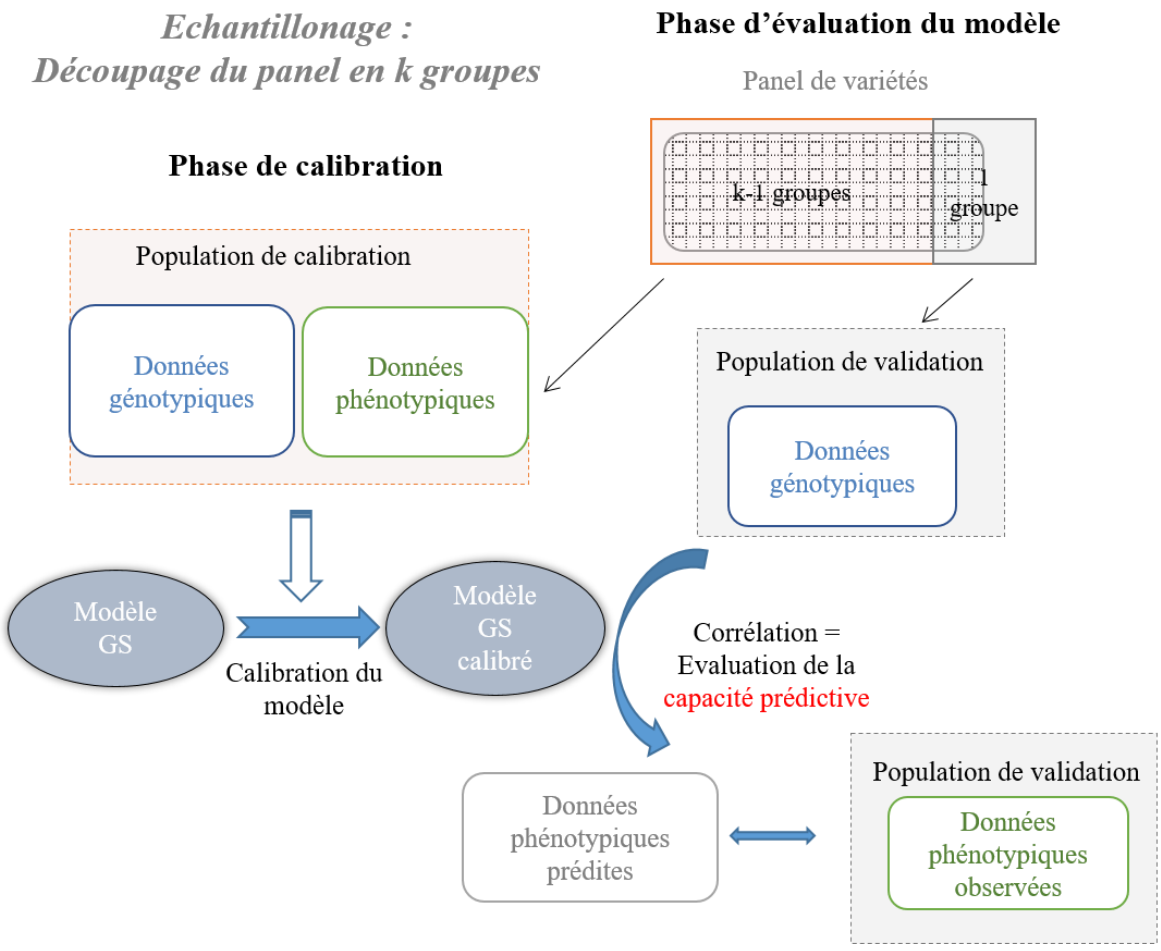

Figure 8: Estimation de la valeur prédictive des modèles de prédiction génomique par validation croisée.

GS : *Genomic selection*

L'estimation des effets de tous les marqueurs peut également être utilisée pour prédire les croisements à faire lors de la première étape des programmes de sélection chez les plantes (Akdemir and Sanchez 2016; Lado et al. 2017; Allier et al. 2019).

b. Modèles de prédiction génomique

La principale difficulté statistique posée par les prédictions génomiques est que le nombre d'effets à estimer est beaucoup plus grand que la taille de la population de référence, aussi le modèle linéaire classique (à effets fixes) ne peut pas être utilisé. Depuis l'introduction des premiers modèles de prédiction génomique (Whittaker et al. 2000; Meuwissen et al. 2001), de nombreux modèles ont été développés (Figure 9).

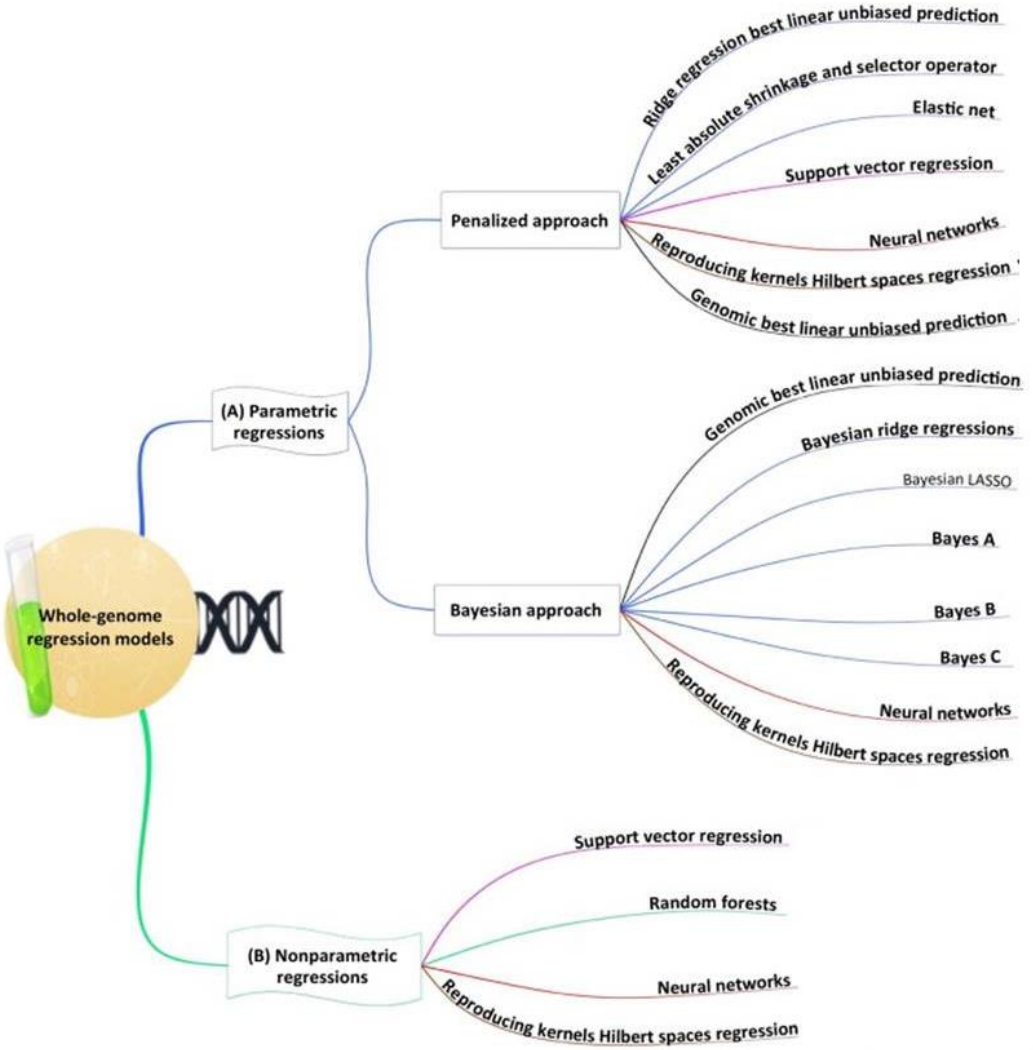

Figure 9 : Diversité des modèles de prédiction génomique (d'après Desta and Ortiz 2014)

La distribution des effets des marqueurs varie en fonction du modèle étudié. Aucun des modèles ne surpasse tous les autres modèles en terme de qualité de prédiction pour toutes les espèces végétales et pour tous les caractères étudiés (Heslot et al. 2012; Pérez and de Los Campos 2014). Ainsi le choix du modèle dépend du caractère à prédire.

Le modèle de prédiction génomique le plus communément utilisé est le *Genomic Best Linear Unbiased Prediction* (GBLUP) (Habier et al. 2013). Ce modèle prédit les effets génétiques des individus en utilisant une matrice d'apparentement, appelée matrice de *Kinship*, calculée à partir des différents génotypes aux marqueurs moléculaires (VanRaden 2008; Endelman and Jannink 2012). Le modèle GBLUP mono-caractère peut être écrit de la façon suivante:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \boldsymbol{\varepsilon} \qquad (1)$$

Avec **y** le vecteur des valeurs phénotypiques de longueur n (avec n le nombre d'individus) ; **β** le vecteur des effets fixes (environnement, année, lieu, groupe génétique, autre phénotype…) ; **X** la matrice de design des effets fixes ; **Z** la matrice de design des effets aléatoires ; **a** le vecteur des effets aléatoires des lignées de longueur n tel que $\mathbf{a} \sim N(0, \mathbf{K}\sigma^2_a)$ avec **K** la matrice de *Kinship* de dimension n x n et σ^2_a la variance génétique additive; et **ε** l'erreur résiduelle telle que $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_\varepsilon)$ avec σ^2_ε la variance résiduelle.

Les modèles de prédiction génomique mono-caractère et intégrant uniquement des effets génétiques additifs ne sont pas les seuls à avoir été développés. En effet, des modèles prenant en compte des effets de dominance (par exemple pour prédire la performance de blés hybrides : Zhao et al. 2013), d'épistasie (Jiang and Reif 2015) et d'interaction génotype x environnement (sur le blé : Burgueño et al. 2012; Ly et al. 2018) ont également été proposés. Des modèles multi-caractères ont aussi été développés (voir partie III).

c. Autres facteurs influençant la performance de la sélection génomique

Le choix du type de modèle utilisé pour réaliser les prédictions génomiques n'est pas le seul facteur ayant un impact sur la précision des prédictions. En effet, l'héritabilité du caractère et l'architecture génétique du caractère étudié, la densité et le type de marqueurs utilisés pour le génotypage, le déséquilibre de liaison entre les marqueurs et les QTL, la taille et de la population d'entraînement ainsi que son degré d'apparentement à la population prédite sont des paramètres qui ont une influence sur la qualité des prédictions. Par exemple, il existe une corrélation positive entre l'héritabilité du caractère et la précision des prédictions (Daetwyler et al. 2008). Par ailleurs, plusieurs études portant sur l'effet de

la densité de marquage moléculaire sur la précision des prédictions génomiques ont montré que la précision augmentait avec le nombre de marqueurs avant d'atteindre un plateau (Calus and Veerkamp 2007; Solberg et al. 2008; de Roos et al. 2009). La taille et la composition de la population d'entraînement sont également des facteurs clés (Desta and Ortiz 2014). En effet, la précision des prédictions génomiques augmente avec la taille de la population d'entraînement (Jannink et al. 2010 ; Lorenz et al. 2011). Plusieurs études ont également montré l'importance de l'apparentement entre la population d'entraînement et la population d'individus prédits pour obtenir des prédictions précises (Habier et al. 2007; Habier et al. 2010; Charmet et al. 2014; Crossa et al. 2014). Cela est dû au fait que plus la population d'entraînement et celle regroupant les individus prédits sont apparentées, plus les déséquilibres de liaison entre marqueurs et QTL observés dans chacune des deux populations sont similaires. De plus, des méthodes basées sur la minimisation de la variance d'erreur de prédiction moyenne (PEVmean) ou sur la maximisation de la moyenne du coefficient de détermination généralisé (CDmean) ont été développées pour optimiser la composition de la population d'entraînement (Rincent et al. 2012; Akdemir et al. 2015; Isidro et al. 2015; Sarinelli et al. 2019). Ces méthodes sont particulièrement intéressantes dans un cas où le budget est limité car elles permettent de réduire la taille de la population d'entraînement tout en limitant la dégradation de la précision des prédictions. Le CD généralisé correspond à la fiabilité attendue des contrastes entre les valeurs génétiques des individus non phénotypés et la moyenne de la population :

$$\mathbf{CD}(\mathbf{c}) = \text{diag}\left[\frac{\boldsymbol{c}'(\boldsymbol{A}-\boldsymbol{\lambda}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}+\ \boldsymbol{\lambda}\boldsymbol{A}^{-1})^{-1})\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}}\right] \qquad (2)$$

Avec **c** un contraste, $\boldsymbol{\lambda}$ le ratio entre la variance résiduelle et la variance additive, **A** la matrice de *Kinship*, **Z** la matrice de design, et $\mathbf{M = I - X(X'X)^{-1}X'}$ avec **I** la matrice identité et **X** la matrice de design (Laloë 1993).

L'approche basée sur le CDmean présente l'avantage de prendre en compte la variance génétique des contrastes entre individus et donc de limiter l'échantillonnage d'individus avec un fort degré d'apparentement. L'optimisation de la composition de la population d'entraînement en utilisant le critère CDmean nécessite de disposer du génotypage des lignées de la population totale et de connaître les variances résiduelles et additives du caractère à prédire.

L'avantage de la sélection génomique par rapport à la sélection phénotypique classique ne dépend pas uniquement de la précision des prédictions. En effet, cet avantage se mesure également en termes de durée des cycles de sélection : Heffner et al. (2010) ont montré que la sélection génomique permet de réduire la durée du schéma de sélection du maïs et ainsi accélérer le gain génétique. Les coûts de génotypage et de phénotypage sont également d'importants paramètres à prendre en compte lorsqu'il s'agit de comparer les deux méthodes de sélection (Riedelsheimer and Melchinger 2013).

3. Mise en pratique de la sélection génomique dans les programmes d'amélioration

a. Premiers résultats d'application de prédictions génomiques en sélection

Dans un premier temps, la sélection génomique a été mise en place dans des programmes d'amélioration des bovins laitiers (Schaeffer 2006; Hayes et al. 2009). L'introduction de la sélection génomique dans les programmes de sélection des bovins laitiers a généré un doublement du progrès génétique annuel (Schaeffer 2006; Garcia Ruiz et al. 2016). L'intérêt des prédictions génomiques pour les bovins laitiers réside dans la possibilité d'évaluer la performance des mâles dès leur plus jeune âge et ainsi économiser le coût et le temps du testage sur descendance. En effet, avec le testage sur descendance la sélection des taureaux n'est réalisée que lorsqu'ils ont en moyenne 7 ans, c'est-à-dire lorsqu'ils ont pu avoir suffisamment de filles qui produisent du lait et qui peuvent ainsi être évaluées. En revanche, les prédictions génomiques peuvent être effectuées à un stade plus précoce et les taureaux sélectionnés en fonction de leur valeur génétique prédite peuvent alors être utilisés pour l'insémination artificielle dès qu'ils sont en âge de procréer.

L'application de la sélection génomique s'étend aujourd'hui à d'autres filières animales telles que les élevages porcins et équins, ceux des autres ruminants, des volailles ou des poissons (Jussiau et al. 2013). La sélection génomique a également été mise en place dans des programmes d'amélioration d'espèces végétales, que ce soit pour des espèces pérennes (palmier à huile, pin maritime, peuplier, etc.) ou pour des espèces annuelles comme le blé tendre (Lin et al. 2014).

b. Utilisation des prédictions génomiques pour l'amélioration du blé tendre

Les premières études qui ont montré l'intérêt des prédictions génomiques chez le blé tendre ont été menées au début des années 2010 (Crossa et al. 2010; Heffner et al. 2011). Elles ont montré que cette méthode de sélection pouvait conduire à une réduction de la durée des cycles de sélection ce qui permet d'accélérer le progrès génétique, tout en réduisant le coût du schéma de sélection.

D'autres études ont ensuite été menées dans le but d'évaluer la capacité des modèles de prédiction génomique mono-caractère pour prédire différents caractères d'intérêt agronomique chez le blé tendre. Ainsi, de nombreuses études ont cherché à prédire le rendement en grain chez le blé tendre ou des composantes du rendement (dont Poland et al. 2012; Lado et al. 2013; Storlie and Charmet 2013; Zhao

et al. 2015; Norman et al. 2017), mais aussi la qualité boulangère ou des paramètres associés (dont Battenfield et al. 2016 ; Guzman et al. 2016 ; Liu et al. 2016), ou les résistances aux maladies (notamment Ornella et al. 2012; Rutkoski et al. 2012; Daetwyler et al. 2014; Rutkoski et al. 2014; Arruda et al. 2015).

Bien que de nombreuses études se soient intéressées aux prédictions génomiques chez le blé tendre, le nombre d'articles scientifiques portant sur la mise en pratique de la sélection génomique dans les schémas de sélection du blé tendre est quant à lui plus restreint. La plupart de ces articles présentent ainsi la qualité des prédictions obtenues mais ne s'interroge pas sur la faisabilité de telles approches ni au coût des schémas de sélection intégrant des prédictions génomiques. Bassi et al. (2016) ont comparé différents schémas de sélection de blé utilisant des prédictions génomiques à différentes générations du schéma. Utiliser des prédictions génomiques à toutes les générations du schéma de sélection (y compris les plus précoces) permet de réduire la durée du cycle, mais nécessite de génotyper un nombre élevé de plantes surtout dans les premières générations. De plus, dans les programmes de sélection français, il est difficile d'imaginer supprimer l'évaluation dans un réseau de plusieurs essais des plantes sélectionnées en dernière année du programme de sélection avant de les soumettre aux essais DHS et VATE (tests coûteux) en vue de leur inscription au catalogue officiel. A l'inverse, n'utiliser les prédictions génomiques qu'à un stade tardif du schéma de sélection permet de limiter le budget alloué au génotypage mais le progrès génétique annuel réalisé est plus faible qu'avec un schéma de sélection uniquement basé sur les prédictions génomiques. Il est également possible de combiner la sélection assistée par marqueurs pour éliminer les plantes ne présentant pas les allèles favorables aux QTL majeurs lors des premières générations du schéma de sélection, puis d'utiliser des prédictions génomiques aux générations suivantes (Heffner et al. 2010).

Le choix du plan de croisements a impact important sur le gain génétique potentiel du programme de sélection car c'est de cette décision que dépendent la moyenne de la valeur génétique ainsi que la variance génétique observées dans la descendance. La valeur des croisements peut être prédite en estimant les effets des marqueurs (Akdemir et Sanchez 2016; Lado et al. 2017; Allier et al. 2019). Par exemple, Lado et al. (2017) ont appliqué la prédiction des croisements avec des modèles de prédiction génomique dans des programmes de sélection de blé tendre.

Actuellement, la plupart des travaux sur les prédictions génomiques dans le domaine végétal portent sur la prédiction d'un seul caractère à la fois. Or de nombreux caractères, potentiellement corrélés, sont mesurés dans les programmes de sélection et les modèles de prédiction pourraient bénéficier de l'information contenue dans les caractères corrélés, soit pour améliorer la précision de prédiction, soit diminuer le coût de la sélection.

III. Prédictions génomiques multi-caractères

1. Diversité des modèles de prédiction génomique multi-caractère

a. Prédictions génomiques avec des caractères corrélés

La performance globale de nouvelles variétés de plantes ou d'animaux dépend de plusieurs caractères qui peuvent être corrélés. Il est possible de prédire chaque caractère indépendamment en utilisant des modèles de prédiction génomique mono-caractère, mais ces analyses univariées n'exploitent pas l'information contenue dans la potentielle corrélation entre les caractères à prédire.

Henderson et Quaas (1976) ont introduit les BLUP multivariés pour analyser simultanément plusieurs caractères. Contrairement aux modèles de prédiction génomique mono-caractères, les modèles de prédiction génomique multi-caractère ont été spécialement conçus pour bénéficier à la fois de l'information contenue dans la corrélation génétique entre les caractères et de celle apportée par l'apparentement entre les individus (Calus and Veerkamp 2011).

Les modèles de prédiction génomique multi-caractère sont particulièrement intéressants en amélioration végétale et animale dans deux cas :

1) Lorsque la performance d'un individu dépend de chacun des caractères corrélés,

2) Lorsque le caractère d'intérêt est difficile ou cher à phénotyper, ou lorsqu'il ne peut être évalué qu'à un stade tardif, mais qu'il est plus simple ou moins coûteux d'évaluer un ou plusieurs autres caractères corrélés au caractère cible.

Dans cette deuxième situation, il existe deux façons distinctes de prédire la valeur du caractère d'intérêt :

1) Le caractère cible est prédit pour de nouveaux candidats qui n'ont été phénotypés pour aucun des caractères corrélés,

2) Le caractère cible est prédit pour un ensemble de candidats qui ont été totalement ou partiellement phénotypés pour un ou plusieurs caractères secondaires.

Depuis l'introduction du concept de prédiction génomique (Whittaker et al. 2000; Meuwissen et al. 2001), de nombreux modèles mono-caractères ont été développés et comparés. Plus récemment, des modèles de prédiction génomique appliqués au contexte multi-caractère ont été proposés.

b. Modèles de prédiction génomique multi-caractère

Calus et Veerkamp (2011) ont présenté trois modèles de prédiction génomique multi-caractère qui supposaient différentes distributions des effets des marqueurs pour estimer les valeurs génétiques des individus. Le premier modèle correspondait à un modèle GBLUP multi-caractère, le deuxième à un modèle BayesCπ multi-caractère et le dernier à un modèle appelé BayesSSVS (*Bayesian Stochastic Search Variable Selection*) multi-caractère. Le modèle BayesSSVS a permis d'analyser des scénarios dans lesquels un nombre limité de gènes avec des effets forts avaient un impact important sur la corrélation génétique entre les caractères. Dans cette étude les différences de performance des différents modèles testés sur des données simulées étaient faibles.

Jia et Jannink (2012) ont également comparé 3 modèles de prédiction génomique multi-caractère: un modèle GBLUP multi-caractère, un modèle BayesA multi-caractère et un modèle BayesCπ multi-caractère. Contrairement à Calus et Veerkamp (2011), Jia et Jannink (2012) ont comparé ces modèles avec des caractères contrôlés par différentes architectures génétiques. Ils ont montré que les modèles bayésiens multivariés étaient plus performants que le modèle GBLUP multi-caractère lorsque le caractère d'intérêt était contrôlé par peu de QTL. En revanche, pour les caractères hautement polygéniques le modèle GBLUP multi-caractère était aussi efficace que les modèles bayésiens multi-caractères testés.

Plusieurs études ont testé des modèles GBLUP multi-caractères sur des données réelles. Ces études ont été menées sur différentes espèces, notamment végétales telles que le pin (Jia and Jannink 2012), le palmier à huile (Marchal et al. 2016), le seigle (Schulthess et al. 2016), le blé (Rutkoski et al. 2016; Hayes et al. 2017; Sun et al. 2017; Michel et al. 2018; Schulthess et al. 2018), le riz (Wang et al. 2017), ou encore le sorgho (Fernandes et al. 2018).

Le modèle GBLUP multi-caractère peut être écrit de la façon suivante:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \boldsymbol{\varepsilon} \qquad (3)$$

Avec $\mathbf{y}$ le vecteur des valeurs phénotypiques de longueur n x i (avec n le nombre d'individus et i le nombre de caractères) ; $\boldsymbol{\beta}$ le vecteur des effets fixes ; $\mathbf{X}$ la matrice de design des effets fixes ; $\mathbf{Z}$ la matrice de design des effets aléatoires ; $\mathbf{a}$ le vecteur des effets aléatoires des lignées de longueur n x i tel que $\mathbf{a} \sim \text{MVN}(0, \boldsymbol{\Sigma}_a \otimes \mathbf{K})$ avec $\mathbf{K}$ la matrice de *Kinship* de dimension n x n et $\boldsymbol{\Sigma}_a$ la matrice de variance-covariance des effets génétiques additifs de dimension i x i ; et $\boldsymbol{\varepsilon}$ l'erreur résiduelle telle que $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \boldsymbol{\Sigma}_\varepsilon \otimes \mathbf{I}_N)$ avec $\mathbf{I}_N$ la matrice identité et $\boldsymbol{\Sigma}_\varepsilon$ la matrice de variance-covariance des effets résiduels de dimension i x i. Le symbole $\otimes$ indique le produit de Kronecker entre deux matrices et l'abréviation MVN correspond à la loi normale multivariée.

Dans le cas d'un modèle bivarié (i=2), $\boldsymbol{\Sigma_a}$ s'écrit de la façon suivante:

$$\begin{pmatrix} \boldsymbol{\sigma_{a1}^2} & \boldsymbol{\sigma_{a12}} \\ \boldsymbol{\sigma_{a12}} & \boldsymbol{\sigma_{a2}^2} \end{pmatrix}$$

Avec $\boldsymbol{\sigma_{a1}}^2$et $\boldsymbol{\sigma_{a2}}^2$ les variances génétiques de chacun des deux caractères et $\boldsymbol{\sigma_{a12}}$ la covariance génétique entre les deux caractères. La matrice de variance-covariance $\boldsymbol{\Sigma_\varepsilon}$ est de même forme que $\boldsymbol{\Sigma_a}$ et intègre $\boldsymbol{\sigma_{\varepsilon 1}}^2$et $\boldsymbol{\sigma_{\varepsilon 2}}^2$ les variances des erreurs résiduelles de chacun des deux caractères et $\boldsymbol{\sigma_{\varepsilon 12}}$ la covariance entre les erreurs résiduelles des deux caractères.

D'autres études ont quant à elles porté sur des modèles bayésiens multi-caractères (y compris chez le pin : Jia et Jannink 2012, et chez le blé : Lado et al. 2018). Par ailleurs, quelques études se sont intéressées à des modèles multi-caractères plus complexes. C'est notamment le cas de Jiang et al. (2015) qui ont développé un modèle bayésien multivarié qui considère que les effets des SNP sont corrélés. Ils ont évalué cette méthode sur des données simulées et des données réelles portant sur des caractères immunologiques chez la souris et ils ont montré que cette méthode permettait d'obtenir des prédictions plus précises qu'un modèle BayesA multi-caractère. Montesinos-Lopez et al. (2016) ont quant à eux étendu un modèle de prédiction génomique multi-caractère à un modèle bayésien à la fois multi-caractère et multi-environnement (le modèle BMTME). Ce modèle prend en compte les corrélations entre les caractères mais aussi le terme d'interaction à trois éléments (caractère x génotype x environnement). Plus récemment, des modèles multi-caractères basés sur des méthodes d'apprentissage automatiques ont été développés dans le but de réduire la demande en ressources informatiques (Montesinos-López et al. 2018).

Etant donné que les modèles de prédiction génomique multi-caractère peuvent nécessiter d'une importante demande en ressources informatiques et de longs temps de calcul, ces deux facteurs sont des paramètres qu'il est important de considérer lors du choix du modèle utilisé dans un contexte multi-caractère.

Ces différents modèles multi-caractères ont permis d'explorer deux façons de prédire la valeur d'un caractère d'intérêt. Dans le premier cas, la valeur du caractère d'intérêt est prédite pour des individus qui n'ont été phénotypés ni pour le caractère prédit ni pour des caractères secondaires corrélés au caractère prédit. Dans le second cas, la totalité ou une partie des individus pour lesquels la valeur du caractère d'intérêt est prédite ont été phénotypés pour des caractères corrélés au caractère d'intérêt.

2. Prédictions génomiques des performances d'individus non phénotypés

a. Principaux facteurs affectant l'apport des prédictions génomiques multi-caractères par rapport aux prédictions mono-caractères

Plusieurs études portant sur les modèles de prédiction génomique multi-caractère ont été conduites en utilisant des jeux de données simulés afin d'identifier les principaux facteurs affectant l'apport de ces modèles par rapport aux modèles de prédiction génomique mono-caractère en terme de gain de qualité de prédiction (Calus and Veerkamp 2011; Jia and Jannink 2012; Hayashi and Iwata 2013; Guo et al. 2014).

Ces études ont mis en évidence le fait que la corrélation génétique entre les caractères pris en compte par le modèle était un des paramètres clés expliquant l'avantage des modèles de prédiction génomique multi-caractère par rapport aux modèles mono-caractères. En effet, plus la corrélation génétique est élevée, plus l'écart de qualité des prédictions est important entre les prédictions obtenues à partir de ces deux types de modèles. Comme attendu, lorsque la corrélation entre les caractères est proche de zéro, les modèles de prédiction génomique multi-caractère ne sont pas plus précis que les modèles mono-caractères (Calus and Veerkamp 2011; Jia and Jannink 2012; Hayashi and Iwata 2013).

La corrélation génétique entre les caractères n'est pas le seul paramètre ayant un impact sur le gain de précision des prédictions apporté par les modèles de prédiction génomique multi-caractère. L'héritabilité des caractères joue également un rôle important. Comme le montre la Figure 10, les modèles multi-caractères améliorent la qualité des prédictions lorsque le caractère d'intérêt a une faible héritabilité et qu'il est corrélé à un autre caractère avec une héritabilité plus élevée. En revanche l'avantage des modèles multi-caractères comparés aux modèles mono-caractères est faible lorsqu'il s'agit de prédire un caractère avec une forte héritabilité (Jia and Jannink 2012; Hayashi and Iwata 2013; Guo et al. 2014). Ces résultats sont particulièrement intéressants en amélioration des plantes puisque de nombreux caractères d'intérêt agronomique, tels que le rendement, ont une faible héritabilité.

Enfin, Jia et Jannink (2012) ont comparé des modèles de prédiction génomique mono-caractère et multi-caractère en utilisant des données simulées avec différentes architectures génétiques. Dans cette étude, les prédictions obtenues avec des modèles multi-caractères étaient plus précises que celles obtenues avec des modèles mono-caractères lorsque le caractère prédit était contrôlé par un nombre limité de gènes majeurs. En revanche, lorsque le caractère prédit était hautement polygénique le gain de qualité de prédiction apporté par les modèles multi-caractères était plus faible (Figure 10).

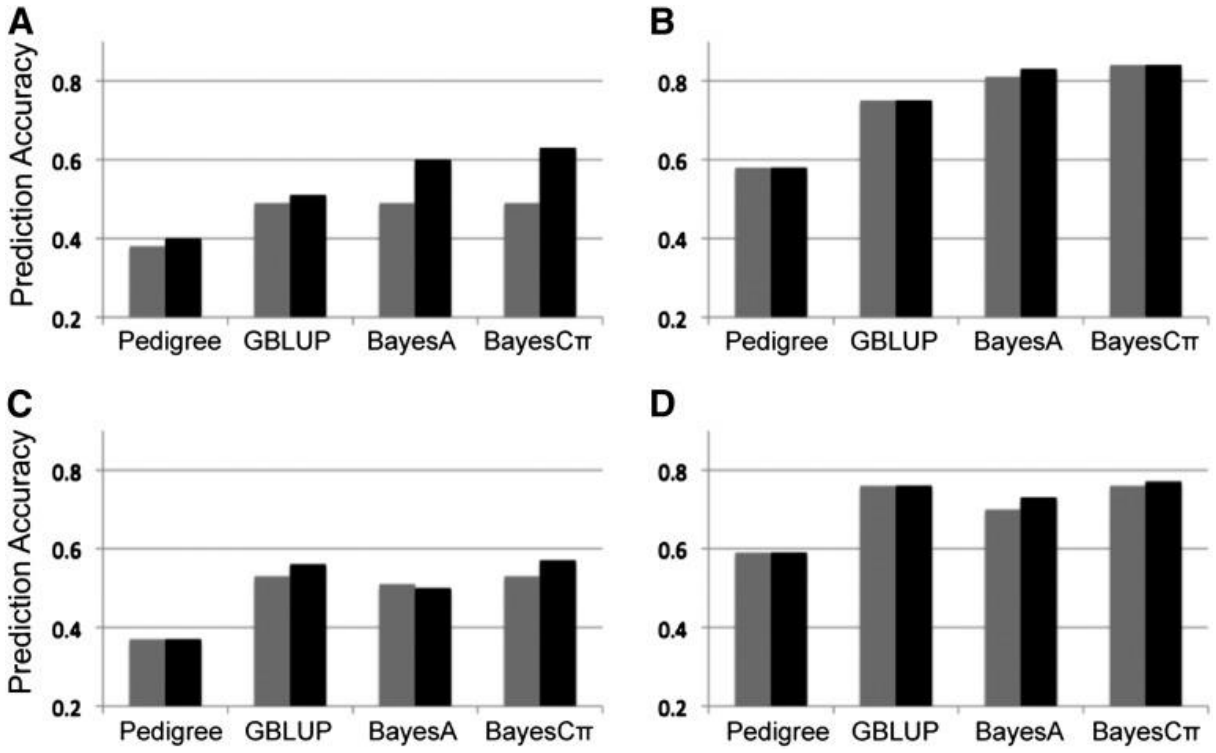

Figure 10: Comparaison des modèles de prédiction génomique (Jia and Jannink 2012)

Les modèles mono-caractères (gris) et multi-caractères (noir) sont évalués pour un caractère 1 (A and C) avec une faible héritabilité ($h^2 = 0.1$) corrélé à un caractère 2 (B and D) avec une forte héritabilité ($h^2 = 0.5$), contrôlés par 20 QTL (A and B) et par 200 QTL (C and D). La corrélation génétique entre les deux caractères est de 0.5 quel que soit l'architecture génétique considérée.

En conclusion, les prédictions génomiques multi-caractères sont d'autant plus intéressantes en termes de gain de précision des prédictions que le caractère d'intérêt est le caractère secondaire sont corrélés, que le caractère secondaire est plus héritable (et moins cher ou plus facile à mesurer que le caractère d'intérêt) et que l'architecture du caractère est faiblement polygénique.

b. Comparaison de modèles de prédiction génomique mono-caractère et multi-caractère pour prédire la performance d'individus non phénotypés avec des données réelles

Les modèles de prédiction génomique multi-caractère ont également été utilisés pour prédire la performance de nouveaux individus non phénotypés en utilisant des jeux de données empiriques. Contrairement aux études portant sur des données simulées, les études portant sur des données réelles ont montré que les avantages des modèles de prédiction génomique multi-caractère en terme de gain de qualité de prédiction étaient faibles ou nuls lorsque les prédictions portaient sur des individus qui

n'avaient été phénotypés ni pour le caractère d'intérêt, ni pour le caractère corrélé. Ces études ont été conduites chez différentes espèces de plantes ; notamment chez le pin (Jia and Jannink 2012), chez le soja (Bao et al. 2015), chez le seigle (Schulthess et al. 2016), chez le maïs (Dos Santos et al. 2016), chez le blé (Lado et al. 2018; Michel et al. 2018; Schulthess et al. 2018), et chez le sorgho (Fernandes et al. 2018). Par ailleurs, des problèmes de convergence ont été observés lors de l'utilisation de modèles de prédiction génomique multi-caractère (Michel et al. 2018).

3. Prédictions génomiques pour des individus phénotypés pour le caractère corrélé

a. Comparaison de modèles de prédiction génomique multi-caractère pour prédire la performance d'individus phénotypés pour le caractère corrélé avec des modèles de prédiction génomique mono-caractère

Il est également possible de prédire un caractère d'intérêt en utilisant un modèle multivarié quand les individus constituant la population de validation, ou du moins une partie d'entre eux, ont été phénotypés pour un ou plusieurs caractère(s) corrélé(s). Plusieurs études portant sur des données réelles ont montré que l'utilisation de modèles multi-caractères dans ce cas permettait d'obtenir des prédictions plus précises que les modèles mono-caractères.

Ce résultat a notamment été illustré dans l'article de Fernandes et al. (2018) et il est présenté dans la Figure 11. Dans cette étude, l'objectif était d'améliorer les prédictions génomiques du rendement en biomasse chez le sorgho qui est un caractère dont le phénotypage est coûteux et laborieux. Six caractères corrélés (l'humidité de la biomasse, la hauteur de plante mesurée à différents stades et l'aire sous la courbe de croissance en hauteur) ont été utilisés dans des modèles multi-caractères car ils étaient moins chers et plus simples à phénotyper que le rendement en biomasse. Lorsque tous les individus (y compris ceux de la population de validation) étaient phénotypés pour la hauteur de plante, la précision des prédictions du rendement en biomasse était améliorée de 50% par rapport à une situation mono-caractère.

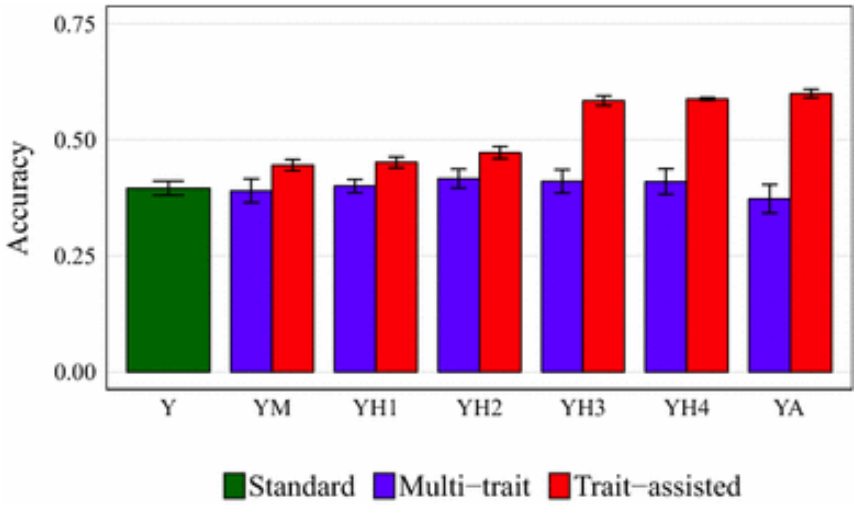

Figure 11 : Comparaison de la précision des prédictions du rendement en biomasse en utilisant des modèles mono-caractères et multi-caractères chez le sorgho (Fernandes et al. 2018)

Les caractères étudiés sont le rendement en biomasse (Y), l'humidité (M), la hauteur de plante à 30 (H1), 60 (H2), 90 (H3) et 120 (H4) jours après la plantation et l'aire sous la courbe de croissance (A). La précision des prédictions correspond à la corrélation entre les valeurs prédites (GEBV) et les moyennes ajustées et est appelée *accuracy*. Le modèle *standard* correspond à un modèle mono-caractère. Le modèle *multi-trait* correspond à un modèle multi-caractère où seuls les individus de la population d'entraînement ont été phénotypés. Le modèle *trait-assisted* correspond à un modèle multi-caractère où tous les individus ont été phénotypés pour le caractère secondaire et seuls les individus de la population d'entraînement ont été phénotypés pour le caractère d'intérêt.

