

PrediCroit : Estimation de la capacité de prédiction génomique de la valeur d'un croisement

Claire OGET-EBRAD^{1,8}, Emmanuel HEUMEZ², Laure DUCHALAIS³, Ellen GOUEMAND-DUGUE⁴, Sandrine BERGES⁵, Patrice WALCZAK⁶, Paul BATAILLON⁷, François-Xavier OURY¹, Jean-Michel ELSEN⁸, Sophie BOUCHET¹

1 - INRAE-Université Clermont-Auvergne, UMR 1095, GDEC, 5 chemin de Beaulieu 63100 Clermont-Ferrand, FRANCE

2 - INRAE-UE 0972, GCIE, 2 chaussée Brunehaut, Estrées Mons, BP50136, 80203 Peronne Cedex, FRANCE

3 - AGRI-Obtentions, chemin de la Petite Minière, 78280 Guyancourt, FRANCE

4 - FLORIMOND DESPREZ VEUVE & FILS, 59242 Cappelle-en-Pévèle, FRANCE

5 - INRAE-UE 1375, PHACC, 5 chemin de Beaulieu 63100 Clermont-Ferrand, FRANCE

6 - INRAE UE 1373, FERLUS, Les Verrines, CS80006, 86600 Lusignan, FRANCE

7 - INRAE UE 0802, APC, 24 chemin de Borde-Rouge, CS52627, 31326 Castanet-Tolosan, FRANCE

8 - INRAE-Université de Toulouse, UMR1388, GenPhySE, Castanet-Tolosan, FRANCE

*Coordinateur du projet : Sophie BOUCHET, sophie.bouchet@inrae.fr

INTRODUCTION

Les prédictions génomiques peuvent être utilisées pour optimiser le choix des croisements, pour un ou plusieurs caractères quantitatifs simultanément (Zhong and Jannink 2007; Lehermeier *et al.* 2017a; Neyhart *et al.* 2019; Danguy des Déserts *et al.* 2023).

L'objectif du FSOV PrediCroit est de tester notre capacité de prédiction de la valeur d'un croisement à partir de marqueurs moléculaires ainsi que la corrélation entre deux caractères dans une descendance. Pour preuve de concept, nous testons ici notre capacité à prédire les meilleurs croisements pour le rendement, la date d'épiaison, la hauteur et le taux de protéine. Nous estimons aussi la corrélation entre le rendement et le taux de protéine. L'objectif du projet est (1) de comparer la moyenne (PM : Parental Mean), l'écart-type (SD : Standard Deviation) et la moyenne des meilleurs

individus (UC : Utility Criterion) dans une descendance (Figure 1) ainsi que la corrélation entre caractères ($r_{G(1,2)}$), estimés analytiquement et les valeurs observées au champ, (2) d'identifier par simulation les facteurs qui influencent notre capacité de prédiction de ces composantes, (3) de tester si une formule analytique d'estimation de SD qui prend en compte l'erreur d'estimation des effets des marqueurs et la composition génomique des parents améliore l'estimation.

Pour ceci, nous avons évalué dans six lieux et quatre années 6 365 descendants issus de 101 croisements avec 55 descendants en moyenne provenant des programmes de sélection INRAE/Agri-obtentions et Florimond-Desprez (FD).

Un article résume une partie des résultats (Oget *et al.* 2024).

MATÉRIEL ET MÉTHODE

Prédiction déterministe de la valeur d'un croisement

Grâce aux prédictions génomiques, il est possible de prédire la moyenne (PM : parental mean) et l'écart type (SD : standard deviation) d'un croisement, ainsi que la moyenne des meilleurs descendants (UC : Usefulness Criterion), la corrélation entre deux caractères ($r_{G(1,2)}$) ainsi que la moyenne des meilleurs individus pour des caractères secondaires (UCcor).

Pour le caractère principal :

D'après Schnell et Utz (1975) :

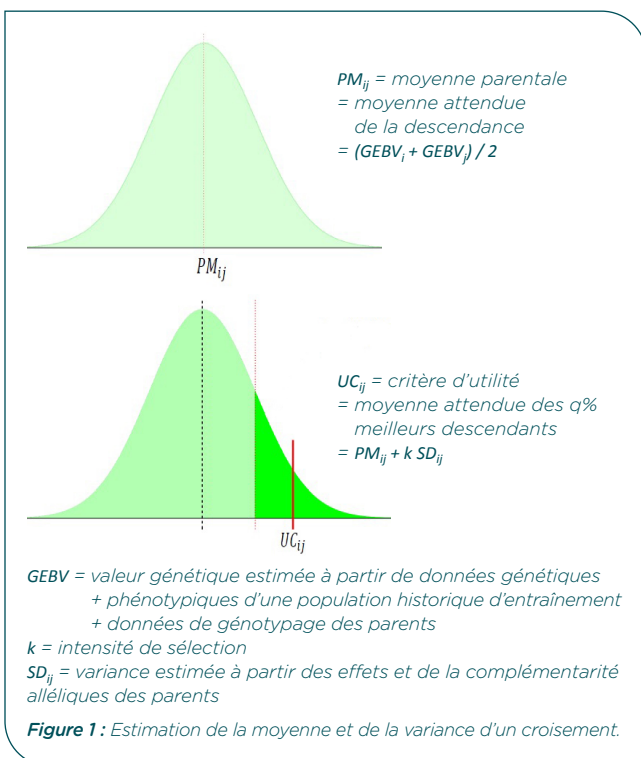
$$UC_1 = PM_1 + iSD_{G(1)} \quad (1)$$

Avec UC_1 la moyenne espérée des meilleurs descendants, PM_1 la moyenne espérée des parents (estimée comme la moyenne des GEBV), i l'intensité de sélection, $SD_{G(1)}$ l'écart type espéré de la descendance pour le caractère principal.

