

[SG-Rendement] Calibration et Implémentation d'un outil de Sélection Génomique Rendement dans un programme de sélection blé tendre

Laure DUCHALAIS *, Simon TEYSSÉDRE ¹, Bruno CLAUSTRES ¹, Christophe MICHELET ¹, Matthieu BOGARD ²

1 - RAGT 2n SAS Rue Emile SINGLA, Site de Bourran – B.P. 3336 – 12033 RODEZ Cedex 9

2 - ARVALIS- Institut de Végétal, 6 chemin de la côte vieille, 31450 BAZIEGES

* *Coordinateur* : Laure DUCHALAIS, lduchalais@ragt.fr

1. Introduction

La situation des producteurs français n'a jamais été aussi difficile. Le prix de la matière première est très variable et le prix des intrants ne cesse d'augmenter, les revenus agricoles sont donc largement impactés.

Le monde agricole est en constante évolution grâce aux nouvelles pratiques et aux nouveaux outils de gestion culturale. Cependant, même si ces évolutions tendent à minimiser les dépenses des agriculteurs, elles ne leurs garantissent pas un revenu approprié. Dans un tel environnement économique, il est donc indispensable de garantir aux agriculteurs des rendements élevés et stables. Malheureusement, une stagnation des rendements de blé tendre a été constatée, depuis 1996. Les gains génétiques et autres améliorations technologiques ne semblent pas suffire à compenser les aléas climatiques observés au cours des deux dernières décennies. La sélection de variétés de blé adaptées aux contraintes futures (climatiques, économiques...) et à l'environnement ciblé est devenue un facteur essentiel pour la durabilité de nos secteurs agricoles.

La disponibilité de milliers de marqueurs moléculaires à l'échelle du génome a rendu l'utilisation de la sélection génomique (SG) possible chez les plantes et les animaux (Meuwissen *et al.*, 2001 ; Bernado et Yu, 2007; de Los Campos *et al.*, 2009, 2010; Crossa *et al.*, 2010, 2011; Pérez *et al.*, 2010). Afin de pouvoir utiliser la SG, des données de phénotypage et de génotypage doivent être collectées sur une population dite d' « entraînement ». Ces données sont ensuite utilisées pour estimer les effets des allèles, et créer ainsi un modèle de prédiction. Ce modèle de prédiction peut ensuite être appliqué aux génotypes des candidats de la population cible afin de calculer leur « Genomic Estimated Breeding Values » (GEBVs). La sélection s'effectue alors sur la base de ces GEBVs.

Les sélectionneurs font face à différents types de problèmes de prédiction. Un premier problème consiste à prédire les valeurs génétiques des lignées nouvellement développées en début de cycle. Ces lignées n'ont généralement que très peu voire aucune donnée de phénotypage. La précision de la prédiction dépend alors beaucoup du degré d'apparentement entre les candidats de sélection et les individus utilisés pour former l'équation de prédiction. Un deuxième problème de prédiction se rapporte à la prévision des performances des lignées avancées qui ont été évaluées dans un ou plusieurs environnements. Ici, la tâche principale est de prédire les performances dans les environnements où la lignée n'a pas encore été évaluée.

Une piste d'amélioration des modèles de SG pour des caractères complexes comme le rendement en grains qui sont sous l'influence de nombreux facteurs génétiques, environnementaux et de leurs interactions, est la prise en compte d'effets génétiques non additifs. En effet, la plupart des études sur la SG font implicitement l'hypothèse que les caractères étudiés sont sous l'influence de QTLs dont les effets s'additionnent. Or, il est raisonnable de penser

que des interactions entre QTLs existent et ont un impact sur ces caractères.

Dans la perspective d'étudier la stabilité des performances des génotypes, il apparaît nécessaire d'introduire dans les modèles de SG des composantes pour prédire les effets environnementaux et les interactions GxE afin d'être en mesure de prédire les performances génotypiques dans des environnements non testés. Des études sur ce sujet ont été réalisées (Jarquin *et al.*, 2014) et ont permis de proposer de tels modèles en utilisant des co-variables calculées sur la base de données météorologiques et agronomiques.

Ce projet vise à démontrer l'efficacité d'utilisation d'un outil de SG rendement dans un programme de sélection de blé tendre et à améliorer les modèles de SG par la prise en compte des effets environnementaux et des interactions GxE.

2. Matériel et méthode

2.1 - Application d'un outil de SG rendement dans un programme de sélection de blé tendre

► Matériel végétal

Population d'entraînement R2n

La population d'entraînement R2n se compose de 264 lignées : 11 variétés commerciales et 253 lignées avancées dans le programme de sélection R2n.

Population cible n°1

La population cible n°1 rassemble 500 lignées au stade Fn au cours de la saison 2014/2015. Ces lignées ont été choisies afin que chacun des croisements réalisés par R2n soit représenté. Pour les croisements aux effectifs les plus faibles, seule une lignée a été choisie. Pour les croisements aux effectifs plus importants, plusieurs lignées ont été choisies aléatoirement en fonction de leur effectif total. Enfin, la totalité des lignées des deux croisements majoritaires (41 lignées pour l'un et 46 lignées pour l'autre) a été sélectionnée afin de pouvoir étudier les prédictions de la SG intra-croisement.

Population cible n°2

La population cible n°2 se compose de 482 lignées au stade Fn+1 au cours de la saison 2015/2016. De précocité intermédiaire à tardive, elles représentent une grande majorité de la diversité disponible à ce stade dans le programme de sélection blé tendre R2n.

► Génotypage

Les extractions d'ADN des lignées appartenant aux différentes populations ont été réalisées au laboratoire de biologie moléculaire chez RAGT Seeds. L'ensemble des échantillons a ensuite été envoyé à l'université de Bristol où ils ont été génotypés avec la puce Axiom 35 K. Les individus et marqueurs avec plus de 20% de données manquantes ont été supprimés. Les marqueurs ayant une fréquence de l'allèle minoritaire (MAF) < 1% ont également été supprimés.

► Phénotypage

Population d'entraînement R2n

Les 264 lignées constituant la population d'entraînement R2n ont été au moins phénotypées pour le rendement sur 5 lieux en 2 répétitions traitées en 2014 et en 2015. Les conduites culturales ont été menées à l'optimum afin de bien évaluer le potentiel de productivité des différentes lignées.

Population cible n°1

Les lignées de la population cible n°1 ont été évaluées pour le rendement traité sur une seule micro-parcelle à Louville-la-Chenard (28) en 2015. Elles ont été implantées sous forme d'essais à 30 entrées dont 4 témoins répétés.