Un gain de précision des prédictions en utilisant des modèles où les individus constituant la population de validation ont été phénotypés pour un ou plusieurs caractère(s) corrélé(s) a également été observé dans des études portant sur le blé tendre. Ces résultats ont notamment été obtenus pour des études portant sur la prédiction du rendement (Rutkoski et al. 2016; Sun et al. 2017; Crain et al. 2018) et sur la prédiction de caractères liés à la qualité boulangère du blé tendre (Lado et al. 2018 ; Michel et al. 2018). Schulthess et al. (2018) ont quant à eux montré que la prédiction de la résistance à la fusariose pouvait être améliorée en utilisant l'information apportée par deux caractères corrélés : la hauteur de plante ainsi que la date d'épiaison.

b. Allocation des ressources en utilisant des prédictions génomiques pour des individus phénotypés pour le caractère corrélé

Lado et al. (2018) ont utilisé un modèle de prédiction génomique multi-caractère pour prédire des caractères liés à la qualité boulangère chez le blé. Ils ont évalué ce modèle avec différentes tailles de population d'entraînement, différents pourcentages d'individus phénotypés pour les caractères corrélés et en utilisant un à trois caractères corrélés. Ce travail a montré que les méthodes de prédiction

génomique multi-caractère pourraient être intéressantes pour optimiser l'allocation des ressources dans un schéma de sélection. En effet, cela pourrait notamment permettre de phénotyper moins d'individus pour le caractère d'intérêt et d'utiliser le budget économisé pour phénotyper le caractère corrélé sur plus d'individus qui font partie de la population d'entraînement et de la population de validation.

Pour conclure, les modèles de prédictions génomiques multi-caractères présentent l'avantage de pouvoir exploiter l'information contenue dans la corrélation entre deux caractères. Les modèles multi-caractères peuvent souffrir d'une demande en capacité de calcul élevée, de temps de calcul longs et de problèmes de convergence. Cependant, plusieurs études utilisant des données réelles ont montré l'intérêt des méthodes de prédiction multi-caractère quand le caractère d'intérêt est coûteux ou difficile à mesurer mais que les individus pour lesquels on cherche à prédire la valeur de ce caractère ont été phénotypés pour un ou plusieurs caractères dont le phénotypage est plus facile à obtenir et qui sont corrélés au caractère d'intérêt. Cette approche est particulièrement intéressante quand le phénotype des caractères corrélés est moins cher ou plus simple à mesurer que le caractère d'intérêt et lorsqu'il peut être obtenu à un stade plus précoce.

Chapitre 2 :

Présentation du matériel étudié

I. Présentation des données phénotypiques

1. Le réseau d'essais

Les résultats présentés dans les chapitres 3 et 4 de ce manuscrit ont été obtenus en étudiant une population de référence composée de 1912 lignées de générations F8 et F9 issues du programme de sélection de blé tendre de l'INRA (devenu INRAE le 01/01/2020) et de sa filiale Agri-Obtentions. La Figure 12 illustre de façon schématique l'organisation du programme de sélection étudié et donne un ordre de grandeur du nombre de lignées présentes à chaque génération.

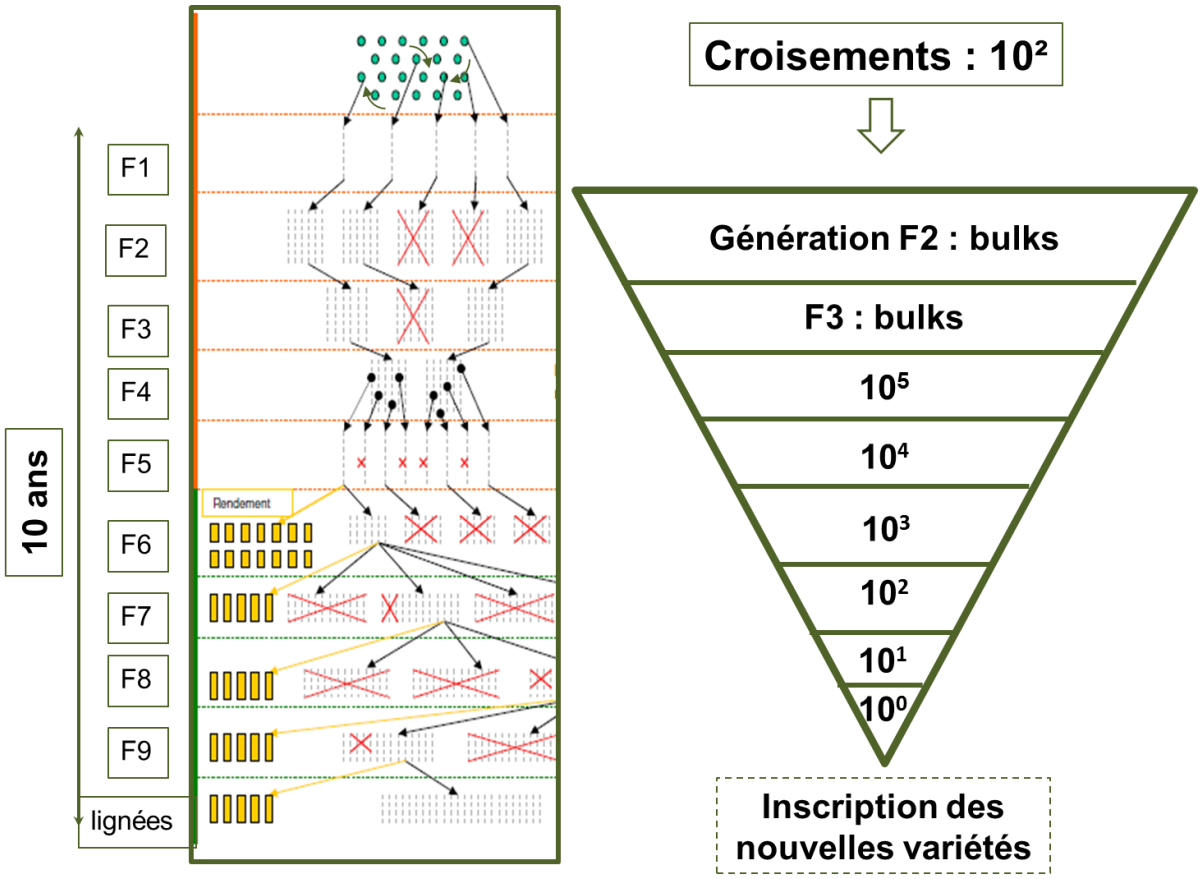

Figure 12 : Organisation du programme de sélection étudié.

Les valeurs indiquées sur la partie droite de la figure correspondent à un ordre de grandeur du nombre de lignées présentes à chaque génération.

Chaque lignée a été évaluée entre 2000 et 2016. Dans chaque essai, des lignées inscrites au catalogue officiel français et faisant partie des lignées les plus cultivées ont été utilisées comme témoins. Le réseau

d'essais dans lequel a été effectué le phénotypage comprend 11 lieux en France (Figure 13). Ce réseau d'essais s'étend d'Est en Ouest de Colmar (dans le Haut-Rhin) à Rennes (en Ille-et-Vilaine). L'essai le plus au Nord se situe dans le département de la Somme et l'essai le plus au Sud est localisé dans le département du Puy-de-Dôme. Bien que le réseau soit constitué de 11 lieux, des essais n'ont pas été réalisés dans chacun de ces lieux chaque année. En effet, en moyenne 8 lieux ont été utilisés par année.

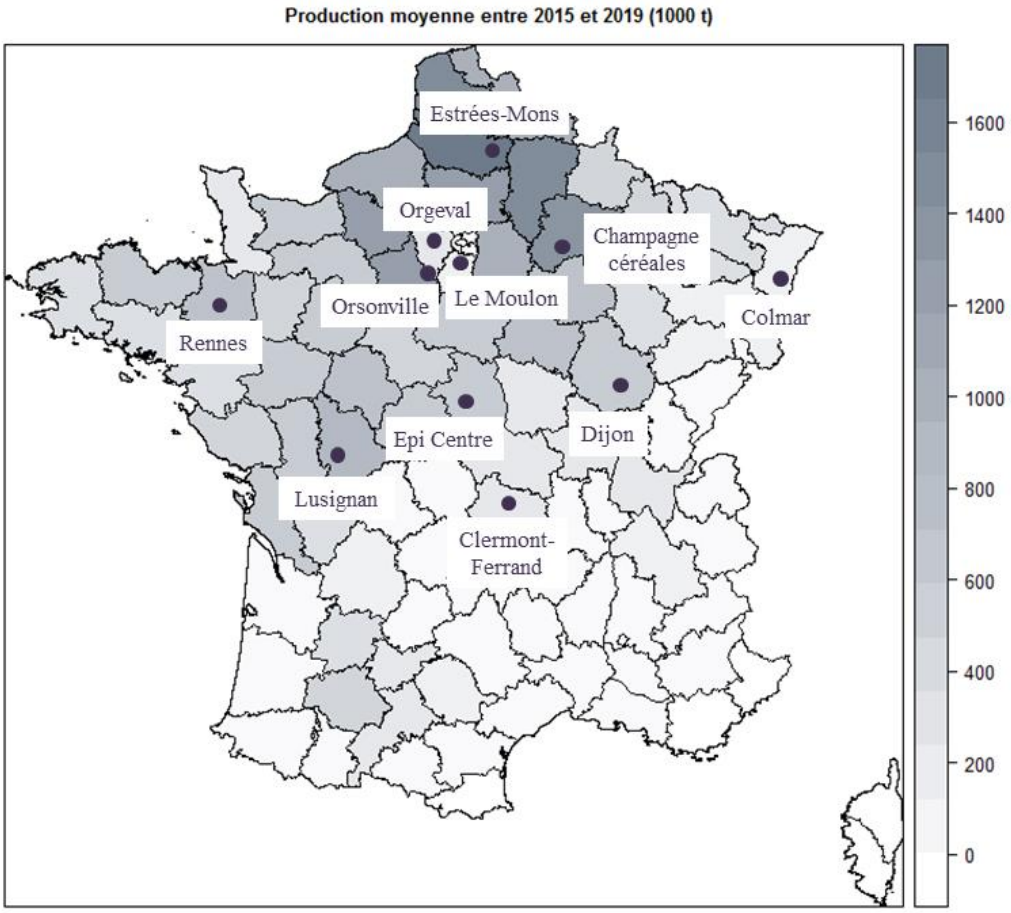

Figure 13 : Localisation des essais

La carte de France sur laquelle sont placés les 11 essais illustre la production moyenne de blé tendre par département entre 2015 et 2019 (Agreste, 2019).

Chaque année, plusieurs types de conduite ont été menés : une conduite traitée, une conduite non traitée, une conduite dite « faible intrant », et des essais spécifiques pour la résistance au froid ou aux maladies. La conduite traitée correspond à une gestion des cultures permettant d'optimiser le rendement. Des herbicides, des fongicides, des insecticides et des engrais minéraux, notamment azotés sont utilisés pour cette conduite. La conduite non traitée correspond quant à elle à une conduite conventionnelle qui ne fait pas usage de traitements fongicides. Pour la conduite dite « faible intrant », la densité de semis est

réduite de 40%, les apports en azote sont également réduits et aucun fongicide ni insecticide n'est utilisé. Seules les données de phénotypage obtenues dans les essais avec une conduite traitée ont été utilisées dans les analyses présentées dans ce manuscrit. Ceci permet d'éviter de prendre en compte les effets des maladies ou du manque d'azote sur la composition du grain et de travailler sur la qualité potentielle en absence de facteurs limitants autres que pédoclimatiques.

2. Les caractères évalués

Dans chaque environnement, correspondant à un lieu et une année donnés, plusieurs caractères d'intérêt agronomique ont été évalués.

Le rendement, la hauteur de plante, la date d'épiaison et la teneur en protéine ont été mesurés dans tous les essais. Les composantes du rendement, telles que le poids de mille grain, ont également été évalués dans certains essais. La teneur en protéine a été estimée par spectroscopie proche infrarouge (*Near-Infrared Reflectance Spectroscopy* ; NIRS). Des notes de résistance à plusieurs maladies (la fusariose, l'oïdium, la rouille brune, la rouille jaune et la septoriose) allant de 1 (note pour les plantes résistantes) à 9 (note pour les plantes très sensibles) ont également été relevées. Cependant, ces notes sont généralement peu indicatives dans les essais avec une conduite traitée puisque des traitements y sont appliqués.

Seules les lignées de la génération F9 ont été évaluées pour des caractères en lien avec la qualité boulangère. L'aptitude à la panification dite française utilisant seulement de la farine, de l'eau et du sel, sans autre additif (sauf parfois de l'acide ascorbique, un antioxydant) est le caractère utilisé en France lors de l'inscription au catalogue officiel pour classer les différentes variétés de blé tendre en fonction de leur qualité boulangère. Avant 2003, le test de panification CNERNA permettait d'obtenir la note de panification. Mais depuis 2003, elle est obtenue grâce à des tests définis par la méthode normalisée NF V03-716, aussi appelée méthode BIPEA. Les résultats portant sur la note de panification présentés dans de manuscrit reposent uniquement sur les valeurs issues des tests BIPEA. Ces tests permettent de mesurer le volume du pain, la note de pâte, la note de pain et la note de mie (la valeur maximale de chacune de ces trois notes est 100). La note finale de panification correspond à la somme de ces trois notes.

D'autres caractères liés à la qualité boulangère ont également été mesurés. C'est notamment le cas des paramètres de l'alvéographe de Chopin : la force boulangère notée W, la ténacité de la pâte notée P, l'extensibilité de la pâte notée L et l'équilibre entre la ténacité et l'extensibilité de la pâte qui correspond au rapport P/L. Ces paramètres permettent d'apprécier les caractéristiques rhéologiques de la pâte. De

plus, d'autres tests, tels que le test de Pelshenke ou le test de sédimentation de Zeleny, ont été réalisés pour certains environnements.

La Table 2 indique le nombre de lignées phénotypées chaque année pour la note de panification (valeurs sur la diagonale). Pour chaque couple d'années, le nombre de lignées communes phénotypées pour la note de panification est également renseigné dans la Table 2.

	2003	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
2003	53													
2004	18	49												
2005	5	16	48											
2006	1	7	20	58										
2007	1	5	7	21	68									
2008	1	3	3	8	27	55								
2009	2	4	2	5	10	24	55							
2010	1	4	2	4	8	13	24	69						
2011	0	1	0	1	3	5	8	18	50					
2012	0	1	0	0	1	3	4	9	22	59				
2013	0	1	0	0	1	3	4	5	13	32	77			
2014	0	1	0	0	1	1	1	3	6	9	30	68		
2015	0	0	0	0	0	0	0	0	1	2	12	28	72	
2016	0	0	0	0	0	0	0	0	1	2	6	16	37	72

Table 2 : Nombre de lignées phénotypées pour la note de panification dans le réseau d'essais.

II. Présentation des données de génotypage

1. Les différents types de données de génotypage

Parmi les lignées phénotypées pour le rendement, 871 lignées ont également été génotypées. Au total, 814 lignées ont été génotypées en utilisant la puce TaBW280K (Rimbert et al. 2018) qui contient 280K SNP de bonne qualité répartis sur l'ensemble du génome du blé (bien que la densité de marquage soit plus faible sur le génome D). Par ailleurs, une puce de 35K SNP indépendants inclus dans TaBW280K et répartis sur l'ensemble du génome a été développée. Les lignées produites après 2015 ont été génotypées avec la puce 35K.

De plus, 200 lignées faisant partie du matériel présenté précédemment ont été génotypées en utilisant 11 marqueurs KASP dérivés de la séquence d'ADN de gènes codant pour les sous-unités gluténines de haut poids moléculaire *Glu-A1*, *Glu-B1* et *Glu-D1* qui ont un impact connu sur la force boulangère. Ces marqueurs ont été décrits dans l'article de Ravel et al. 2020 (lien vers l'article en annexe).

2. Analyses préliminaires

Le chapitre 3 de ce manuscrit se concentre sur l'étude des 398 lignées qui ont été à la fois génotypées et phénotypées pour la note de panification. Etant donné que ces lignées sont issues d'un programme de sélection dont les essais ont été réalisés sur plus de dix années consécutives, il était intéressant de regarder si les lignées les plus anciennes étaient plus proches entre elles d'un point de vue génétique ou bien si l'année d'évaluation des lignées n'avait pas d'impact sur les proximités génétiques entre les lignées. Pour cette étude, seuls les marqueurs présents la puce 35K SNP (et par conséquent également présents sur la puce 280K) ont été utilisés. Les marqueurs ont été filtrés de sorte à ne garder que ceux dont la fréquence de l'allèle mineur était supérieure à 0.05 et dont le pourcentage de données manquantes n'excédait pas 10%. Les données manquantes restantes ont été imputées en utilisant l'algorithme EM (Poland et al. 2012) implémenté dans le paquet rrBLUP (Endelman 2011). Les distances euclidiennes entre individus ont été calculées selon la formule suivante :

$$d = \sqrt{\Sigma(x_i - y_i)^2} \qquad (4)$$

Avec x et y les vecteurs contenant respectivement les données de marquage de chacune des deux lignées pour lesquelles on cherche à calculer la distance. Une analyse en coordonnées principales a ensuite été réalisée (Figure 14).

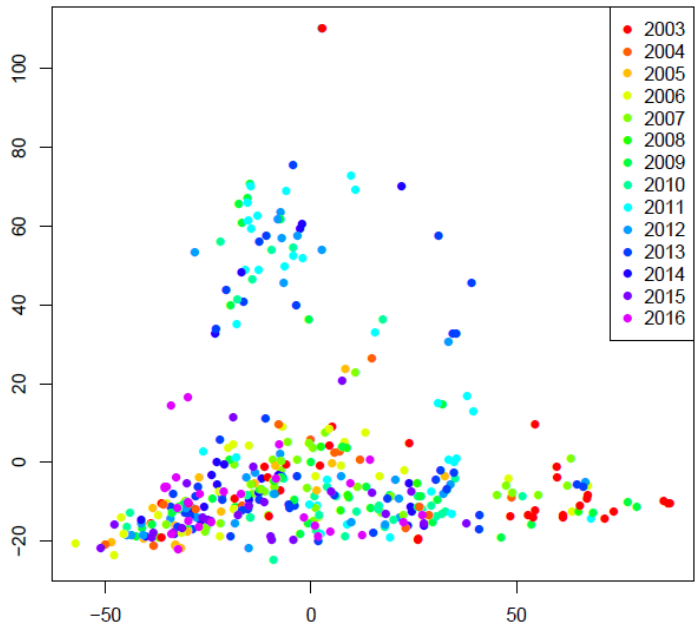

Figure 14 : Représentation de l'analyse en coordonnées principales réalisée sur les distances entre les lignées

La première année à laquelle la lignée a été évaluée pour la note de panification est représentée sur la figure et chaque couleur correspond à une année.

D'après une classification hiérarchique les lignées d'une année donnée sont réparties dans l'ensemble des groupes génétiques. L'ensemble de ces analyses nous laisse penser que les lignées évaluées une année donnée ne sont pas plus proches d'un point de vue génétique de celles évaluées la même année (ou une autre année consécutive) que de celles évaluées plusieurs années avant ou après. Les ré-échantillonnages pour les cross-validations peuvent donc se faire au hasard sur l'ensemble du jeu de données.

III. Prédictions génomiques mono-caractères

L'objectif de la thèse est d'étudier l'intérêt des outils de prédiction génomique pour l'amélioration du blé tendre. Le but de ce paragraphe est de présenter les précisions des prédictions génomiques mono-caractères des principaux caractères mesurés dans les essais en condition de culture dite traitée (le rendement, la hauteur de plante, la précocité, la teneur en protéines, les paramètres de l'alvéographe de Chopin et les résultats des tests BIPEA) car ces résultats ne sont pas abordés dans le reste de ce manuscrit (Table 3). En effet, le troisième chapitre du manuscrit se concentre sur les prédictions génomiques de la note de panification en utilisant différents types de modèles de prédictions génomiques. De plus, dans le quatrième chapitre les prédictions génomiques sont réalisées sur des caractères simulés.

Avant de prédire les caractères, des moyennes ajustées pour chacune des lignées (*Best Linear Unbiased Estimations* ; BLUE) ont été estimées. Un modèle GBLUP mono-caractère a ensuite était utilisé pour réaliser les prédictions en utilisant la matrice de *Kinship* et les BLUE des lignées faisant partie de la population d'entraînement. Les BLUE et les valeurs prédites ont été estimés en utilisant le logiciel ASReml-R (Butler et al. 2009). La méthode de validation croisée a été utilisée afin d'évaluer la précision des prédictions qui correspond à la corrélation de Pearson entre les valeurs prédites et les BLUE (souvent appelé *predictive ability* et notée r).

Caractères		Précision des
Rendement		0.55
Hauteur de plante		0.58
Date d'épiaison		0.51
Teneur en protéine		0.56
Alvéographe de Chopin	**W**	0.57
	P	0.62
	L	0.59
Tests BIPEA	**Volume**	0.37
	Note de pâte	0.36
	Note de pain	0.36
	Note de mie	0.40
	Note de panification	0.38

Table 3 : Précision des prédictions génomiques mono-caractères des principaux caractères

Chapitre 3 :

Apport des prédictions génomiques multi-caractères

I. Préambule

Ce chapitre est présenté sous la forme d'un article scientifique qui a été soumis dans le journal *Theoretical and Applied Genetics* et qui est en cours de révision (modifications mineures demandées). Cet article se concentre sur l'étude de méthodes visant à améliorer la qualité des prédictions génomiques tout en optimisant l'utilisation du budget alloué au phénotypage.

Pour réaliser ces analyses, nous avons pris comme exemple la note de panification, qui est utilisée en France lors de l'inscription de nouvelles variétés et qui permet de classer les variétés en fonction de leur qualité boulangère. Ce caractère est complexe et il ne peut être mesuré qu'à un stade tardif du programme de sélection, lorsqu'il y a suffisamment de grains par lignée pour produire la quantité de farine requise pour effectuer le test de panification.

Nous nous sommes plus particulièrement intéressés à l'utilisation de modèles de prédiction génomique multi-caractère pour améliorer la qualité des prédictions du caractère d'intérêt et réduire le budget alloué au phénotypage.

Afin d'identifier d'autres pistes d'amélioration de l'allocation des ressources, nous avons également étudié l'impact du nombre de mesures phénotypiques par lignée sur la qualité des prédictions ainsi que l'importance des lignées témoins phénotypées dans une large gamme d'environnements.

Des méthodes d'optimisation de la population d'entraînement, utilisée pour calibrer les modèles, ont également été analysées dans un contexte multi-caractère dans le cadre de ce projet.

Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: application to bread making quality

S Ben-Sadoun[1], R Rincent[1], J Auzanneau[2], FX Oury[1], B Rolland[3], E Heumez[4], C Ravel[1], G Charmet[1], S Bouchet[1]

[1] *INRAE - UCA UMR1095, Genetics Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu 63000 Clermont-Ferrand - France*

[2] *Agri-Obtentions, Ferme de Gauvilliers 78660 Orsonville - France*

[3] *INRAE - Agrocampus Ouest Rennes- Université Rennes 1 UMR 1349 IGEPP BP35327, 35653 Le Rheu Cedex, France*

[4] *INRAE UE Lille, 2 chaussée Brunehaut, Estrées-Mons, BP 50136, 80203 Peronne Cedex, France*

Corresponding author:

Sophie Bouchet

INRAE-UCA UMR1095, Genetics Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu 63000 Clermont-Ferrand – France

sophie.bouchet@inra.fr

ORCID: 0000-0001-5868-3359

II. Abstract

Key message **Trait-assisted genomic prediction approach is a way to improve genetic gain by cost unit, by reducing budget allocated to phenotyping or by increasing the program's size for the same budget.**

This study compares different strategies of genomic prediction to optimize resource allocation in breeding schemes by using information from cheaper correlated traits to predict a more expensive trait of interest. We used bread wheat baking score (BMS) calculated for French registration as a case study. To conduct this project, 398 lines from a public breeding program were genotyped and phenotyped for BMS and correlated traits in 11 locations in France between 2000 and 2016.

Single trait (ST), multi-trait (MT) and trait-assisted (TA) strategies were compared in terms of predictive ability and cost. In MT and TA strategies, information from dough strength (W), a cheaper trait correlated to BMS (r = 0.45), was evaluated in the training population or in both the training and the validation sets, respectively. TA models allowed to reduce the budget allocated to phenotyping by up to 65% while maintaining the predictive ability of BMS. TA models also improved the predictive ability of BMS compared to ST models for a fixed budget (maximum gain: +0.14 in cross-validation and +0.21 in forward prediction).

We also demonstrated that the budget can be further reduced by approximately one fourth while maintaining the same predictive ability by reducing the number of phenotypic records to estimate BMS adjusted means. In addition, we showed that the choice of the lines to be phenotyped can be optimized to minimize cost or maximize predictive ability. To do so, we extended the mean of the generalized coefficient of determination (CDmean) criterion to the multi-trait context (CDmulti).

Key words

Multi-trait selection. Trait-assisted selection. Selective phenotyping. Resource allocation optimization. Wheat baking quality. Coefficient of determination (CD)

III. Introduction

Thanks to exponential decrease of genotyping costs, improvement of computing and data storage capacities genomic selection is becoming a powerful tool in plant breeding. Genomic selection uses all marker information to calculate genomic estimated breeding values (GEBVs) for complex traits and selection of candidates is made directly on GEBVs without further phenotyping (Whittaker et al. 2000; Meuwissen et al. 2001). The use of genomic predictions at key steps of a breeding program can improve genetic gain per unit of time and cost by optimizing resource allocation, minimizing phenotypic evaluation (selective phenotyping) of candidates to selection in particular (Heslot et al. 2017).

Several models of genomic predictions were proposed (reviewed by Heffner et al. 2009). The genomic best linear unbiased prediction (GBLUP) model relies on a marker based kinship (similarity) matrix. It is simple to implement and efficient in most cases (Heslot et al. 2015). Other factors also affect the prediction accuracy, including the size and the composition of the training set. Assuming that the number of markers is sufficient, the accuracy of the calibration model strongly depends on congruency between the allelic composition represented in the training population to estimate marker effects, and the allelic composition of the candidates whose performance is to be predicted (Habier et al. 2007). When the prediction uses unrelated populations to train the prediction equations, prediction accuracy actually becomes negligible (Crossa et al. 2014). Methods based on minimizing the mean of the prediction error variance (PEVmean; Rincent et al. 2012; Akdemir et al. 2015; Isidro et al. 2015; Sarinelli et al. 2019) or maximizing the mean of the generalized coefficient of determination (CDmean) of the estimated contrast between candidates and the mean of the panel were proposed to optimize the sampling of key individuals to be phenotyped in unstructured (Rincent et al. 2012) and in structured multi-familial populations (Rincent et al. 2017). In any case, accurate predictions will not be obtained until a large enough pool of individuals that is relevant to the candidates is genotyped and phenotyped with a sufficient precision.

So far, most of genomic prediction studies in plant science focused on single-trait analyses, while applied breeding programs usually address multi-trait selection. We can actually take advantage of correlation between traits of interest and secondary traits to optimize phenotyping, especially if one secondary trait correlated to the targeted trait is easier to phenotype, can be obtained at an earlier stage, or is cheaper. It is possible to predict correlated traits simultaneously using multivariate best linear unbiased prediction (BLUP; Henderson and Quaas 1976). Multi-trait genomic prediction (MT) models benefit from information contained in both genetic correlation between traits and genetic relationship among individuals (Calus and Veerkamp 2011). Two strategies are possible: 1) either the training population is genotyped and phenotyped for both traits and the candidate population is genotyped but

not phenotyped for any of the traits, or 2) the candidates can be partly phenotyped for the trait of interest or the secondary trait. The latter strategy is called trait-assisted genomic selection (TA; Fernandes et al. 2018). Benefits of MT models over single-trait genomic prediction (ST) models were reported in simulated and empical data (Calus and Veerkamp 2011; Jia and Jannink 2012; Hayashi and Iwata 2013; Guo et al. 2014). As for single trait prediction, under a major QTL genetic architecture, Jia and Jannink (2012) found that Bayesian multivariate models (BayesA or BayesCπ) performed better than multi-trait GBLUP model. But for traits with polygenic genetic architecture, multi-trait GBLUP model was equal to the Bayesian multivariate models. MT models can however suffer from a high computational demand, time and some convergence problems (Michel et al. 2018). Obviously, genetic correlation between traits is a key factor determining the MT advantage over ST methods (Calus and Veerkamp 2011; Jia and Jannink 2012; Hayashi and Iwata 2013; Guo et al. 2014). Although MT models improve the predictive ability when the targeted trait has a low heritability and the secondary trait has higher heritability, the advantage of MT models to predict high heritability traits is low (Jia and Jannink 2012; Hayashi and Iwata 2013; Guo et al. 2014).

In bread wheat breeding programs, baking tests or dough rheology measurements are part of official variety evaluation for registration in many countries. Improving baking quality is a key aspect of wheat improvement but criteria may be specific to countries and targeted markets. In France, the most important criterion for baking quality is a normalized (NF V03-716) bread-making score (BMS). It is used to classify registered varieties into end-use quality classes. Varieties presenting high BMS scores are reserved for French baking; they must yield at least 2% higher than reference varieties in registration trials. Varieties with intermediate BMS scores are mainly meant to be exported or used in mixed flours and must yield 4% higher than reference varieties. Varieties with lower BMS can only be used as feed or fuel and must yield 7% more than reference varieties. This BMS score is a complex quantitative trait that integrates several physical measurements of dough and loaf. Its assessment requires a large amount of grains and flour which are not available in the first generations of the breeding program, i.e. in row-nursery or small plots. In addition, BMS evaluation is time-consuming and quite expensive. Therefore, this trait is only evaluated at the very last steps of wheat breeding programs after the number of candidate lines was largely reduced. In French breeding programs, BMS is often estimated at earlier stages using less expensive traits such as the Chopin alveograph parameters which describe the dough rheological characteristics. Dough strength (W), in particular has a correlation of approximately 0.4 to BMS (Oury et al. 2010).

Furthermore, dough strength was shown to be itself influenced by high-molecular-weight glutenin subunits (HMW-GS; reviewed by Shewry 2009) that are cheaper to type and necessitate just a few grains. HMW-GS are encoded by the *Glu-1* loci, named *Glu-A1*, *Glu-B1* and *Glu-D1*, located on the long arm of the homoeologous chromosomes of group 1 (Payne 1987). Each locus comprises the two

tightly linked genes *Glu-1-1* and *Glu-1-2*. To predict the degree of elasticity or tenacity of the dough of candidates to selection, HMW-GS used to be typed by sodium-dodecyl-sulphate polyacrylamide gel electrophoresis (SDS-PAGE) of grain proteins. Several attempts were made to derive molecular markers from glutenin DNA sequences (reviewed by Liu et al. 2012) because they are cheaper and easier to implement routinely. Ravel et al. (2020) recently developed KASP markers that predict the rheological properties of the dough slightly better than SDS-PAGE.

Previous studies concerning multi-trait genomic predictions of baking quality related parameters such as dough rheological traits (Hayes et al. 2017; Michel et al. 2018; Lado et al. 2018) did not focus on BMS. In addition, none of these approaches tested the economic gain achievable with MT and TA models.

In this study, we demonstrate how to decrease the cost of breeding an expensive trait by using the information of a cheap correlated trait. To do so, we used genomic prediction of BMS as a case study. We evaluated different strategies in terms of predictive ability and cost. We compared ST models, univariate genomic prediction models with molecular markers associated with HMW-GSs as fixed effects (ST-glu), MT and TA models using dough strength (W) as a correlated trait. Only individuals of the training set were phenotyped for the trait of interest in both MT and TA models. In MT model, only the individuals of the training set were phenotyped for the correlated trait W. In TA model, the training set and a variable proportion of the validation set were phenotyped for the correlated trait. In addition, we adapted the CDmean criterion (Rincent et al. 2012) to the multi-trait context (CDmulti) in order to optimize the phenotyping strategy for the secondary trait and the trait of interest before collecting any phenotype. CDmulti was used to determine which individuals should be phenotyped for the secondary and/or the trait of interest to optimize predictive ability of the prediction set for the trait of interest.

IV. Materials and Methods

1. Plant material

We analyzed a breeding population of *F8-F9* winter-type bread wheat lines developed by the Institut National de la Recherche en Agriculture, Alimentation et Environnement (INRAE) and its subsidiary company Agri-Obtentions. Each of the 1912 lines were evaluated between 2000 and 2016. Each year, the lines were phenotyped in 7 to 9 locations in France (Supplementary Table S1). In each environment, some registered varieties were used as controls, (6.8 control lines on average for trials in which BMS was phenotyped). Each control line has been evaluated for BMS for several consecutive years, up to 8 years for some of them. Those controls are essential to estimate adjusted means in further analyses. Crop management methods corresponded to high yield objectives (optimized pesticide, fungicide and nitrogen amount) to avoid confusing effects of disease or N-supply limitation.

2. Genotyping data

In total, 814 lines were genotyped using a 280K SNP array (Rimbert et al. 2018). As predictive ability was the same using a subset of 35K markers that minimize Linkage Desequilibrum (LD; result not shown), the 57 breeding lines that were produced after 2015 were genotyped with a 35K SNP array included in the 280K array. Finally, 21210 markers with minor allele frequency superior to 0.05 and less than 10% of missing data were kept for analyses. Missing data were imputed using a kinship-based EM algorithm (Poland et al. 2012), based on the assumption that marker genotypes follow a multivariate normal (MVN) distribution. A genomic relationship matrix K was computed using Endelman and Jannink equation (2012):

$$K = \frac{WW^T}{2\sum(p_k - 1)p_k} \quad (5)$$

where W is a centered N × M marker matrix whose i-th row and k-th column is $w_{ik} = x_{ik} + 1 - 2p_k$ with X_{ik} the genotype of the *i*-th individual for the *k*-th marker as $\{-1,0,1\}$ and p_k the allele frequency at the *k*-th marker. These algorithms were implemented in the A.mat function of the rrBLUP package (Endelman 2011).

Additionally, a subset of 200 lines was genotyped at 11 DNA-based KASP markers derived from the HMW-GS loci *Glu-A1*, *Glu-B1*, and *Glu-D1* (Ravel et al. 2020). We calculated Pearson's correlation between each of the KASP marker and the 21210 SNP markers. We identified 18 SNP markers with correlation higher than 0.9 with at least one of the HMW-GS markers. In the model using those KASP markers as fixed effects, the array of these 18 SNP markers was filtered out. The datasets with the genotyping data are available in the INRA Dataverse repository (https://data.inra.fr/). They can be accessed with the following link https://doi.org/10.15454/AABGO7.

3. Phenotypic data

In total, 871 lines were both genotyped and evaluated for grain yield in 9.9 environments on average. Although yield, height, heading date and protein were recorded for all the *F8* lines, baking and rheological traits were only available in half of the lines that were selected for *F9* evaluation. Protein content was estimated by Near-Infrared-Reflectance. This test was realized according to American Association of Cereal Chemists (AACC 39-10) method, using a wholemeal flour produced on a Cyclotec mill with a 0.8 mm sieve. Chopin alveograph (method AACC 54-30A) was carried out on flour obtained with a Chopin mill or a Brabender Senior mill (the extraction rate of these two mills is about 70%). The dough strength (W), the tenacity (P) and the extensibility (L) of the dough were evaluated with a Chopin Alveolink apparatus (Tripette & Renaud, 92396 Villeneuve-la-Garenne, France). The bread-baking tests were realized according to the AFNOR method NF V03-716 (also called BIPEA test). This baking test gives a measure of loaf volume, a score for dough (out of 100 units), a score for crust (out of 100 units) and a score for crumb (out of 100 units); the sum of these 3 scores giving the final BIPEA score (out of 300 units). This final BIPEA score corresponds to the Bread Making Score (BMS). In total, 398 lines were both genotyped using an array and phenotyped for BIPEA BMS. The dataset with the phenotypic data is available in the INRA Dataverse repository (https://data.inra.fr/). It can be accessed with the following link https://doi.org/10.15454/AABGO7.

4. Statistical analysis of phenotypic data

Adjusted means of the lines were estimated using Best Linear Unbiased Estimations (BLUEs) in ASReml-R library (Butler et al. 2009) using the following model for each trait:

$$y_{ijk} = \mu + g_i + e_j + g_i \times e_j + \varepsilon_{ijk} \qquad (6)$$

where y_{ijk} is the phenotypic value of the i-th genotype in the j-th environment (combination of year and site) and for the k-th repetition, μ is the overall mean, g_i is the fixed effect of the i-th genotype, e_j is the effect of j-th environment, $g_i \times e_j$ is the effect of the interaction between the i-th genotype and the j-th environment, and ε_{ijk} is the residual error for the i-th genotype in the j-th environment and for the k-th repetition. All effects except the genotype effect were considered as random. Correlation between traits was calculated as the Pearson's correlation between BLUEs. Note however that traits linked to baking quality were measured only once in each environment. Therefore, the $g_i \times e_j$ interaction was not included in the model for these traits.