D'après Zhong et Jannink (2007) :

$$SD_G^2 = \sum_{i=1}^L \alpha_i^2 + 2 \sum_{i < j} \frac{1-2c_{ij}}{1+2c_{ij}} \alpha_i \alpha_j \quad (2)$$

Avec SD_G^2 la variance génétique du caractère, α_i et α_j les effets de substitution des allèles des loci i et j , c_{ij} le taux de recombinaison entre les loci i et j .



Pour le caractère secondaire, d'après Neyhart et. al (2019) :

$$SD_{G(1,2)}^2 = \sum_{i=1}^{L(1)} \sum_{j=1}^{L(2)} \frac{1-2c_{i(1)j(2)}}{1+2c_{i(1)j(2)}} \alpha_{i(1)} \alpha_{j(2)} \quad (3)$$

Avec $SD_{G(1,2)}^2$ la covariance génétique attendue entre deux caractères, $c_{i(1)j(2)}$ le taux de recombinaison entre le locus i du caractère 1 et le locus j du caractère 2, $\alpha_{i(1)}$ l'effet de substitution d'allèle du locus i pour le caractère 1, et $\alpha_{j(2)}$ l'effet de substitution d'allèle du locus j pour le caractère 2.

$$r_{G(1,2)} = \frac{SD_{G(1,2)}}{\sqrt{\sigma_{G(1)}^2 \sigma_{G(2)}^2}} \quad (4)$$

Avec $r_{G(1,2)}$ la corrélation génétique entre les caractères 1 et 2, $SD_{G(1)}^2$ et $SD_{G(2)}^2$ les variances génétiques respectives des caractères 1 et 2, $SD_{G(1,2)}$ la racine carrée de la covariance génétique attendue entre deux caractères.

$$UC_2 = PM_2 + iSD_{G(2)}$$

$$UC_{cor} = PM_2 + ir_{G(2)}SD_{G(2)} \quad (5)$$

Avec UC_2 la moyenne espérée des meilleurs descendants, PM_2 la moyenne espérée des parents, i l'intensité de sélection, $SD_{G(2)}$ l'écart type espéré de la descendance pour le caractère secondaire, $r_{G(1,2)}$ la corrélation génétique attendue entre les deux caractères.

La population d'entraînement contient les variétés françaises inscrites ces 20 dernières années ainsi que les lignées du programme de sélection INRAE-AO, soit 2146 lignées au total. Un modèle RR-BLUP et des modèles bayésiens ont été utilisés pour estimer les effets des marqueurs. Les taux de recombinaison c ont été estimés à partir d'un panel de diversité mondial (Danguy et al, 2021).

► Estimation au champ de la valeur d'un croisement

Au total, 101 croisements ont été produits. Les parents ont été génotypés avec une puce 35K.

Tous les partenaires ont évalué au champ les descendance à l'aide d'un dispositif augmenté (20% de répétitions). Deux témoins (RGT_CESARIO et CHEVIGNON) sont répétés dans tous les essais. Les parents des croisements sont également répétés dans les essais à partir de l'année 2. Au total, 9727 parcelles ont été évaluées sur 4 ans.

Les données expérimentales pour le rendement, la teneur en protéines, la hauteur et la précocité ont été ajustées après corrections spatiales. La qualité des données est très satisfaisante et les corrélations très bonnes entre essais.

RÉSULTATS

► Comparaison de la valeur des croisements observée et prédite

Pour le rendement (taux de protéine), les corrélations entre essais (année x lieu) sont de 0.48 (0.43) en moyenne, la répétabilité des mesures de 0.32 (0.52) et l'héritabilités génomiques de 0.53 (0.51) (Table 1).

Caractère	cor_année	rep_plot	h ² _genome
Rendement	0.48	0.32	0.53
Prot	0.43	0.52	0.51
Hauteur	0.79	0.73	0.56
Épiaison	0.88	0.85	0.73

Table 1 : Statistiques expérimentales.

Pour notre population d'entraînement, on remarque que la hauteur et la date d'épiaison sont mieux prédits avec des modèles de prédiction génomique bayésiens imposant une distribution des effets des marqueurs non Gaussienne, suggérant une architecture polygénique avec des QTL majeurs (Table 2).

