

Claire OGET-EBRAD^{1,8}, Emmanuel HEUMEZ², Laure DUCHALAI³, Ellen GOUEMAND-DUGUE⁴, Sandrine BERGES⁵, Patrice WALCZAK⁶, Paul BATAILLON⁷, François-Xavier OURY¹, Jean-Michel ELSSEN⁸, Sophie BOUCHET^{1*}

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

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

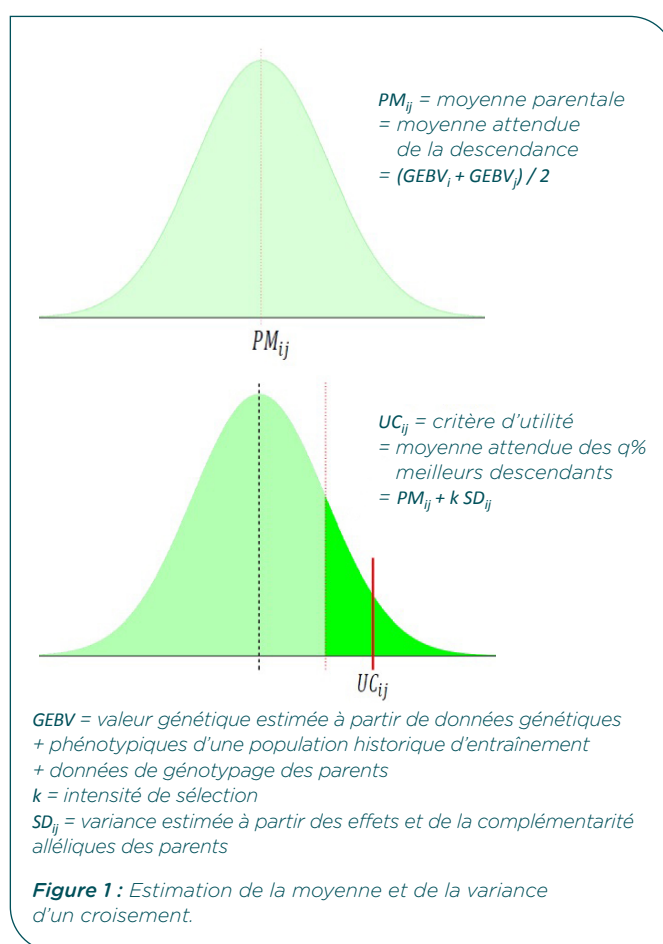
8 - INRAE-Université de Toulouse, UMR1388, GenPhySE, 24 chemin de Borde-Rouge, CS52627, 31326 Castanet-Tolosan, FRANCE

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

INTRODUCTION

L'amélioration variétale nécessite de combiner un maximum d'allèles favorables dans une même variété pour la tolérance aux stress biotiques et abiotiques mais aussi la qualité boulangère. Le choix des plans de croisement est une étape essentielle du programme de sélection. La prise en compte de caractères secondaires en plus du rendement est indispensable.

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 PrediCrop 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 (Fig. 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.



PRÉDICTION ANALYTIQUE DE LA VALEUR D'UN CROISEMENT ET DE LA CORRÉLATION DE DEUX CARACTÈRES

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 :

$$UC_1 = PM_1 + kSD_{G(1)}$$

(Schnell et Utz 1975)

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

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

(Zhong et Jannink, 2007)

Pour le caractère secondaire :

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

(Neyhart et al, 2019)

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

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

$$UC_{cor} = PM_2 + kr_{G(2)}SD_{G(2)}$$

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 .

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.

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.

Avec UC_2 la moyenne espérée des meilleurs descendants, PM_2 la moyenne espérée des parents, k 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 utilisée pour ce projet 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).