Population cible n°2

Les lignées de la population cible n°2 ont été évaluées pour le rendement traité sur un site sans répétition en 2015 et sur 5 sites différents en 2 répétitions en 2016. Elles ont été implantées sous forme d'essais à 30 entrées dont 4 témoins répétés.

Au sein de cette population, 30 individus ont pu être évalués pour le rendement traité en 2017 sur 13 lieux communs français en 2 répétitions.

► Ajustement des données

Dans chaque essai, les effets terrains ont été ajustés selon cette formule :

$$Y_i = \text{mean} + \text{fixed}(\text{variety}) + \text{fixed}(\text{Rep}) + e$$

Pour les modèles de SG sans prise en compte des effets GxE, une moyenne ajustée des variétés sur l'ensemble des lieux a été calculée à partir de tous les \hat{Y}_i précédents et du modèle :

$$\hat{Y} = \text{mean} + \text{fixed}(\text{Loc}) + \text{fixed}(\text{variety}) + \text{random}(\text{variety} \times \text{loc}) + e$$

Avec :

$$\text{Variety} \times \text{loc} \sim N(0, \sigma_{ge}^2)$$

► Test de différents modèles de SG sur la population cible n°1 et application du meilleur modèle sur cette dernière

Modèles testés

Différents modèles de sélection génomique ont été testés sur la population cible n°1 : un modèle G-BLUP prenant en compte uniquement les effets d'additivité, un modèle G-BLUP prenant en compte à la fois les effets d'additivité et d'épistasie, un modèle RR-BLUP utilisant GS3 (Legarra *et al.*, 2013), un modèle RKHS utilisant BGLR (50 000 itérations et 20 000 burnin période) et un modèle bayésien utilisant BayesR avec 4 classes (Rvarg*c (0.0001, 0.001, 0.01) ; 50 000 itérations avec 20 000 burnin périodes) ont été appliqués.

Validation croisée

Afin de valider l'efficacité des différents modèles testés, une validation croisée de type k-fold a été réalisée avec k=10% des données. Cette validation croisée a permis d'identifier le meilleur modèle de SG qui a ensuite été appliqué pour réaliser les prédictions.

► Calculs des GEBV₂₀₁₄ et des GEBV₂₀₁₄₊₂₀₁₅ à l'aide du meilleur modèle défini sur la population cible n°1

Une première GEBV dite GEBV₂₀₁₄ a été calculée à partir de toutes les données de phénotypage disponibles sur la population d'entraînement R2N avant la récolte 2014 (récolte 2014 incluse) puis une seconde GEBV dite GEBV₂₀₁₄₊₂₀₁₅ a été calculée à partir de toutes les données de phénotypage disponibles avant la récolte 2015 (récolte 2015 incluse) sur la population cible n°1.

► Application d'une sélection phénotypique classique sur la population cible n°1

Les sélectionneurs de R2n ont appliqué une sélection classique sur la première population cible. Les lignées trop sensibles, pas assez productives ou n'ayant pas été appréciées par les sélectionneurs, ont été éliminées du programme de sélection. A l'inverse, les lignées retenant leur attention ont été conservées et à nouveau phénotypées sur 5 lieux en 2 répétitions en 2016 pour le rendement traité.

► Application du meilleur modèle de SG de la population cible n°1 sur la population cible n°2

Le meilleur des 5 modèles testés sur la population cible n°1 a été retenu pour être appliqué sur la population cible n°2.

► Calculs des GEBV₂₀₁₄ et des GEBV₂₀₁₄₊₂₀₁₅ à l'aide du meilleur modèle sur la population cible n°2

Une première GEBV dite GEBV₂₀₁₄ a été calculée à partir de toutes les données de phénotypage disponibles sur la population d'entraînement R2N avant la récolte 2014 (récolte 2014 incluse) puis une seconde GEBV dite GEBV₂₀₁₄₊₂₀₁₅ a été calculée à partir de toutes les données de phénotypage disponibles avant la récolte 2015 (récolte 2015 incluse).

► Validation de l'utilisation d'un modèle de SG rendement dans un programme de sélection blé tendre

Etude des corrélations entre GEBVs et rendements traités

Les coefficients de détermination ont été calculés entre GEBVs et moyenne des rendements traités 2015 pour les populations cibles n°1 et n°2, entre GEBVs et moyenne des rendements traités 2016 pour la population cible n°2, entre GEBVs et moyenne des rendements traités 2017 pour les 30 individus de la population cible n°2 ayant été poursuivis l'année suivante.

Comparaison de l'efficacité de la sélection génomique et de la sélection phénotypique ou des 2 combinées

Pour la population cible n°1, les pools de lignées suivants ont été créés afin de pouvoir valider ou invalider l'utilisation d'un outil de SG rendement dans un programme de sélection :

- Un pool « Bottom 50 GS- » correspondant aux lignées possédant une valeur de rendement 2015 intermédiaire et une valeur de GEBV₂₀₁₄ se situant dans les 50 GEBVs les plus mauvaises.
- Un pool « Bottom 50 PS- » correspondant aux lignées possédant une valeur de GEBV₂₀₁₄ intermédiaire et se situant dans les 50 plus mauvaises lignées pour le rendement traité en 2015 qui ont été sélectionnées pour passer dans l'année supérieure.
- Un pool « Top 50 GS+ » correspondant aux lignées possédant une valeur de rendement 2015 intermédiaire et une valeur de GEBV₂₀₁₄ se situant dans les 50 GEBVs les plus élevées.
- Un pool « Top 50 PS+ » correspondant aux lignées possédant une valeur de GEBV₂₀₁₄ intermédiaire et se situant dans les 50 meilleures lignées pour le rendement traité en 2015 qui ont été sélectionnées pour passer dans l'année supérieure.
- Un pool « Top 50 PS+ and Top 50 GS+ » correspondant aux lignées se situant à la fois dans les 50 meilleures lignées en terme de GEBV₂₀₁₄ et dans les 50 meilleures lignées en terme de rendement traité 2015.
- Un pool « Bottom 50 PS- and 50 GS- » correspondant aux lignées se situant à la fois dans les 50 plus mauvaises lignées en terme de GEBV₂₀₁₄ et dans les 50 moins bonnes lignées en terme de rendement traité 2015 ayant été sélectionnées pour passer dans l'année supérieure.
- Un pool « Bottom 50 PS- and 50 GS+ » correspondant aux

lignées se situant à la fois dans les 50 meilleures lignées en terme de $GEBV_{2014}$ et dans les 50 moins bonnes lignées en terme de rendement traité 2015 ayant été sélectionnées pour passer dans l'année supérieure.