Modèle	Rendement	Prot	Hauteur	Épiaison
BayesA	0.62	0.6	0.6	0.81
rrBLUP	0.62	0.52	0.6	0.65

Table 2 : Capacité de prédiction des GEBV en cross-validation pour différents modèles de prédiction génomique (60% training population / 40% Validation population)

Nous avons comparé les composantes de valeur de croisement (PM, SD et UC) prédites et observées pour tous les caractères. Pour UC, nous avons comparé la valeur moyenne des 7% et 10% meilleurs descendants (Figure 2).

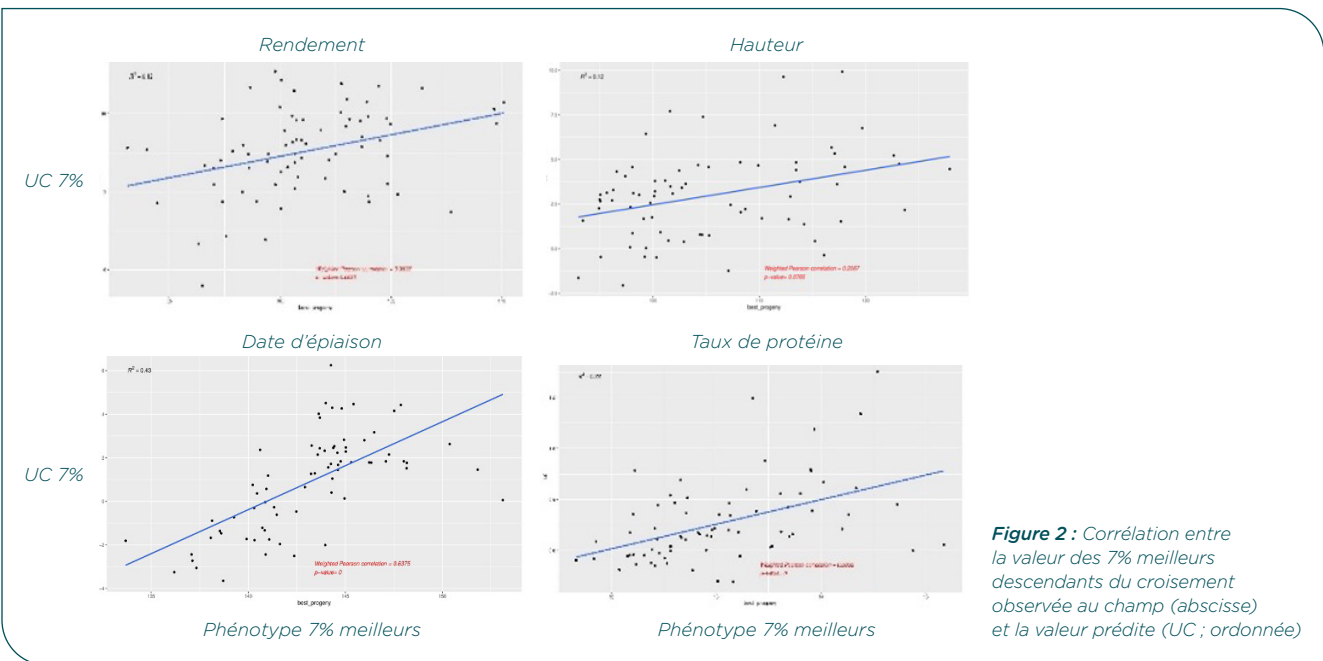
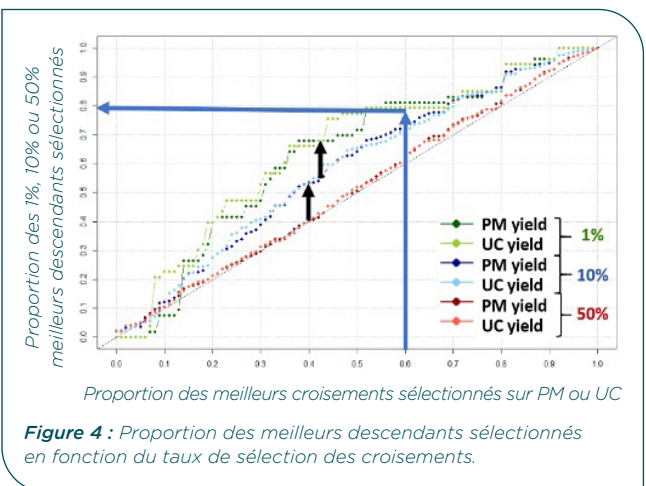
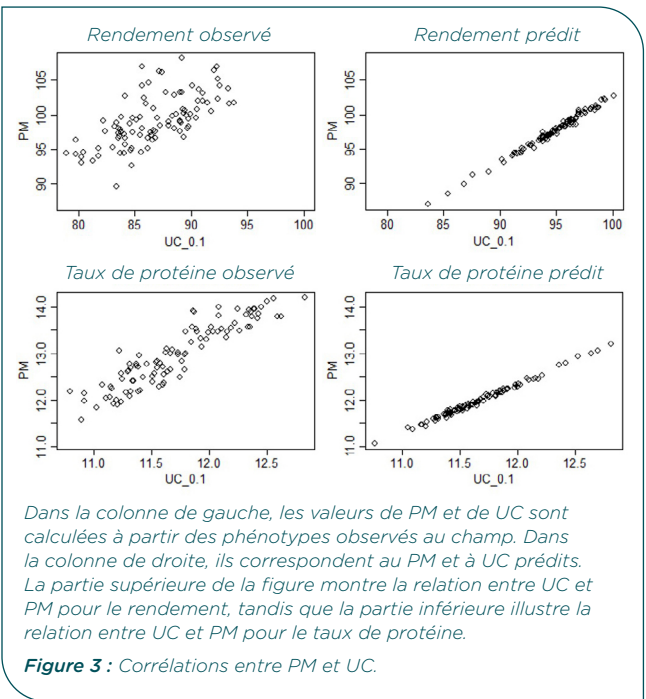