PARAMÈTRES QUI FONT VARIER LA CAPACITÉ DE PRÉDICTION : SIMULATIONS

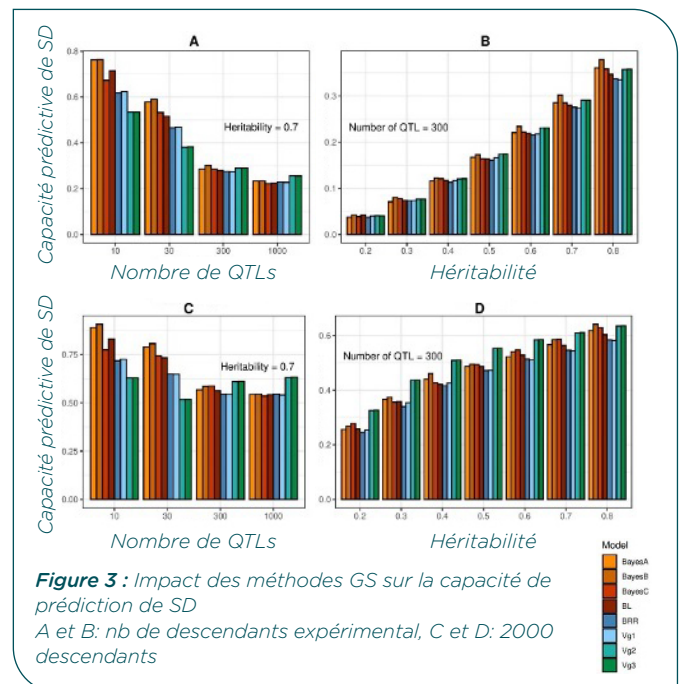
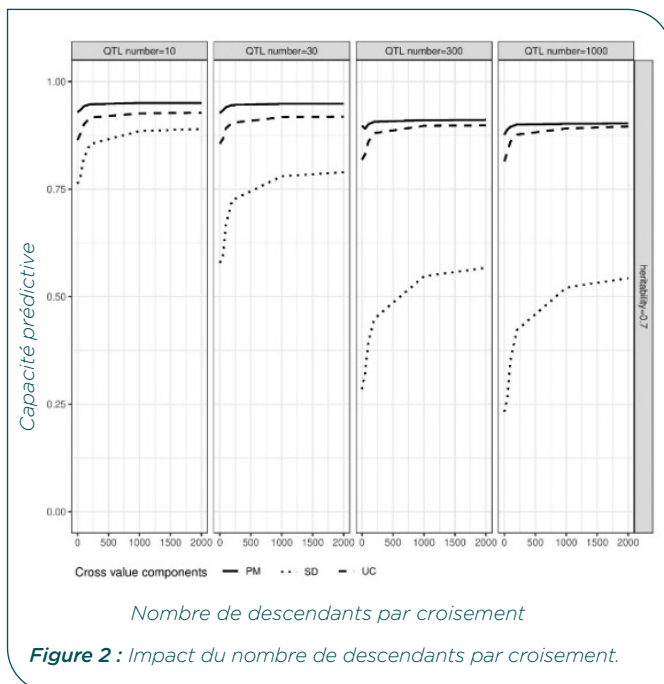
Nous avons estimé les composantes de la valeur d'un croisement (PM, SD, UC) en utilisant une formule analytique ou en simulant les descendances de 73 croisements expérimentaux.

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 positionnés aléatoirement le long de la carte génétique (Danguy des Déserts et al. 2021). Leurs effets suivent une distribution normale $N(0,1)$. Dans les simulations, nous avons utilisé soit les effets des marqueurs vrais (TRUE scenario), où l'erreur d'estimation des effets des marqueurs serait négligeable (tailles de populations d'entraînement et de descendances infinies, modèle de prédiction parfait), soit des effets de marqueurs estimés (ESTIMATED scenario).

Pour chaque scénario, nous avons comparé les valeurs des trois composantes dans les descendances 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. La capacité de prédiction augmente avec la taille de la descendance (Figure 2). On observe deux seuils de taille importants, 250 et 1000 qui correspondent à des sauts en termes de capacité de prédiction.

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 (Figure 3). Dans le cas contraire, ce sont les modèles bayésiens qui produisent les meilleures estimations.



PRÉDICTION EXPÉRIMENTALE DE LA VALEUR D'UN CROISEMENT ET DE LA CORRÉLATION DE DEUX CARACTÈRES

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.

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'héritabilité génomiques de 0.53 (0.51) (Table 1).

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).

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.

Modèle	Rendement	Prot	Hauteur	Épiaison
BayesA	0.62	0.6	0.6	0.81
rrBLUP	0.62	0.52	0.6	0.65
cor (UC,pheno)	0.45	0.52	0.54	0.74

Table 2 : Capacité de prédiction de la population d'entraînement.