The estimation of adjusted means can affect the predictive ability of the genomic prediction models. Therefore, we evaluated the impact of the number of phenotypic records used to calculate BLUEs on the predictive ability. We especially investigated the importance of control lines that were phenotyped in all locations during several successive years in the estimation of BLUEs and the predictive ability, by comparing two different cases of BLUE calculations:

1) In the first case, all phenotypic records for control lines and one to four phenotypic records (in four different trials) for the other lines were used to estimate BLUEs for BMS and W;
2) In the second case, one to four phenotypic records for each line (the control and the other lines) were used to estimate BLUEs for BMS and W.

For each scenario, the phenotypic records were randomly sampled 10 times.

The variance components of the traits were assessed with ASReml-R using the model (6) but in this case all the effects (including genotypic effects) were considered as random.

Repeatability was estimated at the plot and design levels as:

$$\text{plot repeatability} = \frac{\sigma_G^2}{\sigma_G^2+\sigma_{GxE}^2+\sigma_\varepsilon^2} \qquad (7)$$

$$\text{design repeatability} = \frac{\sigma_G^2}{\sigma_G^2+\sigma_{GxE}^2/t+\sigma_\varepsilon^2/(t\times r)} \qquad (8)$$

Where σ^2_G; σ^2_{GxE} and σ^2_ε are the genotype, genotype x environment and residual variances respectively, t and r are the average number of trials per genotype and replicates in each trial per genotype, respectively. The variance components and repeatability were estimated using lines which were both phenotyped and genotyped. For the traits linked to rheological quality, σ^2_{GxE} term was removed from equations (7) and (8).

5. Genomic prediction models

a. Single-trait GBLUP models

First, a single-trait GBLUP model was used to obtain genomic predictions for different traits:

$$y = X\beta + Za + \varepsilon \qquad (9)$$

Where y is a N x 1 vector of BLUEs obtained in the first step with N the number of lines; β is the vector of fixed effects; X is the corresponding design matrix for the fixed effect; Z is the design matrix for the random effects; a is the N x 1 vector of lines random effects with the additive genetic variance σ^2_a and $a \sim N(0, K\sigma^2_a)$; and ε is the residual error with the residual variance σ^2_ε and $\varepsilon \sim N(0, I\sigma^2_\varepsilon)$.

Three types of single-trait models were performed. The first one was a simple ST model in which the vector of fixed effects β modelled the grand mean. The second model, called ST-glu, included the 11 HMW-GS markers as fixed effect. The third model (ST-glu_stepwise) included as fixed effect the HMW-GS markers that explained a significant genetic variance. They were selected using a stepwise regression model implemented in the stepAIC function from the MASS package (Venables and Ripley 2002).

b. Multi-trait GBLUP models

The same genomic relationship matrix K was used for fitting the multi-trait model:

$$y = X\beta + Za + \varepsilon \qquad (10)$$

Where y corresponds to an N × 2 vector of BLUEs for two traits, BMS which is the trait of interest and W which is a correlated trait less expensive to measure, β is the vector of fixed effects (the grand mean in this study); X is the corresponding design matrix for the fixed effect; a is the vector of N × 2 line effects with the corresponding random effect design matrix Z and $a \sim MVN(0, \Sigma_a \otimes K)$ with the variance–covariance matrix Σ_a of the form:

$$\begin{pmatrix} \sigma^2_{a1} & \sigma_{a12} \\ \sigma_{a12} & \sigma^2_{a2} \end{pmatrix}$$

where σ_{a1}^2 and σ_{a2}^2 are the genetic variance of the first and second trait, respectively, and σ_{a12} is the genetic covariance between both traits. The variance of the residual effect followed $\varepsilon \sim MVN(0, \Sigma_\varepsilon \otimes I)$ where I is the identity matrix of dimension N x N and Σ_ε has the same form than Σ_a. $\otimes$ indicate the Kronecker product operator between matrices.

We investigated two ways of predicting the value of the trait of interest using multivariate models: either we predicted it for un-phenotyped individuals (MT model), or we predicted it for a set of individuals that were fully or partly phenotyped for the correlated trait (TA model). In both MT and TA scenarios, measurements of the targeted trait were only available for the training population.

6. Model validation

To compare the different models, predictions were evaluated using 5-fold cross-validations. This method consists in randomly splitting the data set in 5 folds of equal size and using 4 folds as the training set in order to predict performance of the lines from the fifth fold. The same procedure is applied to the four other folds and it is repeated 20 times. Models were tested using the lines which were both genotyped and phenotyped and which were not considered as control lines.

The predictive ability was calculated as the mean of the 100 Pearson's correlations between predicted values and BLUEs. The standard deviation of the predictive ability was also calculated.

7. Cost evaluation

Budget allocated to phenotyping was obtained using the following equation:

$$B = Cp_{BMS} \times N_{BMS} + Cp_W \times N_W \qquad (11)$$

Where N_{BMS} and N_W are the number of lines phenotyped for BMS and for W respectively; and Cp_{BMS} (150€) and Cp_w (20€) are the cost of phenotyping one line for BMS and for W respectively.

In the equation (11) Cp_{BMS} and Cp_w are numeric constants but N_{BMS} and N_W are variable depending on the experimental design. It means that for a fixed budget there is several ways to allocate it between lines that are phenotyped for BMS or W, in the training set and in the validation set.

Phenotyping cost for W corresponds to a minimum cost when we do it internally at INRA. BMS must be evaluated by certified companies and the price can vary between providers and depending on the volume we phenotype each year. In order to evaluate the impact of the cost ratio between W and BMS on the predictive ability of BMS, we made this ratio vary (ratio = 1/2, 1/3, 1/4, 1/5, 1/10 and 1/20). We performed this analysis with three different budgets, with Cp_{BMS}=150€, with a MT approach and a TA approach where all lines are phenotyped for W (Supplementary Table S2).

8. Multi-trait CDmean (CDmulti) and optimization algorithm

The objective was to optimize the choice of individuals from the training set that should be phenotyped for the trait of interest BMS and for the correlated trait W to predict the individuals from the validation set. For that purpose, we extended the CDmean criterion (Rincent et al. 2012) to a multi-trait CDmean criterion (CDmulti). We first computed the prediction error variance (PEV), using the following equations (Lynch and Walsh 1998):

$$\begin{bmatrix} X^T({\Sigma_\varepsilon}^{-1} \otimes I)X & X^T({\Sigma_\varepsilon}^{-1} \otimes I)Z \\ Z^T({\Sigma_\varepsilon}^{-1} \otimes I)X & Z^T({\Sigma_\varepsilon}^{-1} \otimes I)Z + ({\Sigma_a}^{-1} \otimes K^{-1}) \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T({\Sigma_\varepsilon}^{-1} \otimes I)y \\ Z^T({\Sigma_\varepsilon}^{-1} \otimes I)y \end{bmatrix} \quad (12)$$

where y is the vector of phenotypes, β is a vector of fixed effects (in our case it is the intercept), a is a vector of random genetic values, X and Z are the corresponding design matrices for the fixed and random effects respectively, K is the kinship matrix, Σ_a is the genetic variance–covariance matrix between traits, and Σ_ε is the residual variance-covariance matrix between traits.

With the following notations:

$$\begin{bmatrix} X^T({\Sigma_\varepsilon}^{-1} \otimes I)X & X^T({\Sigma_\varepsilon}^{-1} \otimes I)Z \\ Z^T({\Sigma_\varepsilon}^{-1} \otimes I)X & Z^T({\Sigma_\varepsilon}^{-1} \otimes I)Z + ({\Sigma_a}^{-1} \otimes K^{-1}) \end{bmatrix}^{-1} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \quad (13)$$

we have:

$$PEV(\hat{a}) = diag(C_{22}) \quad (14)$$

The multi-trait CD was defined using the following equation:

$$CD_{multi}(\hat{a}) = 1 - \frac{PEV(\hat{a})}{diag(\Sigma_a \otimes K)} \quad (15)$$

CD_{multi} is the expected reliability (before phenotyping) for each trait and each individual. It takes values between 0 and 1. The closer to 1 the criterion is, the more reliable the predictions are expected to be.

We optimized the sampling by maximizing the mean of the CD_{multi} ($\overline{CD}_{multi}$) of individuals in the validation set for BMS, the trait of interest. For this a simple exchange algorithm was used. The initialization step consisted in randomly sampling calibration lines for BMS and/or W phenotyping and the corresponding $\overline{CD}_{multi}$ was calculated for the validation set for BMS. For each next step, one of the two traits (BMS or W) was randomly chosen. Then a random exchange between one individual in the calibration set for the chosen trait and one individual which was not in the calibration set was realized. If this new calibration set resulted in a higher$\overline{CD}_{multi}$, the new calibration set was accepted, otherwise the exchange was rejected. This procedure was repeated 2000 times to reach an optimum.

The genetic and residual variance-covariance matrices between traits (Σ_a and Σ_ε) needed to be estimated from phenotypic data. Because CD_{multi} was maximized before getting phenotypic values, matrices were estimated using an independent dataset. To do so, we used 201 lines phenotyped for W and BMS by the Groupement d'étude et de contrôle des variétés et des semences (GEVES) and genotyped using the same set of SNP markers.

We tested this optimization algorithm for several scenarios (Figure 15).

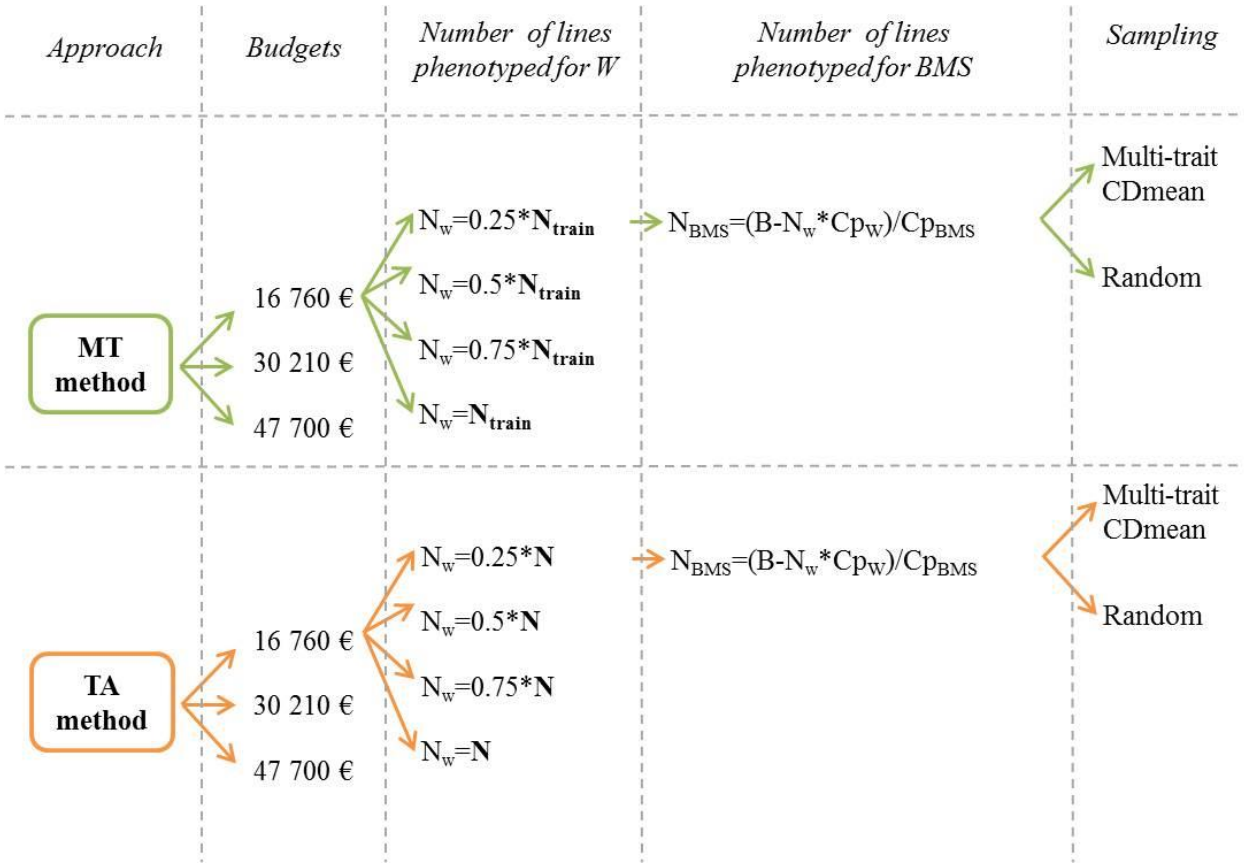

Figure 15 : Definition of each scenario.

N_W, N_{BMS}: Number of lines phenotyped for W and BMS, respectively. N: Total number of lines (number of lines in the training and the validation sets). N_{train}: Number of lines in the training set.

The optimization algorithm was performed with two different conditions:

1) in the first case, only individuals of the training set could be phenotyped for BMS and W (MT approach),

2) in the second case, only individuals from the training set could be phenotyped for BMS, while any of the individuals (from the training or the validation sets) could be phenotyped for the secondary trait W (TA approach).

In order to be able to compare GP models for different scenarios, lines from the validation set were never considered phenotyped for BMS, so that BMS predictive ability was always calculated on the same set of individuals.

We performed the optimization algorithm using different sample sizes that should be phenotyped for BMS and/or for W (Table 4).

		MT		TA	
Budget	**% of lines phenotyped for W**	**N_W**	**N_{BMS}**	**N_W**	**N_{BMS}**
47 700€	100	318	275	398	264
	75	238	286	298	278
	50	159	296	199	291
	25	79	307	99	304
30 210€	100	318	159	398	148
	75	238	169	298	161
	50	159	180	199	174
	25	79	190	99	188
16 760€	100	318	69	398	58
	75	238	80	298	72
	50	159	90	199	85
	25	79	100	99	98

Table 4 : Number of selected individuals in each scenario.

N_W, N_{BMS}: Number of lines phenotyped for W and BMS, respectively

We defined three budgets:

1) The most expensive budget allows to phenotype all the lines from the training set for BMS (47 700€),

2) The intermediate budget allows to phenotype all the lines from the training set for W and half of the lines for BMS (30 210€),

3) the less expensive budget allows to phenotype a quarter of the lines from the training set for BMS and three quarter of the lines from the training set for W (16 760€).

For a fixed budget, equation (11) has two unknowns (N_{BMS}, N_W), and there are many different possible combinations of individuals and traits to phenotype. We tested this optimization algorithm with 4 resource allocation strategies for each of the three budgets and under both MT and TA conditions. The resource allocation strategies correspond to a proportion (25, 50, 75 and 100%) of the training set (MT approach) or of the total population (TA approach) that is phenotyped for W. Then the N_{BMS}, the number of lines phenotyped for BMS, was calculated using the following equation obtained from equation (11):

$$N_{BMS} = \frac{B-(Cp_W x N_W)}{Cp_{BMS}} \quad (16)$$

Training and validation sets were defined using 5-fold cross-validations. For each scenario (MT or TA approach, one budget and one resource allocation), we tested 20 folds (4 replicates x 5-fold cross-validations). In addition, for each fold and each scenario the $\overline{CD}_{multi}$ optimization algorithm was repeated 5 times, which means that for each scenario the procedure has been performed 100 times. The BMS BLUEs from the training set and W BLUEs from the training set (MT) and part of the validation set (TA) were used to predict BMS in the validation set. We compared predictive abilities obtained from optimized ($\overline{CD}_{multi}$) and random training sets (the mean predictability of 100 random sets for each scenario and each validation set).

Furthermore, the most expensive budget (47 700€) allowed to perform a ST approach in which all the individual from the training set were phenotyped for BMS. Therefore, we also compared predictive ability obtained for optimized training sets within each scenario with the predictive ability obtained with a ST model for BMS (calculated for each cross-validation that has been tested with the optimization algorithm).

In order to mimic real life situation, we also used a forward prediction strategy. In this case, the lines evaluated in 2003-2013 composed the training set and were used to predict the lines evaluated from 2014 to 2016.

The script of the optimization algorithm is available from the corresponding author on request. In addition, we extended the generalized CD of contrasts to the multi-trait context to adapt the optimization criterion to specific prediction objectives (see Appendix).

V. Results

1. Trait variation and variance components

Variance components and repeatabilities of yield, phenology and traits linked to bread making quality are presented in Table 5.

Trait		Mean	Var	σ^2_E	σ^2_G	σ^2_{GxE}	σ^2_e	Plot repeatability	Design repeatability
Grain yield		91.34	269.67	236.06	14.05	26.55	9.27	0.28	0.84
Plant height		87.85	94.27	50.85	30.31	5.87	4.94	0.74	0.95
Heading date		140.98	63.04	45.47	8.08	$7{,}72 \times 10^{-4}$	1.93	0.81	0.98
Protein content		11.17	1.89	1.4	0.26	0.21	0.1	0.46	0.91
Chopin alveograph Parameters	**Dough strength:W**	205.63	3844.01	788.98	2030.03	---	1045.85	0.66	0.92
	Tenacity: P	77.46	695.31	166.1	413.3	---	117.67	0.78	0.95
	Extensibility: L	85.05	936.17	254.77	465.46	---	282.5	0.62	0.9
	P/L ratio	1.16	0.74	0.14	0.34	---	0.24	0.59	0.89
BIPEA Test	**Dough score**	79.24	215.21	27.78	69.5	---	120.13	0.37	0.77
	Crust score	50.18	483.88	96.94	145.1	---	249.32	0.37	0.77
	Crumb score	93.98	61.16	15.13	12.16	---	36.53	0.25	0.66
	Loaf volume	1482.44	48629.99	10647.39	15815.57	---	23975.46	0.4	0.79
	BMS	223.4	1276.6	141.13	466.23	---	688.96	0.4	0.8

Table 5 : Summary statistics, variance components and repeatabilities for the main traits and traits linked to bread making quality

σ^2_G; σ^2_{GxE} and σ^2_ε are the genotype, genotype x environment and residual variances respectively.

Plot repeatability ranged from 0.25 for crumb score to 0.81 for heading date. Design repeatability ranged from 0.66 for crumb score to 0.98 for heading date. Pearson's correlations between BLUEs of each trait were estimated (Supplementary Figure S1). As expected, yield was negatively correlated with protein content (r = -0.63). The Chopin alveograph parameters were also correlated with each other. Indeed, Tenacity (P) was highly negatively correlated with Extensibility (L) (r = -0.6), while W and P were highly positively correlated (r = 0.63). Loaf volume, dough, crust and crumb scores are components of the BIPEA test which mechanically explains their correlation with BMS (the final BIPEA score).

As W is the trait that is the more correlated to BMS (r = 0.45) and its plot repeatability and design repeatability is higher (0.66 and 0.92 respectively) than for BMS (0.40 and 0.80 respectively), we used it as a secondary trait in MT and TA models in further analyses.

2. Contribution of glutenin (HMW-GS) markers to genomic prediction predictive ability

The 11 KASP markers derived from HMW-GS loci *Glu-A1*, *Glu-B1*, and *Glu-D1* explained 31% and 8% of the phenotypic variation for W and BMS, respectively. The stepwise regression identified three markers explaining 30% of the W genetic variance and 4 markers explaining 8% of the BMS genetic variance, thus nearly as much as the 11 markers used together.

In order to compare ST, ST-glu and ST-glu_stepwise models, only the 200 lines that were phenotyped and genotyped with the 35K array and the KASP markers derived from HMW-GS loci were used for genomic predictions (Figure 16). For ST model with 200 lines, predictive ability was 0.49 for W and 0.28 for BMS. ST-glu and ST-glu_stepwise models had almost the same predictive ability for both traits (0.55 for W and 0.28 for BMS). They both improved the predictive ability of W but not that of BMS.

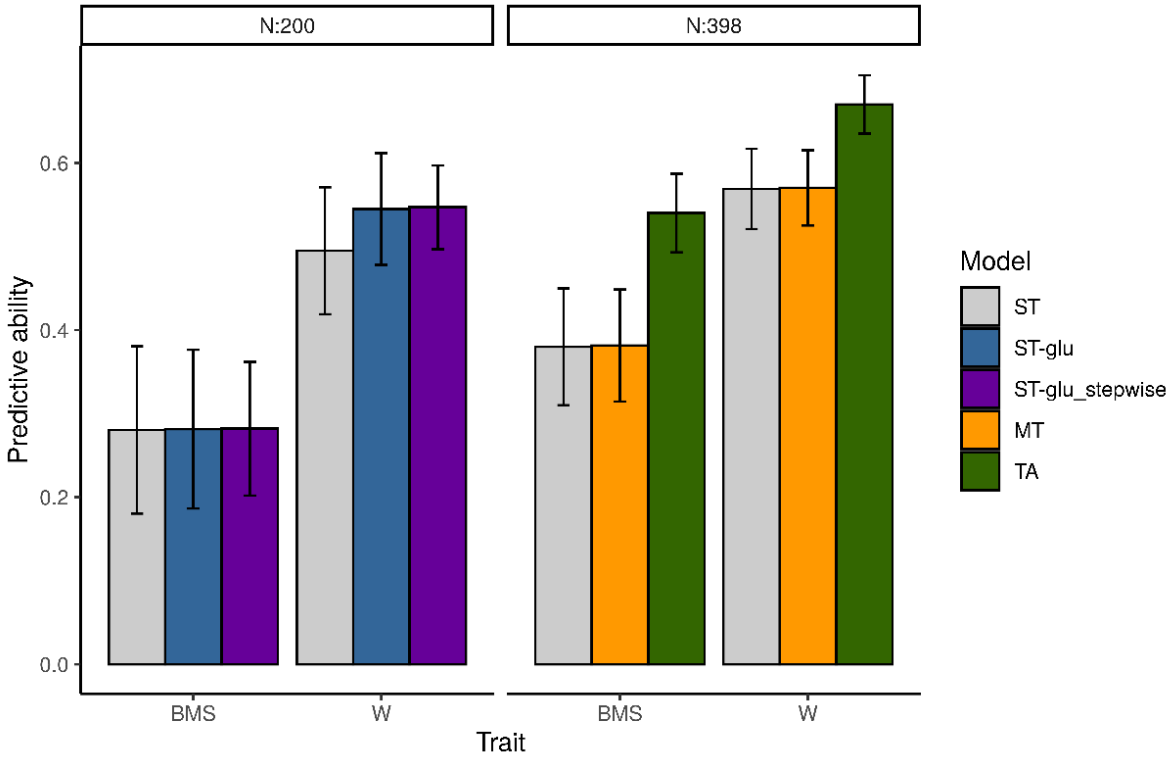

Figure 16: Contribution of HMW-GS markers to predictive ability of BMS and W and comparison of predictive ability using single-trait (ST), multi-trait (MT) and trait assisted (TA) genomic prediction models.

N: Number of lines. For the prediction of W, BMS was used as the correlated traits in bivariate models (MT and TA). For the prediction of BMS, W was used as the correlated traits in bivariate models (MT and TA). Error bars stand for standard deviations.

3. Single-trait versus multi-trait and trait-assisted genomic prediction models

Three models (ST, MT and TA) were tested using the 398 lines which were both genotyped and phenotyped for W and BMS. For ST model, predictive ability was 0.57 for W and 0.38 for BMS (Figure 16). Although MT model did not improve the quality of BMS predictions, TA model improved it (+0.15). In addition, we predicted W using BMS as correlated trait. MT model did not improve the quality of W predictions, unlike TA model. Besides, the gain in predictive ability generated by TA model was lower for W (+0.10) than for BMS (+0.15).

We compared MT and TA approaches with three different budgets and four different resource allocation strategies (Figure 17). TA method performed better in terms of predictive ability than MT method in each scenario in cross-validation and in forward prediction. For each of the three budgets, the predictive ability for BMS was higher when all the lines (of the training set in MT case and of the total population

in TA case) were phenotyped for the less expensive trait, W. The intermediate budget with at least 75 % of lines phenotyped for W was almost as accurate (0.5) than the expensive budget with 100% of lines phenotyped for W (0.51). This means that we can save 36% of the budget with this phenotyping strategy.

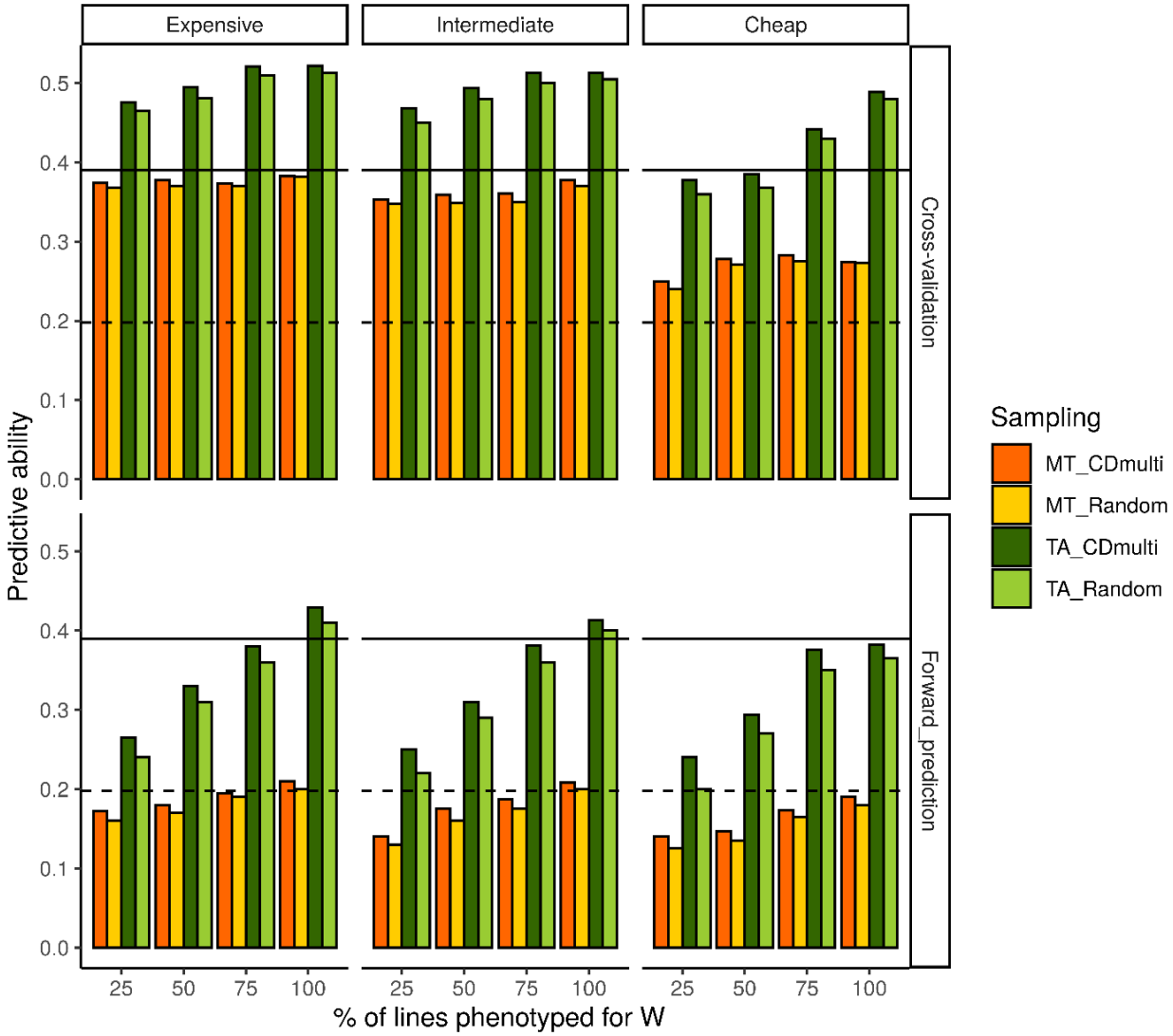

Figure 17 : Comparison of predictive ability for BMS in each scenario.

Results obtained with the 5-fold cross-validation method on the top. Results obtained when only the most recent lines are predicted on the bottom (Forward prediction). Solid lines correspond to predictive ability obtained using the ST model in cross-validation scenario. Dashed lines correspond to predictive ability obtained using the ST model in forward prediction scenario. Expensive, intermediate and cheap correspond to the three budgets. MT_CDmulti, TA_CDmulti: MT and TA approaches using the CDmulti to sample lines to phenotype, respectively. MT_Random, TA_Random: MT and TA approaches using a random sampling, respectively.

We used as a reference the predictive ability obtained with a ST model in which all the individuals from the training set were phenotyped for BMS (horizontal line in Figure 17). The cost of this ST approach is equal to the most expensive budget. For the intermediate budget, when all the lines from the training

set were phenotyped for W, the predictive ability of BMS with MT model was close to the predictive ability obtained with ST model and the budget was reduced by 36%. In addition, for each budget, the predictive ability for BMS obtained using the TA model was higher than the predictive ability obtained with the ST model when at least three quarters of the lines were phenotyped for W. This means that TA approach improved the predictive ability compared to ST and MT approaches (maximum gain: +0.14 in cross-validation and +0.21 in forward prediction when all the lines were phenotyped for W) and allowed to reduce the cost by up to 65%.

In addition, the ST model to predict BMS for the lines evaluated between 2014 and 2016 using older lines (evaluated from 2003 to 2013) as training set, showed a lower predictive ability than in cross-validation scenarios (-0.2). MT approach slightly improved the predictive ability compared with ST model, but it remained under 0.22. TA approach greatly improved this predictive ability (maximum predictive ability with random sampling: 0.41) whichever the budget. In addition, the predictive ability reached with TA approach in forward prediction was closed to the predictive ability obtained with ST approach in cross-validation scenario.

4. Selective phenotyping using Multi-trait CDmean criterion

To test whether we could minimize phenotyping or improve predictive ability by optimizing the choice of individuals to phenotype for BMS and/or W, we used a multi-trait CDmean criterion ($\overline{CD}_{multi}$) in an optimization algorithm. The multi-trait CD criterion allowed to improve slightly but systematically the predictive ability (+0.013 on average) for BMS compared to random sampling. The gain in terms of predictive ability allowed by the multi-trait CD criterion was slightly higher for TA approach in forward prediction strategy (+0.023 on average).

On average, the gain in predictive ability generated by TA model compared to MT model was +0.13 in cross-validation scenarios and +0.15 in forward prediction strategy. Note that for TA method using the multi-trait CDmean criterion, the number of lines phenotyped for W in the validation set was higher compared to random sampling (Supplementary Figure S2).

5. Impact of the cost ratio between W and BMS on BMS predictive ability

We evaluated the impact of the cost ratio between W and BMS on BMS predictive ability for MT and TA approaches. To do so, we compared predictive ability of BMS in several scenarios (Supplementary Table S2) by varying the cost ratio (ratio = 1/2, 1/3, 1/4, 1/5, 1/10 and 1/20). We fixed the budget allocated to phenotyping (the three budgets described above), the cost of phenotyping one line for BMS cost for one line (Cp_{BMS}=150€) and the percentage of the lines phenotyped for W (100% of the training set in MT approach and 100% of the total population in TA approach).

When the cost difference between the two traits was small and the budget allocated to phenotyping was low, the budget was not sufficient to phenotype each line for W (Supplementary Table S2). Therefore, only the results obtained with the intermediate and the expensive budget were shown in the Figure 18.

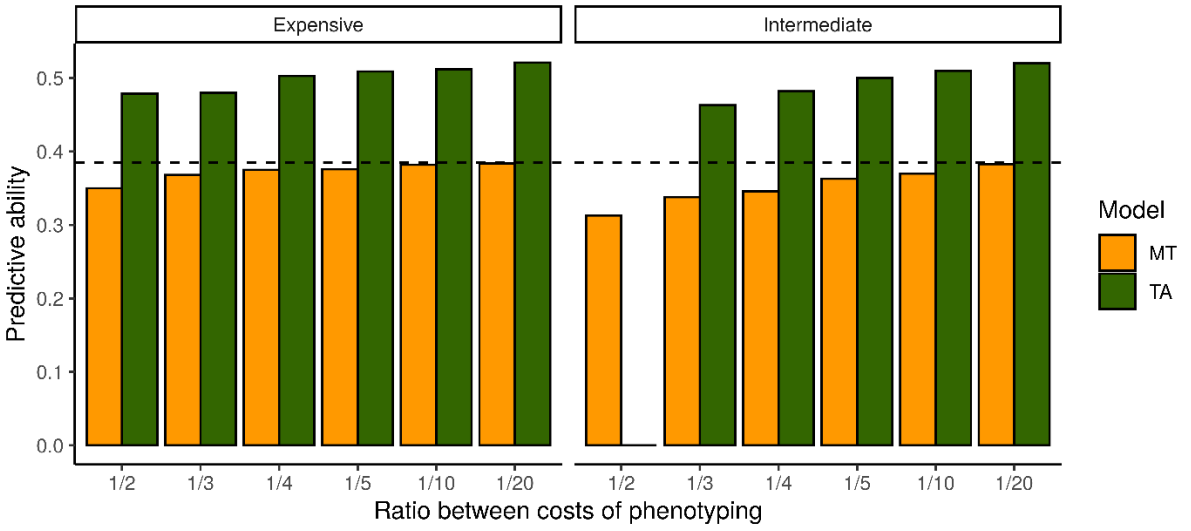

Figure 18: Impact of cost ratio between W (cheap) and BMS (expensive) traits on the predictive ability of BMS using multi-trait (MT) and trait assisted (TA) genomic prediction models.

The dashed line corresponds to predictive ability obtained using ST model. Ratio between phenotyping costs: $\frac{Cp_W}{Cp_{BMS}}$ with Cp_W W cost for one line and Cp_{BMS}BMS cost for one line. "Expensive" and "intermediate" correspond to the two budgets.

When the cost difference between the two traits increased, the number of lines phenotyped for BMS increased as well, and consequently the predictive ability of BMS was improved. Indeed for the expensive budget (which corresponds to the budget needed to phenotype each line of the training set for BMS), the predictive ability of BMS ranged from 0.35 when the ratio was 1/2 to 0.38 when the ratio

was 1/20 for MT approach and ranged from 0.48 when the ratio was 1/2 to 0.52 when the ratio was 1/20 for TA approach. For the intermediate budget, the same trends were found. We could not evaluate the predictive ability of BMS with the TA model when the ratio between the costs of phenotyping of each trait was 1/2 because N_{BMS} was too small and it led to convergence model issues. Even if the interest of using a TA approach rather than a ST approach was slightly lower when the cost difference decreased, TA model still performed better than ST approach when the ratio was 1/2 and for a fixed budget (gain in predictive ability: +0.09).

6. Impact of the number of phenotypic records on predictive ability

We evaluated the impact of the number of phenotypic records to estimate BMS and W BLUEs on BMS predictive ability for ST, MT and TA models (Figure 19). The number of phenotypic records for each line (except the control lines) ranged from 1 to 4. The three models were tested using the 322 lines which were both genotyped and phenotyped for W and for BMS in at least 4 environments.

We also evaluated the impact of control lines (evaluated 7 environments on average for BMS) to estimate BLUEs on BMS predictive ability. To do so, we compared two cases regarding controls: 1) all the phenotypic records of control lines were used to estimate BLUEs in order to have a more connected design; 2) we sampled the same number of control lines as candidate lines to calculate BLUEs.

For the three models, the predictive ability increased with the number of phenotypic records when we used all controls. The highest gain in terms of predictive ability was when the traits were evaluated in two environments compared to one (+ 0.11, + 0.08 + 0.07 for ST, MT, TA respectively). In addition, the predictive ability for using three records was slightly higher compared to two (+ 0.05 on average) and almost similar using four records (+ 0.02 on average) whichever the model used. For each scenario, the predictive ability was higher when all the phenotypic records of control lines were used to calculate BLUEs (+0.04, +0.04 +0.03 for ST, MT, TA respectively for 2 records). Note that the budget allocated to phenotyping can be significantly lower (-50 or 25% for 2 or 3 records). In addition, when both traits were measured in more than one environment, MT did not allow to improve predictive ability of BMS compared to ST model. But TA model always performed better than MT and ST models whichever the number of phenotypic records.

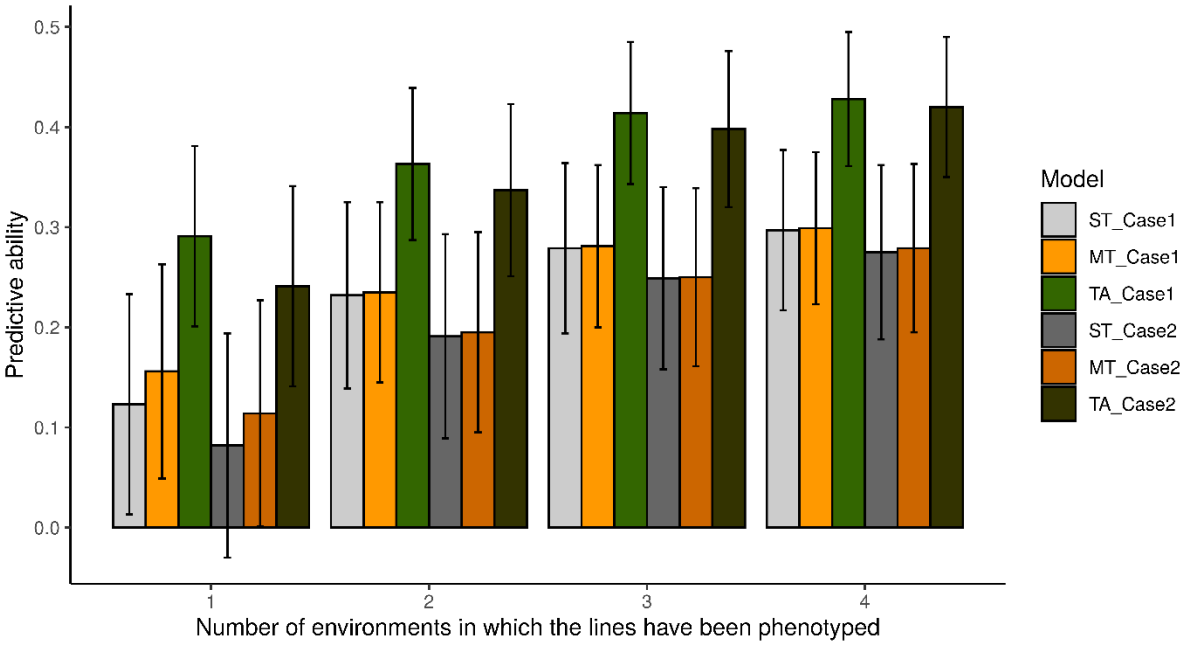

Figure 19: Impact of the number of phenotypic records on predictive ability of BMS

Case 1: all the phenotypic records of control lines were used to estimate BLUEs in order to have a more connected design. Case 2: phenotypic records of control lines were sampled in the same way as candidate lines to calculate BLUEs. Error bars stand for standard deviations.