Figure 2 : Corrélation entre la valeur des 7% meilleurs descendants du croisement observée au champ (abscisse) et la valeur prédite (UC ; ordonnée)

Les moyennes de descendance observées et prédites (PM) sont significativement corrélées : 0.38, 0.63, 0.51 and 0.91, pour le rendement, le taux de protéine, la hauteur et la date d'épiaison respectivement. Les écart-types (SD) observés et prédits sont significativement corrélés pour la hauteur et la date d'épiaison (0.59 et 0.38, respectivement) mais pas pour le rendement et le taux de protéine (0.01 et 0.13, respectivement). Mais les valeurs prédites et observées de l'UC sont significativement corrélées (0.45, 0.52, 0.54 and 0.74 pour le rendement, le taux de protéine, la hauteur et la date d'épiaison, respectivement).

On remarque que pour la hauteur et la date d'épiaison, SD est mieux prédit avec des modèles de prédiction génomique bayésiens imposant une distribution des effets des marqueurs non Gaussienne (0.63-0.64 et 0.46-0.49, respectivement) qu'avec des modèles imposant une distribution Gaussienne (0.53-0.54 et 0.33-0.34, respectivement), suggérant une architecture polygénique avec des QTL majeurs.

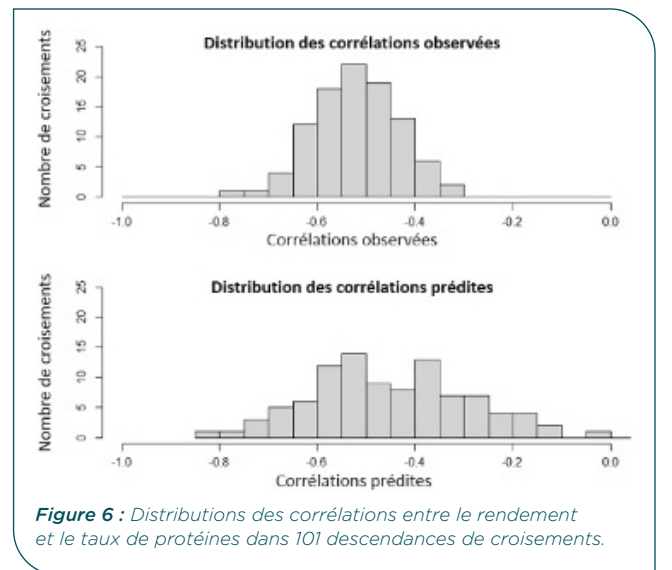
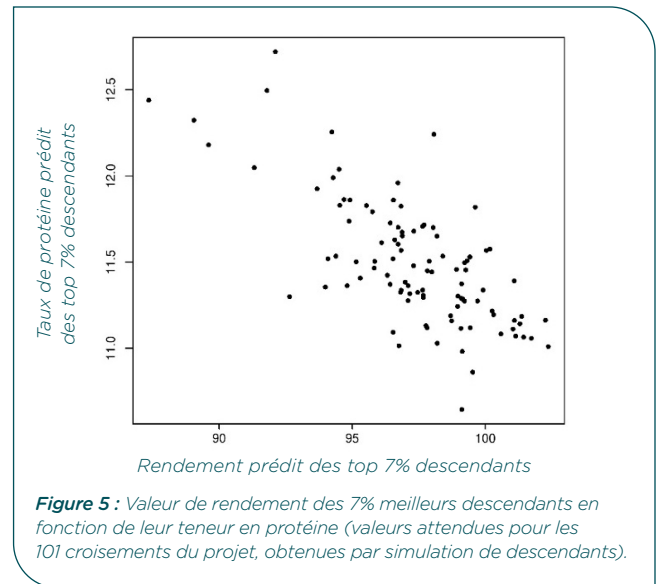
On remarque que dans ces croisements élites, la corrélation entre PM et UC est très forte, pour les paramètres observés au champ et pour les valeurs prédites en particulier (Figure 3).



En triant les croisements sur PM ou UC (Figure 4), on peut obtenir 80% des 1% meilleurs descendants en faisant seulement 60% des croisements. Par contre UC n'est pas plus efficace que PM.