- Un pool « Top 50 PS+ and Bottom 50 GS- » correspondant aux lignées se situant à la fois dans les 50 moins bonnes lignées en terme de $GEBV_{2014}$ et dans les 50 meilleures lignées en terme de rendement traité 2015.

Pour la population cible n°2, les pools de lignées suivants ont été créés afin de pouvoir valider ou invalider l'utilisation d'un outil de SG rendement dans un programme de sélection :

- Un pool « GS- » correspondant aux lignées possédant une valeur de rendement 2015 intermédiaire et une valeur de $GEBV_{2014+2015}$ se situant dans les 100 GEBVs les plus mauvaises.
- Un pool « PS- » correspondant aux lignées possédant une valeur de $GEBV_{2014+2015}$ intermédiaire et se situant dans les 100 plus mauvaises lignées pour le rendement traité en 2015 qui ont été sélectionnées pour passer dans l'année supérieure.
- Un pool « GS+ » correspond aux lignées possédant une valeur de rendement 2015 intermédiaire et une valeur de $GEBV_{2014+2015}$ se situant dans les 100 GEBVs les plus élevées.
- Un pool « PS+ » correspondant aux lignées possédant une valeur de $GEBV_{2014+2015}$ intermédiaire et se situant dans les 100 meilleures lignées pour le rendement traité en 2015 qui ont été sélectionnées pour passer dans l'année supérieure.
- Un pool « PS+ et GS+ » correspondant aux lignées se situant à la fois dans les 100 meilleures lignées en terme de $GEBV_{2014+2015}$ et dans les 100 meilleures lignées en terme de rendement traité 2015.
- Un pool « PS- et GS- » correspondant aux lignées se situant à la fois dans les 100 plus mauvaises lignées en terme de $GEBV_{2014+2015}$ et dans les 100 moins bonnes lignées en terme de rendement traité 2015 ayant été sélectionnées pour passer dans l'année supérieure.
- Un pool « PS- et GS+ » correspondant aux lignées se situant à la fois dans les 100 meilleures lignées en terme de $GEBV_{2014+2015}$ et dans les 100 moins bonnes lignées en terme de rendement traité 2015 ayant été sélectionnées pour passer dans l'année supérieure.
- Un pool « PS+ et GS- » correspondant aux lignées se situant à la fois dans les 100 moins bonnes lignées en terme de $GEBV_{2014+2015}$ et dans les 100 meilleures lignées en terme de rendement traité 2015.
- Un pool « Empty » correspondant aux lignées présentant un comportement intermédiaire à la fois pour la $GEBV_{2014+2015}$ et pour le rendement traité 2015, pool par rapport auquel il sera intéressant de comparer tous les autres pools pour conclure sur l'efficacité de l'une ou l'autre des 2 méthodes de sélection.

2.2 - Prise en compte des effets environnementaux et des interactions GxE dans les modèles de SG

► Caractérisation des sites expérimentaux et calcul de co-variables environnementales

Sur chacun des sites où les populations cibles et d'entraînement ont été implantées, des analyses de sol ont été réalisées afin de connaître la composition du sol (%limon, %argile, % sable, % de Matière Organique, taux de cailloux).

Des indicateurs pédoclimatiques ont été calculés pour les différentes

lignées utilisées dans ce projet à l'aide du modèle écophysologique développé par ARVALIS Institut du végétal. Dans un premier temps, un rattachement des lignées à une variété paramétrée pour le modèle a été effectué. Pour cela, la date d'épiaison moyenne de chaque lignée a été calculée puis chaque lignée a été rattachée au témoin présentant la date d'épiaison la plus proche.

Enfin, des calculs d'indicateurs (somme de températures moyennes, nombre de jours de gel, avec des températures échaudantes ou un déficit hydrique...) ont été réalisés pour différentes périodes de développement. Ces indicateurs ont été utilisés comme co-variables dans les modèles de GS incorporant un terme d'interaction GxE (Jarquin *et al.* 2014).

► Introduction d'interaction GxE dans nos modèles de sélection génomique

Le modèle G-BLUP, classiquement utilisé en sélection génomique (MA), sert à évaluer la valeur génétique additive d'une lignée. Lorsque l'on souhaite estimer la valeur génétique totale d'une lignée dans un environnement particulier, la valeur génétique additive n'est alors pas suffisante car d'autres termes interviennent : les déviations dues aux effets d'épistasie ; les interactions avec l'environnement (GxE). Si l'extension du modèle classique (MA) à la prise en compte de l'épistasie (MAI) est directement possible via l'utilisation des marqueurs, l'extension à la prise en compte des interactions GxE (MAIE) nécessite la caractérisation de nos lieux d'essais par des co-variables environnementales.

En utilisant les co-variables décrites précédemment, nous nous sommes intéressés à la capacité de prédiction des modèles MA, MAI et MAIE pour prédire : a) la valeur génétique moyenne totale des lignées ; b) la valeur génétique totale des lignées dans un environnement spécifique. Une validation croisée sur la population d'entraînement R2n avec 50 répétitions a été réalisée pour évaluer la capacité de prédiction des différents modèles, en supprimant aléatoirement 15% des lignées à chaque répétition et en prédisant leur valeur génétique pour le a) ou leur valeur génétique dans un environnement spécifique pour le b).

2.3 - Calibration de modèles de SG spécifiques à différents scénarios environnementaux

► Données de phénotypage et de génotypage

Les données de phénotypage et de génotypage (TaBW420K ; Rimbart *et al.* 2018) pour un panel de 426 variétés de blé tendre testées dans 46 environnements (lieu x année x conduite) en France ont été utilisées pour évaluer l'intérêt de calibrer des modèles de SG spécifiques à une population d'environnements donnée. Ces essais visaient à évaluer la réponse des variétés à des conditions climatiques et agronomiques variées (fertilisation azotée optimale versus limitée, condition irriguée versus pluviale). Les données de génotypage ont été analysées afin d'éliminer les marqueurs génétiques présentant un taux de données manquantes > 20%, un taux d'hétérozygotie > 10% et une fréquence de l'allèle minoritaire < 5%. La matrice de génotypage ainsi obtenue comprenait 170 000 marqueurs SNP.

► Caractérisation et regroupement des essais

Le modèle de culture CHN développé par Arvalis a été utilisé afin de calculer des indices de stress azoté et hydrique dans chaque essai en tenant compte des données climatiques, de l'itinéraire technique (date et densité de semis, apport d'azote, irrigation) et des caractéristiques du sol (texture, volume exploitable par les racines). L'indice de stress azoté est en lien direct avec l'INN tandis que l'indice de stress hydrique dépend de la fraction d'eau transpirable du sol. Ces deux indices

journaliers varient entre zéro (stress maximal) et un (absence de stress) et réduisent la production de biomasse et l'augmentation de la surface foliaire.