CORRÉLATION DES COMPOSANTES OBSERVÉES ET PRÉDITES

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 4).

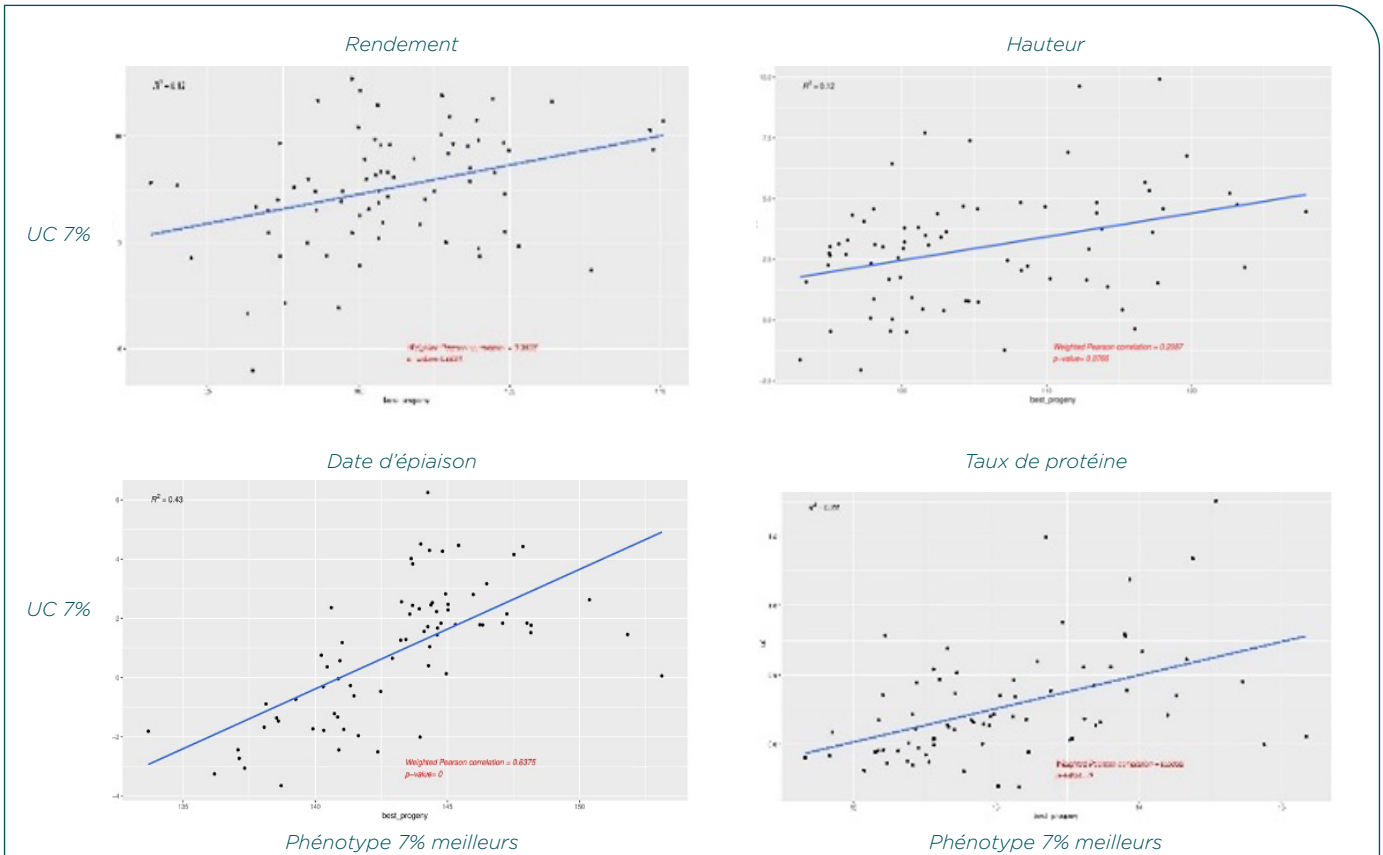
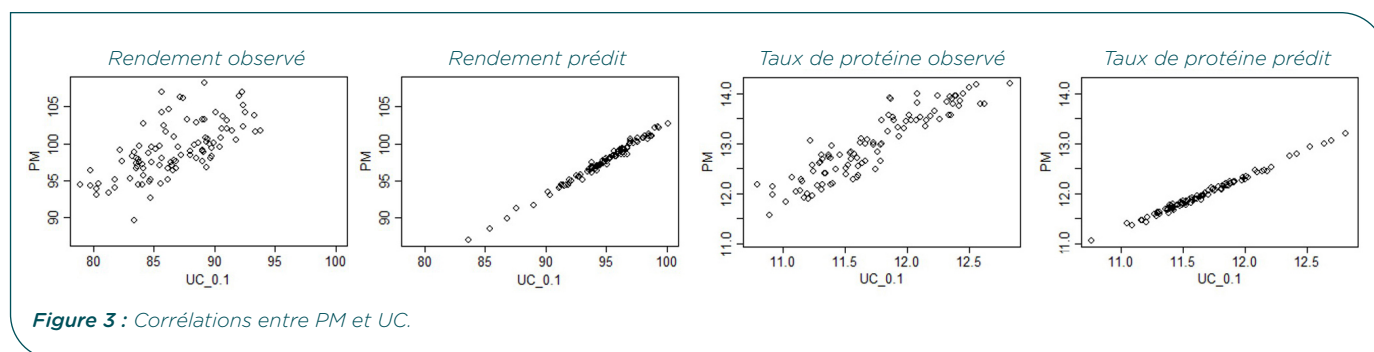