► Comparaison de la corrélation rendement / taux de protéine observée et prédite

Les corrélations rendement/taux de protéine observées et prédites sont négatives (Figure 5). Nous observons cependant une grande variabilité des corrélations entre teneur en protéine et rendement suivant les croisements, d'après les prédictions et au champ (Figure 6).



Malgré une corrélation entre corrélations observées et prédites nulle, la valeur des meilleurs descendants pour le caractère secondaire (UCcor) est bien prédite (Table 3).

Cor (cor(obs), cor(pred))	rrBLUP	BayesA
Hauteur/Rdt	0.27	0.16
Taux de protéine/Rdt	0.39	0.38
Date d'épiaison/Rdt	0.61	0.65

Table 3 : Corrélation entre valeurs observées et prédites de la la valeur des meilleurs individus en rendement pour le caractère secondaire.

► Corrélation entre la proportion de descendants GPD+ des croisements observée et UCor

La proportion de descendants GPD+ (GPD>6) (valeur observée) est corrélée à la prédiction UCor (teneur en protéine des meilleurs individus en rendements) ($r=0.6$).

Si on ne s'intéresse qu'au 20% meilleurs croisements en rendement, le nombre de GPD+ par croisement est corrélé à la corrélation rendement/taux de protéine prédite.

Certains des parents de croisements sont des blés améliorants. La plupart sont des blés BPS ou BP. Les blés améliorants semblent être des variétés génétiquement très différentes, en particulier pour les allèles liés aux protéines (Figure 6).

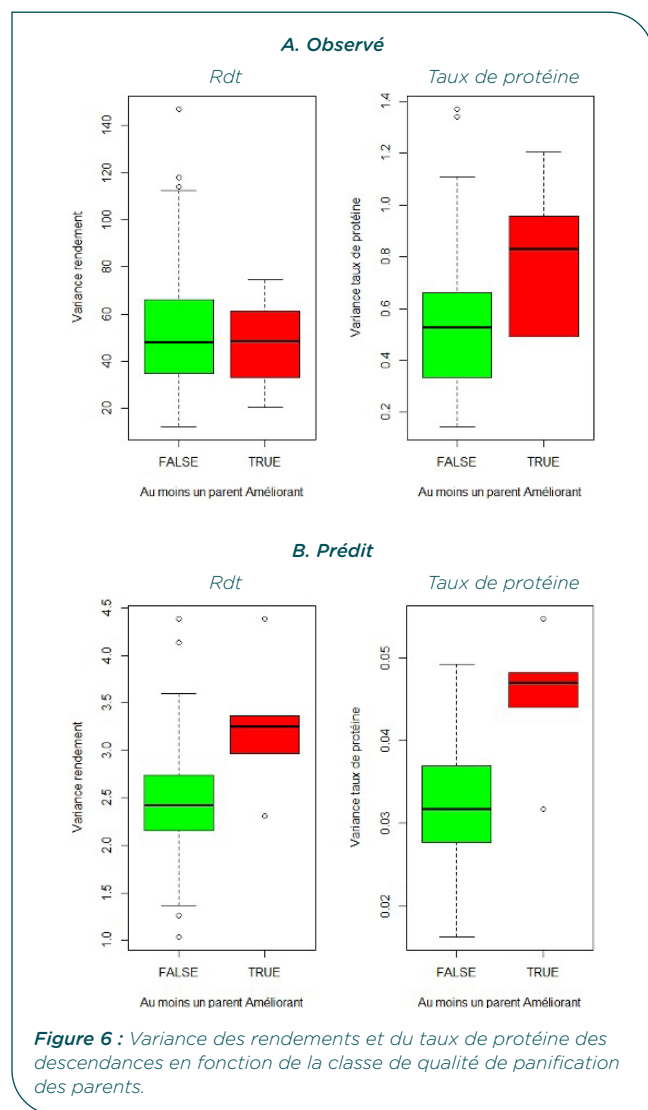


Figure 6 : Variance des rendements et du taux de protéine des descendance en fonction de la classe de qualité de panification des parents.

Par contre les descendants GPD+ avec un bon rendement viennent de croisements BPSxBPS ou BPSxBP en majorité, rarement avec un parent améliorant.

En conclusion, on prédit très bien la valeur des meilleurs croisements avec PM. Il y a un problème de shrinkage dans les prédictions des SD qui sont déjà faibles dans le matériel élite. SD contribue très peu à UC par rapport à PM. La faible capacité prédictive de SD provient aussi (1) de la taille limitée de la population

d'entraînement, (2) d'une diversité de croisements limitée, il ne s'agit que de « bons » croisements, (3) des tailles de descendance limitées, ne permettant pas de réaliser leur potentiel de variance, limitant la probabilité de produire des « outliers ».