L'ensemble des simulations ont été réalisées pour la variété Soissons présente dans l'ensemble des essais. Les indices de stress azoté et hydrique journaliers ont été centrés sur la date d'épiaison et des valeurs moyennes ont été calculées sur des fenêtres de 100°C jours de -1000 à +600°C jours autour de la date d'épiaison. La dynamique de ces indices a ensuite été analysée et les essais ont été regroupés en k groupes correspondant à différents scénarios environnementaux à l'aide d'un algorithme des k -médianes selon la méthode proposée par Chenu *et al.* (2011).

► Modèle de prédiction

Une matrice d'apparentement a été calculée à partir de la matrice de génotypage avec le package rrBLUP (Endelman, 2011) dans R (R Core Team, 2016). Le modèle de prédiction utilisé correspondait au modèle 5 de Jarquin *et al.* (2014) et peut être formulé ainsi :

$$Y_{ij} = \mu + w_{ij} + g_i + gw_{ij} + \varepsilon_{ij}$$

Avec Y_{ij} le rendement en grains du génotype i dans l'essai j , w_{ij} l'effet des co-variables environnementales pour le génotype i dans l'essai j , g_i l'effet du génotype i , et gw_{ij} l'interaction entre le génotype i et les co-variables environnementales dans l'essai j et ε_{ij} l'erreur résiduelle. Dans ce modèle, les termes w , g , gw , ε suivent une loi normale et sont indépendants. En revanche, seuls les résidus sont distribués de manière aléatoire et les autres termes présentent une structure dérivée de matrices de similarités. Ces dernières permettent de prédire les performances des variétés dans des environnements non testés.

► Evaluation des capacités prédictives par validation croisée

Une procédure de validation croisée de type "leave-one-out" a été utilisée afin d'estimer les capacités prédictives du modèle calibré sur différents jeux de données pour prédire les performances des variétés dans des environnements non-testés. Trois jeux de données ont été considérés :

- Le set 1 correspond à l'ensemble des essais excepté l'essai j appartenant au groupe k
- Le set 2 correspond à l'ensemble des essais excepté les essais appartenant au groupe k
- Le set 3 correspond à l'ensemble des essais appartenant au groupe k excepté l'essai j

Dans chaque cas, les rendements en grains prédits pour les variétés présentes dans l'essai j ont été calculés puis la corrélation entre rendements prédits et observés a été estimée. Cette procédure a été répétée pour les essais de 1 à j . Enfin, les coefficients de corrélations obtenus à l'aide du modèle calibré sur ces trois jeux de données ont été comparés pour chaque groupe d'essais.

Dans ce cadre, le set 1 correspond à la situation où aucune caractérisation environnementale et aucun regroupement d'essai n'a été effectué et où l'ensemble des données disponibles sont utilisées pour calibrer le modèle de SG. Le set 2 correspond à la situation où la base de données de calibration ne contient pas d'essai correspondant au scénario environnemental k (par exemple, absence d'essai en condition de stress hydrique ou de fertilisation azotée limitante). Le set 3 quant à lui représente la situation où les performances des variétés dans l'essai k sont prédites à l'aide d'un modèle calibré spécifiquement pour le scénario environnemental k . Les calibrations ont été réalisées à l'aide du package BGLR (Perez et De los Campos, 2014) dans R.

3. Résultats et Discussion

3.1 - Application d'un outil de SG rendement dans un programme de sélection de blé tendre

► Génotypage

Après suppression des individus avec plus de 20% de données manquantes : 246 individus de la population d'entraînement R2N ont été génotypés à l'aide de 19 740 SNPs polymorphes de la puce 35 K, 481 individus de la population cible n°1 ont été génotypés à l'aide de 19 740 SNPs polymorphes de la puce 35 K et 479 individus de la population cible n°2 ont été génotypés à l'aide de 18 705 SNPs polymorphes de la puce 35 K.

Après suppression des marqueurs ayant une MAF trop faible ou trop de données manquantes, 13 793 SNPs ont été considérés pour les analyses de SG.

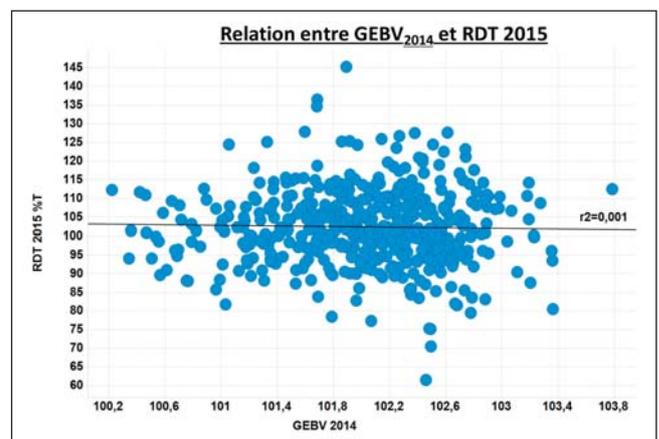
► Test de différents modèles de sélection génomique sur la population cible n°1

Pour les 5 modèles testés, les coefficients de détermination entre GEBVs et moyennes ajustées par validation croisée se sont avérés relativement faibles ($0.14 < r^2 < 0.19$). C'est néanmoins le modèle avec le r^2 maximal qui a été retenu pour la suite des analyses, c'est-à-dire un modèle G-BLUP prenant en compte à la fois les effets d'épistasie et les effets d'additivité.

► Validation de l'utilité d'utiliser un modèle de sélection génomique rendement dans un programme de sélection blé tendre

Etude des corrélations entre GEBVs et rendements traités Sur la population cible n°1

Le graphe ci-dessous montre la relation entre GEBV₂₀₁₄ (en abscisse) et rendements obtenus en 2015 (en ordonnée) pour l'ensemble des individus de la population cible n°1.



Le r^2 entre GEBV₂₀₁₄ et le rendement mesuré en 2015 est alors nul ($r^2 = 0.001$).