Figure 4 : Corrélations entre la valeur des 7% meilleurs descendants du croisement observée et prédite.

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 écarts-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 dans ces croisements élités, la corrélation entre PM et UC est très forte, pour les paramètres observés au champ et pour les valeurs prédites en particulier (Figure 5). En triant les croisements sur PM ou UC, on peut obtenir 80% des 1% meilleurs descendants en faisant seulement 60% des croisements. Par contre UC n'est pas plus efficace que PM.

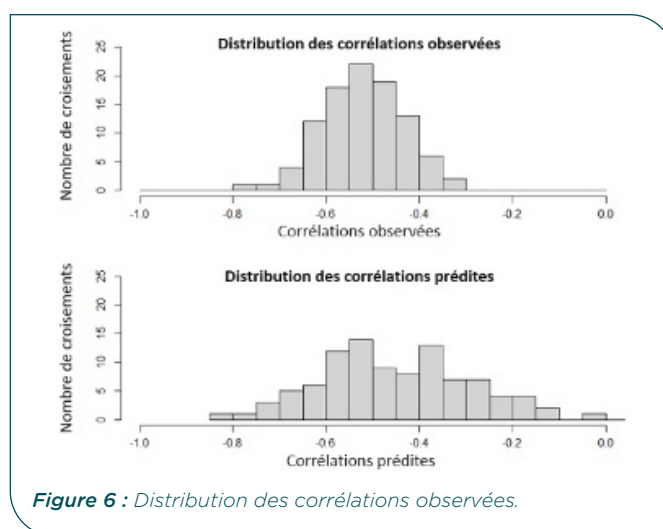


Les corrélations rendement/taux de protéine observées et prédites sont négatives. Nous observons cependant une grande variabilité des corrélations entre teneur en protéine et rendement suivant les croisements, d'après les prédictions et au champ (Figure 6).

La valeur prédite des meilleurs descendants pour le caractère secondaire (UCcor) est corrélée à la valeur observée au champ (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 valeur des meilleurs individus en rendement pour le caractère secondaire.



La proportion de descendants GPD+ (GPD>6) (avec un taux de protéine au champ supérieur à l'attendu à ce niveau de rendement) est corrélée à la prédiction UCcor (teneur en protéine des meilleurs individus en rendements) (r=0.6).

CONCLUSION

Nous prédisons 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 élit. 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 limitée, ne permettant pas de réaliser leur potentiel de variance, limitant la probabilité de produire des « outliers ». Vg3 permet d'améliorer les prédictions lorsque le nombre de QTL>=300 et la taille de la population est grande.

Pour la prédiction des corrélations, notre premier objectif est d'améliorer la compréhension du déterminisme génétique conjoint du rendement et du taux protéine, la proportion de QTL pléiotropes, en liaison, en phase ou répulsion, de manière à optimiser les plans de croisements.

Les résultats sont publiés dans TAG (Oget et al, 2024).



RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.fr



semae



FLORIMOND
DESPREZ