► Capacité de prédiction attendue en fonction de la taille des descendance : simulations

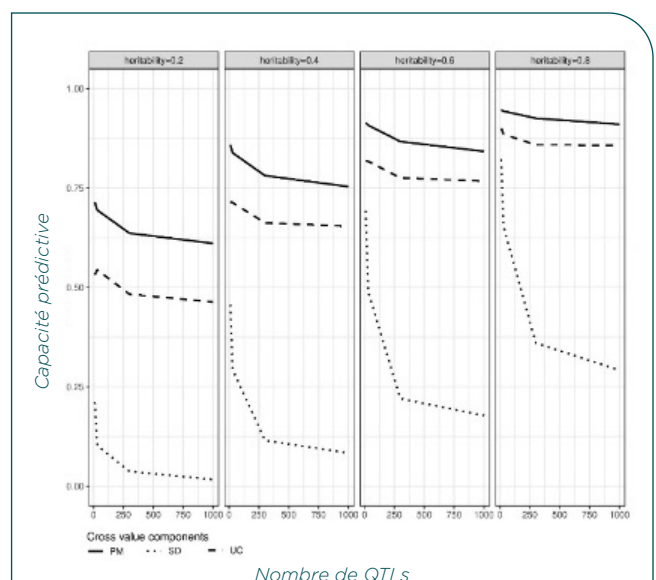
Nous avons estimé les composantes de la valeur d'un croisement (PM, SD, UC) en utilisant une formule analytique ou en simulant les descendance de 73 croisements expérimentaux.

Nous avons simulé différents scénarios, en faisant varier le nombre de QTL (de 10 à 1,000), l'héritabilité du caractère (de 0.2 à 0.8), et la taille de la descendance (de la taille expérimentale de 55 en moyenne à 2,000). Nous avons simulé 20 caractères pour chaque nombre de QTL. Les QTL ont été positionné aléatoirement le long de la carte génétique (Danguy des Déserts *et al.* 2021). Leurs effets suivent une distribution normale $N(0,1)$. Ces effets des marqueurs vrais ont été ajustés pour que la variance des TBV soit similaire à la variance des GEBV observée pour le rendement dans la TP.

Pour les effets des marqueurs, nous avons utilisé soit les effets des marqueurs vrais (TRUE scenario) où l'erreur d'estimation des effets des marqueurs serait négligeable (les tailles de TP et de la descendance sont infinies et le modèle de prédiction parfait), soit des effets de marqueurs estimés (ESTIMATED scenario) considérant que la taille de la TP est limitée.

Pour chaque scénario, nous avons comparé les valeurs des trois composantes dans les descendance simulées avec les valeurs estimées avec la formule analytique.

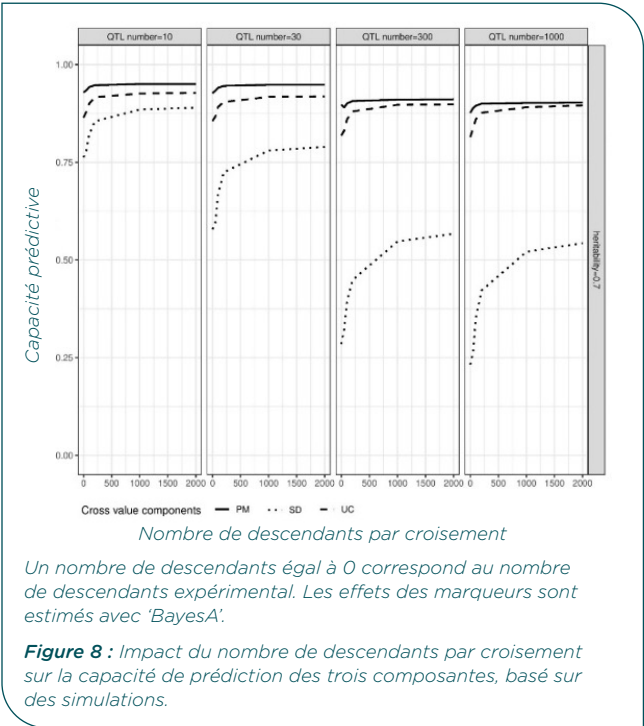
La capacité de prédiction augmente avec l'héritabilité et diminue avec le nombre de QTL, en particulier SD (Figure7).



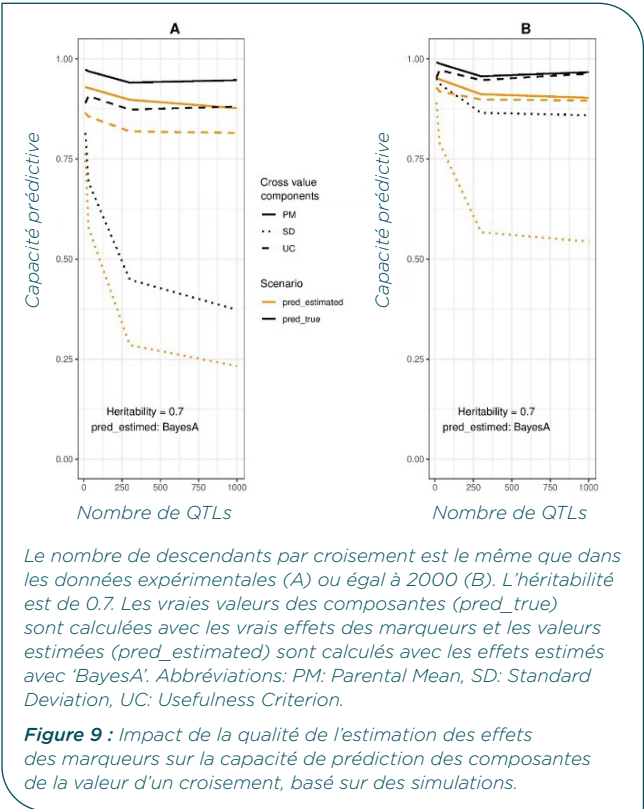
Le nombre de descendants par croisement est le même que dans les données expérimentales, les effets des marqueurs sont estimés avec un modèle de prédiction génomique 'BayesA'. PM: Parental Mean, SD: Standard Deviation, UC: Usefulness Criterion.