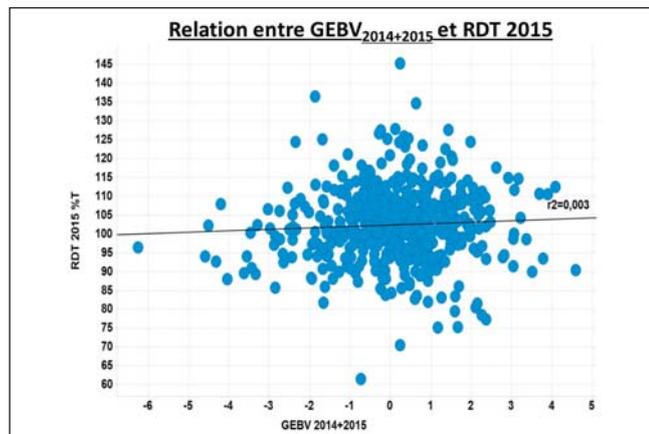
Les prédictions par SG intra-croisement ont ensuite été étudiées. Pour le premier croisement, composé de 41 lignées, le r^2 entre GEBV₂₀₁₄ et le rendement mesuré 2015 est également très bas avec une valeur de 0.008. Pour le second croisement composé de 46 lignées, le r^2 chute alors à 0.

Avec ce modèle et les données de phénotypage disponibles après la récolte 2014, il est donc impossible de prédire le potentiel de rendement traité d'une lignée que ce soit aussi bien au sein d'un croisement ou entre lignées issues de plusieurs croisements différents.

Dans un second temps, les valeurs de rendement obtenues en

2015 et les $GEBV_{2014+2015}$ sur l'ensemble de la population cible n°1 ont été comparées.

Le graphe ci-dessous montre la relation entre $GEBV_{2014+2015}$ (en abscisse) et rendements obtenus en 2015 (en ordonnée) pour l'ensemble des individus de la population cible n°1.



Le r^2 entre $GEBV_{2014+2015}$ et le rendement mesuré en 2015 est alors de 0.003.

Pour le croisement contenant 41 lignées, le r^2 s'améliore de manière significative puisqu'il atteint alors une valeur de 0.101. Pour le croisement contenant 46 lignées, le r^2 reste trop faible ($r^2 = 0.024$).

Une année de phénotypage supplémentaire de la population d'entraînement n'a donc pas permis l'amélioration de notre modèle de prédiction mise à part pour les prédictions intra croisement de celui composé de 41 lignées.

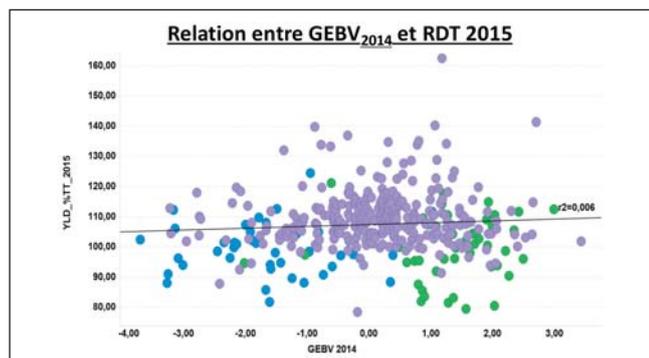
Les individus de la population cible n°1 ont été phénotypés pour le rendement traité que sur une seule parcelle d'expérimentation en 2015, et donc nécessairement que dans un seul environnement. Cela peut expliquer la difficulté à prédire le rendement par SG.

Sur la population cible n°2

Dans un premier temps, les valeurs de rendement obtenues en 2015 et 2016 ont été comparées aux $GEBV_{2014}$ sur l'ensemble de la population cible n°2.

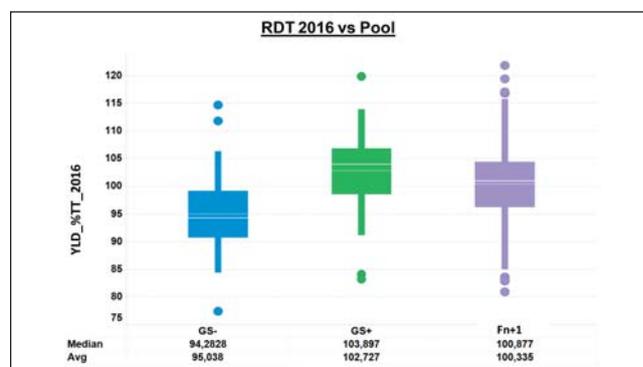
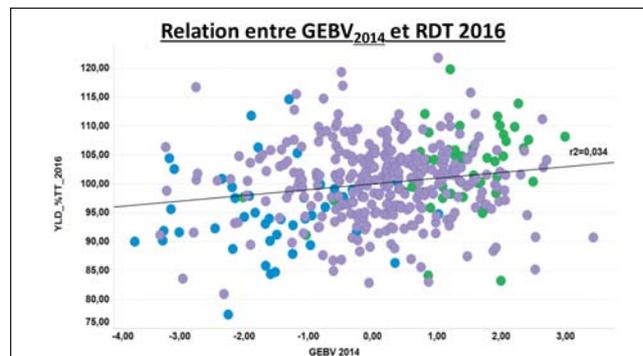
Dans tous les graphiques ci-dessous, les individus de la 2^{ème} population cible seront représentés par 3 couleurs différentes. En bleu, les individus qui ont été choisis pour passer au stade F_{n+1} en raison de leur faible $GEBV_{2014}$ pour le rendement indépendamment de toute sélection phénotypique (pool GS-); en vert, les individus qui ont été choisis pour passer au stade F_{n+1} en raison de leur bonne $GEBV_{2014}$ pour le rendement indépendamment de toute sélection phénotypique (pool GS+) et en violet les 328 individus restant de la population cible n°2 qui ont été choisis pour passer au stade F_{n+1} uniquement sur la base de leur phénotypage.

La corrélation entre $GEBV_{2014}$ et rendement 2015 est alors du même niveau que celle obtenue sur la 1^{ère} population cible. Le R^2 est de 0.006 (graphe ci-dessous).



Une fois encore cette absence de prédiction peut s'expliquer par le fait que la population cible n°2 a été phénotypée sur une seule parcelle d'expérimentation en 2015 et donc nécessairement que dans un seul environnement.

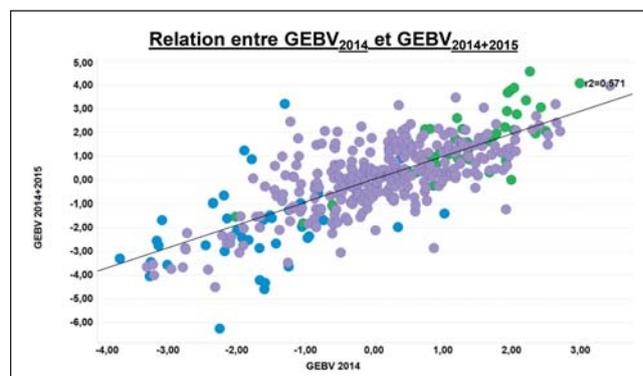
Comme le montre ensuite le graphe ci-dessous, la corrélation entre $GEBV_{2014}$ et rendement 2016 augmente significativement par rapport à celle obtenue entre $GEBV_{2014}$ avec les rendements 2015 tout en demeurant assez faible ($r^2 = 0.034$).