VI. Discussion

This study focused on the comparison of different strategies to optimize resource allocation in breeding programs by using information from a cheap correlated trait to predict an expensive trait of interest. We used winter bread wheat baking quality (BMS) as a case study. Single-trait (ST), multi-trait (MT) and trait-associated (TA) strategies were compared in terms of predictive ability and cost. In multivariate approaches, information from dough strength (W) which is correlated (r = 0.45) to BMS and is less expensive to phenotype, was used. In MT analyses, only lines of the training set were phenotyped for W, whereas lines of the training and the validation sets were evaluated for W in TA analyses. In addition, we proposed and tested a multi-trait CDmean criterion (CDmulti) to optimize the choice of the set of lines to be phenotyped in the multi-trait context. To conduct this study, a public French breeding population of winter bread wheat lines was analyzed.

1. Ability of glutenin (HMW-GS) markers to predict BMS

In order to complement scarce phenotyping or pedigree information, Lande and Thompson (1990) proposed to rank individuals using molecular scores calculated as the sum of the effects of their alleles at selected QTLs. In this strategy, a threshold has to be chosen to determine the list of loci that significantly contribute to the trait variation. As it has been shown that allele effects were often overestimated in QTL detection (Beavis 1998), and that the infinitesimal model (Fisher 1918) that considers many loci of small effects was the best model to explain the variation of many quantitative traits such as yield, Meuwissen et al. (2001) proposed instead to use all available independent markers. To do so, phenotypic values are regressed in a linear model on all markers simultaneously, considering them as random effects with the same small genetic variance. The genetic values of individuals are estimated by best linear unbiased prediction. The critical difference with the Lande and Thompson approach is that this method does not require setting a significance threshold for the loci selected for trait prediction since all the markers are used in the model. In some cases however, explicitly modeling major QTLs as fixed effects in genomic prediction models can improve prediction accuracy in simulated data (Bernardo 2014) and experimental bread wheat data (Arruda et al. 2016; Michel et al. 2018; Sarinelli et al. 2019). In bread wheat, HMW-GS are important determinants of rheological properties of dough (reviewed by Anjum et al. 2007). Ravel et al. (2020) developed a set of KASP markers derived from glutenin DNA sequences to identify HMW-GS alleles. They showed that these markers detected more alleles and explained more rheological traits variation than SDS-PAGE method, which was

commonly used in plant breeding. In our study, we integrated 11 of these HMW-GS markers as fixed effects into ST models. It improved the predictive ability of W compared to ST model. These results are consistent with the conclusion of Michel et al. (2018) who found that prediction accuracy was increased for several dough rheological traits when *Glu-1* loci markers were including as fixed effects into the RR-BLUP model. But this model did not improve the predictive ability of BMS. This result might be due to the low percentage of genetic variance of BMS explained by the HMW-GS markers (8%). Bernardo (2014) actually demonstrated that only QTLs that explained at least 10% of the genetic variance could improve prediction accuracy.

2. Prediction of unphenotyped individuals (MT)

Our results were consistent with some previous studies on experimental data in which MT models did not improve the predictive ability for individuals that were neither phenotyped for the trait of interest nor for the secondary trait, as in pine tree (Jia and Jannink 2012), soybean (Bao et al. 2015), rye (Schulthess et al. 2016), maize (Dos Santos et al. 2016), wheat (Michel et al. 2018; Lado et al. 2018; Schulthess et al. 2018) and sorghum (Fernandes et al. 2018).

Previous studies showed that MT could be beneficial in terms of accuracy compared to ST model when a part of the individuals of the training set were not phenotyped for the targeted trait but the correlated trait was evaluated for a larger number of individuals (Guo et al. 2014; Michel et al. 2018). We found a similar trend in our study. In fact, when only half of the lines in the training set were phenotyped for BMS but all the lines (training and validation sets) were phenotyped for W, the predictive ability was similar compared to a ST model in which all the lines from the training set were phenotyped for BMS, but the budget was reduced by one third.

In the INRAE-Agri-Obtentions bread wheat breeding program, *F8* and *F9* lines are evaluated in 9 environments in France on average. BMS is evaluated in 7 environments on average for control lines and in 4 environments for most of the other lines. Those measurements are very expensive (150€) and in this study, we tested if we could decrease the number of records (or the number of environments) without affecting the predictive ability of BMS. As expected, predictive ability increased with the number of phenotypic records. The higher gap of predictive ability was observed when 2 phenotypic records instead of 1 were used to predict BMS with ST model (+0.08). However, with 3 phenotypic records instead of 4, the predictive ability was little affected (+0.02), but the budget allocated to phenotyping was reduced by 25%. The predictive ability was higher (+0.03) when all the phenotypic records of control lines were used to calculate BLUEs. Controls are essential in the breeding program in order to take into account *GxE* and it is important to evaluate them in a large number of environments.

By constrast, the number of records for other lines could be decreased with little effect on the predictive ability. Although phenotypic records were selected randomly for each line, we could further optimize the experimental design (Heslot et al, 2017) and control lines distribution in particular.

3. Prediction when (part of) the candidates are phenotyped for the correlated trait (TA)

TA method consists of predicting a trait using a multi-trait model when at least some of the candidates are phenotyped for one or several correlated traits. Several studies using experimental data demonstrated that TA models perform better than ST and MT models in terms of accuracy. TA models using high-throughput phenotyping for instance improved the prediction accuracy of bread wheat grain yield by up to 70% (Rutkoski et al. 2016; Sun et al. 2017; Crain et al. 2018). TA models also improved bread wheat baking quality related parameters using protein content (Michel et al. 2018) or dough rheological traits (Lado et al. 2018) as correlated traits. Predictive ability of *Fusarium* head blight severity in hybrid bread wheat was improved using plant height and heading date as correlated traits (Schulthess et al. 2018). Wheat is not the only specie for which TA approach has been tested, Fernandes et al. (2018) showed that TA models increased prediction accuracy by up to 50% when using plant height as correlated trait to predict yield in sorghum.

Our results were consistent with these studies, since predictive ability using TA model was higher than ST and MT predictive abilities whichever the studied scenario.

For each studied budget (low, intermediate, high), the predictive ability of BMS was higher when all lines (training and validation) were phenotyped for the less expensive trait, W. It seems more important to phenotype a maximum of lines with a highly heritable trait than just a few lines for an expensive and less heritable trait. For similar budget, TA allowed higher predictive ability than ST (maximum gain: +0.14). For similar predictive ability (approximately 0.38), TA allowed reducing the budget by up to 65%. So TA is a way to improve genetic gain by cost unit, either by reducing phenotyping cost and budget or increasing the program's size for the same budget.

We showed that the gain in predictive ability increased with the difference of phenotyping cost between the expensive and the cheap trait. These results highlighted the advantage of using a correlated trait as cheap as possible in multi-variate genomic prediction models. Nevertheless, TA model was still interesting compared to ST model with the smallest cost difference (ratio = 1/2) tested in this study (gain

in predictive ability: +0.09). In addition, even if it has no practical interest, we tried to predict W (plot repeatability = 0.66) using BMS (plot repeatability = 0.4) information and TA model, to compare the effect of the secondary trait's heritability (approximated by plot repeatability) on predictive ability. The gain was inferior to those when we predicted BMS using W. This is congruent with the hypothesis that the higher the heritability of the secondary trait compared to the trait of interest, the higher the gain in predictability (Jia and Jannink 2012; Hayashi and Iwata 2013; Guo et al, 2014).

4. Forward prediction

In order to study a situation more similar to the one the breeders are dealing with, we investigated a forward prediction strategy. ST predictive ability was almost divided by two when the oldest lines were used to predict the most recent lines compared to a cross-validation scenario. This discrepancy in predictive ability cannot be attributed to differences in the size of training sets. Indeed, the size of training set in cross-validation scenario was 318 lines vs 309 in the case of forward prediction. Storlie and Charmet (2013) also reported that forward prediction in wheat is less accurate than cross-validation. Since the youngest lines are not differently (or less) genetically related to the oldest lines than the oldest lines with each other (results not shown), the most likely explanation is that new candidates are evaluated in new environments (years and locations) that have never been observed. Although BLUEs were used to correct the environmental main effect, they did not correctly (efficiently) account for *GxE* interactions. Using TA by phenotyping some candidates with a correlated trait, we could take into account *GxE* in the new environment and better predict the candidates. This result suggests that TA model are particularly interesting to predict lines evaluated in new environments.

Moreover, a genomic Bayesian multi-trait and multi-environment model was developed by Montesinos-Lopez et al (2016) to take into account trait x genotype x environment interaction that could be interesting to test in that context.

5. Optimization of the training set based on a multi-trait CD criterion (CDmulti)

The size and the composition of the training population are key parameters of the accuracy of genomic prediction models. Optimizing the training set by selecting the most predictive individuals instead of using a random sampling leads to an increase of the predictive ability (Rincent et al. 2012; Akdemir et

al. 2015; Isidro el al. 2015; Rincent et al. 2017; Sarinelli et al. 2019). In this study, we specifically developed a multi-trait CDmean criterion (CDmulti), which slightly but systematically improved the predictive ability for BMS. The gain of predictive ability was a bit higher for TA approach than MT approach, in particular with the forward prediction. This may be due to the fact that $\overline{CD}_{multi}$leads to sample more lines to phenotype for W in the validation set (more related lines) than in the training set (Supplementary Figure S2). In this example the use of $\overline{CD}_{multi}$resulted in small increase of predictive ability, but we can suppose that it could be more valuable in datasets with different genetic structure, or in the situation of lower budget in which the choice of the phenotyped lines would be essential. Note that the criterion CDmulti can be adapted to particular prediction objectives (eg: optimize predictive abilities within families as in Rincent et al. 2017) by using the corresponding contrasts (see Appendix for the generalized version of CDmulti).

6. Conclusion

In this study, we addressed the question of resource allocation, which is hardly ever mentioned in multi-trait studies. We demonstrated how to decrease the breeding cost of an expensive trait by using the information of a cheap correlated trait using trait-assisted (TA) genomic predictions. TA models are useful tools for breeders to optimize resource allocation while maintaining the predictive ability for the traits of interest. In our study, it reduced the budget allocated to phenotyping by up to 65% for Bread Making Score (BMS). TA models can also be used to improve the predictive ability of the trait of interest (+0.14 using cross-validation and +0.21 using forward predictions) for a fixed budget. The less expensive and labor-intensive, and the more heritable the correlated trait is compared to the trait of interest, the more TA could be useful in a breeding program.

In addition, we highlighted the importance of control lines that are phenotyped in a large number of environments and which allow to decrease the number of phenotypic records (3 instead of 4) for the other lines with a small effect on BLUEs estimation.

The multi-trait CD criterion (CDmulti) specifically developed in this study allowed a slight but systematic improvement of BMS predictive ability. It would be interesting to evaluate this criterion in a larger and more diverse population. The tools and approaches developed in this study and exemplified using BMS prediction of bread wheat can be used in any species when relevant cheap correlated traits are available.

Author contribution statement

JA, FXO, BR and EH designed the field trials and collected the phenotypic data. RR supported in statistical analysis and in developing the multi-trait CDmean algorithm. CR developed KASP markers derived from HMW-GS loci. SBS analyzed the data and wrote the manuscript. SB and GG guided through the study and helped improving the manuscript. All authors approved the final manuscript.

Acknowledgements

The authors thank the genotyping platform GENTYANE at INRA Clermont-Ferrand (gentyane.clermont.inra.fr) which has conducted the genotyping. The work in experimental units by INRA (Clermont-Ferrand, Dijon, Estrées-Mons, Lusignan, Le Moulon and Rennes) and in Agri-Obtentions is also gratefully acknowledged.

Doctoral work of SBS was funded by a grant from the Auvergne-Rhônes-Alpes region and from the European Regional Development Fund (FEDER).

This work was supported by the Breedwheat project thanks to the funding from the French Government managed by the National Research Agency (ANR) in the framework of Investments for the Future (ANR-10-BTBR-03) France AgriMer and the French Fund to support Plant Breeding (FSOV).

Compliance with ethical standards

Conflict of interest

The authors declare no conflict of interest

VII. Conclusion

Cette étude nous a permis de comparer différentes stratégies d'optimisation de l'allocation des ressources dans un programme de sélection en utilisant l'information d'un caractère qui est à la fois corrélé au caractère d'intérêt et qui est moins cher à phénotyper. La note de panification (BMS) a été choisie comme caractère à prédire pour réaliser ces analyses. Il s'agit d'un caractère important pour les sélectionneurs car il est utilisé en France lors de l'inscription de nouvelles variétés pour évaluer la qualité boulangère des variétés. De plus, nous disposions du phénotypage d'un caractère corrélé au caractère à prédire BMS (r = 0.45) et moins coûteux à phénotyper : la force boulangère (W), un paramètre de l'alvéographe de Chopin.

Nous avons comparé des modèles de prédiction génomique mono-caractère (ST, ST-glu et ST-glu_stepwise) et multi-caractère (MT et TA). Les modèles ST-glu et ST-glu_stepwise prenaient en compte en effet fixe les données de génotypage de marqueurs dérivés des séquences d'ADN de gènes connus pour avoir un effet sur les caractéristiques rhéologiques de la pâte. Nous n'avons pas obtenu de prédictions génomiques plus précises pour BMS en utilisant les modèles ST-glu et ST-glu_stepwise à la place d'un modèle ST. Dans les modèles TA, les lignées de la population de validation (ou au moins une partie d'entre elles) étaient phénotypées pour W, le caractère le moins cher. Nous avons montré que la méthode TA permettait de réduire le budget alloué au phénotypage (cette réduction pouvant atteindre 65% du budget) tout en maintenant une précision des prédictions génomiques de BMS équivalente ou supérieure à celle obtenue avec le modèle ST avec le budget le plus élevé. Par ailleurs, pour un budget alloué au phénotypage fixe, l'approche TA permettait d'améliorer la précision des prédictions génomiques de BMS. Ce gain pouvait atteindre +0.14 dans le cas d'une méthode de validation croisée et +0.21 lorsque les lignées les plus anciennes servaient à calibrer le modèle permettant de prédire la performance des lignées les plus récentes (prédictions *forward*).

Nous avons également étudié l'impact du nombre de mesures phénotypiques par lignée sur la précision des prédictions. Nous avons montré que le budget alloué au phénotypage pouvait être réduit d'environ un quart sans réduire la précision des prédictions de BMS en réduisant le nombre d'essais sans lesquels est évaluée chaque lignée. Nous avons également noté l'importance des lignées témoins, qui sont phénotypées dans un grand nombre d'environnement, pour maintenir une précision des prédictions satisfaisante.

Par ailleurs, nous avons étendu le critère du CDmean à un contexte multi-caractère (CDmulti) afin d'optimiser le choix des lignées à phénotyper pour chacun des deux caractères afin d'améliorer la précision des prédictions et / ou de réduire le budget alloué au phénotypage. Nous avons montré que

cette méthode permettait d'obtenir des prédictions légèrement plus précises pour BMS dans l'ensemble des scénarios testés (les différents scénarios correspondent à différentes valeurs de budget et à plusieurs façons de répartir le budget entre le phénotypage de BMS et celui de W).

L'approche TA ainsi que l'utilisation du CDmulti peuvent être généralisées à d'autres espèces et à d'autres caractères. Cependant, il est important de noter que les avantages apportés par l'approche TA, que ce soit en termes d'amélioration de la précision des prédictions ou en termes de réduction des coûts liés au phénotypage, dépendent de la corrélation entre les caractères considérés et du rapport entre les coûts de phénotypage de chacun des deux caractères.

Chapitre 4 :

Comparaison de schémas de sélection simulés

I. Préambule

Ce chapitre est présenté sous la forme d'un article scientifique qui en cours de préparation.

Cet article porte la comparaison différents schémas de sélection de blé tendre simulés afin d'évaluer l'apport des prédictions génomique en termes de gain génétique à court et plus long terme et en termes d'évolution de la diversité génétique pour un budget donné. Contrairement au chapitre précédent, le budget défini dans cet article intègre le coût de chacune des étapes du schéma de sélection (coûts des croisements, de production des haploïdes doublés, de multiplication des semences, de génotypage et d'évaluation au champ).

Nous avons simulé deux types de schémas : un schéma de sélection phénotypique (PS) avec deux étapes de sélection basées sur des essais en parcelles, et un schéma de sélection combinée phénotypique et génomique (GPS) avec une première étape de sélection génomique et une seconde étape de sélection basée sur des essais en parcelles. Dans le cas des schémas PS, les descendants étaient obtenus en croisant de façon aléatoire les meilleures lignées issues du cycle précédent. Pour les schémas GPS, nous avons comparé deux stratégies pour obtenir les croisements : soit ils étaient obtenus de la même façon que dans les schémas PS, soit ils étaient optimisés en utilisant un critère d'utilité basé sur les données de génotypage des parents.

Nous avons défini 36 scénarios faisant varier différents paramètres (le budget alloué au programme de sélection, l'intensité de sélection à chaque étape, le coût de génotypage et de l'héritabilité du caractère) afin d'évaluer l'impact de ces paramètres sur l'évolution du gain génétique et de la diversité génétique pour les différents types de schémas PS et GPS.

Nous avons développé un *pipeline* pour réaliser ces simulations et ces analyses.

.

Integration of genomic selection into bread wheat breeding schemes: a simulation pipeline including economic constraints

S Ben-Sadoun[1], Furgeray-Scarbel A[2], FX Oury[1], E Heumez[3], B Rolland[4], J Auzanneau[5], G Charmet[1], Lemarie S[2], S Bouchet[1]

[1] *INRAE - UCA UMR1095, Genetics Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu 63000 Clermont-Ferrand - France*

[2] *Laboratoire d'Economie Appliquée de Grenoble (GAEL), 241 Rue des Résidences, 38400 Saint-Martin-d'Hères - France*

[3] *INRAE UE Lille, 2 chaussée Brunehaut, Estrées-Mons, BP 50136, 80203 Peronne Cedex, France*

[4] *INRAE - Agrocampus Ouest Rennes- Université Rennes 1 UMR 1349 IGEPP BP35327, 35653 Le Rheu Cedex, France*

[5] *Agri-Obtentions, Ferme de Gauvilliers 78660 Orsonville - France*

Corresponding author:

Sophie Bouchet

INRAE-UCA UMR1095, Genetics Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu 63000 Clermont-Ferrand – France

sophie.bouchet@inra.fr

ORCID: 0000-0001-5868-3359

II. Abstract

Little effort has been made to optimize the allocation of resources using genomic predictions to maximize the economic return to investment in terms of genetic gain per unit time and money in breeding programs. We built a simulation pipeline designed to become a decision tool to help breeders adjusting breeding schemes according to their either short or long term objectives. We used it to explore different scenarios in order to investigate under which conditions (at what step of the breeding program) genomic predictions could improve genetic gain. We compared 36 scenarios for an identical budget per year, varying strategy (phenotypic selection PS or genomic selection + phenotypic selection: GPS), budget, trait heritability, relative selection intensity at two key steps and genotyping cost. With GPS strategy, we also optimized mating using genomic predictions. The pipeline is developed in the R environment. We illustrate it here using bread wheat historical data, with 3 cycles of 5 years selection. We showed that GPS selection using mating optimization improved significantly genetic gain for all scenarios while GPS and PS had similar efficiency.

Key words:

Breeding scheme, genomic prediction, mating optimization, economic constraint, bread wheat.

III. Introduction

The objective of bread wheat breeding programs is to develop new varieties that outperform current varieties in terms of yield, adaptation, resistance to biotic and abiotic stresses, and/or end use qualities. A great challenge in plant breeding is to improve the genetic gain per unit of time with a limited budget. To meet this goal, the optimization of resource allocation appears to be a key point. In addition, breeders must find a compromise between short-term genetic gain and the conservation of genetic diversity within the breeding program in order to guarantee long-term genetic gain (Jannink 2010).

Furthermore, the exponential decrease of genotyping costs, the improvement of computing tools, data storage capacities and algorithms efficiency, and the development of new statistical methods have led to the development of a new powerful approach to optimize breeding schemes: genomic selection (GS). GS is a marker-based selection method that uses thousands to millions of molecular markers evenly spread along the genome to predict the genetic value of candidates to selection (Whittaker et al. 2000; Meuwissen et al. 2001). According to the breeder's equation (Lush 1937), GS could improve genetic gain by (i) shortening the breeding cycle by replacing field evaluation with genotyping at juvenile stage (reviewed by Crossa et al. 2017) or introgressing exotic alleles (for example using bi-parental populations with 3 steps of GS accelerated selection cycles in greenhouse per year in maize instead of one year of phenotypic selection; Bernardo et al. 2009; Comb et al. 2013), (ii) decreasing the budget allocated to field evaluation by optimizing the number of genotypes and replicates per environment in the experimental design (Heslot et al. 2017) and by this way increasing the size of the breeding program (number of crosses or progenies per cross), (iii) increasing genetic variance by optimizing mating to cumulate favourable alleles (Zhong and Jannink 2007; Akdemir and Sanchez 2016; Lehermeier et al. 2017; Allier et al. 2019), (iv) increasing accuracy of selection. Several factors influence the accuracy of GS. These factors include trait architecture and trait heritability (Daetwyler et al. 2008), training set size and composition (Jannink et al. 2010; Lorenz et al. 2011), congruency between the allelic composition represented in the training population to estimate marker effects, and the allelic composition of the candidates whose performance is to be predicted (Habier et al. 2007; Charmet et al. 2014; Crossa et al. 2014), marker density (Calus and Veerkamp 2007; Solberg et al. 2008; de Roos et al. 2009), and statistical model for estimation of marker effects (Desta and Ortiz 2014). .

Most of previous researches on GS in plant breeding focused on the prediction accuracy of unphenotyped lines. In bread wheat, studies evaluated the prediction accuracy of grain yield (including Poland et al. 2012; Lado et al. 2013; Storlie et Charmet 2013; Zhao et al. 2015; Norman et al. 2017), traits linked to bread making quality (including Battenfield et al. 2016; Guzman et al. 2016; Liu et al. 2016; Hayes et al. 2017; Lado et al. 2018; Michel et al. 2018), or disease resistances (including Ornella

et al. 2012; Rutkoski et al. 2012; Daetwyler et al. 2014; Arruda et al. 2015). At the breeding program scale, some simulation work show the interest of GS compared to classical phenotypic evaluation in terms of genetic gain. For example, Rutkoski et al. 2014 showed that GS accuracies were sufficiently high (GS accuracy twice higher than PS accuracy) to achieve greater gain from selection per unit time compared with phenotypic selection for bread wheat adult stem rust resistance. But they did not compare the two strategies at fixed cost. Riedelsheimer and Melchinger (2013) compared the efficiency of PS and GS for grain yield assuming a single selection cycle and a given budget using a biparental population of maize DH lines. They showed that with large GxE interactions and under limited resources, it was beneficial to use an index combining PS and GS to maximize genetic gain. They also noticed that DH price was a limiting factor for large genetic gain.

Simulation studies enable the comparison of a wide range of scenarios that would not be possible to test experimentally. They allow to evaluate an unlimited number of selection cycles (long-term selection) in a short amount of time with a cost limited to data storage.

In this study we compared the genetic gain and the evolution of genetic diversity in two types of simulated breeding schemes: one called Phenotypic Selection (PS) with two steps of selection based on field trials, and one called Genomic and Phenotypic Selection (GPS) that combines a first step of genomic selection and a second step of selection based on field trials. We explored different scenarios in order to investigate under which conditions GPS was more cost-effective than PS. We compared scenarios for a given budget.

IV. Materials and methods

We developed a pipeline to simulate and compare breeding programs in terms of genetic gain and level of diversity for a given budget (Figure 20).

1. Data set

The pipeline was tested using a real breeding population of 700 winter-type bread wheat lines developed by the Institut National de la Recherche en Agriculture, Alimentation et Environnement (INRAE) and its subsidiary company Agri-Obtentions. These lines were selected between 2000 and 2013. Each line was genotyped using a 280K SNP array (Rimbert et al. 2018). In order to limit computing time, we used a subset composed of 12119 SNP evenly spread along the genetic map.

2. Simulation of the breeding programs

We compared two types of breeding schemes: one called Phenotypic Selection (PS) with two steps of selection based on field trials, and one called Genomic and Phenotypic Selection (GPS) that combines a first step of genomic selection and a second step of selection based on field trials (Figure 1). Both breeding programs were designed using doubled haploids (DHs) to reduce the time required for inbred development. We counted three years for cross, DH from F1 productions and seed multiplication, one year of PS (or GPS) selection and a last year of phenotypic selection. For both PS and GPS approaches, a breeding cycle lasts five years. We simulated three successive cycles.

We start the simulation with N_{ref} lines (700 in our example) that have been genotyped and phenotyped. For GPS, the first vector of marker effects is calculated using a ridge regression on those phenotypes. The database of phenotypes is incremented at each phenotyping step so that the vector of marker effects is updated at each cycle. At the first step of each cycle, N_c crosses are produced among the N_p best individuals from step 4 of previous cycle or N_{ref} individuals for the first cycle. In PS scheme, crosses are obtained by randomly mating the N_p individuals. For the GPS strategy, we can choose between two mating strategies. The first method is random as for PS scheme. The second method optimizes the complementarity between parental alleles (Daetwyler et al. 2015). We calculate the value of a cross as the value of the individual that would have inherited the best chromosomes of its parents.

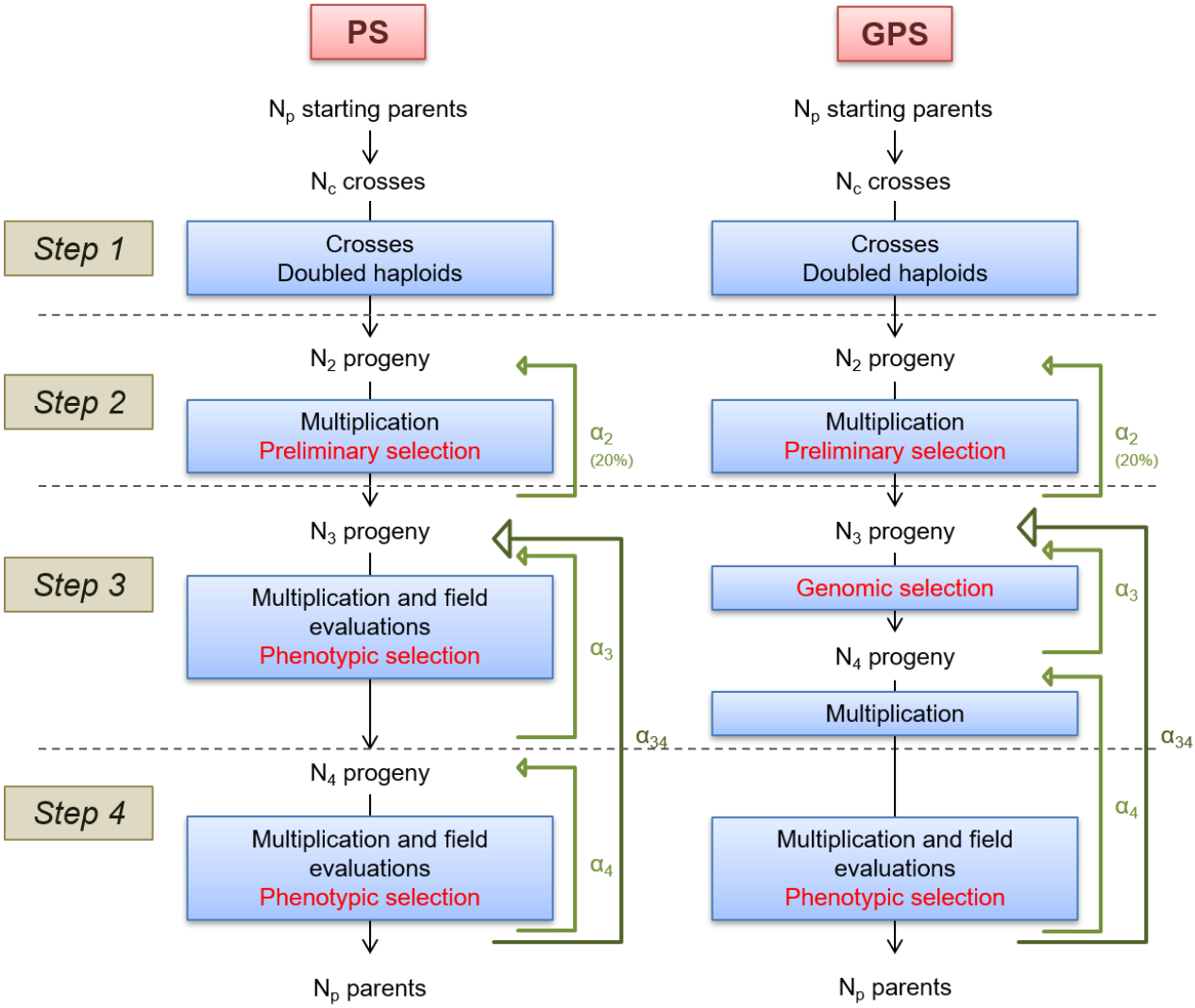

Figure 20: PS and GPS breeding schemes.

N_p and N_c: number of parents and crosses respectively. N_2, N_3 and N_4: number of progenies at the beginning of steps 2, 3 and 4 respectively. α_2, α_3 and α_4: selection rate on steps 2, 3 and 4 respectively. α_{34}: global selection rate on steps 3 and 4.

At step 2, N_2 progeny are multiplied for a one trial cost. We considered that the selection was not for the targeted traits but for multiple agronomic visual criteria, disease in particular. So the selection was not modelled, it was considered random for the targeted trait.

At step 3, N_3 progeny are simulated.

$$N_3 = N_2\alpha_2 \qquad (17)$$

with α_2 the selection rate at step 2 ($\alpha_2 = 0.2$ in this study).

Each chromosome of each progeny is either parental or recombined. If recombined, the number of cross-overs on each chromosome is sampled from a Poisson distribution. The cross-overs are distributed uniformly along the genetic map.

In PS scheme, these individuals are evaluated in the field and we consider a multiplication cost corresponding to the seed production of five trials. Selection is based on that phenotypic data with an intensity of selection α_3. In GPS scheme, the selection is based on GEBVs calculated as the cross product between the vector of marker effects and the matrix of genotypes of progenies, excluding markers to which QTLs were assigned. N_4 ($N_3\alpha_3$) progeny are multiplied, the cost of seed production corresponding to ten trials. The last step is a phenotypic selection of N_p parents for the next cycle in both PS and GPS schemes with a selection rate α_4.

Instead of fixing α_3 and α_4, we compared different rates λ of selection intensity between steps 3 and 4, so that λ is defined as:

$$\alpha_3 = {\alpha_{34}}^{\lambda} \qquad (18)$$

and

$$\alpha_4 = {\alpha_{34}}^{1-\lambda} \qquad (19)$$

With:

$$\alpha_{34} = \alpha_3\alpha_4 = \frac{N_P}{N_2\alpha_2} \qquad (20)$$

If λ is higher than 0.5, the intensity of selection is higher at step 3 than step 4, the opposite if λ is lower than 0.5. If $\lambda = 0.5$, the intensity of selection is equal at steps 3 and 4.

3. Costs modelling

We defined C_x the cost of each operation X of the breeding program (Table 6). The cost of an operation may depend on the step where the operation was done. The step is specified as an exponent. For example, we assumed that field evaluations were realized in one trial at step 2, five trials at step 3 and ten trials at step 4, which explains the different costs of seed multiplication and evaluation.

Notation	Definition	Value
$\mathbf{C_C}$	Cost for one cross	100€
$\mathbf{C_{DH}}$	Cost for each doubled haploid	37€
$\mathbf{C_M^2}$	Cost for multiplication for one line at step 2	10€
$\mathbf{C_M^3}$	Cost for multiplication for one line at step 3	50€
$\mathbf{C_M^4}$	Cost for multiplication for one line at step 4	100€
$\mathbf{C_P^3}$	Cost for field evaluation of one line at step 3	100€
$\mathbf{C_P^4}$	Cost for field evaluation of one line at step 4	500€
$\mathbf{C_G}$	Cost to genotype one line	10€ or 37€ depending on the scenario

Table 6 : Operation costs

We defined the total cost of the PS scheme using the following equation:

$$CT_{PS}(N_P, N_C, \lambda, n_2) = K[N_C C_C + N_2(C_{DH} + C_M^2 + \alpha_2(C_M^3 + C_P^3)) + N_4(C_M^4 + C_P^4)] \quad (21)$$

with K the number of cycle (K = 3 in this study) and n_2 the progeny number per cross at the beginning of step 2. In our study, n_2 was considered uniform. But note that the pipeline offers the possibility to make it proportional to cross value (Usefulness Criterion).

Thanks to equation 4, we have

$$N_4 = \frac{N_P}{\alpha_4} = \frac{N_P}{\alpha^{1-\lambda}} = \frac{N_P}{\left(\frac{N_P}{N_2\alpha_2}\right)^{1-\lambda}} = N_p^{\lambda}(N_2\alpha_2)^{1-\lambda} \quad (22)$$

Therefore, the total cost of the PS scheme can be defined as follows:

$$CT_{PS}(N_P, N_C, \lambda, n_2) = K\left[N_C C_C + N_2(C_{DH} + C_M^2 + \alpha_2(C_M^3 + C_P^3) + N_p^{\lambda}(N_2\alpha_2)^{1-\lambda}(C_M^4 + C_P^4)\right] \quad (23)$$

To define the total cost of the GPS scheme (CT_{GPS}), we introduced the genotyping cost of the reference population composed of N_{ref} lines. Note that cost of field evaluation at the third step was replaced by the cost of genotyping and that only N_4 lines that were selected at step 3 were multiplied for GPS (instead of N_3 for PS).

We defined CT_{GPS} using the following equation:

$$CT_{GPS}(N_P, N_C, \lambda, n_2) = N_{ref}C_G + K\left[N_C C_C + N_2\left(C_{DH} + C_M^2 + \alpha_2 C_G\right) + N_p^{\lambda}(N_2\alpha_2)^{1-\lambda}(C_M^3 + C_M^4 + C_P^4)\right] \quad (24)$$

As genotyping cost is inferior to phenotyping cost, the number of progenies will be larger in GPS strategy. As we fix the total cost, either the number of crosses or the number of progenies per cross will be different between PS and GPS schemes. For GPS scheme, we tested a scenario called GPS.N_C with the same number of crosses and a different number of progenies n_2 per cross and a scenario called GPS.n_2 with the same number of progenies per cross and a different number of crosses.

4. Trait simulation

For this study, 20 random samples of 100 SNPs were assigned as QTL with additive effects following a normal distribution. The favorable allele was attributed at random to one of the two SNP alleles, so that coupling and repulsion associations also occur at random. The entry-mean heritability (h^2) was set to either 0.2, 0.4 or 0.7.

In the pipeline, phenotypic values of lines are obtained by adding normally distributed error to the genotypic values. In our study, the residual variance was 80% (h^2 = 0.2), 60% (h^2 = 0.4) or 30% (h^2 = 0.7) of phenotypic variance.

5. Genomic prediction models

Selection at step 3 based on field trials in PS was replaced by a step of genomic selection in GPS. The true genotypic value (TBV) or Genomic Estimated Breeding Value (GEBV) of each line are calculated as the sum of simulated allelic effects or estimated allelic effects across all loci respectively.

Marker effects are estimated using a single-trait RR-BLUP model implemented in the ‘rrBLUP’ R package (Endelman 2011). The matrix of genotypes and the vector of phenotypes are incremented at each step 4 so that prediction equations (marker effects) were re-evaluated at each cycle of the GPS breeding program.

In order to predict the value of each potential cross, we estimated a usefulness criterion corresponding to the value of the best possible progeny according to Daetwyler et al. (2015):

$$UC_{ij} = \sum_{k=1}^{nchr} \max(GEBV_{ki}, GEBV_{kj}) \qquad (25)$$

With UC_{ij} the usefulness criterion of the cross between the i-th and the j-th parents, and k the chromosome number. This UC assumes chromosome being inherited without recombination, which is true for half of the chromatids in a single meiosis, as it is the case for doubled haploids from F1.