Figure 7 : Impact de l'architecture génétique sur la capacité de prédiction des trois composantes de la variance, basé sur des simulations.

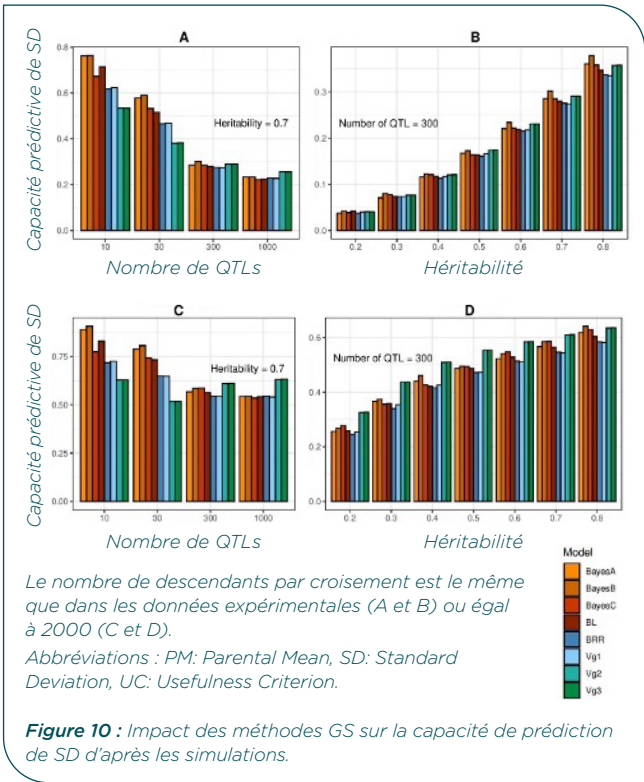
La capacité de prédiction augmente avec la taille de la descendance (Figure 8). On observe deux seuils de taille importants, 250 et 1000 qui correspondent à des sauts en termes de capacité de prédiction.



Nous avons comparé un scénario où les effets sont parfaitement estimés (« TRUE ») ou estimés à partir d'une base de données (« ESTIMATED »). Alors que PM et UC sont toujours correctement estimés quelque soit le scénario, SD n'est bien estimé (accuracy > 0.4) que lorsque les effets des marqueurs sont parfaitement estimés (Figure 9A) ou la taille de la population est grande (Figure 9B).



D'après les simulations, la capacité de prédiction augmente avec l'héritabilité, la taille de la descendance et diminue avec le nombre de QTL, en particulier pour SD. SD est très impacté par la qualité d'estimation des effets des marqueurs. Une formule qui prend en compte l'erreur d'estimation des effets des marqueurs (Vg3) permet d'améliorer les prédictions lorsque le nombre de QTL >= 300 et la taille de la population est grande (2000 ici) (Figure 10 C et D). Dans le cas contraire, ce sont les modèles bayésiens qui produisent les meilleures estimations.



Nous validons ainsi notre capacité de prédiction d'UC malgré une mauvaise corrélation de SD avec les données expérimentales. Cette capacité de prédiction dépend de l'architecture du caractère et de la taille de la descendance. Autrement dit, pour réaliser le potentiel de variance estimé analytiquement pour un croisement donné, il faut produire une taille de descendance suffisante. Un article scientifique détaille ces résultats (Oget-Ebrad et al. 2024).

DISCUSSION

La corrélation entre l'UC et la moyenne des meilleures lignées au champ est bonne (Table 4) même si la taille de nos populations expérimentales (55 descendants en moyenne) ne nous permet pas de valider notre capacité de prédiction de l'écart-type dans la descendance.

	Rendement	Prot	Hauteur	Épiaison
GEBV accuracy	0.62	0.6	0.6	0.81
cor(UC / phenotypes)	0.45	0.52	0.54	0.74

Table 4 : Statistiques de prédictions.

La capacité de prédiction de l'UC pour la date d'épiaison nous permet d'avoir confiance en nos modèles de prédiction génomique. La capacité de prédiction est moins bonne pour le rendement, sans doute à cause du GxE, mais reste intéressante d'un point de vue gain génétique (Danguy *et al.*, 2023).

La prédiction est moins bonne pour la hauteur de plante, sans doute parce que la variance pour ce caractère est très faible dans le matériel élite et la précision de mesure d'environ 5 cm. Nous pensons avoir de meilleures prédictions de la hauteur dans des panels plus variés. En effet, il a déjà été rapporté plusieurs fois que la capacité de prédiction de la variance de la descendance, et de la valeur d'un croisement, dépend du rapport $\text{var}(\text{sd}) / \text{var}(\text{PM})$ pour l'ensemble des descendances.