Le graphe ci-dessus, montre une différence significative ($p = 1.39 \times 10^{-7}$) entre les 3 pools. Le pool vert « GS+ » est significativement meilleur que le pool violet « F_{n+1} » qui est lui-même significativement meilleur que le pool GS-.

Malgré l'absence de prédiction de la SG sur le rendement 2016, cette figure montre qu'il aurait été plus performant d'utiliser les valeurs de $GEBV_{2014}$ plutôt que les valeurs de rendement 2015 (pour rappel : 1 seule parcelle d'expérimentation) pour sélectionner le matériel à phénotyper en 2016.

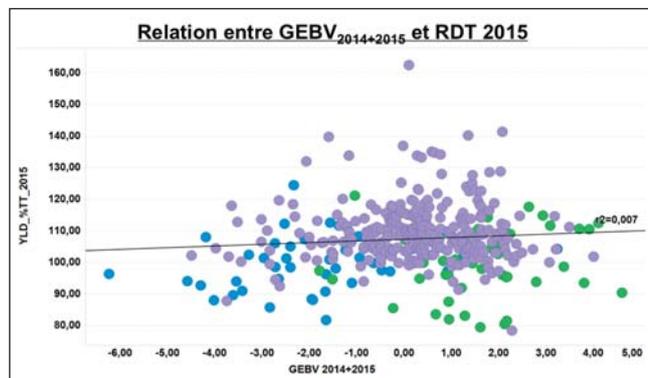
Le graphe ci-dessous montre la corrélation entre $GEBV_{2014}$ et $GEBV_{2014+2015}$. La corrélation est bonne ($r^2 = 0.571$) mais grâce au jeu de couleur nous pouvons tout de même voir que certains individus à faible $GEBV_{2014}$ ne sont plus si mauvais en $GEBV_{2014+2015}$ et inversement. Il est donc intéressant de voir quelles sont les nouvelles corrélations entre $GEBV_{2014+2015}$ et rendement 2015 puis rendement 2016.



Les valeurs de rendement obtenues en 2015, 2016 et 2017 ont

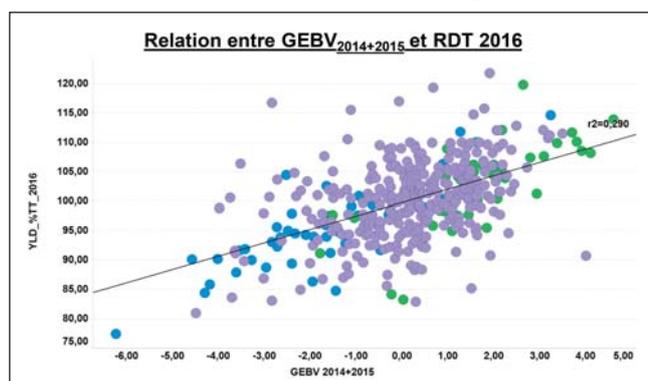
donc été comparées aux $GEBV_{2014+2015}$ sur l'ensemble de la population cible n°2.

Pour l'année 2015, la corrélation entre $GEBV_{2014+2015}$ et rendement est quasi-nulle ($r^2 = 0.007$) comme le montre le graphe ci-dessous.

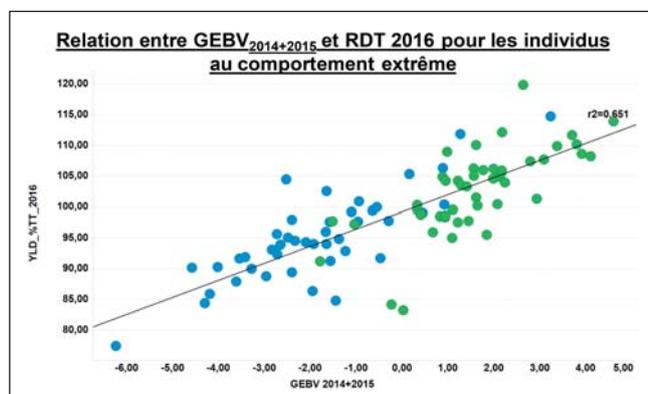


Malgré le phénotypage supplémentaire réalisé en 2015 sur la population d'entraînement R2n, les prédictions ne s'améliorent pas. Une fois encore cette absence de prédiction peut s'expliquer par le fait que la population cible n°2 a été phénotypée sur une seule parcelle d'expérimentation en 2015.

Pour l'année 2016, la corrélation entre $GEBV_{2014+2015}$ et rendement est assez élevée ($r^2 = 0.290$) comme le montre le graphe ci-dessous.

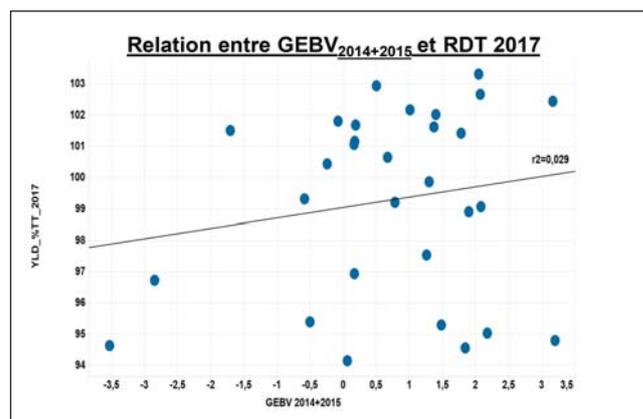


Dans ce cas, la multiplication des environnements dans lesquels la population cible n°2 a été testée et l'ajout d'environnements supplémentaires dans lesquels la population d'entraînement R2n a été phénotypée a un impact très intéressant sur les qualités de prédiction et le r^2 obtenu est largement satisfaisant. En ne s'intéressant qu'aux individus aux comportements extrêmes identifiés sur la base des $GEBV_{2014+2015}$, on peut alors obtenir un r^2 relativement fiable ($r^2 = 0.651$).



Ces résultats confirment donc la possibilité de pouvoir identifier les individus les plus prometteurs sur la base de leur GEBV et de pouvoir éliminer les individus les plus mauvais sur la base de leur GEBV sans prendre de risque.