6. Simulations of different scenarios

We evaluated the impact of several parameters on the final genetic gain. To do so, we simulated breeding programs with two total costs for 15 years (three cycles of five years; CT ∈{22 500 000€, 45 000 000€}, i.e. CT per year of 1.5 or 3M euros), two genotyping cost (CG ∈{37€ , 10€}), three relative selection intensity λ between step 3 and step 4 (λ ∈{ 0.25, 0.5, 0.75}), and three levels of heritability of the trait (h^2 ∈{0.2, 0.4, 0.7}). Description of each scenario is available in supplementary Table S3.

It led to the evaluation of 36 scenarios for five different strategies (PS, GPS.N_C: fixed number of crosses, GPS.n_2: fixed number of progenies per cross, GPSopt.N_C: optimized mating design with a fixed number of crosses, GPSopt.n_2: optimized mating design with a fixed number of progenies per cross). For each strategy / scenario, we tested 20 simulated traits (100 QTLs randomly sampled for each simulated trait) to evaluate the variance due to different QTL positions.

For each strategy / scenario / trait, we ran the algorithm 10 times to estimate the variance due to mendelian sampling only.

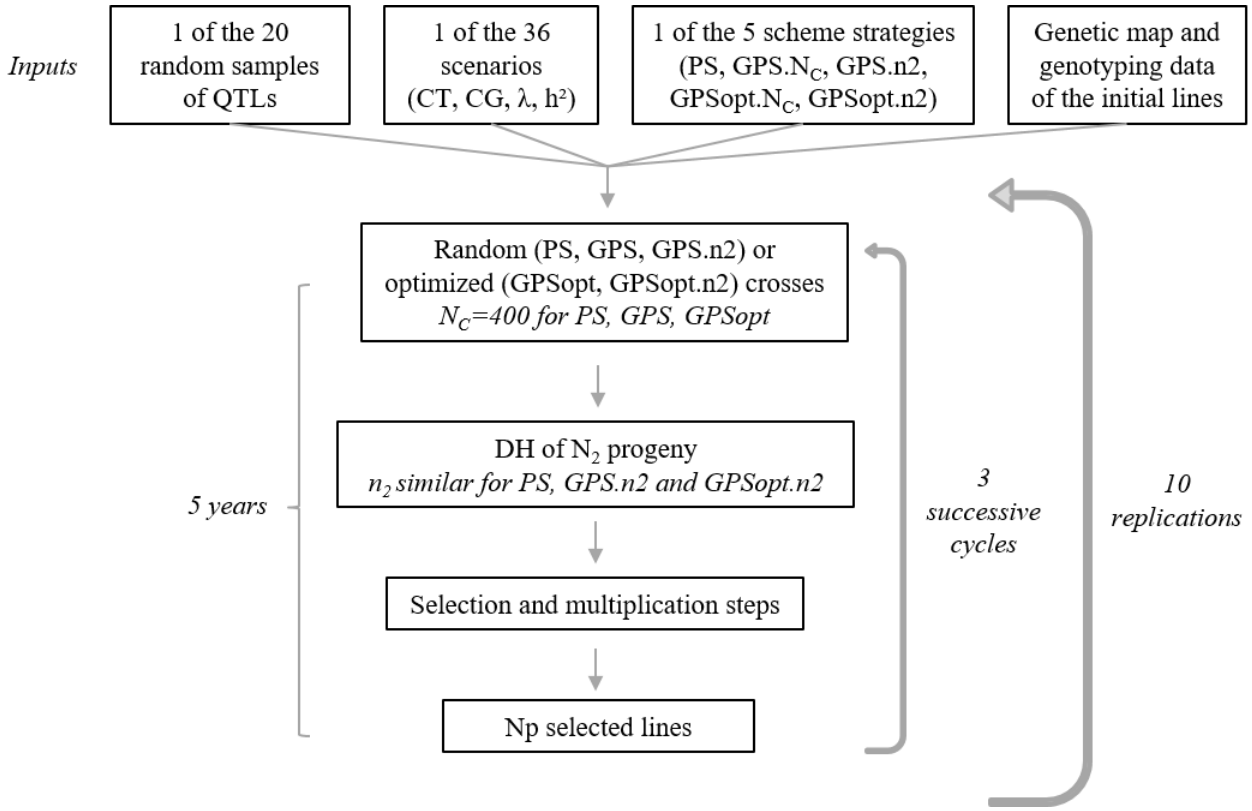

Figure 21 : Description of the simulations.

CT, CG: Total cost and genotyping cost respectively. NC: Number of crosses. λ: Relative selection rate at step 3 compared to step 4. h^2: Heritability. DH: Doubled haploids. N_2: Progeny number at the beginning of step 2.

As illustrated in Figure 21, the algorithm used to run the simulations required several input data: one vector including true marker effects for the simulated trait, one scenario (combination of CT, C_G, λ and h^2 values), the strategy (PS, GPS.N_C, GPS.n2, GPSopt.N_C, GPSopt.n2), a matrix with chromosome number and genetic position in columns 1 and 2 respectively, a matrix of genotypes with N_{ref} rows, a vector of phenotypes of N_{ref} length. The first step of the algorithm consisted of obtaining crosses: by randomly coupling the best individuals (PS, GPS.N_C and GPS.n2) or by optimizing crosses (GPSopt.N_C and GPSopt.n2). For PS, GPS.N_C and GPSopt.N_C the number of crosses was similar. DHs were obtained at the following step for N_2 progeny. N_2 was similar for PS, GPS.n2 and GPSopt.n2. Then, selection and multiplication steps were simulated depending on the type of breeding scheme (PS or GPS) as described in Figure 20. These steps resulted in selecting N_p lines that were used as parents for crosses in the second cycle. For each simulation, we performed three cycles of five years (i.e. 15 years for one breeding scheme).

7. Analysis of simulation results

We evaluated the genetic gain achieved for each strategy and scenario by evaluating the difference between the true genetic value (TBV) of the 200 best lines at the end of the third cycle and the TBV of the starting reference population. The variance components of the input parameters, i.e. strategy, total cost, trait heritability, QTL sampling, and relative intensity of selection at step 3 compared to step 4 (λ), were tested by ANOVA analyses conducted with the linear model (lm) function of R (R Core Team, 2016). F tests were considered significant at $\alpha = 0.05$. For GPS strategies, we also studied the impact of genotyping cost on genetic gain. For the different strategies, genetic gains for pairwise scenarios were compared by the Student-Newman-Keuls test function from the R library Agricolae. Means were judged significantly different when the P-values of the Student-Newman-Keuls test were < 0.05. We also represented the evolution of the cumulative genetic gain (the difference between the TBV of the 200 best lines at the end of the cycle and the TBV of the starting reference population) and the evolution of genetic gain at each cycle (the difference between the TBV of the 200 best lines at the end of the cycle and the TBV of the 200 best lines at the end of the previous cycle).

Furthermore, we analysed the evolution of genetic diversity over successive cycles for each strategy and scenario. To do so, we measured the percentage of alleles present in both the reference population and the 200 best progenies of the last cycle. As for genetic gain comparison between strategies / scenarios, the significance of the variance components of the different parameters were tested by ANOVA.

For GPS strategy, a usefulness criterion was used to select best crosses. Note that in our example, we did not constrain parents to any contribution limitation. Therefore, we studied the contribution of initial parents for each strategy and scenario. To do so, we focused on the pedigree of lines selected at the end of the cycle and we recorded the average number of progenies to which each line of the reference population contributed as parents.

V. Results

1. Population size under different strategies and scenarios

For each strategy (PS, GPS.N_C, GPS.n2, GPSopt.N_C, GPSopt.n2) and scenario (CT, C_G, λ), we calculated the number of progenies at each step and scenarios. Results are shown in Table 7.

Breeding	Average annual	C_G	λ	N_c	N_2	N_3	N_4
PS	1 500 000€	---	0.25	400	62 317	12 463	4 436
			0.5	400	82 710	16 542	1 819
			0.75	400	92 056	18 411	620
	3 000 000€	---	0.25	400	133 186	26 637	7 841
			0.5	400	173 744	34 749	2 636
			0.75	400	188 511	37 702	741
GPS.N_C and GPSopt.N_C	1 500 000€	10€	0.25	400	80 738	16 148	5 387
			0.5	400	122 798	24 560	2 216
			0.75	400	143 022	28 604	692
		37€	0.25	400	75 666	15 133	5 131
			0.5	400	111 716	22 343	2 114
			0.75	400	128 921	25 784	674
	3 000 000€	10€	0.25	400	176 686	35 337	9 692
			0.5	400	262 291	52 458	3 239
			0.75	400	294 270	58 854	828
		37€	0.25	400	164 883	32 977	9 203
			0.5	400	237 977	47 595	3 085
			0.75	400	265 198	53 040	807
GPS.n2 and GPSopt.n2	1 500 000€	10€	0.25	517	80 595	16 119	5 380
			0.5	592	122 448	24 490	2 213
			0.75	620	142 581	28 516	691
		37€	0.25	484	75 569	15 114	5 126
			0.5	539	111 486	22 297	2 112
			0.75	559	128 633	25 727	674
	3 000 000€	10€	0.25	530	176 514	35 303	9 685
			0.5	603	261 908	52 382	3 237
			0.75	624	293 817	58 763	828
		37€	0.25	495	164 767	32 953	9 198
			0.5	548	237 724	47 545	3 084
			0.75	562	264 903	52 981	807

Table 7 : Population size under different strategies and scenarios.

CT: annual cost of the breeding program. CG: individual genotyping cost. λ: Relative selection rate at step 3 compared to step 4. N_C, N_2, N_3, N_4: Number of crosses and number of progenies at the beginning of steps 2, 3 and 4 respectively

When the budget CT was doubled, the number of progenies was 1.9 times superior at each step on average (for given values of C_G and λ). When the cost to genotype one line (C_G) decreased from 37 to 10 euros, the number of progenies at each step was 1.1 times superior at each step on average (for given values of CT and λ).

In addition, for given values of CT and C_G, the higher the value of λ was, the higher the population size in step 2 and the smaller the population size in step 4.

2. Cost repartition between operations

We estimated the percentage of the total budget allocated to each operation (crosses, DH, multiplication, field experiment and genotyping) under different strategies (PS, GPS.N_C, GPS.n2, GPSopt.N_C, GPSopt.n2) and scenarios (CT, C_G, λ). Results are shown in Table 8.

For each strategy, the production of DHs was the major expense. Indeed, for PS strategy between 30.8% and 46.5% of the global budget was allocated to DH production. For schemes with a step a genomic selection, this percentage was higher (between 37.3% and 72.6%).

Multiplication steps required on average between 20 and 25% of the global budget. Since one field evaluation was more expensive at step 4 than at step 3, scenarios with a higher number of lines in the last step (λ<0.5) used more money for field evaluations.

For strategies using genotyping data, the genotyping required between 2.2% and 13.1% of the global budget. The part of budget allocated to genotyping was higher when the cost of genotyping was 37€.

Breeding scheme	Average annual CT	C_G	λ	% percentage of the budget allocated to				
				crosses	DH	multiplication	field experiment	Genotyping
PS	1 500 000€	---	0.25	0.5	30.8	22.5	46.2	0
			0.5	0.5	40.8	24.5	34.2	
			0.75	0.5	45.4	25.3	28.7	
	3 000 000€	---	0.25	0.3	32.9	23.0	43.8	
			0.5	0.3	42.8	24.9	32.0	
			0.75	0.3	46.5	25.6	27.6	
GPS.N_C and GPSopt.N_C	1 500 000€	10€	0.25	0.5	39.9	21.5	35.9	2.2
			0.5	0.5	60.6	20.8	14.8	3.3
			0.75	0.5	70.6	20.5	4.6	3.8
		37€	0.25	0.5	37.3	20.4	34.2	7.6
			0.5	0.5	55.1	19.1	14.1	11.2
			0.75	0.5	63.6	18.6	4.5	12.8
	3 000 000€	10€	0.25	0.3	43.6	21.4	32.3	2.4
			0.5	0.3	64.7	20.7	10.8	3.5
			0.75	0.3	72.6	20.4	2.8	3.9
		37€	0.25	0.3	40.6	20.2	30.7	8.2
			0.5	0.3	58.7	18.9	10.3	11.8
			0.75	0.3	65.4	18.5	2.7	13.1
GPS.n2 and GPSopt.n2	1 500 000€	10€	0.25	0.7	39.7	21.5	35.9	2.2
			0.5	0.8	60.4	20.8	14.7	3.3
			0.75	0.8	70.3	20.4	4.6	3.9
		37€	0.25	0.6	37.3	20.3	34.2	7.6
			0.5	0.7	55.0	19.1	14.1	11.1
			0.75	0.7	63.5	18.5	4.5	12.8
	3 000 000€	10€	0.25	0.3	43.5	21.5	32.3	2.4
			0.5	0.4	64.6	20.7	10.8	3.5
			0.75	0.4	72.5	20.4	2.8	3.9
		37€	0.25	0.3	40.6	20.2	30.7	8.2
			0.5	0.4	58.6	18.9	10.3	11.8
			0.75	0.4	65.3	18.5	2.7	13.1

Table 8 : Percentage of the total budget allocated to each operation.

DH: Doubled haploids. CT: Total cost. C_G: Genotyping cost. λ: Relative intensity of selection at step 3 compared to step 4.

3. Proportion of genetic gain variance explained by each parameter

We evaluated the impact of input parameters (i.e. strategy, total cost, trait heritability, QTL sampling, and λ) on genetic gain that were defined as the difference between the true genetic value (TBV) of the 200 best lines at the end of the third cycle and the TBV of the N_{ref} initial lines. To do so, we performed ANOVA analyses (Table 9) and we considered $\alpha = 0.05$ the significance threshold.

Trait	Factor	F	P value	% of SS
Genetic gain	**Strategy**	1899.7	< 2e-16	16.1
	λ	244.3	< 2e-16	0.7
	h^2	14 028.5	< 2e-16	39.6
	Average annual CT	780.3	< 2e-16	2.2
	QTL sampling	290.2	< 2e-16	0.8

Table 9 : Impact of input parameters on the genetic gain (ANOVA results).

h^2: trait heritability. λ: relative intensity of selection at step 3 compared to step 4. CT: Total cost. C_G: genotyping cost (37€). % of SS: (Sum of squares)/(Total sum of squares)

We found that each of the five factors had a significant effect on the final genetic gain, yet their contribution to explained variance varied a lot. The largest impact was due to trait heritability: the more heritable, the highest the genetic gain (average genetic gain was 13.64, 16.69 or 19.24 when the heritability was 0.2, 0.4 or 0.7, respectively), which is obviously expected. The strategy was the second most significant parameter. The strategy and the trait heritability explained 16.1% and 39.6% of the genetic gain variance, respectively. Note that doubling the total cost had a rather low impact on the final genetic gain (+1.07 on average). In addition, λ and QTL sampling had a low impact on the final genetic gain. For GPS strategies, we also studied the impact of genotyping cost on genetic gain and we concluded that decreasing genotyping cost from 37 to 10 in GPS strategies had no significant effect on the genetic gain.

We did the same analysis with the genetic gain obtained for various selection intensity at last step (for the 10, 50, 100 and 200 best lines) and we obtained the same conclusions (results are available in Table S4).

4. Significance of genetic gain difference between strategies

For each scenario, we compared the genetic gain for the different strategies (PS, GPS.N_C, GPS.n2, GPSopt.N_C, GPSopt.n2) using the Student-Newman-Keuls test. Means were considered as significantly different when P-values were below 0.05.

GPS.n2 and GPSopt.n2 allowed to do 156 more crosses, on average, compared to GPS.N_C and GPSopt.N_C. Whatever the scenario, the genetic gain for GPS.N_C (GPSopt.N_C) and GPS.n2 (GPSopt.n2) was not significantly different. In addition, the genetic gain for GPS.N_C and GPS.n2 was not significantly superior to the genetic gain for PS whatever the scenario. Even if the difference was not significant, the genetic gain for GPS.N_C and GPS.n2 was larger than genetic gain for PS when $\lambda = 0.25$ (i.e. when the intensity of selection is superior in step 4 where phenotypic selection is applied compared to step 3 where genomic selection is applied). In contrast, when $\lambda = 0.75$, the genetic gain was larger but not significantly for PS compared to GPS.N_C and GPS.n2.

Furthermore, genetic gain was always significantly larger for GPSopt.N_C and GPSopt.n2 than for PS, GPS.N_C and GPS.n2 whatever the values of the CT, λ, C_G and h^2. The difference between the genetic gain for GPSopt.N_C or GPSopt.n2 and the genetic gain for PS was even larger when the selection rate was higher at the step 4 than the step 3 ($\lambda = 0.25$). Indeed, the average differences between both strategies were +3.94 for $\lambda = 0.25$, +2.90 for $\lambda = 0.5$ and +1.46 for $\lambda = 0.75$.

5. Evolution of the genetic gain over cycles

As expected, we observed an increase of the 200 best lines mean TBV over the cycles for every scenario / strategy. Figure 22 illustrates the cumulative genetic gain at the end of each cycle with annual CT = 3 000 000€ and C_G = 37€. Results for other values of CT and C_G are available in supplementary material (Figures S3, S4 and S5). In addition, we noticed that the genetic gain (the difference between the TBV of the 200 best lines at the end of the cycle and the TBV of the 200 best lines at the end of the previous cycle) tended to decrease over the cycles in particular for GPSopt.N_C and GPSopt.n2 strategies. Figure 23 illustrates these results with annual CT = 3 000 000€ and C_G = 37€. Results for other values of CT and C_G are available in supplementary material (Figures S6, S7 and S8).

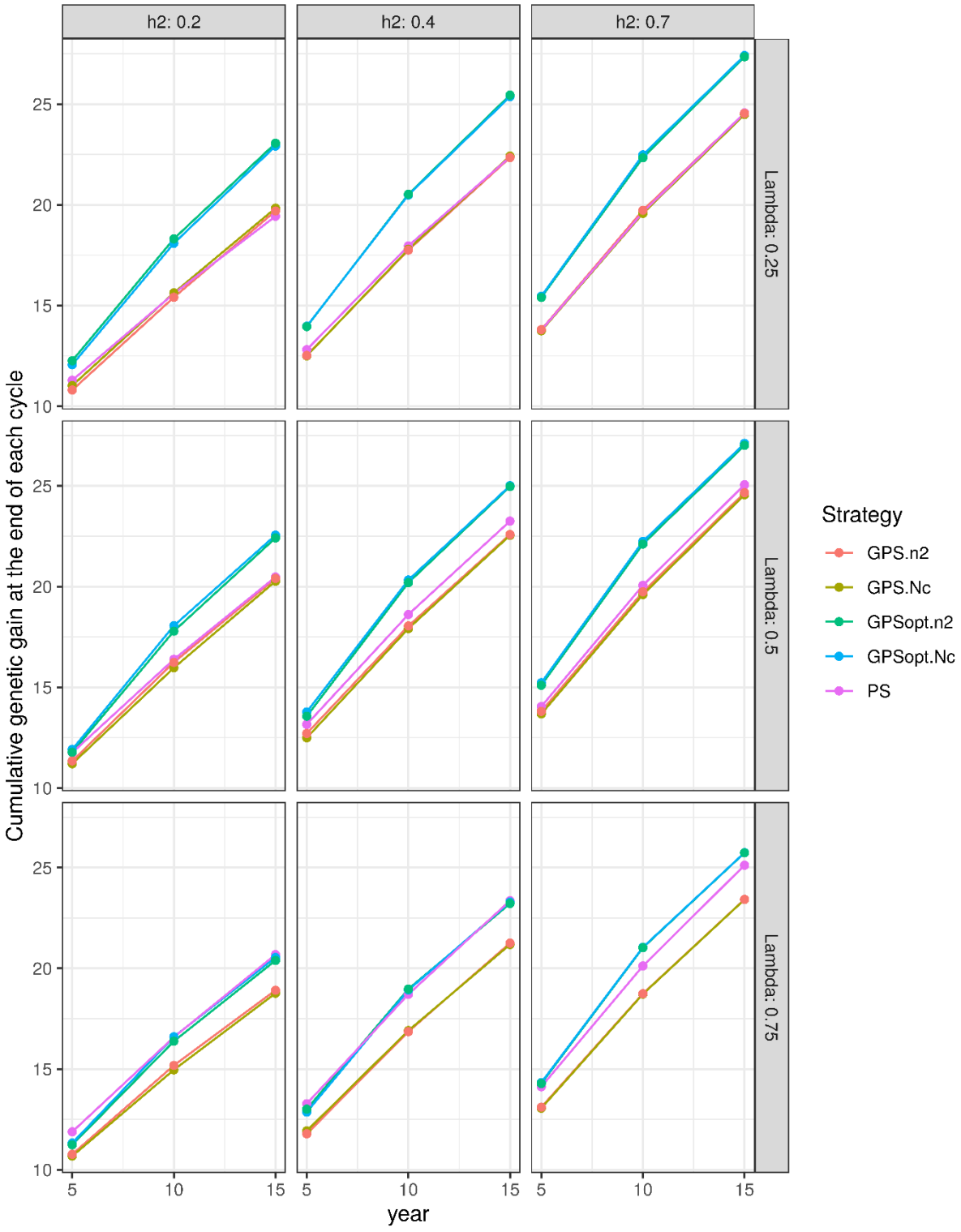

Figure 22: Evolution of the cumulative genetic gain at the end of cycle.

Annual total cost (CT) = 3 000 000€ and genotyping cost (C_G) = 37€.

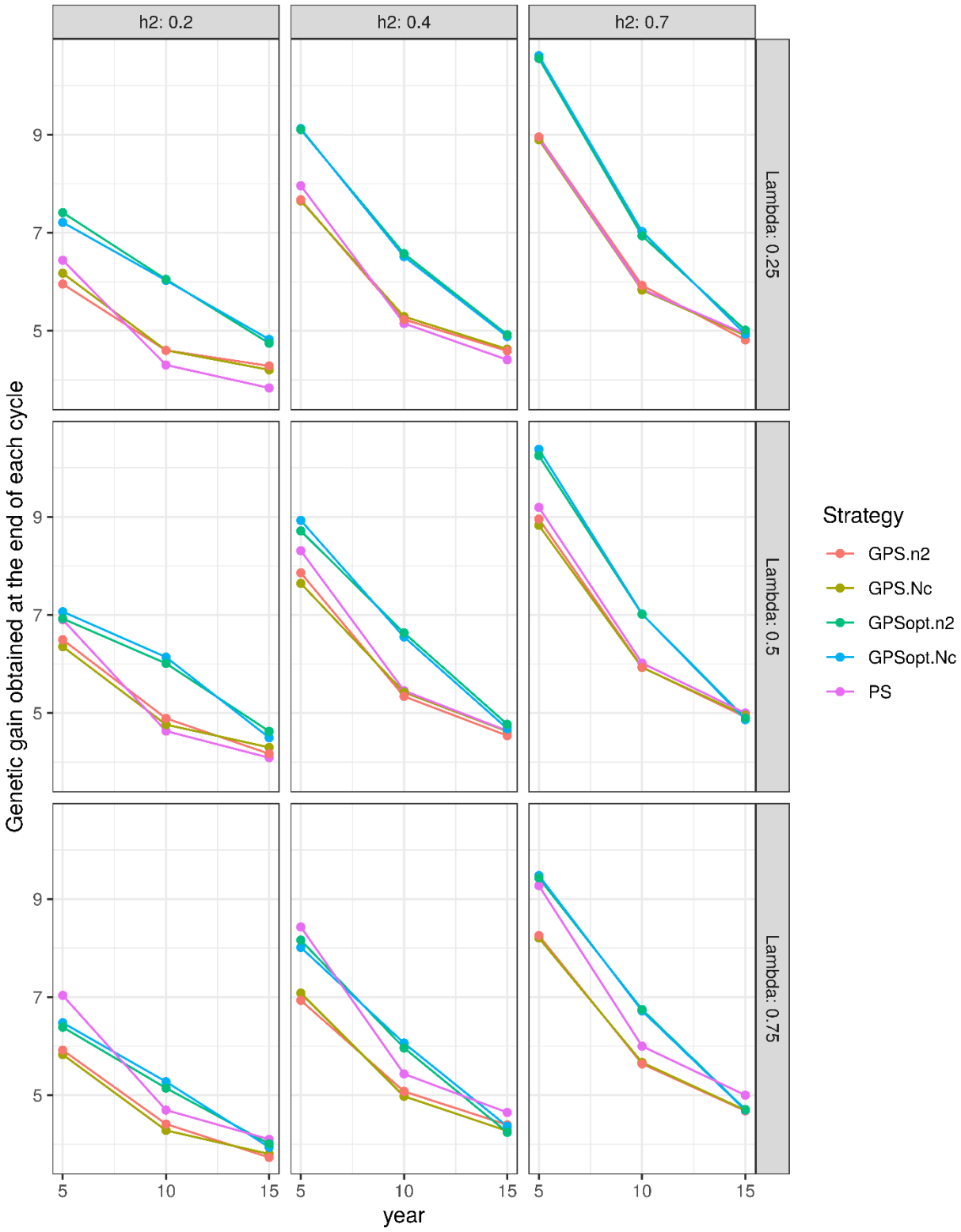

Figure 23: Evolution of the genetic gain achieved at the end of each cycle.

Annual total cost (CT) = 3 000 000€ and genotyping cost (C_G) = 37€.

6. Proportion of genetic diversity variance explained by each parameter and evolution of genetic diversity

We evaluated the impact of the several factors on the percentage of alleles present in both the reference population and the 200 best progenies of the last cycle. To do so, we performed ANOVA analyses (Table 10) and we considered $\alpha = 0.05$ the significance threshold.

Trait	Factor	F	P value	% of SS
% of alleles	**Strategy**	15 248.92	< 2e-16	67.8
	λ	6 705.78	< 2e-16	9.9
	h^2	123.75	< 2e-16	0.2
	Average annual CT	515.46	< 2e-16	0.8
	QTL sampling	6.68	9.7e-3	1.9e-2

Table 10 : Impact of input parameters on the percentage of alleles (ANOVA results).

h^2: trait heritability. λ: relative intensity of selection at step 3 compared to step 4. CT: Total cost. C_G: genotyping cost (37€). % of SS: (Sum of squares)/(Total sum of squares)

We found that each of the five factors had a significant effect on the variance of the percentage of alleles in the reference population present in the 200 best progenies of the last cycle, even if trait heritability, annual total cost and QTL sampling (the simulated trait) explained less than 1% of the variance. Strategy and λ explained 67.8% and 9.9% of the variance, respectively.

We observed a decrease of polymorphism over the cycles for all scenarios and strategies. This decrease was more pronounced for GPS breeding schemes (GPS.N_C, GPS.n2, GPSopt.N_C and GPSopt.n2 strategies) than for PS. In addition, when crosses were optimized, the loss of genetic diversity was even larger. Figure 24 illustrate these results with CT = 3 000 000€ and C_G = 37€ (Figures S9, S10 and S11 for other values of CT and C_G).

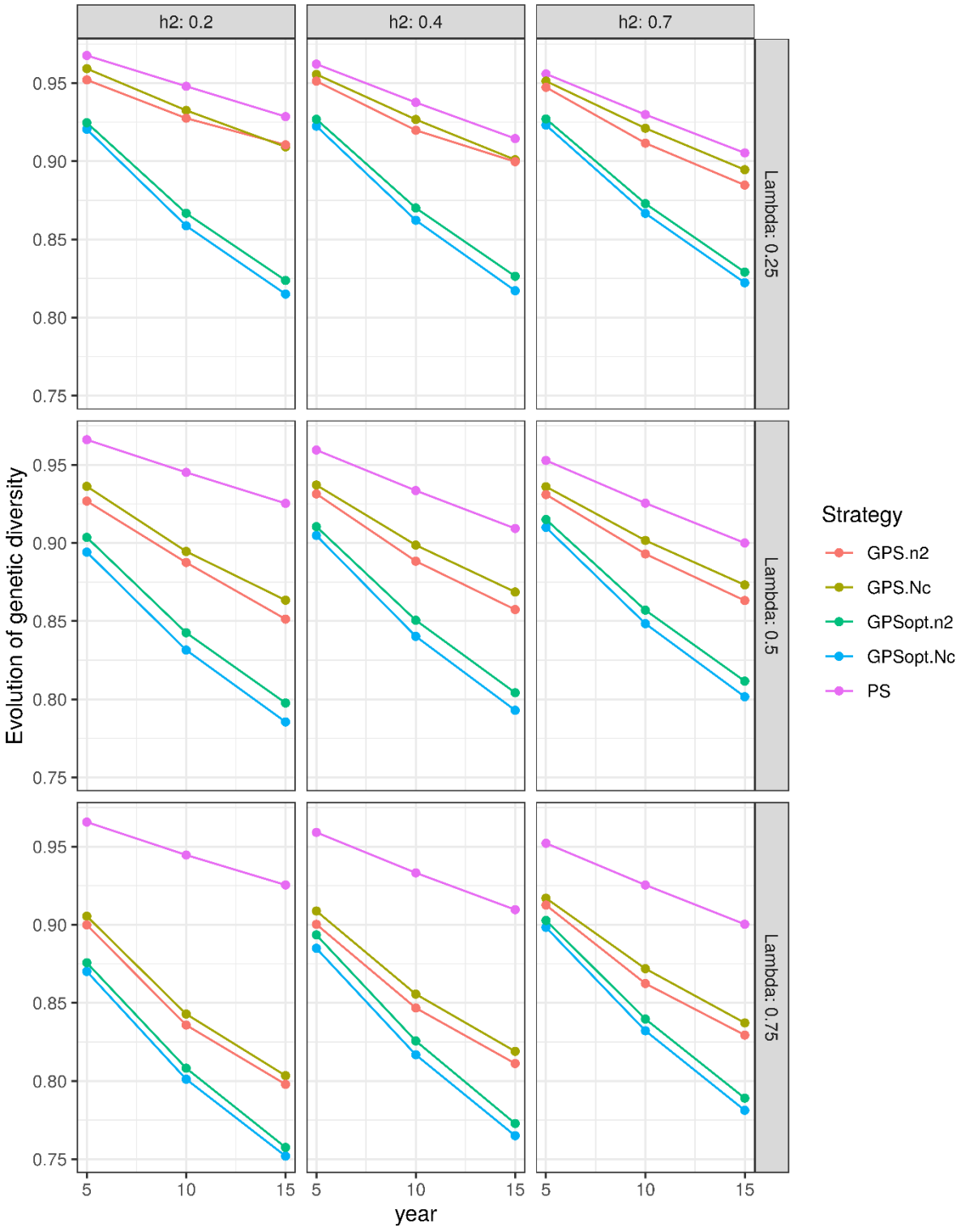

Figure 24: Evolution of genetic diversity over the cycles.

Annual total cost (CT) = 3 000 000€ and genotyping cost (C_G) = 37€.

7. Parental contributions

We analysed the pedigree of the Np lines selected at the end of the last cycle. We calculated the number of parental lines that contributed to the progenies.

For each simulation of the GPSopt.N_C strategy, on average 63 different initial lines were used as parents. It is important to notice that the number of lines varied widely between simulations since the standard deviation was 20.3. For each simulation of the GPSopt.n2 strategy, on average 67 different initial lines were used as parents. In this case the number of lines also varied widely (sd = 21.5). The average number of lines that contributed to the progenies was higher for strategies using random mating (PS, GPS.N_C and GPS.n2) than for strategies using a usefulness criterion to select best crosses (GPSopt.N_C and GPSopt.n2). The average number of lines that contributed to the progenies was 93 different initial lines for each simulation of the PS, GPS.N_C and GPS.n2 strategies.

VI. Discussion

This study focused on two types of breeding schemes: one called Phenotypic Selection (PS) with two steps of selection based on field trials, and the other one called Genomic and Phenotypic Selection (GPS) that combines a first step of genomic selection and a second step of selection based on field trials. We finalize the cycle by phenotypic selection because most plant breeders provide field data for candidates to official variety registration. In order to compare PS and GPS schemes with a fixed budget, GPS could either have the same number of crosses than PS (GPS.N_C) or it could have the same number of progenies per cross at the beginning of step 2 (GPS.n2). In addition, we explored two methods to obtain crosses for GPS schemes: by random mating of the best lines (GPS.N_C or GPS.n2 strategies) or by optimization using a usefulness criterion (GPSopt.N_C or GPSopt.n2 strategies). We tested several scenarios by varying cost of the breeding scheme, genotyping cost, trait heritability and selection intensities at each step of selection, in order to investigate under which conditions GPS was more efficient than PS. We evaluated each strategy by comparing the genetic gain of lines selected at the end of the breeding program and the evolution of genetic diversity throughout the selection stages.

1. Comparison of genetic gain of selected lines at the end of the breeding programs

The objective of bread wheat breeding programs is to develop new varieties that outperform existing varieties. In other words, the aim of the breeders is to create new varieties which have a higher True Breeding Value (TBV) than the varieties already available on the market, with the hope that the realized phenotype will also be superior. As expected, in this study the TBV of the 200 best lines increased over the cycles in each scenario and for all strategies.

The comparison of the genetic gain at the end of each simulated breeding scheme (i.e. at the end of the third cycle) enabled us to study major factors affecting the genetic gain. Indeed, we highlighted the fact that trait heritability and the strategy of the breeding program are the two main factors affecting the value of the genetic gain. The importance of trait heritability on the genetic gain was consistent with the breeder's equation (Lush 1937). To a lesser extent, the relative intensity of selection at step 3 compared to step 4 (λ) and the QTL sampling had also a significant effect on the genetic gain variance. In this study, we only simulated traits controlled by 100 QTLs randomly sampled from 12119 real genomic

markers. However, it would be interesting to evaluate the impact of the trait architecture on the efficiency of the different breeding programs for example by varying the number of sampled QTLs, the distribution of their effect or possible QTLxQTL interactions. We simulated breeding programs with two values of total cost, and we found that the total cost had a low but significant effect on the genetic gain variance. This result showed that doubling the global budget did not lead to a strong increase of the genetic gain (+6% on average). Furthermore, the genotyping cost had no significant effect on the genetic gain variance. This may be due to DH expense that is so high that varying the other costs has low impact on progeny number.

The comparison of the final genetic gain of each simulated breeding scheme also enabled us to identify under which conditions the use of genotyping data in the breeding program was the most interesting. We proposed two ways to use genotyping data in breeding programs: 1) genotyping information was only used in genomic predictions (GPS.N_C and GPS.n2 strategies), or 2) genotyping information was used in a usefulness criterion to optimize crosses and in genomic predictions (GPSopt.N_C and GPSopt.n2 strategies). In this study, we found that only strategies with optimized crosses would be significantly superior to PS in terms of average genetic gain of the 200 best-selected lines. The advantage given by these strategies was more pronounced when the selection was more intense at the step of selection based on field evaluation than at the step of genomic selection. We highlighted the importance of optimizing crosses. However, the optimization could be further improved. It is not realistic to consider that we can get the very best possible progeny with a limited progeny size. We will include in the second version a usefulness criterion that takes into account recombination rates across the genome for a limited number of progenies and simulate them by a stochastic process (Lehermeier et al. 2017; Müller et al. 2018). In addition, the number of progenies per cross was fixed in this study but it could be worthwhile to adjust the number of progenies of each cross depending on the predicted value of the cross.

2. Evolution of genetic diversity

Genetic diversity is a key parameter in plant breeding since it has an impact on long term genetic gain (according to breeder's equation; Lush 1937). However, it has been shown in both experimental study (in wheat: Rutkoski et al. 2014) and by stochastic simulations (Jannink 2010; Lin et al. 2016) that GS accelerates the loss of diversity compared to phenotypic selection due to the rapid fixation of regions of the genome with an effect on the trait of interest. Our results were consistent with previous studies. Indeed, we observed a faster decrease of the percentage of alleles in GPS strategies compared to PS schemes whatever the simulated scenario. This loss of alleles was even faster for strategies with optimized crosses (GPSopt.N_C and GPSopt.n2). In order to reduce this loss, we could place additional

weight on low-frequency favourable alleles as recommended by Jannink (2010). Since the loss of diversity is particularly fast when crosses were optmized, a second option would be to define a maximum number of crosses for which each parent could contribute in order to avoid having too many lines with one (or two) common parent. We will also implement UC that give the possibility to control final diversity by estimating simultaneously the desired parental contribution, in order to guarantee long-term genetic gain (UCPC, Allier et al. 2019). In this study, we considered that only lines from the previous generation could be mated. This assume no introduction of diversity in the breeding scheme and this is not realistic. All breeders introduce external bread wheat material at each cycle in their crosses. In order to simulate more realistic breeding programs, it would be important to give the possibility to add parents from an external pool (for example lines selected by other breeders).

3. Resource allocation

A great challenge in plant breeding is to improve and accelerate the genetic gain with a limited budget. To meet this goal, breeders have to evaluate the best ways to allocate their budget.

In this study, we found that production of doubled haploids (DHs), used to reduce the time required for inbred development (Maluszynski et al. 2013), was the major expense in the simulated breeding programs. Indeed, up to 46.5% and 72.6% of the global budget were allocated to DH production in PS and GPS schemes respectively. Therefore, it could be interesting to simulate breeding schemes based on other breeding methods than DH production (for example using Single Seed Descent; SSD) that could be economically more interesting and produce higher variability in progeny.

We also noticed that field evaluation accounted for a significant part of the budget (up to 46.2% for PS and up to 35.9% for GPS strategies). In addition, the percentage of the global budget allocated to field evaluation was higher for the most interesting scenarios in terms of genetic gain, i.e. the scenarios with the higher number of lines at step 4. To reduce phenotypic cost, each line could be phenotyped in a relevant subset of trials. We could optimize the experimental design in order to decrease the number of lines observed in each environment without decreasing selection accuracy (Heslot et al. 2017). The idea is to observe all alleles but not systematically all lines in all environments.