Estimer la corrélation dans les croisements à fort potentiel de rendement permet de choisir les meilleurs croisements pour rendement et protéine. Par contre UC n'apporte rien par rapport à PM. Il s'agit de matériel élite, les variances prédites sont très shrinkées, les croisements ne sont pas très variés et surtout les tailles de descendances sont très faibles.

Dans un programme de sélection, pour réaliser le potentiel d'un croisement, il faudrait essayer d'évaluer de grandes descendances pour les UC exceptionnels, plus tôt dans les générations, donc avec un typage très peu cher. Pour l'instant, seule la variance inter croisement est exploitée, pas la variance intra croisement dans les programmes de sélection INRAE-AO. Un seul descendant par croisement maximum est évalué pour le rendement à partir de la F6.

Pour la prédiction des corrélations, notre premier objectif est d'améliorer la compréhension du

déterminisme génétique conjoint du rendement et du taux protéine en utilisant un maximum de populations et d'environnements grâce à une analyse meta-QTL. Nous souhaitons mieux comprendre le nombre de QTLs pléiotropes et en DL, ceux qui sont en répulsion de phase en particulier dans le matériel élite. Nous essaierons d'imputer les données de génotypage 35K avec la 420K et utiliserons tous les marqueurs pour mieux estimer les variances.

Nous allons tester notre capacité de prédiction de SD et de la corrélation entre deux caractères dans un matériel plus varié par simulation. Nous testerons des scénarios avec un typage moins cher en F4 pour choisir un plus grand nombre de descendants à évaluer en F6 pour certains croisements à UC exceptionnel. Nous évaluerons le gain génétique pour les deux caractères sur le long terme.

Nous comparerons des scénarios de sélection pour le rendement et le taux de protéine en sélectionnant soit sur le phénotype, la GEBV, PM_1 et PM_2 , UC_1 et UC_2 , UC_1 et UCcor en partant de matériel élite ou plus variés, pour évaluer la possibilité de sélectionner pour le rendement avec moins de contrainte sur le taux de protéine grâce aux prédictions génomiques.

En effet, le matériel élite se distingue du matériel plus ancien par :

- une proportion d'allèles favorables pour le rendement (taux de protéine) supérieure (inférieure),
- un nombre d'allèles favorables pour le rendement en répulsion avec les allèles favorables à la teneur en protéine supérieur,
- une corrélation rendement/taux de protéine plus négative.

Références bibliographiques

Bijma P, Wientjes YCJ, Calus MPL. 2020. Breeding Top Genotypes and Accelerating Response to Recurrent Selection by Selecting Parents with Greater Gametic Variance. *Genetics*. 214(1):91. <https://doi.org/10.1534/genetics.119.302643>

Danguy des Déserts A et al. 2023. Comparison of genomic-enabled cross selection criteria for the improvement of inbred line breeding populations. *G3 Genes|Genomes|Genetics*. jkad195. <https://doi.org/10.1093/g3journal/jkad195>

Danguy des Déserts A, Bouchet S, Sourdille P, Servin B. 2021. Evolution of Recombination Landscapes in Diverging Populations of Bread Wheat. *Genome Biology and Evolution*. 13(8) <https://doi.org/10.1093/gbe/evab152>. <https://doi.org/10.1093/gbe/evab152>

Heslot N, Jannink J-L, Sorrells ME. 2013. Using Genomic Prediction to Characterize Environments and Optimize Prediction Accuracy in Applied Breeding Data. *Crop Science*. 53(3):921-933. <https://doi.org/10.2135/cropsci2012.07.0420>

Lehermeier C, Teyssèdre S, Schön C-C. 2017a. Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. *Genetics*. 207(4):1651-1661

Lehermeier C, Teyssèdre S, Schön C-C. 2017b. Genetic Gain Increases by Applying the Usefulness Criterion with Improved Variance Prediction in Selection of Crosses. *Genetics*. 207(4):1651-1661. <https://doi.org/10.1534/genetics.117.300403>

Neyhart JL, Lorenz AJ, Smith KP. 2019. Multi-trait Improvement by Predicting Genetic Correlations in Breeding Crosses. *G3 Genes|Genomes|Genetics*. 9(10):3153-3165. <https://doi.org/10.1534/g3.119.400406>

Oget-Ebrad C et al. 2024. Validation of cross-progeny variance genomic prediction using simulations and experimental data in winter elite bread wheat. *Theor Appl Genet*. 137(10):226. <https://doi.org/10.1007/s00122-024-04718-6>

Schnell FW, Utz HF. 1975. Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter.

Wellmann R. 2019. Optimum contribution selection for animal breeding and conservation: the R package optiSel. *BMC Bioinformatics*. 20(1):25. <https://doi.org/10.1186/s12859-018-2450-5>

Zhong S, Jannink J-L. 2007. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics*. 177(1):567-576