Le graphe ci-dessous montre la corrélation entre $GEBV_{2014+2015}$ et rendement traité 2017 pour 30 lignées de la population cible conservées en 2017.



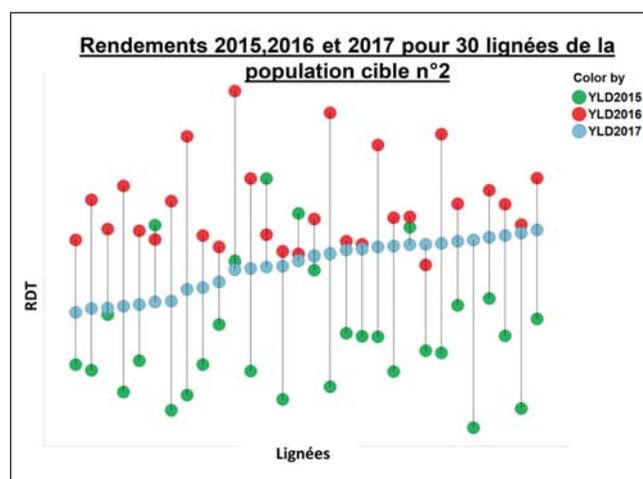
Contrairement à 2016, le r^2 entre rendement traité 2017 et $GEBV_{2014+2015}$ est relativement faible ($r^2 = 0.029$ en 2017 vs $r^2 = 0.306$ en 2016). Le faible effectif (30 individus en 2017 vs 482 en 2016) peut expliquer une variation de cette importance.

Le tableau ci-dessous montre les corrélations entre année pour le rendement traité pour ces 30 lignées.

	YLD_%TT_2015	YLD_%TT_2016	YLD_%TT_2017
YLD_%TT_2015	1		
YLD_%TT_2016	- 0.113015425	1	
YLD_%TT_2017	0.090454022	0.097875168	1

Il n'existe donc aucune corrélation entre les 3 années où des données de rendement ont pu être collectées.

L'effet année est donc très important et le classement variétal est complètement chamboulé d'une année sur l'autre comme le montre le graphe ci-dessous.

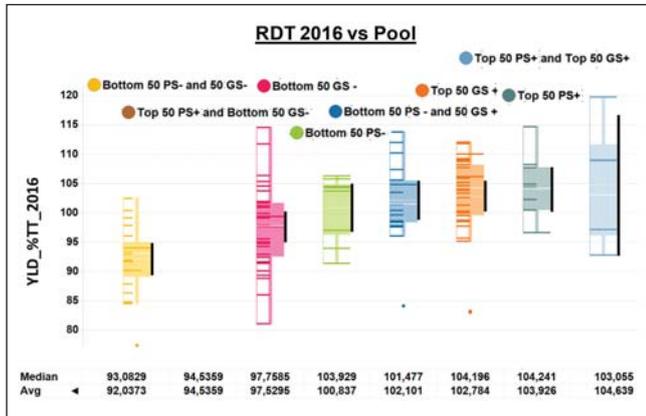


L'année 2016 a été impactée par un effet microdochium très important, cela peut donc expliquer les différences de comportement variétal observées. Par contre, l'absence de corrélation entre 2015 et 2017 ne semble pouvoir être expliquée que par le faible nombre d'individus et par le fait que le matériel avait été phénotypé sur un seul lieu en 2015 vs 13 en 2017.

Comparaison de l'efficacité de la sélection génomique vs celle de la sélection phénotypique ou des 2 combinées

Sur la population cible n°1

Le graphe ci-dessous montre l'impact de l'appartenance des lignées de la population cible n°1 aux différents pools sur le rendement traité 2016.



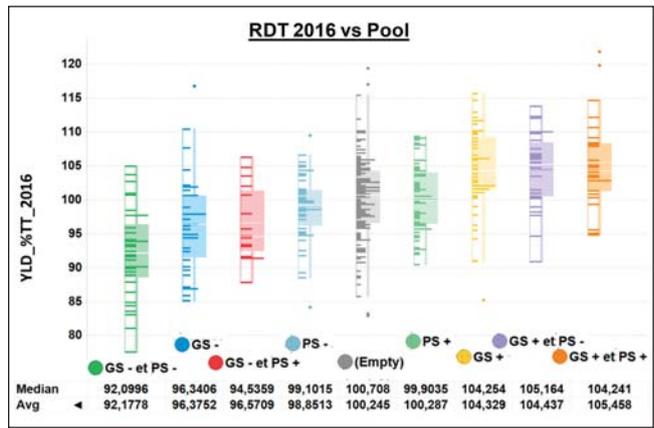
Le nombre de lignées par pool est dans l'ensemble assez faible. En raison d'un effectif trop faible, nous ne pourrions pas conclure sur les 2 pools marron et bleu ciel : « Top 50 PS+ and Bottom 50 GS – » et « Top 50 PS+ and Top 50 GS+ ». Par contre, pour les autres pools une légère tendance se dessine. Le pool orange « Top 50 GS+ » et le pool gris « Top 50 PS+ » ont des médianes équivalentes à 104% des témoins et des moyennes assez proches avec un léger avantage pour le pool « Top 50 PS+ » (103.9% vs 102.7%). Le pool jaune « Bottom 50 PS- and 50 GS – » arrive en dernière position à 92% des témoins en moyenne. Le pool rose « Bottom 50 GS- » donne une large gamme de réponse sur le rendement traité 2016. Ce pool est en effet un des plus mauvais en terme de moyenne (97.5% des témoins) pour le rendement 2016 mais un certain nombre de lignées performe très bien. Sur les 30 lignées appartenant à ce pool, 5 d'entre elles ont été gardées pour passer dans l'année supérieure indépendamment de leur faible GEBV mais elles ont des valeurs de rendement 2016 toutes inférieures à 104% des témoins et même une valeur de 81% des témoins pour la plus mauvaise. Les 4 lignées au dessus de 105% des témoins sont particulièrement intéressantes et il aurait été dommage de les éliminer sur la base des résultats obtenus par SG. Elles n'auraient néanmoins pas été sélectionnées non plus sur la base du phénotype. Il n'aurait donc pas été commis d'erreur trop importante en éliminant les individus aux GEBV₂₀₁₄ les plus faibles.

Enfin, le pool bleu foncé « Bottom 50 PS- and 50 GS+ » est lui aussi assez informatif. Une lignée se détache vers le bas à 84% des témoins et plusieurs individus se détachent au-dessus de 105% des témoins. Sans la sélection génomique, la totalité de ces lignées auraient été éliminée alors que certaines se révèlent intéressantes en termes de potentiel de rendement traité. Se servir des 2 méthodes de sélection aurait donc été un plus puisque ce matériel n'aurait pas été semé en prenant compte des résultats de SG pour ces lignées.