In the study, we focused on single-trait selection. However, real breeding programs most often deal with simultaneous improvement of several traits that can be correlated with each other. That is why simulations of breeding programs with several selection objectives would be more realistic. To do so, it would be important to simulate traits controlled by QTLs with pleiotropic effects. For example, Jia and

Jannink (2012) sampled QTL effects on two phenotypic traits from a standard bivariate normal distribution with a correlation of 0.5. In addition, a selection index could be used to select several traits more efficiently than by selecting each trait individually (Hazel 1943). The selection index would depend on the relative economical weight of each selected traits. Akdemir et al. (2019) proposed an approach to multi-trait breeding based on a multi-optimization framework by setting optimal compromise solutions identified by an effective and complete search procedure in order to help the breeders make the best choice.

VII. Conclusion

Afin d'évaluer l'impact des prédictions génomiques dans des schémas de sélection de blé tendre, nous avons analysé l'évolution du gain génétique ainsi que l'évolution de la diversité génétique au sein de différents programmes de sélection simulés.

Deux types de schémas ont été simulés dans cette étude : un schéma de sélection phénotypique (PS) avec deux étapes de sélection basées sur des essais en parcelles, et un schéma de sélection combinée phénotypique et génomique (GPS) pour lequel la première étape de sélection est remplacée par une étape de sélection génomique. Nous avons également étudié l'apport des prédictions génomiques lorsqu'elles sont utilisées pour prédire les valeurs des croisements et ainsi optimiser le plan de croisements (stratégies GPSopt). Ces différentes stratégies ont été comparées pour un budget annuel moyen fixé.

La simulation de 36 scénarios faisant varier le budget alloué au programme de sélection, l'intensité de sélection relative entre les deux étapes de sélection, le coût de génotypage d'une lignée et l'héritabilité du caractère nous a permis d'identifier les principaux facteurs ayant un effet sur l'évolution du gain génétique et sur l'évolution de la diversité génétique. Nous avons ainsi montré que l'héritabilité du caractère et la stratégie simulée (Schéma PS ou GPS, avec ou sans optimisation des croisements) avaient un impact significatif sur le gain génétique obtenu après 3 cycles de sélection. En revanche, le fait de doubler le budget total du programme de sélection ou de diminuer le coût de génotypage (de 37€ à 10€) n'avait qu'un faible impact sur le gain génétique. De plus, nous avons montré que la stratégie simulée avait un impact important sur la perte de diversité allélique au cours des cycles de sélection successifs. Nous avons constaté que cette perte de diversité était plus rapide pour les schémas utilisant des prédictions génomiques, en particulier pour ceux où les croisements étaient optimisés, par rapport aux schémas où les descendants étaient issus de croisements aléatoires entre les meilleures lignées issues du cycle précédent.

Grâce à cette étude nous avons également pu identifier des situations pour lesquelles les prédictions génomiques (que ce soit pour prédire la performance des individus ou la valeur des croisements) étaient les plus favorables. En effet, nous avons remarqué que l'écart entre le gain génétique obtenu dans des schémas PS et celui obtenu dans des schémas GPSopt était plus élevé lorsque l'intensité de sélection était plus forte lors de la seconde étape de sélection ($\lambda = 0.25$).

Nous avons développé un *pipeline* pour réaliser ces simulations et ces analyses. Afin de simuler des schémas de sélection plus réalistes et plus sophistiqués plusieurs pistes d'amélioration de ce *pipeline* sont envisagées. L'objectif à terme est (1) de permettre la simulation de schémas de sélection basés sur

de la sélection par filiation monograine (*Single Seed Descent* ; SSD) plutôt que des schémas reposant sur la production d'haploïdes doublés (HD), (2) d'intégrer des critères d'utilité plus sophistiqués prenant en compte les taux de recombinaison, (3) de simuler une introduction de matériel génétique extérieur à chaque génération pour maintenir un potentiel de gain génétique sur le long terme , et (4) de simuler le progrès génétique de plusieurs caractères simultanément.

Discussion générale

Cette thèse a permis d'étudier l'intérêt des prédictions génomiques pour les schémas de sélection du blé tendre. La première partie de ce projet s'est concentrée sur l'analyse de méthodes visant à améliorer la qualité des prédictions génomiques des valeurs génétiques individuelles sans pour autant augmenter le budget alloué au phénotypage. Nous nous sommes plus particulièrement intéressés à l'utilisation de modèles de prédiction génomique multi-caractère et à des méthodes d'optimisation de la population d'entraînement, utilisée pour calibrer les modèles. La simulation de schémas de sélection du blé tendre nous a ensuite permis d'étudier l'impact des prédictions génomiques des valeurs génétiques individuelles et des valeurs des croisements sur l'évolution de la valeur génétique moyenne des candidats et de la diversité génétique au sein des programmes de sélection.

Cette partie conclut le manuscrit en identifiant les apports et les limites des résultats obtenus au cours de la thèse, et en discutant des perspectives à court et à plus long terme de cette étude.

I. Amélioration de modèles de prédiction génomique

1. Apport des modèles multi-caractères

Les prédictions génomiques des valeurs génétiques individuelles consistent à prédire la performance moyenne (ou dans une gamme d'environnements cibles) de candidats à la sélection sans qu'ils ne soient phénotypés pour le caractère d'intérêt. Les modèles de prédiction génomique reposent sur l'estimation simultanée des effets des marqueurs moléculaires répartis sur l'ensemble du génome (Whittaker et al. 2000; Meuwissen et al. 2001). Actuellement, la majeure partie des travaux sur les prédictions génomiques dans le domaine végétal portent sur des modèles mono-caractères. Or ces analyses univariées n'exploitent pas l'information contenue dans les corrélations entre les différents caractères évalués au cours du programme de sélection.

Nous avons étudié l'apport des modèles de prédiction génomique multi-caractère en vue d'améliorer la précision des prédictions et de diminuer l'effort de phénotypage à l'échelle du programme de sélection. Pour cette étude nous avons pris comme exemple la note de panification, un caractère complexe utilisé en France lors de l'inscription de nouvelles variétés afin de classer les variétés en fonction de leur qualité boulangère. Nous disposions également de données de phénotypage pour la force boulangère (W), un

caractère modérément corrélé à la note de panification (r = 0.45) et dont le phénotypage est moins coûteux. Ce caractère a été utilisé dans les modèles de prédiction génomique multi-caractère que nous avons étudiés. Nous nous sommes intéressés à deux types de modèles multi-caractères :

1) Dans le cas du premier modèle, nous ne disposions de données de phénotypage que pour les lignées de la population d'entraînement pour calibrer le modèle (modèle *multi-trait*; MT).

2) Pour la calibration du second modèle, toutes les lignées (y compris celles de la population de validation) pouvaient être phénotypées pour le caractère corrélé W, mais seules celles de la population d'entraînement pouvaient être phénotypées pour la note de panification (modèle *trait-assisted*; TA).

Lorsque les lignées prédites n'avaient pas été phénotypées (ni pour le caractère d'intérêt, ni pour le caractère corrélé), nous n'avons pas observé d'amélioration de la précision des prédictions en utilisant un modèle multi-caractère plutôt qu'un modèle mono-caractère. D'autres études portant sur des données expérimentales avaient trouvé des résultats similaires (notamment Bao et al. 2015; Lado et al. 2018; Schulthess et al. 2018).

Par ailleurs, nous avons montré que le fait de phénotyper une partie des lignées de la population de validation pour le caractère secondaire corrélé pouvait permettre d'améliorer la précision des prédictions de la note de panification par rapport à des prédictions mono-caractères. Ces résultats étaient concordants avec ceux disponibles dans la littérature (Rutkoski et al. 2016; Sun et al. 2017; Crain et al. 2018; Fernandes et al. 2018; Lado et al. 2018; Michel et al. 2018; Schulthess et al. 2018). Nous avons évalué la précision des prédictions en utilisant deux méthodes de validation : en réalisant des validations croisées et en réalisant des prédictions dites *forward*. Cette seconde stratégie consistait à prédire les lignées les plus récentes en utilisant un modèle calibré grâce aux données collectées sur les lignées plus anciennes, stratégie qui se rapproche davantage des attentes du sélectionneur. Le gain de précision engendré par l'approche TA était particulièrement important pour cette seconde approche (gain pouvant atteindre +0.21 dans le cas des prédictions *forward*).

L'avantage économique de différentes approches utilisant des modèles de prédiction génomique multi-caractère a rarement été testé ou discuté dans les études déjà publiées. L'une des originalités du travail réalisé durant la thèse a été de prendre en compte le budget alloué au phénotypage lors de la comparaison des différentes approches de prédiction génomique. Nous avons alors montré que l'approche TA permettait de réduire le budget alloué au phénotypage tout en améliorant la qualité des prédictions génomiques par rapport à une approche mono-caractère. Cette réduction du budget alloué au phénotypage pouvait être causée par une diminution du nombre de lignées phénotypées pour le caractère le plus cher et/ou par une réduction du nombre d'essais dans lesquels les lignées avaient été évaluées.

De plus, le budget alloué au phénotypage pourrait être davantage réduit en optimisant les designs expérimentaux (Heslot et al. 2017).

Cette approche semble donc être une méthode prometteuse pour les sélectionneurs afin d'optimiser l'allocation des ressources dans les programmes de sélection. De plus, l'approche TA peut être généralisée à d'autres espèces et à d'autres caractères. Les avantages apportés par cette approche en termes d'amélioration de la précision des prédictions et en termes de réduction des coûts liés au phénotypage dépendent de la corrélation entre les caractères et du rapport entre les coûts de phénotypage de chacun des deux caractères.

2. Optimisation de la population d'entraînement pour améliorer la qualité des prédictions avec un budget limité

Plusieurs études ont montré que la taille et la composition de la population d'entraînement ont un impact sur la précision des prédictions génomiques. En effet, la précision des prédictions augmente avec la taille de la population d'entraînement (Jannink et al. 2010; Lorenz et al. 2011) et avec le degré d'apparentement entre la population d'entraînement et la population de validation (Habier et al. 2007; Habier et al. 2010; Charmet et al. 2014; Crossa et al. 2014).

Lorsque le budget alloué au phénotypage est limité et qu'il est donc impossible de phénotyper toutes les lignées, il est intéressant d'optimiser la composition de la population d'entraînement (c'est-à-dire de déterminer quelles sont les lignées à phénotyper en priorité) afin de maximiser la précision des prédictions. Des méthodes basées sur la minimisation de la variance d'erreur de prédiction moyenne (PEVmean) et sur la maximisation de la moyenne du coefficient de détermination généralisé (CDmean) ont été développées pour répondre à cet objectif (Rincent et al. 2012; Akdemir et al. 2015; Isidro et al. 2015; Sarinelli et al. 2019). Le coefficient de détermination généralisé correspond à la fiabilité attendue des contrastes entre les valeurs génétiques des individus non phénotypés et la moyenne de la population. Par rapport au PEVmean, le CDmean présente l'avantage de prendre en compte la variance génétique des contrastes entre individus et donc de limiter l'échantillonnage d'individus avec un fort degré d'apparentement. Ces méthodes d'optimisation nécessitent de disposer du génotypage des lignées de la population totale afin de calculer la matrice de *Kinship*, et de connaître les variances résiduelles et additives du caractère à prédire.

Durant la thèse, nous avons étendu le critère CDmean (Rincent et al. 2012) au contexte multi-caractère avec un critère que nous avons appelé CDmulti et qui permet d'optimiser le choix des individus à phénotyper pour chacun des caractères. L'optimisation utilisant le critère CDmulti est basée sur les

données de génotypage des lignées de la population totale et sur les matrices de variance-covariance génétiques et résiduelles entre les caractères.

Nous avons évalué l'intérêt du critère CDmulti pour améliorer la qualité des prédictions de la note de panification avec un budget alloué au phénotypage fixé. L'estimation des matrices de variance-covariance génétiques et résiduelles entre la note de panification et la force boulangère a été réalisée en utilisant un jeu de données indépendant correspondant aux données d'évaluation des lignées inscrites au Catalogue officiel (évaluations réalisées par le GEVES). Dans notre étude, l'utilisation du critère CDmulti à la place d'un échantillonnage aléatoire des lignées à phénotyper a conduit à une légère amélioration de la qualité des prédictions de la note de panification (en moyenne le gain de précision des prédictions était de +0.013) dans la plupart des scenarios testés (scénarios faisant varier le budget alloué au phénotypage et le pourcentage de lignées phénotypées pour la force boulangère). Nous avons évalué le critère CDmulti dans une population constituée de lignées issue d'un programme de sélection dont la diversité était limitée. Il serait intéressant de savoir si l'utilisation du critère CDmulti induirait une amélioration de la précision des prédictions plus nette pour une population présentant une plus grande diversité génétique.

3. Autres moyens d'améliorer les modèles de prédiction génomique

Nous avons montré qu'intégrer le phénotypage d'un caractère corrélé dans des modèles de prédiction génomique pouvait permettre de prédire de façon plus précise le caractère d'intérêt. Cependant il ne s'agit pas de la seule manière d'améliorer les prédictions génomiques.

L'une des façons d'améliorer les prédictions est d'intégrer en effet fixe dans les modèles des marqueurs en déséquilibre de liaison avec des gènes connus pour avoir un impact sur le caractère prédit (Bernardo 2014). Nous avons testé cette approche en intégrant en effet fixe des marqueurs dérivés des séquences d'ADN des gènes *Glu-A1*, *Glu-B1* et *Glu-D1* qui ont un impact connu sur les capacités rhéologiques de la pâte à pain. Nous avons constaté que cette méthode permettait d'obtenir des prédictions plus précises de la force boulangère mais qu'elle ne permettait pas d'améliorer la qualité des prédictions de la note de panification. Cette observation était probablement due au fait que les marqueurs dérivés des séquences des gènes *Glu-A1*, *Glu-B1* et *Glu-D1* n'expliquaient que 8% de la variance génétique de la note de panification. Or Bernardo (2014) a montré que seuls les marqueurs expliquant au moins 10% de la variance génétique du caractère d'intérêt peuvent entrainer une amélioration de la qualité des prédictions s'ils sont en effet fixe dans le modèle.

Les modèles de prédiction génomique peuvent également être complexifiés en intégrant des effets non-additifs, comme des effets de dominance (Technow et al. 2012; Zhao et al. 2013; Wang et al. 2014), des effets d'épistasie (Su et al. 2012; Muñoz et al. 2014; Jiang and Rief 2015) ou des effets d'interactions génotype x environnement (Burgueño et al. 2012; Cuevas et al. 2016; Jarquín et al. 2017; Ly et al. 2018).

Par ailleurs, l'information du pedigree des candidats à la sélection peut être prise en compte dans les modèles de prédiction génomique (Burgueño et al. 2012), bien que des marqueurs denses sur tout le génome soient supposés capturer davantage d'information. L'utilisation conjointe des données de marquage moléculaire et du pedigree pourrait permettre d'élargir la taille de la population d'entraînement en y intégrant des lignées qui auraient été phénotypées mais pas génotypées. Il est également possible d'inclure des données de transcriptomique dans les modèles de prédiction génomique, mais le choix des organes/stades/conditions de prélèvement complexifie l'approche, qui s'écarte alors du concept de sélection génomique. Schrag et al. (2018) ont montré que la prédiction de la performance d'hybrides chez le maïs était meilleure avec des modèles combinant des données de génotypage et des données de transcriptomique plutôt qu'avec des modèles ne prenant en compte qu'un des deux types de données. Cependant l'étude de l'expression des gènes engendre des coûts supplémentaires qu'il faudrait évaluer afin d'estimer l'intérêt de ces modèles d'un point de vue économique, et pose aussi la question de la stabilité de ces expressions selon les environnements.

Le phénotypage à haut débit a connu un essor important au cours des dernières décennies (Araus and Cairns 2014). De nouveaux outils portant de nombreux capteurs ont été développés afin de caractériser les variétés et les conditions environnementales dans les champs, de manière précise, rapide et non destructive. L'amélioration des techniques de phénotypage à haut débit et des méthodes de traitements des données ainsi collectées constitue une opportunité d'augmenter la taille des populations d'entraînement afin d'obtenir des prédictions génomiques plus fiables. Les données issues de méthodes de phénotypage à haut débit peuvent être utilisées dans des modèles de prédiction génomique multi-caractère. Des études ont montré l'intérêt de d'utiliser des données de phénotypage haut débit (en particulier des mesures de la température de la canopée et de l'indice de végétation par différence normalisée) pour prédire le rendement en grain chez le blé tendre avec des modèles multi-caractères (Rutkoski et al. 2016; Sun et al. 2017; Crain et al. 2018). Par ailleurs, Rincent et al. (2018) ont présenté une méthode de prédiction prometteuse pour l'amélioration végétale : la sélection phénomique. Cette méthode a pour principe d'estimer les similarités entre les individus non pas à partir des données de génotypage mais à partir de données de spectroscopie proche infrarouge (*Near Infrared Reflectance Spectroscopy* ; NIRS) sur différents organes (feuilles, grains, etc.). Cette approche présente l'avantage d'utiliser des mesures peu coûteuses et non destructives. Certes les spectres obtenus sont influencés par les conditions environnementales, mais Rincent et al. (2018) montrent qu'ils contiennent un signal génétique suffisant pour générer des prédictions aussi bonnes que les marqueurs pan-génomiques chez

le blé à condition que les lignées de la population d'entraînement aient été phénotypées dans les mêmes environnements que les individus de la population de validation.

II. Mise en place des prédictions génomiques dans des programmes de sélection du blé tendre

1. Comparaison de schémas de sélection avec ou sans étape de prédiction génomique

De nombreuses études se sont concentrées sur l'évaluation de la qualité des prédictions génomiques et ont permis d'identifier les principaux facteurs affectant la qualité des prédictions génomiques et de développer des modèles de prédiction plus fiables. ont porté sur les prédictions génomiques mono-caractères de caractères d'intérêt agronomique chez le blé tendre, comme le rendement ou des composantes du rendement (Poland et al. 2012; Dawson et al. 2013; Lado et al. 2013; Storlie and Charmet 2013; Zhao et al. 2015; Norman et al. 2017), la qualité boulangère ou des paramètres associés (Battenfield et al. 2016; Guzman et al. 2016; Liu et al. 2016), ou encore des résistances aux maladies (Ornella et al. 2012; Rutkoski et al. 2012; Arruda et al. 2015). Ces études ont souligné l'intérêt des prédictions génomiques pour l'amélioration du blé tendre. Battenfield et al. (2016), par exemple, ont montré que le gain génétique obtenu avec des approches de sélection génomique pouvait être jusqu'à 2.7 fois supérieur à celui obtenu grâce à de la sélection phénotypique.

En revanche, le nombre d'articles scientifiques portant sur la mise en pratique de la sélection génomique dans les schémas de sélection du blé tendre est plus restreint. Or l'efficacité de la sélection génomique ne repose pas uniquement sur la qualité des prédictions. En effet, le gain génétique à court et long terme, l'évolution de la diversité génétique et le coût global du schéma de sélection intégrant des prédictions génomiques sont des facteurs qu'il est important de prendre en compte lors de l'évaluation de l'apport des prédictions génomiques dans les programmes de sélection. Il n'est pas réaliste de tester l'impact de toutes les combinaisons de paramètres (budget, nombre de croisements, nombre de descendants, taux de sélection à chaque étape, etc.) pouvant varier dans un programme réel. Mais il est possible de simuler des schémas de sélection en intégrant ou non des prédictions génomiques, et de faire varier de nombreux paramètres afin d'évaluer leur impact relatif.

De plus, les prédictions génomiques peuvent être utilisées à différentes étapes du programme de sélection (pour prédire les meilleurs croisements, dans les premières étapes de sélection pour éliminer les plus mauvais individus, à la fin du programme de sélection pour identifier les meilleurs candidats pour participer à la génération suivante, pour choisir le sous-ensemble d'individus à phénotyper, etc.). Il est donc intéressant de simuler des schémas de sélection intégrant des étapes de prédictions génomiques à différentes générations afin d'identifier les étapes et les conditions pour lesquelles les

prédictions génomiques améliorent le plus le gain génétique ou économique du programme de sélection. Bassi et al. (2016) ont comparé différents schémas de sélection de blé utilisant des prédictions génomiques à différentes étapes du schéma. Ils ont notamment montré que l'utilisation des prédictions génomiques à toutes les générations du schéma de sélection (y compris lors des générations les plus précoces lors desquelles le nombre de variétés est le plus élevé) permettait de réduire la durée du cycle, mais que cela nécessitait d'allouer une part importante du budget au génotypage. Ils ont également montré qu'à l'inverse l'utilisation de prédictions génomiques à un stade tardif du schéma de sélection permettait de limiter le budget alloué au génotypage mais que le progrès génétique annuel réalisé était plus faible qu'avec un schéma de sélection uniquement basé sur les prédictions génomiques.

Nous avons analysé l'évolution du gain génétique ainsi que l'évolution de la diversité génétique au sein de schémas de sélection lorsqu'une étape de sélection phénotypique est remplacée par de la sélection génomique. Pour cette étude nous avons simulé deux types de schémas : un schéma de sélection phénotypique (PS) avec deux étapes de sélection basées sur des essais en parcelles, et un schéma de sélection combinée phénotypique et génomique (GPS) avec une première étape de sélection génomique et une seconde étape de sélection basée sur des essais en parcelles. Nous avons également simulé deux méthodes pour obtenir les croisements dans le cas des schémas GPS : soit ils étaient obtenus en croisant de façon aléatoire les meilleures lignées issues du cycle précédent (méthode aussi utilisée pour les schémas PS), soit ils étaient optimisés en utilisant un critère d'utilité, accessible seulement pour des lignées génotypées, donc pour le schéma GPS. Les schémas simulés ont été définis de telle sorte que le budget annuel moyen de chacun des schémas soit équivalent. Le budget a été estimé en prenant en compte le coût de chacune des étapes du schéma de sélection (coûts des croisements, de production des haploïdes doublés, de multiplication des semences, de génotypage et d'évaluation au champ). Afin d'évaluer l'impact du budget alloué au programme de sélection, de l'intensité de sélection à chaque étape, du coût de génotypage et de l'héritabilité du caractère sur les différents types de schémas PS et GPS, 36 scénarios faisant varier ces différents paramètres, ont été simulés pendant 3 cycles de sélection successifs correspondant chacun à une durée de 5 ans. Nous avons montré que le gain génétique était significativement plus élevé pour les schémas de sélection GPS où les croisements ont été optimisés (GPSopt) dans chaque scénario évalué. L'avantage des stratégies utilisant des prédictions génomiques était particulièrement marqué lorsque l'intensité de sélection durant l'étape de sélection génomique était plus faible que celle à l'étape de sélection phénotypique. Nous avons également constaté que la perte de diversité allélique était plus rapide pour ces schémas de sélection GPSopt que pour les schémas PS et GPS où les descendants étaient obtenus en croisant les lignées parentales au hasard. D'autres études avaient déjà souligné que l'utilisation de prédictions génomiques pouvaient accentuer la réduction de la diversité génétique à long terme (Jannink 2010; Rutkoski et al. 2014; Lin et al. 2016).

Afin de simuler des schémas de sélection plus réalistes et plus sophistiqués plusieurs pistes d'amélioration sont envisagées. La conception du plan de croisements est un élément crucial pour le sélectionneur car c'est de cette décision que dépendent la moyenne de la valeur génétique ainsi que la variance génétique observées dans la descendance. Le plan de croisements a donc un impact important sur le gain génétique potentiel du programme de sélection. C'est pour cette raison que la première piste envisagée pour améliorer les simulations consiste à intégrer des critères d'utilité plus complexes qui pourront notamment prendre un compte les taux de recombinaison entre marqueurs ou la contribution des différents parents pour assurer un gain génétique sur le long terme (Zhong and Jannink 2007; Lehermeier et al. 2017; Allier et al. 2019). Par ailleurs, nous n'avons simulé que des schémas de sélection où il n'y avait pas d'introduction de diversité au cours des cycles de sélection. Or cette hypothèse n'est pas réaliste puisque dans les programmes de sélection réels les sélectionneurs ont la possibilité de croiser les lignées élites qu'ils ont développées avec du matériel provenant d'autres obtenteurs, ou issu des nombreuses collections de ressources génétiques mondiales, souvent publiques. Ainsi, simuler des schémas de sélection où il serait possible d'introduire du matériel végétal provenant d'autres programmes de sélection à chaque cycle serait plus proche de la réalité et pourrait permettre de maintenir un gain génétique potentiel sur le long terme. Nous pourrions également simuler des schémas basés sur de la sélection par filiation monograine (méthode aussi appelée *Single Seed Descent* ; SSD) plutôt que des schémas reposant sur la production d'haploïdes doublés car nous avons montré que la production d'haploïdes doublés pouvait représenter plus de la moitié du budget total du programme de sélection (avec les coûts dont nous disposions, correspondant à une sous-traitance de cette activité). Enfin, la sélection variétale consiste à créer des variétés qui combinent plusieurs caractères d'intérêt agronomique ; c'est pourquoi il est indispensable de simuler des programmes de sélection portant sur l'amélioration simultanée de plusieurs caractères.

2. Simulation de schémas de sélection multi-caractères

Certains caractères à sélectionner peuvent être corrélés positivement ou négativement entre eux. Les corrélations négatives existant entre les différents caractères à sélectionner constituent un enjeu pour les sélectionneurs. Pour étudier des schémas de sélection prenant en compte des caractères corrélés il faut simuler des caractères contrôlés par des QTL avec des effets pléiotropiques. Pour cela il est notamment possible d'utiliser une distribution normale à deux variables pour échantillonner les effets des QTL (Jia and Jannink 2012) et ajouter éventuellement des QTL spécifiques à chaque caractère (ce qui diminue alors leur corrélation génétique). Par ailleurs, plusieurs objectifs de sélection peuvent être envisagés dans le cas de schémas multi-caractères et il est important de déterminer les objectifs avant de réaliser

les simulations. L'objectif du sélectionneur peut par exemple être de sélectionner plusieurs caractères (potentiellement corrélés négativement) de façon simultanée. Un indice de sélection a été proposé par Smith (1936) et Hazel (1943) pour optimiser la sélection de plusieurs caractères simultanément. L'objectif du sélectionneur peut également consister à améliorer un caractère sans dégrader la performance des variétés pour le second caractère. L'indice de sélection (sous contrainte) de Kempthorne et Nordskog (1959) a été développé pour ce type de situations. Pour finir, des schémas de sélection multi-caractères avec un unique caractère d'intérêt mais utilisant dans les modèles de prédiction l'information apportée par des caractères corrélés moins coûteux peuvent également être envisagés. Nous avons effectivement montré l'intérêt de telles approches pour optimiser l'allocation des ressources en prenant l'exemple de la qualité boulangère. Pour les schémas visant à améliorer plusieurs caractères, un poids peut être attribué à chacun des caractères en fonction de leur importance économique.

Une unique solution optimale peut être obtenue lorsque l'objectif est d'optimiser l'amélioration d'un seul caractère. En revanche, il existe plusieurs solutions de compromis lorsque l'optimisation porte sur plusieurs caractères. L'enjeu est d'optimiser de façon simultanée plusieurs fonctions (chaque fonction correspondant à un objectif de sélection) jusqu'à ce qu'une solution de compromis optimale soit trouvée. Akdemir et al. (2019) ont proposé un outil d'aide à la décision pour les sélectionneurs afin d'optimiser des schémas de sélection intégrant plusieurs objectifs de sélection. Cette approche a été évaluée sur des données expérimentales (sur le blé et sur l'orge) ainsi que sur des données simulées. Les budgets des programmes de sélection et les enjeux techniques liés à chaque espèce étant très variables, il est difficile de proposer un unique algorithme qui serait applicable à toutes les situations (Wellmann 2019).

III. Autres apports des marqueurs moléculaires en sélection végétale

Le développement des techniques de marquage moléculaire de plus en plus rapides et performantes a permis l'essor des prédictions génomiques des valeurs génétiques individuelles et des valeurs des croisements. Cependant les prédictions génomiques ne constituent pas la seule façon de bénéficier de ces nouvelles technologies en amélioration des plantes.

Les marqueurs moléculaires peuvent notamment être utilisés pour identifier des variétés qui combinent des allèles favorables à différents loci contrôlant la variation des caractères d'intérêt ou pour introgresser des allèles favorables d'une lignée « donneuse » dans une lignée « receveuse » ayant une forte valeur agronomique. Cette méthode est appelée sélection assistée par marqueurs (SAM). La SAM est efficace pour introgresser un allèle favorable (Frisch et al. 1999). Chetelat et al. (1995), par exemple, ont montré qu'il était possible introgresser un allèle favorable au niveau d'un gène majeur contrôlant la concentration en sucrose dans le fruit chez la tomate. Afin de d'introgresser plusieurs allèles favorables, des stratégies de combinaison de gènes ont été proposées (Hospital and Charcosset 1997 ; Charmet et al. 1999 ; Servin et al. 2004). Cependant, l'introgression d'allèles au niveau de nombreux loci qui contrôlent des caractères polygéniques tels que le rendement se révèle délicate (Hospital and Charcosset 1997).

Chez le blé tendre, la SAM a été utilisée dans les programmes de sélection pour identifier les allèles favorables au niveau de gènes de résistance à des maladies telles que la rouille du blé (Miedaner and Korzun 2012), et pour accumuler plusieurs gènes de résistances dans une même variété (Liu et al. 2000). Elle peut également servir à identifier les allèles favorables au niveau des gènes *Glu-A1*, *Glu-B1* et *Glu-D1* qui contrôlent les sous-unités des gluténines de haut poids moléculaires et qui ont un impact sur les capacités rhéologiques de la pâte. Durant ma thèse, j'ai eu l'occasion de travailler sur 17 marqueurs moléculaires dérivés des séquences d'ADN des gènes *Glu-A1*, *Glu-B1* et *Glu-D1*. Ces marqueurs ont été présentés dans la publication de Ravel et al. (2020) où je figure en tant que co-auteur (le lien vers la publication est disponible en annexe). Ils constituent un outil pour les sélectionneurs permettant d'identifier à un stade précoce les allèles favorables des gènes *Glu-A1*, *Glu-B1* et *Glu-D1*.

Par ailleurs, Heffner et al. (2010) ont montré qu'il était possible de combiner la SAM et les prédictions génomiques dans un même schéma de sélection. La SAM est alors utilisée lors des premières étapes du schéma de sélection afin d'éliminer les variétés ne présentant pas les allèles favorables aux QTL, puis des étapes de sélections génomiques sont réalisées afin de sélectionner les meilleurs candidats.

L'élargissement des connaissances concernant le génome des plantes et le déterminisme génétique des caractères d'intérêt a rendu possible le développement de nouvelles techniques de sélection végétale (*New Breeding Techniques* ; NBT). Ces nouvelles techniques font notamment référence à des techniques émergentes d'édition du génome, telles que la technique CRISPR-Cas9 (Jinek et al. 2012), qui visent à modifier une région précise du génome.

Références bibliographiques

Abecassis J, Bergez J-E (2009) Les filières céréalières - Organisation et nouveaux défis, Editions Quae

Ahmad M (2000) Molecular marker-assisted selection of HMW glutenin alleles related to wheat bread quality by PCR-generated DNA markers. Theor Appl Genet 101:892–896. https://doi.org/10.1007/s001220051558

Akdemir D, Beavis W, Fritsche-Neto R, et al. (2019). Multi-objective optimized genomic breeding strategies for sustainable food improvement. Heredity, 122(5), 672-683.

Akdemir D, Sánchez JI (2016) Efficient Breeding by Genomic Mating. Front Genet. 2016 Nov 29;7:210. doi: 10.3389/fgene.2016.00210. PMID: 27965707; PMCID: PMC5126051.

Akdemir D, Sanchez JI, Jannink J-L (2015) Optimization of genomic selection training populations with a genetic algorithm. Genet Sel Evol 47:38. https://doi.org/10.1186/s12711-015-0116-6

Allier A, Moreau L, Charcosset A, et al (2019) Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression. G3 (Bethesda) 9:1469–1479. https://doi.org/10.1534/g3.119.400129

Anjum FM, Khan MR, Din A, et al (2007) Wheat Gluten: High Molecular Weight Glutenin Subunits—Structure, Genetics, and Relation to Dough Elasticity. J Food Sci 72:R56–R63. https://doi.org/10.1111/j.1750-3841.2007.00292.x

Araus J L, Cairns J E (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends in plant science, 19(1), 52-61.

Arruda MP, Brown PJ, Lipka AE, et al (2015) Genomic selection for predicting *Fusarium* head blight resistance in a wheat breeding program. The Plant Genome, 8(3).

Arruda MP, Lipka AE, Brown PJ, et al (2016) Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). Mol Breeding 36:84. https://doi.org/10.1007/s11032-016-0508-5

Arumuganathan K, Earle E D (1991) Nuclear DNA content of some important plant species. Plant molecular biology reporter, 9(3), 208-218.

Balfourier F, Roussel V, Strelchenko P, et al (2007) A worldwide bread wheat core collection arrayed

in a 384-well plate. Theor Appl Genet 114:1265–1275. https://doi.org/10.1007/s00122-007-0517-1

Bao Y, Kurle JE, Anderson G, et al (2015) Association mapping and genomic prediction for resistance to sudden death syndrome in early maturing soybean germplasm. Mol Breeding, 35(6), 128.

Bariana HS, Miah H, Brown GN, et al (2007) Molecular mapping of durable rust resistance in wheat and its implication in breeding. In: Buck HT, Nisi JE, Salomón N (eds) Wheat Production in Stressed Environments. Springer Netherlands, Dordrecht, pp 723–728

Bassi FM, Bentley AR, Charmet G, et al (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Science 242:23–36. https://doi.org/10.1016/j.plantsci.2015.08.021

Battenfield SD, Guzmán C, Gaynor RC, et al (2016) Genomic Selection for Processing and End-Use Quality Traits in the CIMMYT Spring Bread Wheat Breeding Program. Plant Genome 9:. https://doi.org/10.3835/plantgenome2016.01.0005

Beavis WD (1998) QTL analyses: power, precision, and accuracy, pp. 145–162 in Molecular Dissection of Complex Traits, edited by Paterson A. H.. CRC Press, New York.

Belderok B, Mesdag J, Donner DA (2000) The wheat grain. In: Belderok B, Mesdag J, Donner DA, Donner DA (eds) Bread-making quality of wheat: A Century of breeding in Europe. Springer Netherlands, Dordrecht, pp 15–20

Bergez JA et J-E (2009) Les filières céréalières. Editions Quæ

Bernardo R (2009) Genomewide Selection for Rapid Introgression of Exotic Germplasm in Maize. Crop Science 49:419–425. https://doi.org/10.2135/cropsci2008.08.0452

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

Bogard M, Ravel C, Paux E et al. (2014) Predictions of heading date in bread wheat (Triticum aestivum L.) using QTL-based parameters of an ecophysiological model. Journal of experimental botany. 65. 10.1093/jxb/eru328.

Bordes JJ, Grand Ravel CC, Le Gouis JJ, et al (2011) Use of a global wheat core collection for association analysis of flour and dough quality traits. Journal of Cereal Science 54:137–147. https://doi.org/10.1016/j.jcs.2011.03.004

Bordes JJ, Ravel C, Jaubertie J.P. et al. (2013) Genomic regions associated with the nitrogen limitation response revealed in a global wheat core collection. Theor Appl Genet 126, 805–822. https://doi.org/10.1007/s00122-012-2019-z

Brisson N, Gate P, Gouache D, et al (2010) Why are wheat yields stagnating in Europe? A comprehensive data analysis for France. Field Crops Research 119:201–212. https://doi.org/10.1016/j.fcr.2010.07.012

Buerstmayr H, Ban T, Anderson J A (2009) QTL mapping and marker-assisted selection for *Fusarium* head blight resistance in wheat: A review. Plant Breeding. 128. 1 - 26. 10.1111/j.1439-0523.2008.01550.x.

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. Crop Science 52:707–719. https://doi.org/10.2135/cropsci2011.06.0299

de Bustos A, Rubio P, Soler C et al. (2001) Marker assisted selection to improve HMW-glutenins in wheat. Euphytica **119,** 69–73 https://doi.org/10.1023/A:1017534203520

Butler DG, Cullis BR, Gilmour AR and Gogel BJ (2009) ASReml-R reference manual. The State of Queensland, Department of Primary Industries and Fisheries, Brisbane.

Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics 124:362–368. https://doi.org/10.1111/j.1439-0388.2007.00691.x

Calus MP, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol 43:26. https://doi.org/10.1186/1297-9686-43-26

Charmet G, Robert N, Perretant MR, et al (2001) Marker assisted recurrent selection for cumulating QTLs for bread-making related traits. Euphytica 119:89–93. https://doi.org/10.1023/A:1017577918541

Charmet G, Storlie E, Oury FX, et al (2014) Genome-wide prediction of three important traits in bread wheat. Mol Breed 34:1843–1852. https://doi.org/10.1007/s11032-014-0143-y

Charmet G, Robert N, Perretant MR, et al (1999) Marker-assisted recurrent selection for cumulating additive and interactive QTLs in recombinant inbred lines. Theor Appl Genet 99:1143–1148. https://doi.org/10.1007/s001220051318

Chetelat RT, DeVerna JW, Bennett AB (1995) Introgression into tomato (Lycopersicon esculentum) of the L. chmielewskii sucrose accumulator gene (sucr) controlling fruit sugar composition. Theoret Appl Genetics 91:327–333. https://doi.org/10.1007/BF00220895

Combs E, Bernardo R (2013) Genomewide selection to introgress semidwarf maize germplasm into U.S. Corn Belt Inbreds. Crop Science 53:1427–1436 https://doi.org/10.2135/cropsci2012.11.0666

Crain J, Mondal S, Rutkoski J, et al (2018) Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. Plant Genome 11. https://doi.org/10.3835/plantgenome2017.05.0043

Crossa J, Campos G de los, Pérez P, et al (2010) Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186:713–724. https://doi.org/10.1534/genetics.110.118521

Crossa J, Pérez P, Hickey J, et al (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48–60. https://doi.org/10.1038/hdy.2013.16

Crossa J, Pérez-Rodríguez P, Cuevas J, et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science, 22(11), 961-975.

Cuevas J, Crossa J, Soberanis V, et al (2016) Genomic prediction of genotype× environment interaction kernel regression models. The plant genome, 9(3).

Daetwyler HD, Bansal UK, Bariana HS, et al (2014) Genomic prediction for rust resistance in diverse wheat landraces. Theor Appl Genet 127:1795–1803. https://doi.org/10.1007/s00122-014-2341-8

Daetwyler HD, Hayden MJ, Spangenberg GC, Hayes BJ (2015) Selection on Optimal Haploid Value Increases Genetic Gain and Preserves More Genetic Diversity Relative to Genomic Selection. Genetics 200:1341–1348. https://doi.org/10.1534/genetics.115.178038

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLOS ONE 3:e3395. https://doi.org/10.1371/journal.pone.0003395

Dawson JC, Endelman JB, Heslot N, et al (2013) The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Research 154:12–22. https://doi.org/10.1016/j.fcr.2013.07.020

De Rochambeau H (1992) Le progrès génétique et sa réalisation dans les expériences de sélection. INRA Productions Animales hs:83–86

de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of Genomic Predictions Across Multiple Populations. Genetics 183:1545–1553. https://doi.org/10.1534/genetics.109.104935

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601. https://doi.org/10.1016/j.tplants.2014.05.006

Eagles H, Bariana H, Ogbonnaya F, et al (2001) Implementation of markers in Australian wheat breeding. Australian Journal of Agricultural Research 1349–1356. https://doi.org/10.1071/AR01067

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

Endelman JB and Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. G3 2:1405-1413. doi: 10.1534/g3.112.004259

Feillet P (2000) Le grain de blé - Composition et utilisation, Librairie Quae

Feldman M (1995) Wheats. Triticum spp. (Gramineae–Triticinae). In: Smaut J, Simmonds NW (eds) Evolution of crop plants. Longman Scientific and Technical, London, pp 184–192

Fernandes SB, Dias KOG, Ferreira DF, Brown PJ (2018) Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor Appl Genet 131:747–755. https://doi.org/10.1007/s00122-017-3033-y

Fisher RA (1919) XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of The Royal Society of Edinburgh 52:399–433. https://doi.org/10.1017/S0080456800012163

Frisch M, Bohn M, Melchinger AA (1999) Minimum Sample Size and Optimal Positioning of Flanking Markers in Marker-Assisted Backcrossing for Transfer of a Target Gene. Crop Science 39:cropsci1999.0011183X003900040003x. https://doi.org/10.2135/cropsci1999.0011183X003900040003x

Gallais A (2015) Comprendre l'amélioration des plantes, Editions Quae, Collection Synthèses

Gallais A (2018) Histoire de la génétique et de l'amélioration des plantes, Editions Quae

García-Ruiz A, Cole JB, VanRaden PM, et al (2016) Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci USA 113:E3995-4004. https://doi.org/10.1073/pnas.1519061113

Gill BS, Appels R, Botha-Oberholster A-M, et al (2004) A Workshop Report on Wheat Genome Sequencing: International Genome Research on Wheat Consortium. Genetics 168:1087–1096. https://doi.org/10.1534/genetics.104.034769

Guo G, Zhao F, Wang Y, et al (2014) Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet 15:30. https://doi.org/10.1186/1471-2156-15-30

Guzman C, Peña RJ, Singh R, et al (2016) Wheat quality improvement at CIMMYT and the use of genomic selection on it. Appl Transl Genom 11:3–8. https://doi.org/10.1016/j.atg.2016.10.004

Habier D, Fernando RL, Dekkers JCM (2007) The impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177:2389–2397. https://doi.org/10.1534/genetics.107.081190

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194:597–607. https://doi.org/10.1534/genetics.113.152207

Habier D, Tetens J, Seefried F-R, et al (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42:5. https://doi.org/10.1186/1297-9686-42-5

Hayashi T, Iwata H (2013) A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics 14:34. https://doi.org/10.1186/1471-2105-14-34

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443. https://doi.org/10.3168/jds.2008-1646

Hayes BJ, Panozzo J, Walker CK, et al (2017) Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. Theor Appl Genet 130:2505–2519. https://doi.org/10.1007/s00122-017-2972-7

Hazel LN (1943) The Genetic Basis for Constructing Selection Indexes. Genetics 28:476–490

Heffner EL, Jannink J-L, Sorrells ME (2011) Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. The Plant Genome 4:65–75. https://doi.org/10.3835/plantgenome2010.12.0029

Heffner EL, Lorenz A, Jannink, J-L, Sorrells M (2010). Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Science - CROP SCI. 50. 10.2135/cropsci2009.11.0662.

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic Selection for Crop Improvement. Crop Sci 49:1–12. https://doi.org/10.2135/cropsci2008.08.0512

Henderson CR, Quaas RL (1976) Multiple Trait Evaluation Using Relatives' Records. J Anim Sci 43:1188–1197. https://doi.org/10.2527/jas1976.4361188x

Heslot, N, Feoktistov V (2017) Optimization of selective phenotyping and population design for genomic prediction. BioRxiv, 172064.

Heslot N, Jannink J-L, Sorrells ME (2015) Perspectives for Genomic Selection Applications and Research in Plants. Crop Sci 55:1–12. https://doi.org/10.2135/cropsci2014.03.0249

Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Science 52:146–160. https://doi.org/10.2135/cropsci2011.06.0297

Hospital F, Charcosset A (1997) Marker-Assisted Introgression of Quantitative Trait Loci. Genetics 147:1469–1485

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

Jia Y, Jannink J-L (2012) Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. Genetics 192:1513–1522. https://doi.org/10.1534/genetics.112.144246

Jiang J, Zhang Q, Ma L et al (2015) Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. Heredity, 115(1), 29-36.

Jinek M, Chylinski K, Fonfara I, et al (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821. https://doi.org/10.1126/science.1225829

Jannink J-L (2010) Dynamics of long-term genomic selection. Genet Sel Evol 42:35.

https://doi.org/10.1186/1297-9686-42-35

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics, 9(2), 166-177.

Jarquín D, Lemes da Silva C, Gaynor R C, et al (2017) Increasing genomic-enabled prediction accuracy by modeling genotype× environment interactions in Kansas wheat. The plant genome, 10(2).

Jiang Y, Reif JC (2015) Modeling Epistasis in Genomic Selection. Genetics 201:759–768. https://doi.org/10.1534/genetics.115.177907

Jussiau R, Rigal J, Papet A (2013) Amélioration génétique des animaux d'élevage. In: Educagri éditions.

Kempthorne O, Nordskog AW (1959) Restricted selection indices. Biometrics 15, 10-19

Lado B, Battenfield S, Guzmán C, et al (2017) Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs. The Plant Genome 10:. https://doi.org/10.3835/plantgenome2016.12.0128

Lado B, Matus I, Rodríguez A, et al (2013) Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 (Bethesda) 3:2105–2114. https://doi.org/10.1534/g3.113.007807

Lado B, Vázquez D, Quincke M, et al (2018) Resource allocation optimization with multi-trait genomic prediction for bread wheat (*Triticum aestivum* L.) baking quality. TAG Theor Appl Genet 131:2719–2731. https://doi.org/10.1007/s00122-018-3186-3

Laloë D (1993) Precision and information in linear models of genetic evaluation. Genet Sel Evol 25(6), 557.

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Le Gouis J, Bordes J, Ravel C, et al (2012) Genome-wide association analysis to identify chromosomal regions determining components of earliness in wheat. Theor Appl Genet 124:597–611. https://doi.org/10.1007/s00122-011-1732-3

Lehermeier C, Teyssèdre S, Schön C-C (2017) Genetic Gain Increases by Applying the Usefulness Criterion with Improved Variance Prediction in Selection of Crosses. Genetics 207:1651–1661. https://doi.org/10.1534/genetics.117.300403

Lev-Yadun S, Gopher A, Abbo S (2000) The Cradle of Agriculture. Science 288:1602–1603. https://doi.org/10.1126/science.288.5471.1602

Lin Z, Cogan N O, Pembleton L W, et al (2016) Genetic gain and inbreeding from genomic selection in a simulated commercial breeding program for perennial ryegrass. The Plant Genome, 9(1).

Lin Z, Hayes BJ, Daetwyler HD (2014) Genomic selection in crops, trees and forages: a review. Crop Pasture Sci 65:1177–1191. https://doi.org/10.1071/CP13363

Liu Y, He Z, Appels R, Xia X (2012) Functional markers in wheat: current status and future prospects. Theor Appl Genet 125:1–10. https://doi.org/10.1007/s00122-012-1829-3

Liu J, Liu D, Tao W, et al (2000) Molecular marker-facilitated pyramiding of different genes for powdery mildew resistance in wheat. Plant Breeding 119:21–24. https://doi.org/10.1046/j.1439-0523.2000.00431.x

Liu G, Zhao Y, Gowda M, et al (2016) Predicting Hybrid Performances for Quality Traits through Genomic-Assisted Approaches in Central European Wheat. PLoS One 11:. https://doi.org/10.1371/journal.pone.0158635

Lorenz AJ, Chao S, Asoro FG, et al (2011) Genomic Selection in Plant Breeding. Knowledge and Prospects. Advances in agronomy 110:77–123. https://doi.org/10.1016/B978-0-12-385531-2.00002-5

Lush J L, 1937 Animal Breeding Plans, Collegiate Press Inc, Ames, IA.

Ly D, Huet S, Gauffreteau A, et al (2018) Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. Field Crops Research 216:32–41. https://doi.org/10.1016/j.fcr.2017.08.020

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer.

MacRitchie F (1999) Wheat proteins: characterization and role in flour functionality. Cereal Foods World, 44, 188-193.

Maluszynski M, Kasha K, Forster BP et al. (2013). Doubled haploid production in crop plants: a manual. Springer Science & Business Media.

Marchal A, Legarra A, Tisné S et al (2016). Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis Jacq*.) progeny tests. Molecular breeding, 36(1), 2.

Marcussen T, Sandve SR, Heier L, et al (2014) Ancient hybridizations among the ancestral genomes of bread wheat. Science 345:1250092. https://doi.org/10.1126/science.1250092

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564. https://doi.org/10.1073/pnas.74.2.560

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

Michel S, Kummer C, Gallee M, et al (2018) Improving the baking quality of bread wheat by genomic selection in early generations. Theor Appl Genet 131:477–493. https://doi.org/10.1007/s00122-017-2998-x

Miedaner T, Korzun V (2012) Marker-assisted selection for disease resistance in wheat and barley breeding. Phytopathology, 102(6), 560-566.

Montesinos-López OA, Montesinos-López A, Crossa J, et al (2016) A genomic Bayesian multi-trait and multi-environment model. G3- Genes Genom Genet 6(9), 2725-2744.

Montesinos-López OA, Montesinos-López, A, Crossa, J et al (2018) Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3, 8(12), 3829-3840.

Moreau L, Lemarié S, Charcosset A, Gallais A (2000) Economic efficiency of one cycle of marker-assisted selection. Crop Science 40:329–337. https://doi.org/10.2135/cropsci2000.402329x

Mujeeb-Kazi A, Sitch LA (1989) Review of advances in plant biotechnology, 1985-88. Int. Rice Res. Inst.

Müller D, Schopp P, Melchinger AE (2018) Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. G3 (Bethesda) G3.200091.2018

Muñoz P R, Resende M F, Gezan S A, et al (2014) Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics, 198(4), 1759-1768.

Murphy LR, Santra D, Kidwell K, et al (2009) Linkage maps of wheat stripe rust resistance genes yr5 and yr15 for use in marker-assisted selection. Crop Science 49:1786–1790. https://doi.org/10.2135/cropsci2008.10.0621

Norman A, Taylor J, Tanaka E, et al (2017) Increased genomic prediction accuracy in wheat breeding using a large Australian panel. Theor Appl Genet 130:2543–2555.

https://doi.org/10.1007/s00122-017-2975-4

Ornella L, Singh S, Perez P, et al (2012) Genomic Prediction of Genetic Values for Resistance to Wheat Rusts. The Plant Genome 5:136–148. https://doi.org/10.3835/plantgenome2012.07.0017

Osborne TB (1907) The proteins of the wheat kernel. Washington, D.C. : Carnegie Institution of Washington

Oury F-X, Chiron H, Faye A, et al (2010) The prediction of bread wheat quality: joint use of the phenotypic information brought by technological tests and the genetic information brought by HMW and LMW glutenin subunits. Euphytica 171:87. https://doi.org/10.1007/s10681-009-9997-1

Paterson AH, Damon S, Hewitt JD, et al (1991) Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics 127:181–197

Payne PI (1987) Genetics of Wheat Storage Proteins and the Effect of Allelic Variation on Bread-Making Quality. Ann Rev Plant Physiol 38:141–153. https://doi.org/10.1146/annurev.pp.38.060187.001041

Pérez P, Campos G de los (2014) Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198:483–495. https://doi.org/10.1534/genetics.114.164442

Poland J, Endelman J, Dawson J, et al (2012) Genomic Selection in Wheat Breeding using Genotyping by-Sequencing. The Plant Genome 5:103–113. https://doi.org/10.3835/plantgenome2012.06.0006

Poland JA, Rife TW (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome 5:92–102. https://doi.org/10.3835/plantgenome2012.05.0005

Raina S (1997) Doubled haploid breeding in cereals. Plant breeding reviews, 15, 141-186.

Ravel C, Faye A, Ben-Sadoun S, et al (2020) SNP markers for early identification of high molecular weight glutenin subunits (HMW-GSs) in bread wheat. Theor Appl Genet. https://doi.org/10.1007/s00122-019-03505-y

Riedelsheimer C, Melchinger AE (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor Appl Genet 126:2835–2848. https://doi.org/10.1007/s00122-013-2175-9

Riley R, Chapman V (1958) Genetic Control of the Cytologically Diploid Behaviour of Hexaploid

Wheat. Nature 182:713–715. https://doi.org/10.1038/182713a0

Rimbert H, Darrier B, Navarro J, et al (2018) High throughput SNP discovery and genotyping in hexaploid wheat. PLOS ONE 13:e0186329. https://doi.org/10.1371/journal.pone.0186329

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. https://doi.org/10.1007/s00122-017-2956-7

Rincent R, Charpentier J P, Faivre-Rampant P, et al (2018) Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. G3: Genes, Genomes, Genetics, 8(12), 3961-3972.

Rincent R, Laloë D, Nicolas S, 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. https://doi.org/10.1534/genetics.112.141473

Rousset M, Bonnin I, Remoué C, et al (2011) Deciphering the genetics of flowering time by an association study on candidate genes in bread wheat (*Triticum aestivum* L.). Theor Appl Genet 123:907. https://doi.org/10.1007/s00122-011-1636-2

Rutkoski J, Benson J, Jia Y, et al (2012) Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. The Plant Genome 5:51–61. https://doi.org/10.3835/plantgenome2012.02.0001

Rutkoski JE, Poland JA, Singh RP, et al (2014) Genomic selection for quantitative adult plant stem rust resistance in wheat. The Plant Genome https://doi.org/10.3835/plantgenome2014.02.0006

Rutkoski J, Poland J, Mondal S, et al (2016) Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. G3: Genes Genom Genet 6:2799–2808. https://doi.org/10.1534/g3.116.032888

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467

Santos JPR dos, Vasconcellos RC de C, Pires LPM, et al (2016) Inclusion of Dominance Effects in the Multivariate GBLUP Model. PLOS ONE 11:e0152045. https://doi.org/10.1371/journal.pone.0152045

Sarinelli JM, Murphy JP, Tyagi P, 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. https://doi.org/10.1007/s00122-019-03276-6

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123:218–223. https://doi.org/10.1111/j.1439-0388.2006.00595.x

Schrag T A, Westhues M, Schipprack W, et al (2018). Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics, 208(4), 1373-1385.

Schulthess AW, Wang Y, Miedaner T, et al (2016) Multiple-trait and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. Theor Appl Genet 129:273–287. https://doi.org/10.1007/s00122-015-2626-6

Schulthess AW, Zhao Y, Longin CFH, Reif JC (2018) Advantages and limitations of multiple-trait genomic prediction for Fusarium head blight severity in hybrid wheat (*Triticum aestivum* L.). Theor Appl Genet 131:685–701. https://doi.org/10.1007/s00122-017-3029-7

Sears ER, Okamoto M (1958) Intergenomic chromosome relationships in hexaploid wheat. Proc Int Congress Genet 2:258–259

Servin B, Martin OC, Mézard M, Hospital F (2004) Toward a Theory of Marker-Assisted Gene Pyramiding. Genetics 168:513–523. https://doi.org/10.1534/genetics.103.023358

Shewry PR (2009) Wheat. Journal of Experimental Botany 60:1537–1553. https://doi.org/10.1093/jxb/erp058

Shewry PR, Halford NG (2002) Cereal seed storage proteins: structures, properties and role in grain utilization. J Exp Bot 53:947–958. https://doi.org/10.1093/jexbot/53.370.947

Shewry PR, Halford NG, Lafiandra D (2003) Genetics of wheat gluten proteins. Adv Genet 49:111–184. https://doi.org/10.1016/s0065-2660(03)01003-4

Shewry PR, Hey SJ (2015) The contribution of wheat to human diet and health. Food Energy Secur 4:178–202. https://doi.org/10.1002/fes3.64

Shewry PR, Tatham AS, Forde J, et al (1986) The classification and nomenclature of wheat gluten proteins: A reassessment. Journal of Cereal Science 4:97–106. https://doi.org/10.1016/S0733-5210(86)80012-1

Simmonds NW (1995) The relation between yield and protein in cereal grain. Journal of the Science of Food and Agriculture 67:309–315. https://doi.org/10.1002/jsfa.2740670306

Smith HF (1936) A Discriminant Function for Plant Selection. Annals of Eugenics 7:240–250. https://doi.org/10.1111/j.1469-1809.1936.tb02143.x

Snape JW (1989) Doubled haploid breeding: theoretical basis and practical applications. In: Mujeeb-Kazi A, Sitch LA (eds) Review of advances in plant biotechnology, 1985-88.

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454. https://doi.org/10.2527/jas.2007-0010

Storlie E, Charmet G (2013) Genomic Selection Accuracy using Historical Data Generated in a Wheat Breeding Program. Plant Genome 6. https://doi.org/10.3835/plantgenome2013.01.0001

Su G, Christensen O F, Ostersen T, et al (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS one, 7(9).

Sun J, Rutkoski JE, Poland JA, et al (2017) Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. Plant Genome 10:. https://doi.org/10.3835/plantgenome2016.11.0111

Surget A, Barron C (2005) Histologie du grain de blé. Industrie des Céréales 145, 3-7

Tadesse W, Inagaki M, Tawkaz S, (2012) Recent advances and application of doubled haploids in wheat breeding. African Journal of Biotechnology, 11(89), 15484-15492.

Technow F, Riedelsheimer C, Schrag T A, et al (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet. 125(6), 1181-1194.

Venables W N, Ripley B D (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

Wang C, Prakapenka D, Wang S, et al (2014) GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. BMC bioinformatics,

15(1), 270.

Wang X, Li L, Yang Z, et al (2017) Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. Heredity, 118(3), 302-310.

Watson JD, Crick FHC (1953) The Structure of Dna. Cold Spring Harb Symp Quant Biol 18:123–131. https://doi.org/10.1101/SQB.1953.018.01.020

Wellmann R (2019) Optimum contribution selection for animal breeding and conservation: the R package optiSel. BMC Bioinformatics 20:25. https://doi.org/10.1186/s12859-018-2450-5

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genetics Research 75:249–252. https://doi.org/10.1017/S0016672399004462

Wilhelm EP, Boulton MI, Al-Kaff N, et al (2013) Rht-1 and Ppd-D1 associations with height, GA sensitivity, and days to heading in a worldwide bread wheat collection. Theor Appl Genet 126:2233–2243. https://doi.org/10.1007/s00122-013-2130-9

Wrigley CW, Shepherd KW (1973) Electrofocusing of grain proteins from wheat genotypes. Ann N Y Acad Sci 209:154–162. https://doi.org/10.1111/j.1749-6632.1973.tb47526.x

Zhao Y, Li Z, Liu G, et al (2015) Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc Natl Acad Sci USA 112:15624–15629. https://doi.org/10.1073/pnas.1514547112

Zhao Y, Zeng J, Fernando R, Reif JC (2013) Genomic Prediction of Hybrid Wheat Performance. Crop Science 53:802–810. https://doi.org/10.2135/cropsci2012.08.0463

Zhong S, Jannink J-L (2007) Using Quantitative Trait Loci Results to Discriminate Among Crosses on the Basis of Their Progeny Mean and Variance. Genetics 177:567–576. https://doi.org/10.1534/genetics.107.075358

Liste des annexes

Annexes du chapitre 3

Table S1 Location of trials

Table S2 Variation of the cost ratio between W and BMS and its impact on the predictive ability of BMS

Figure S1 Pearson correlations between main traits and traits linked to bread making quality

Figure S2 Average number of lines of the validation set to be phenotyped for W for each resource allocation scenario

Appendix

Annexes du chapitre 4

Table S3 Description of the 36 scenarios

Table S4 Impact of input parameters on the genetic gain for the 10, 50, 100 and 200 best lines (ANOVA results).

Figures S3, S4, S5 Evolution of the cumulative genetic gain at the end of cycle for several values of CT and C_G

Figures S6, S7, S8 Evolution of the genetic gain obtained at the end of cycle for several values of CT and C_G

Figures S9, S10, S11 Evolution of the genetic diversity for several values of CT and C_G

Autre annexe

Ravel C, Faye A, Ben Sadoun S, Ranoux M, et al. (2020). « SNP markers for early identification of high molecular weight glutenin subunits (HMW-GSs) in bread wheat. » Theoretical and Applied Genetics, 1-20.

Name of trial	GPS coordinates
Champagne-Céréales	48°52'N/4°09'E
Clermont-Ferrand	45°46'N/3°04'E
Colmar	48°03'N/7°20'E
Dijon	47°19'N/5°04'E
Epi Centre	47°05'N/2°25'E
Estrées-Mons	49°52'N/3°00'E
Le Moulon	48°42'N/2°09'E
Lusignan	46°25'N/0°11'E
Orgeval	48°55'N/1°58'E
Orsonville	48°27'N/1°52'E
Rennes	48°06'N/1°40'W

Table S1: Location of trials

Budget	Method	W cost per line (€)	BMS cost per line (€)	Ratio between W and BMS costs	N_W	N_{BMS}	Predictive ability of BMS and sd
47700	TA	75	150	1/2	398	119	0.48 ± 0.06
		50	150	1/3	398	185	0.48 ± 0.06
		37.5	150	1/4	398	218	0.50 ± 0.06
		30	150	1/5	398	238	0.51 ± 0.05
		15	150	1/10	398	278	0.51 ± 0.05
		7.5	150	1/20	398	298	0.52 ± 0.04
	MT	75	150	1/2	318	159	0.35 ± 0.12
		50	150	1/3	318	212	0.37 ± 0.11
		37.5	150	1/4	318	238	0.37 ± 0.11
		30	150	1/5	318	254	0.38 ± 0.11
		15	150	1/10	318	286	0.38 ± 0.10
		7.5	150	1/20	318	302	0.38 ± 0.10
30 210	TA	75	150	1/2	398	2	NA
		50	150	1/3	398	68	0.46 ± 0.06
		37.5	150	1/4	398	101	0.48 ± 0.06
		30	150	1/5	398	121	0.50 ± 0.06
		15	150	1/10	398	161	0.51 ± 0.05
		7.5	150	1/20	398	181	0.52 ± 0.05
	MT	75	150	1/2	318	42	0.31 ± 0.12
		50	150	1/3	318	95	0.34 ± 0.12
		37.5	150	1/4	318	121	0.36 ± 0.12
		30	150	1/5	318	137	0.36 ± 0.11
		15	150	1/10	318	169	0.37 ± 0.11
		7.5	150	1/20	318	185	0.38 ± 0.11
16760	TA	75	150	1/2	223	NA	NA
		50	150	1/3	335	NA	NA
		37.5	150	1/4	398	12	NA
		30	150	1/5	398	32	0.45 ± 0.07
		15	150	1/10	398	71	0.48 ± 0.07
		7.5	150	1/20	398	91	0.50 ± 0.06
	MT	75	150	1/2	223	NA	NA
		50	150	1/3	318	5	NA
		37.5	150	1/4	318	32	0.23 ± 0.16
		30	150	1/5	318	48	0.26 ± 0.15
		15	150	1/10	318	79	0.31 ± 0.12
		7.5	150	1/20	318	95	0.32 ± 0.12

Table S2 : Variation of cost ratio between W (cheap) and BMS (expensive) and its impact on the predictive ability of BMS

N_W, N_{BMS}: Number of lines phenotyped for W and BMS, respectively. Sd: Standard deviation. N_{BMS} is not available when the budget is not enough to phenotype all the lines for W. Predictive ability is not available (NA) when N_{BMS} is a negative number and when N_{BMS} is small because of convergence model issues.

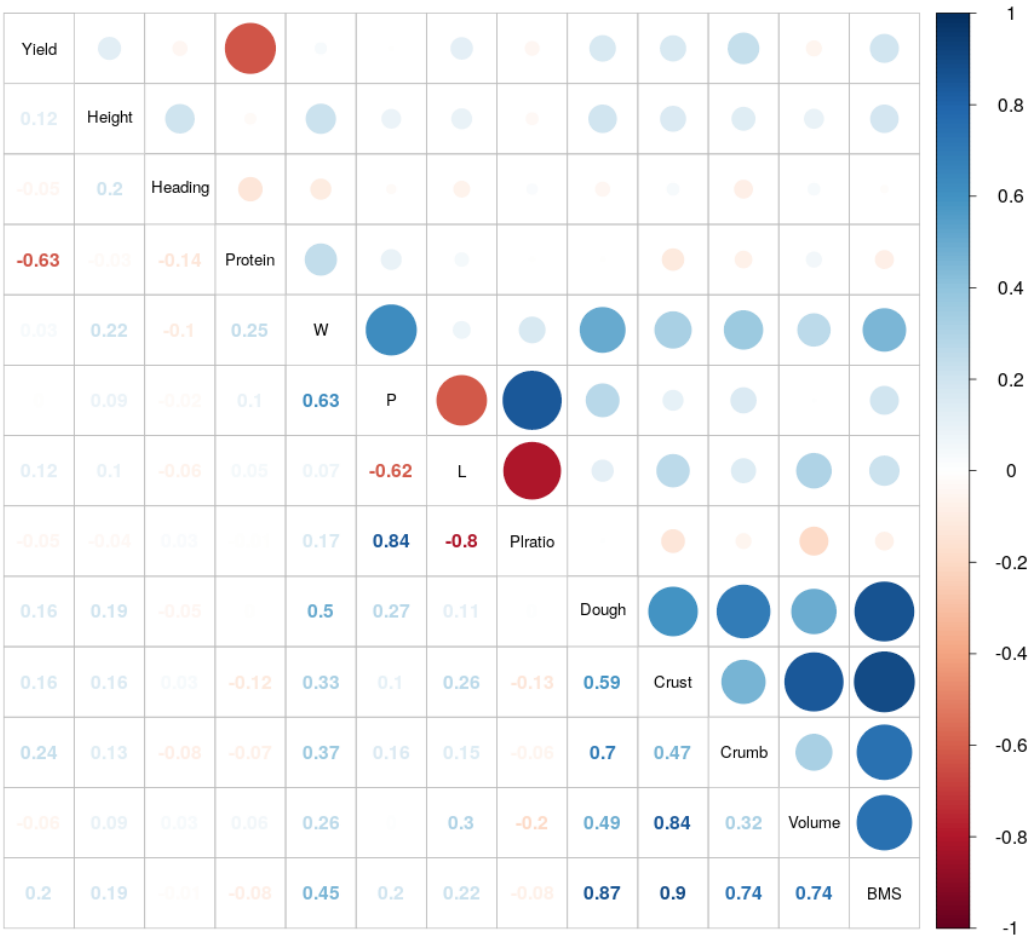

Figure S1: Pearson correlations between main traits and traits linked to bread making quality

Yield: grain yield. Height: plant height. Heading: heading date. Protein: protein content. W: dough strength. P: Tenacity. L: Extensibility. Plratio: ratio Tenacity / Extensibility (P/L). Dough, Crust and Crumb: scores for dough, crust and crumb, respectively (obtained with the BIPEA test). Volume: dough volume. BMS: Bread Making Score (final BIPEA score)

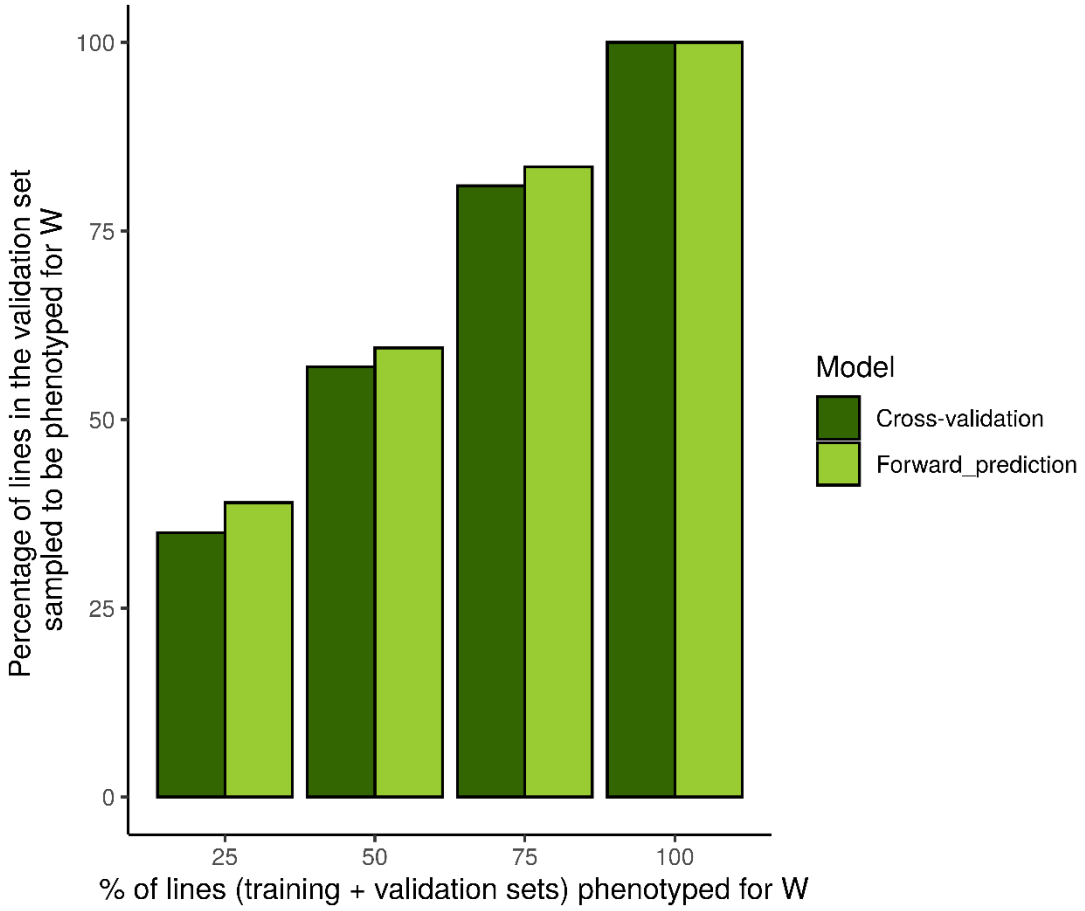

Figure S2: Average number of lines of the validation set to be phenotyped for W for each resource allocation scenario

Appendix

We extended here the generalized CD (Laloë 1993) to the multi-trait context. The objective is to compute the expected reliability (before phenotyping) in a multi-trait context for contrasts corresponding to the prediction objectives. To give an example, it is clear that if the objective is to accurately predict the difference (contrasts) between families, or to focus on predictions within families, the contrasts to be considered will be different and the optimal calibration sets as well. By defining contrasts corresponding to the prediction objectives, one can adapt the CDmulti criterion to the specific prediction objectives (see Rincent et al. 2017 for more details in the single trait context).

For a given contrast c, the generalized Prediction Error Variance (PEV(c)) in the multi-trait context is equal to:

$$PEV(c) = c^T (Z^T M Z + (\Sigma_a{}^{-1} \otimes K^{-1}))^{-1} c \qquad (26)$$

With

$$M = (\Sigma_\varepsilon{}^{-1} \otimes I) - (\Sigma_\varepsilon{}^{-1} \otimes I) X (X^T (\Sigma_\varepsilon{}^{-1} \otimes I) X)^{-1} X^T (\Sigma_\varepsilon{}^{-1} \otimes I) \qquad (27)$$

where X and Z are the design matrices for the fixed and random effects respectively, K is the kinship matrix, Σ_a is the genetic variance–covariance matrix between traits, and Σ_ε is the residual variance-covariance matrix between traits. $\otimes$ indicate the Kronecker product operator between matrices.

The generalized multi-trait CD for a given contrast c is equal to:

$$CDmulti\ (c) = \frac{c^T((\Sigma_a \otimes K) - \left(Z^T M Z + (\Sigma_a{}^{-1} \otimes K^{-1})\right)^{-1})c}{c^T (\Sigma_a \otimes K) c} \qquad (28)$$

As a reminder, a contrast is a vector whose elements sum to 0 and indicating the difference in which we are interested. For example, if we are interested in accurately predicting the difference between individual 1 and individual 2, the contrast to consider will be: $c^T = [1, -1, 0, 0, 0, 0, \dots]$.

Scenario	CT (€)	Average annual CT (€)	C_G (€)		λ	h^2
1	22500000	1500000		37	0.75	0.7
2	45000000	3000000		37	0.75	0.7
3	22500000	1500000		37	0.5	0.7
4	45000000	3000000		37	0.5	0.7
5	22500000	1500000		37	0.25	0.7
6	45000000	3000000		37	0.25	0.7
7	22500000	1500000		37	0.75	0.4
8	45000000	3000000		37	0.75	0.4
9	22500000	1500000		37	0.5	0.4
10	45000000	3000000		37	0.5	0.4
11	22500000	1500000		37	0.25	0.4
12	45000000	3000000		37	0.25	0.4
13	22500000	1500000		37	0.75	0.2
14	45000000	3000000		37	0.75	0.2
15	22500000	1500000		37	0.5	0.2
16	45000000	3000000		37	0.5	0.2
17	22500000	1500000		37	0.25	0.2
18	45000000	3000000		37	0.25	0.2
19	22500000	1500000		10	0.75	0.7
20	45000000	3000000		10	0.75	0.7
21	22500000	1500000		10	0.5	0.7
22	45000000	3000000		10	0.5	0.7
23	22500000	1500000		10	0.25	0.7
24	45000000	3000000		10	0.25	0.7
25	22500000	1500000		10	0.75	0.4
26	45000000	3000000		10	0.75	0.4
27	22500000	1500000		10	0.5	0.4
28	45000000	3000000		10	0.5	0.4
29	22500000	1500000		10	0.25	0.4
30	45000000	3000000		10	0.25	0.4
31	22500000	1500000		10	0.75	0.2
32	45000000	3000000		10	0.75	0.2
33	22500000	1500000		10	0.5	0.2
34	45000000	3000000		10	0.5	0.2
35	22500000	1500000		10	0.25	0.2
36	45000000	3000000		10	0.25	0.2

Table S3: Description of the 36 scenarios

Trait	Factor	F	P value	% of SS
Genetic gain 10	**Strategy**	759.2	< 2e-16	9.3
	λ	404.9	< 2e-16	1.7
	h²	6 664.9	< 2e-16	27.5
	CT	405.9	< 2e-16	1.7
	QTL sampling	132.2	< 2e-16	0.5
Genetic gain 50	**Strategy**	1 230.6	< 2e-16	12.9
	λ	378.1	< 2e-16	1.3
	h²	9 455.0	< 2e-16	32.9
	CT	574.5	< 2e-16	2
	QTL sampling	219.5	< 2e-16	0.8
Genetic gain 100	**Strategy**	1 475.2	< 2e-16	14.3
	λ	325.0	< 2e-16	1
	h²	10 942.2	< 2e-16	35.3
	CT	650.0	< 2e-16	2.1
	QTL sampling	271.3	< 2e-16	0.9
Genetic gain 200	**Strategy**	1 899.7	< 2e-16	16.1
	λ	244.3	< 2e-16	0.7
	h²	14 028.5	< 2e-16	39.6
	CT	780.3	< 2e-16	2.2
	QTL sampling	290.2	3.22e-12	0.8

Table S4: Impact of input parameters on the genetic gain for the 10, 50, 100 and 200 best lines (ANOVA results).

h^2: trait heritability. λ: relative intensity of selection at step 3 compared to step 4. CT: Total cost. C_G: genotyping cost (37€). % of SS: (Sum of squares)/(Total sum of squares)

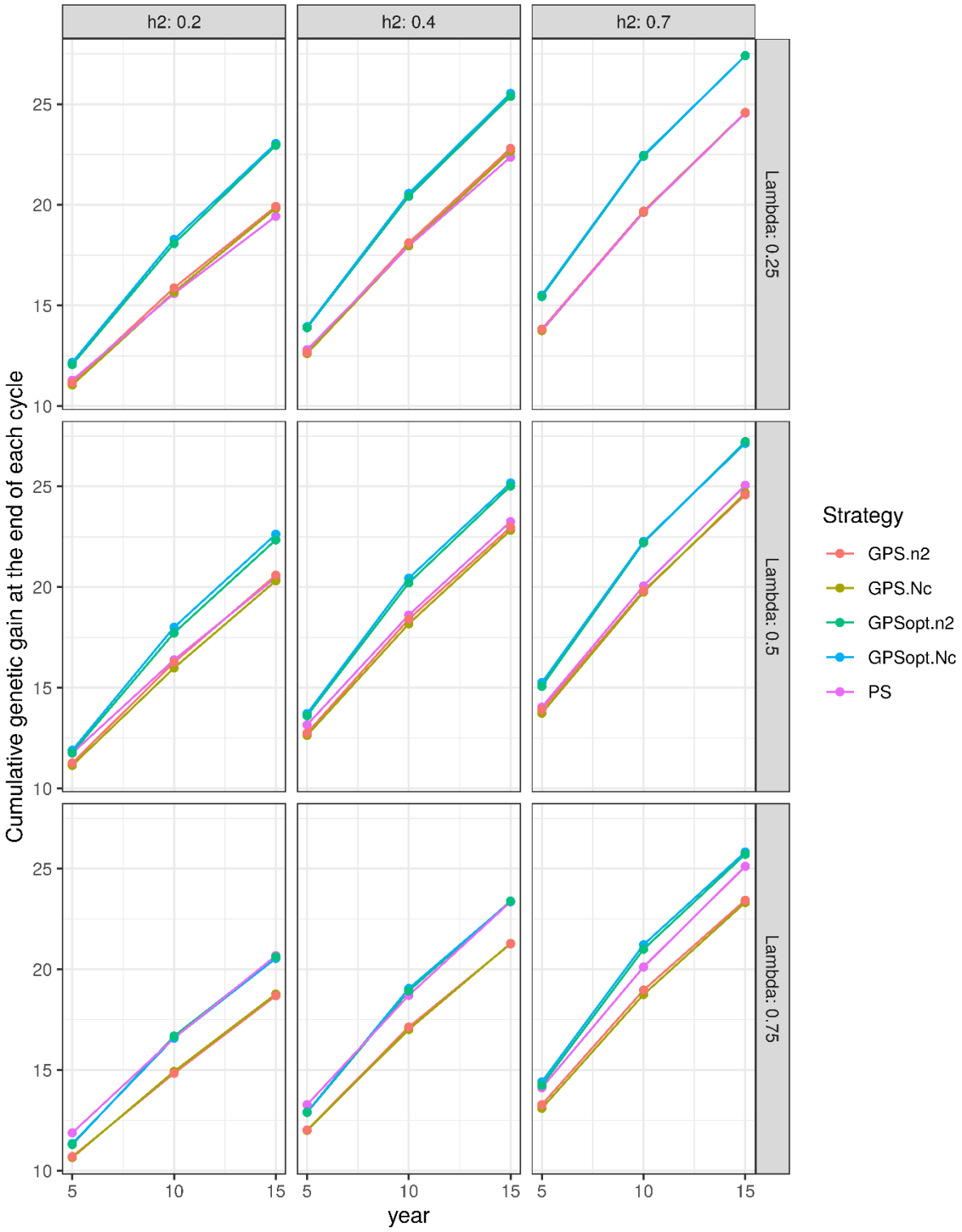

Figure S3: Evolution of the cumulative genetic gain at the end of cycle.

Annual total cost (CT) = 3 000 000€ and genotyping cost (C_G) = 10€.

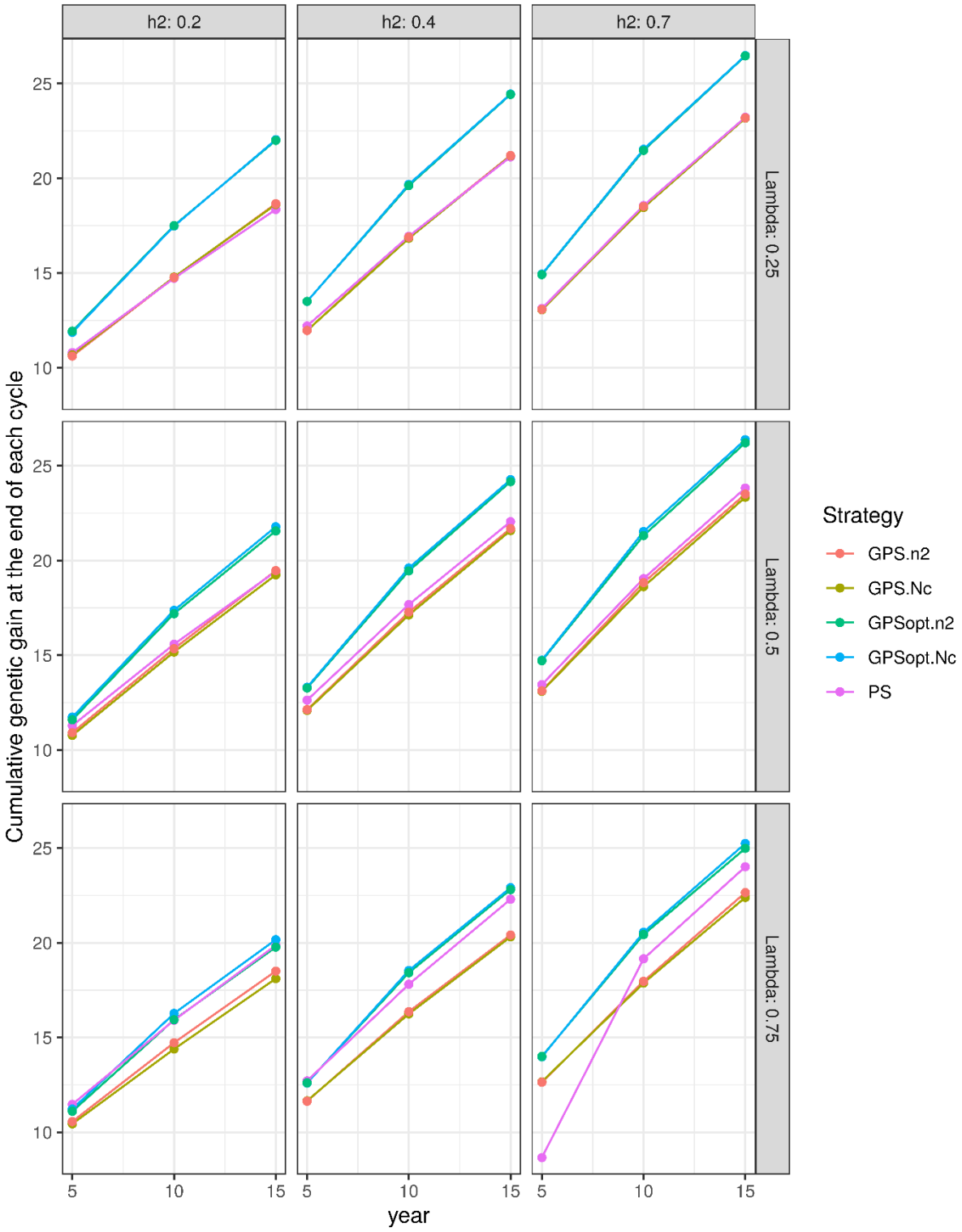

Figure S4: Evolution of the cumulative genetic gain at the end of cycle.

Annual total cost (CT) = 1 500 000€ and genotyping cost (C_G) = 37€.

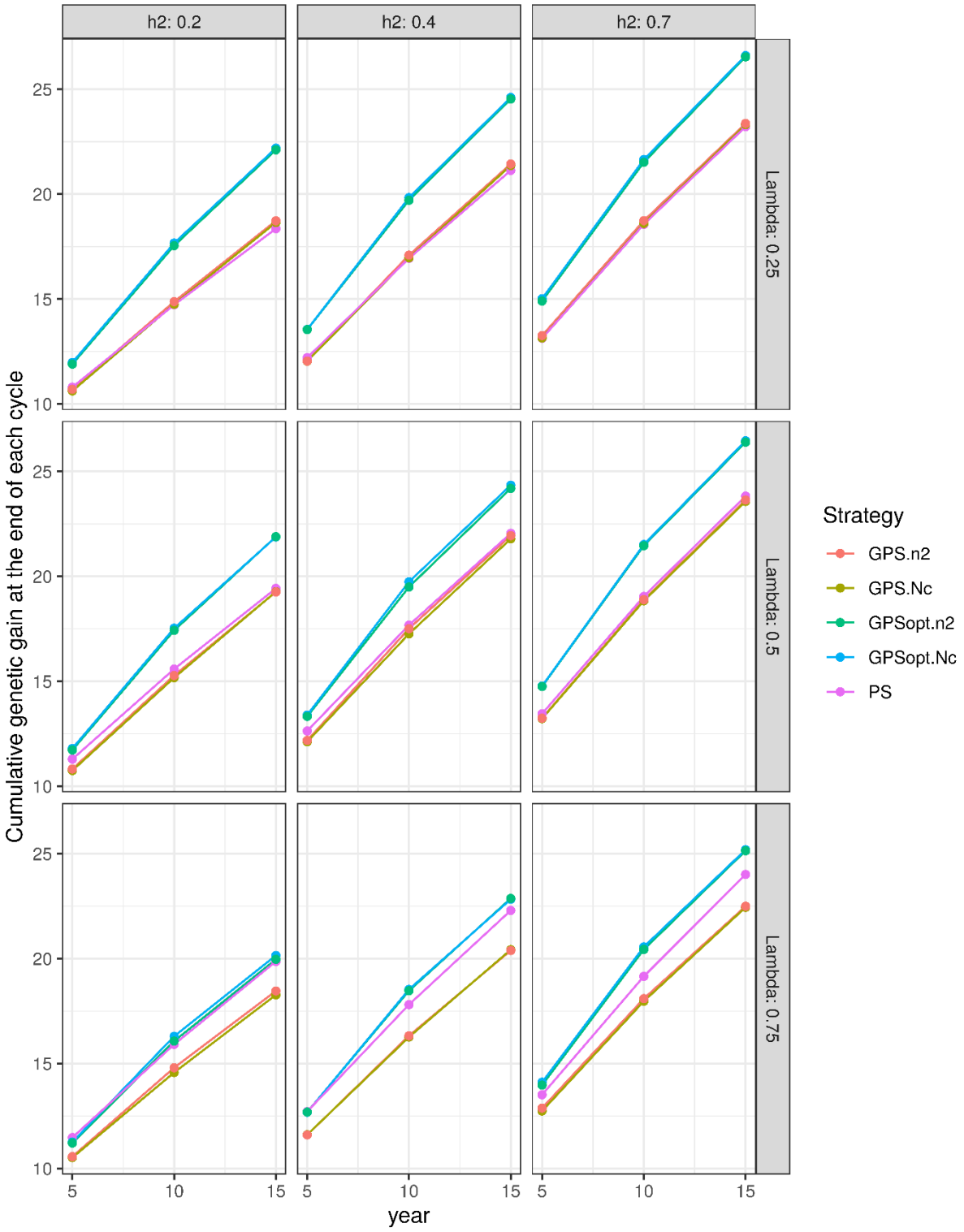

Figure S5: Evolution of the cumulative genetic gain at the end of cycle.

Annual total cost (CT) = 1 500 000€ and genotyping cost (C_G) = 10€.

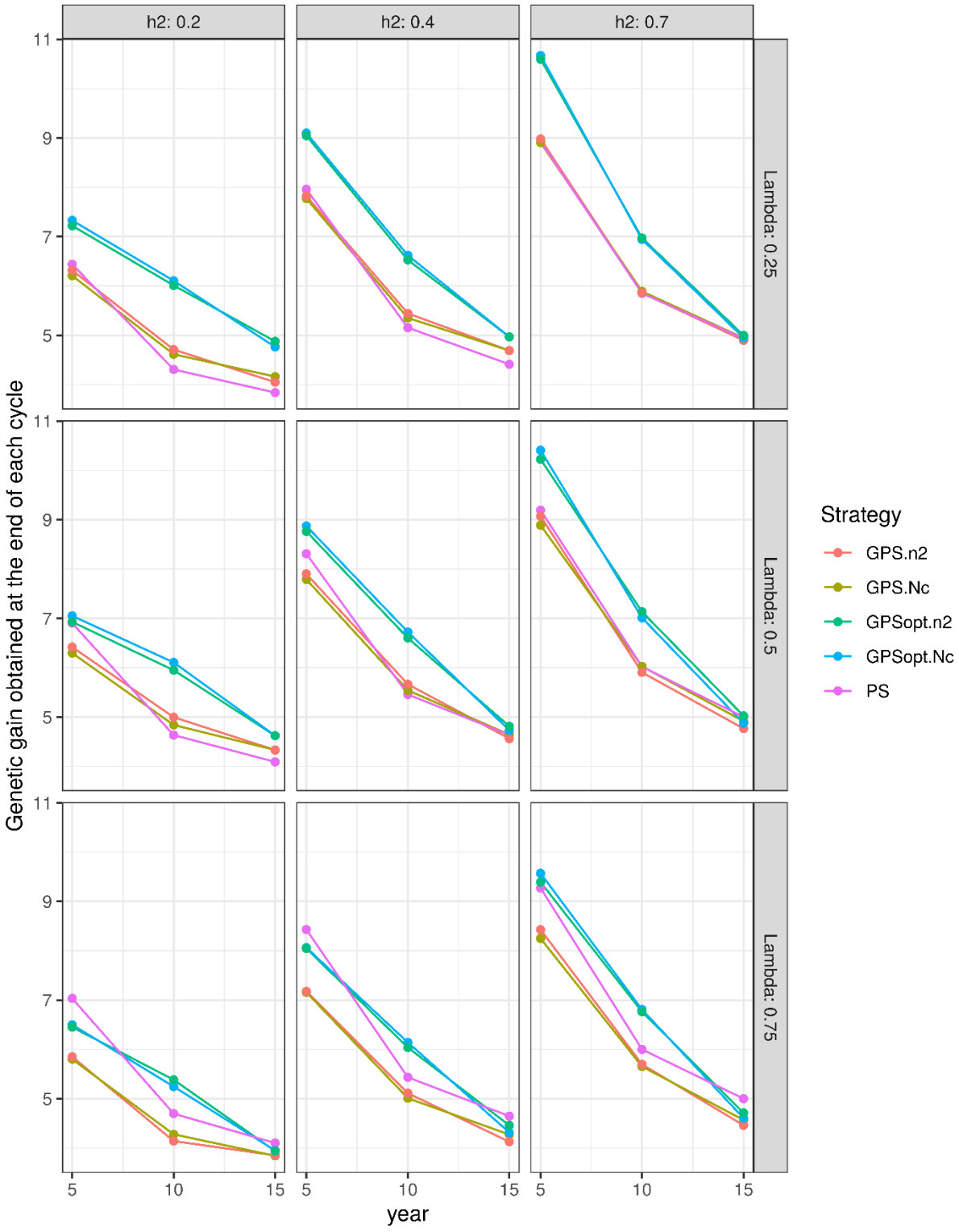

Figure S6: Evolution of the genetic gain achieved at the end of cycle

Annual total cost (CT) = 3 000 000€ and genotyping cost (C_G) = 10€.

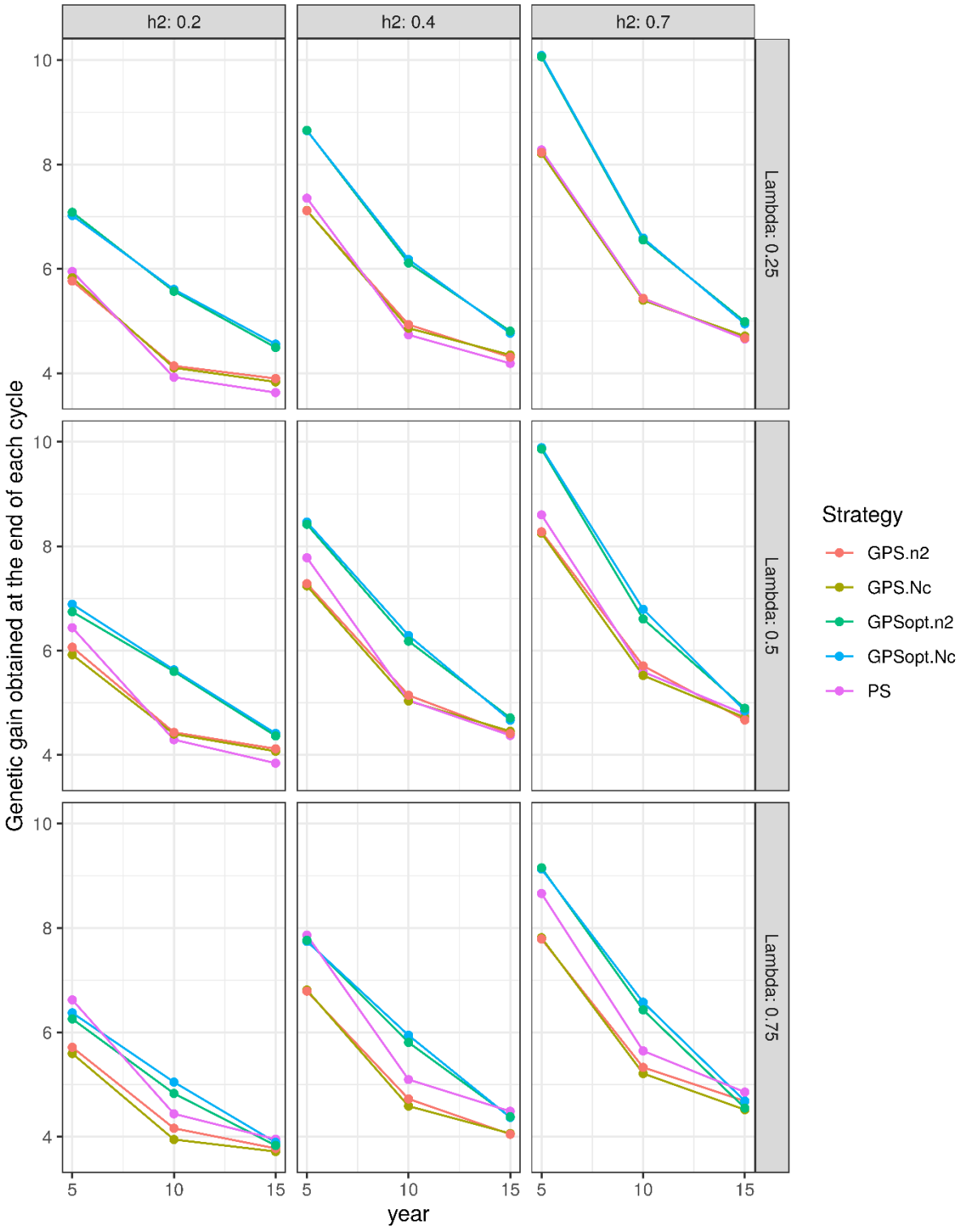

Figure S7: Evolution of the genetic gain achieved at the end of cycle

Annual total cost (CT) = 1 500 000€ and genotyping cost (C_G) = 37€.

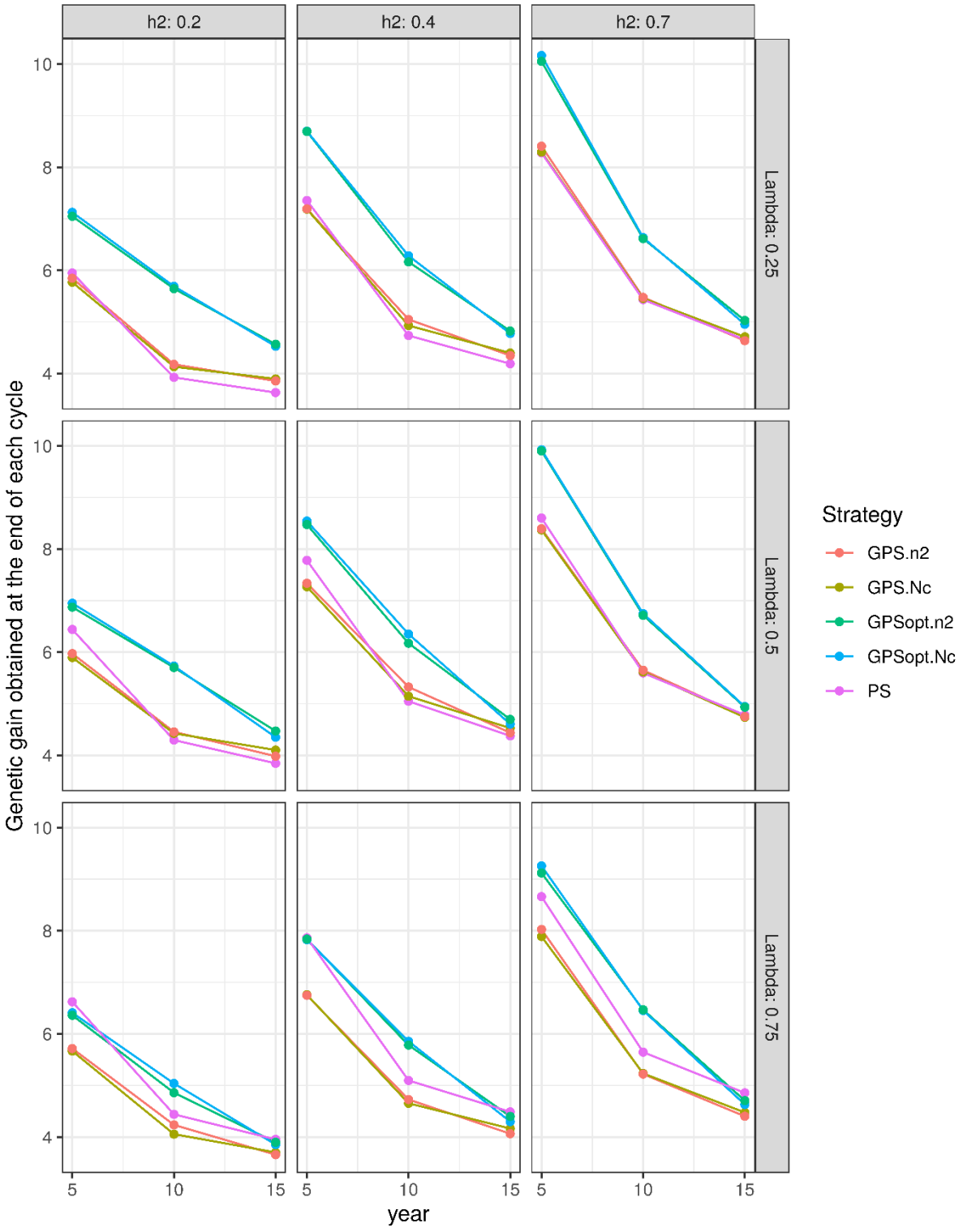

Figure S8: Evolution of the genetic gain achieved at the end of cycle

Annual total cost (CT) = 1 500 000€ and genotyping cost (C_G) = 10€.

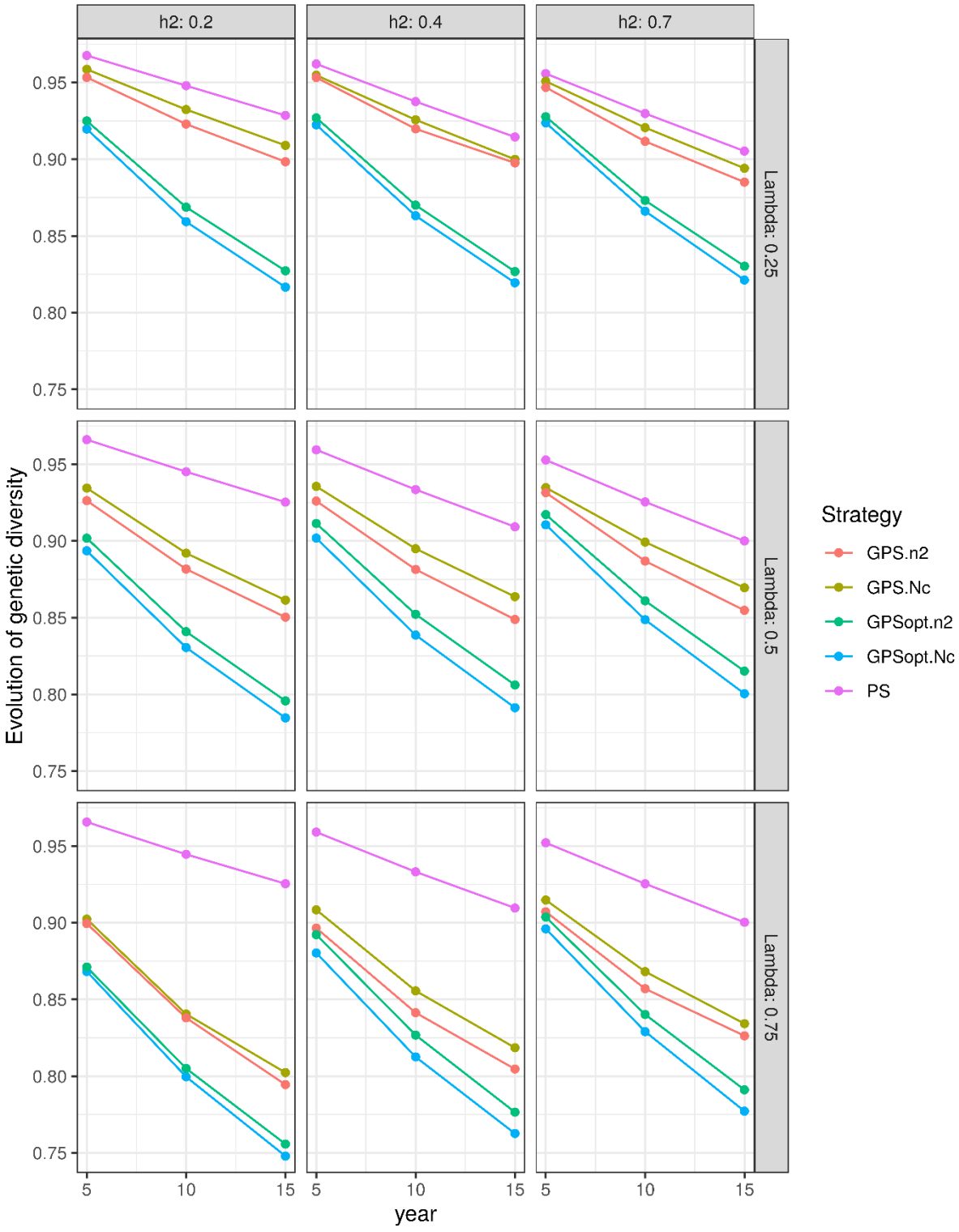

Figure S9: Evolution of the genetic diversity.

Annual total cost (CT) = 3 000 000€ and genotyping cost (C_G) = 10€.

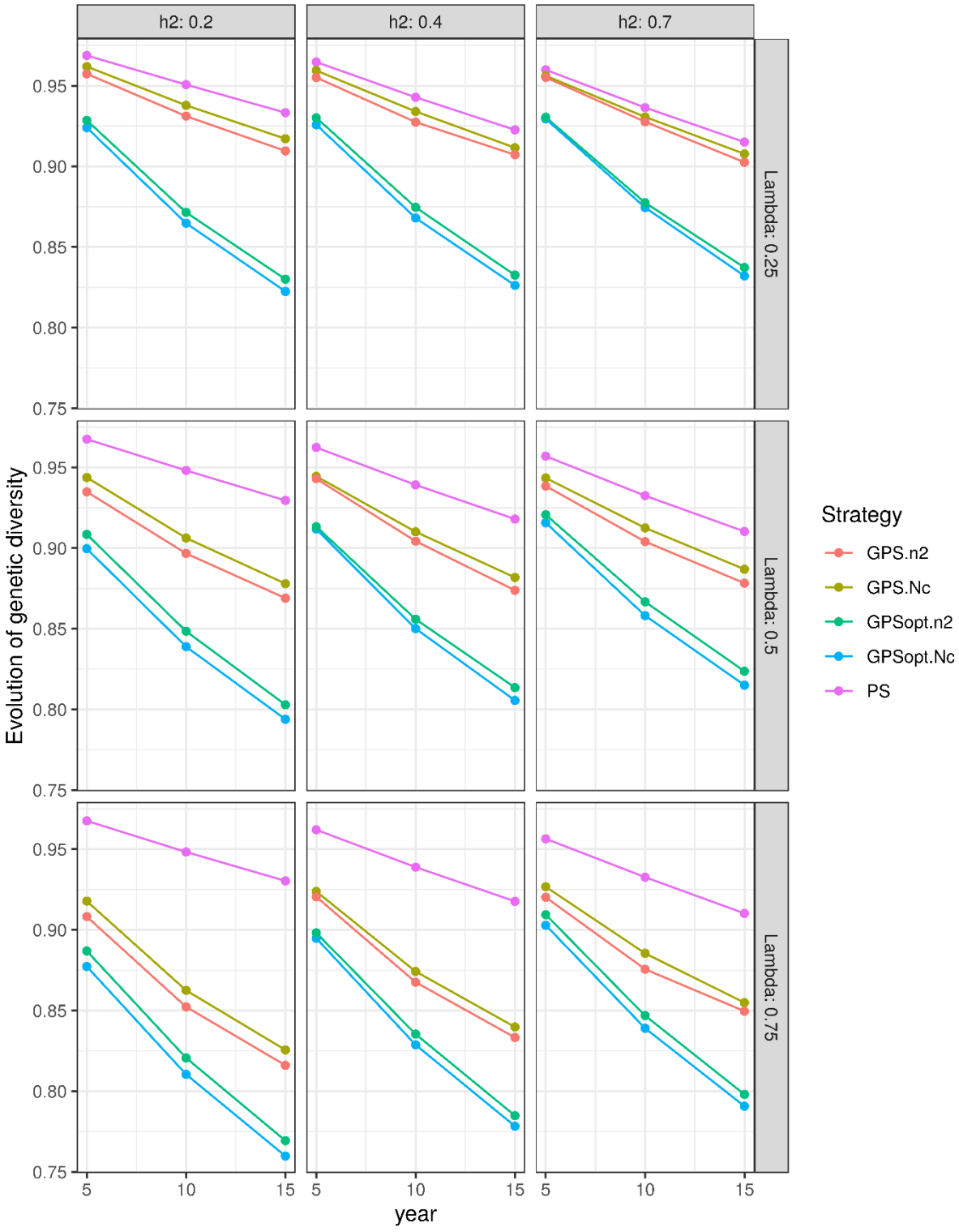

Figure S10: Evolution of the genetic diversity.

Annual total cost (CT) = 1 500 000€ and genotyping cost (C_G) = 37€.

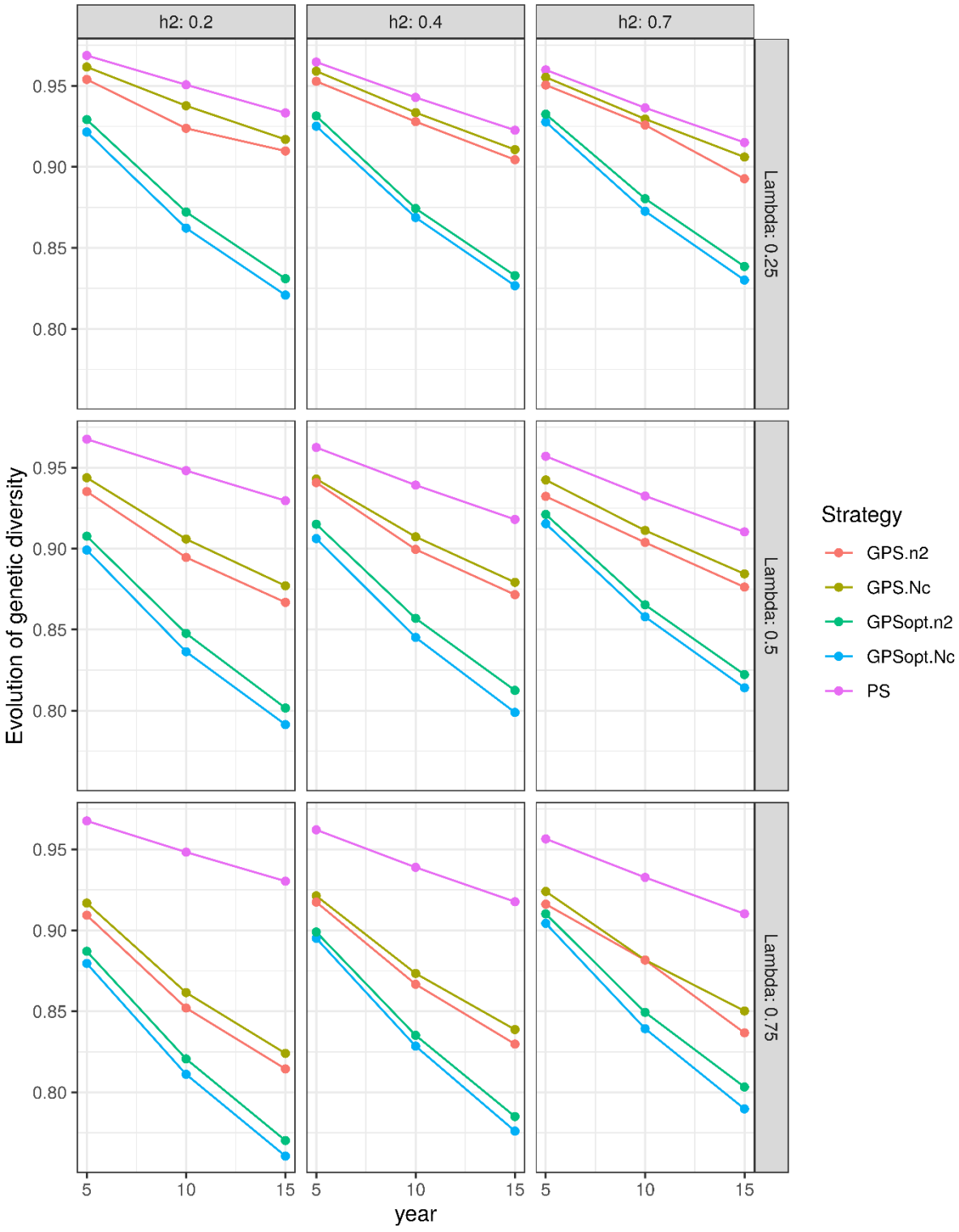

Figure S11: Evolution of the genetic diversity.

Annual total cost (CT) = 1 500 000€ and genotyping cost (C_G) = 10€.

Ravel C, Faye A, Ben Sadoun S, Ranoux M, et al. (2020). « SNP markers for early identification of high molecular weight glutenin subunits (HMW-GSs) in bread wheat. » Theoretical and Applied Genetics, 1-20.

https://link.springer.com/article/10.1007/s00122-019-03505-y

Optimisation du schéma de sélection chez le blé tendre : apport des prédictions génomiques et des caractères corrélés

Résumé

Chez le blé tendre, la sélection variétale consiste à créer de nouvelles variétés regroupant plusieurs caractères d'intérêt agronomique. L'objectif de la thèse était d'étudier l'apport des prédictions génomiques pour optimiser les programmes de sélection chez le blé tendre. Dans un premier temps, nous avons analysé des méthodes visant à améliorer la précision les prédictions génomiques de la qualité boulangère sans pour autant augmenter le budget alloué au phénotypage. Nous nous sommes plus particulièrement intéressés à l'utilisation de modèles de prédiction génomique multi-caractère et à des méthodes permettant d'optimiser le choix des lignées à phénotyper en priorité. Nous avons montré que les prédictions génomiques multi-caractères étaient particulièrement intéressantes lorsque les lignées de la population de validation, ou au moins une partie d'entre elles, avaient été phénotypées pour la force boulangère, caractère corrélé à la qualité boulangère et dont le phénotypage est moins coûteux. En effet, cette approche permettait de réduire le budget alloué au phénotypage sans diminuer la précision des prédictions de la qualité boulangère. Grâce à l'analyse de schémas de sélection simulés, nous avons constaté que pour un budget fixe le gain génétique était plus élevé pour les schémas de sélection intégrant des prédictions génomiques utilisées pour prédire la valeur génétique individuelle de candidats à la sélection et pour prédire la valeur des croisements afin d'optimiser le plan de croisement. En revanche, la perte de diversité génétique était plus intense et plus rapide dans ce type de schémas de sélection. Des pistes de recherches complémentaires et des méthodes permettant d'améliorer les simulations de schémas de sélection ont été suggérées.

Mots clés : Prédictions génomiques, Multi-caractère, Optimisation économique, Simulations, Qualité boulangère, Blé tendre

Abstract

Bread wheat breeding consists in creating new varieties which combine several traits of agronomic interest. The objective of the phD was to study the contribution of genomic predictions in order to optimize bread wheat breeding programs. First, we analyzed methods aiming at improving the genomic prediction accuracy of bread-making quality without increasing the budget allocated to phenotyping. More specifically, we focused on multi-trait genomic prediction models and methods that deal with the optimization of the choice of lines to phenotype. We showed that multi-trait genomic predictions could be particularly interesting when lines of the validation set, or at least a part of them, were phenotyped for dough strength, which is correlated to bread-making quality and which is cheaper to phenotype. Indeed, this approach allowed to reduce the budget allocated to phenotyping without decreasing the genomic prediction accuracy of bread-making quality. Thanks to the analysis of simulated breeding schemes, we noticed that for a given budget the genetic gain was superior for breeding schemes integrating genomic predictions used to predict the performance of new candidates for selection or to predict the crosses value in order to optimize the crosses design. However, the loss of genetic diversity was more intense in this type of breeding schemes. Additional research strategies and methods to improve the simulations of breeding programs were also suggested.

Key words: Genomic predictions, Multi-trait, Economical optimization, Simulations, Bread-making quality, Bread wheat