Sur la base de ces résultats, appliquer une stratégie de sélection génomique aurait été plus performant qu'appliquer une simple stratégie de sélection phénotypique dans un seul environnement.

Sur la population cible n°2

Le graphe ci-dessous montre l'impact de l'appartenance des lignées de la population cible n°2 aux différents pools sur le rendement traité 2016.



Le nombre de lignées par pool est dans l'ensemble satisfaisant afin de pouvoir comparer les pools les uns par rapport aux autres.

Trois pools se distinguent pour leurs bons résultats et permettent de mettre en évidence l'efficacité de la sélection génomique pour distinguer les individus les plus prometteurs. Les pools jaune « GS+ » (moyenne à 104.3% des témoins), violet « GS+ et PS- » (moyenne à 104.4% des témoins) et orange « GS+ et PS+ » (moyenne à 105.5% des témoins) se distinguent de tous les autres et sont significativement bien meilleurs que le pool gris « Empty » de référence (moyenne à 100.2% des témoins). En terme de moyenne, le pool « GS+ et PS+ » est légèrement meilleur que les 2 autres, ceci confirme l'intérêt de travailler en combinant à la fois sélection phénotypique et SG.

Deux pools sont équivalents au pool de référence gris « Empty ». Il s'agit des pools bleu clair « PS- » et vert clair « PS+ » avec des moyennes proches de 99-100% des témoins. Ces résultats mettent une fois de plus en évidence la faible corrélation obtenue entre rendements 2015 et 2016 et donc un fort effet année.

Enfin, 3 pools se distinguent pour leurs faibles moyennes de rendement traité en 2016 et permettent de mettre en évidence l'efficacité de la sélection génomique pour distinguer les individus les moins prometteurs. Il s'agit des pools bleu foncé « GS- » (moyenne à 96.3% des témoins), vert foncé « GS- et PS- » (moyenne à 92.2% des témoins) et rouge « GS- et PS+ » (moyenne à 96.6% des témoins). Le pool vert foncé « GS- et PS- » est significativement plus mauvais que les 2 autres pools. Une fois encore, combiner SG et sélection phénotypique présente donc un intérêt.

Sur la base de ces résultats, appliquer une stratégie de sélection génomique aurait été plus performant qu'appliquer une simple stratégie de sélection phénotypique. Combiner Sélection génomique et Sélection phénotypique semble être un très bon compromis pour identifier les individus les plus prometteurs et pour éliminer ceux qui le sont le moins.

3.2 - Prise en compte des effets environnementaux et des interactions GxE dans les modèles de SG

Le tableau ci-dessous montre les composantes de la variance et de l'héritabilité pour chacun des modèles testés.

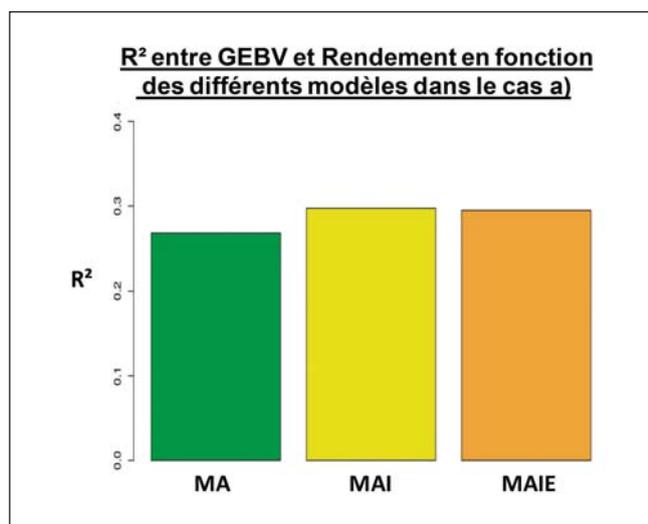
Modèle	Additif	Epistasie	Genetic	GxEC	GxE	Résiduelle	r	r_env	H ²
MA	10.49	NA	11.47	NA	11.09	63.92	0.13	0.26	0.6
MAI	6.08	3.61	10.1	NA	10.88	63.75	0.12	0.25	0.58
MAIE	4.93	3.33	8.26	19.84	0.14	59.38	0.09	0.32	0.5

Les résultats des variances estimées par les différents modèles montrent une répétabilité du caractère étudié (à savoir le rendement) autour de 0.10, suggérant une corrélation au carré entre une seule observation phénotypique (1 plot) et la vraie

valeur génétique totale (qui est inconnue) de 0.1. De la même manière, on trouve une répétabilité dans un environnement donné de 0.25 suggérant une corrélation au carré entre une seule observation phénotypique et la valeur génétique totale dans l'environnement testé de 0.25. Enfin, l'héritabilité estimée au sens large du dispositif est de l'ordre de 0.55 suggérant une corrélation au carré entre la moyenne des observations phénotypiques d'une lignée et la vraie valeur génétique totale de 0.55.

Concernant les différences entre les modèles, le modèle MAIE, qui intègre les co-variables environnementales, capture mieux les effets GxE que les modèles MA et MAI (une résiduelle plus faible et un effet GxE total plus important).

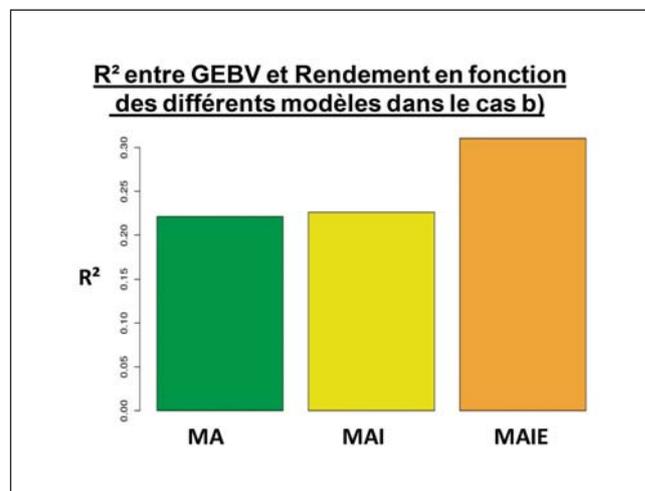
En termes de prédiction des valeurs génétiques moyennes, cas du a), il semble que l'ajout d'épistasie dans le modèle améliore les prédictions mais pas l'ajout du GxE comme le montre la figure ci-dessous.



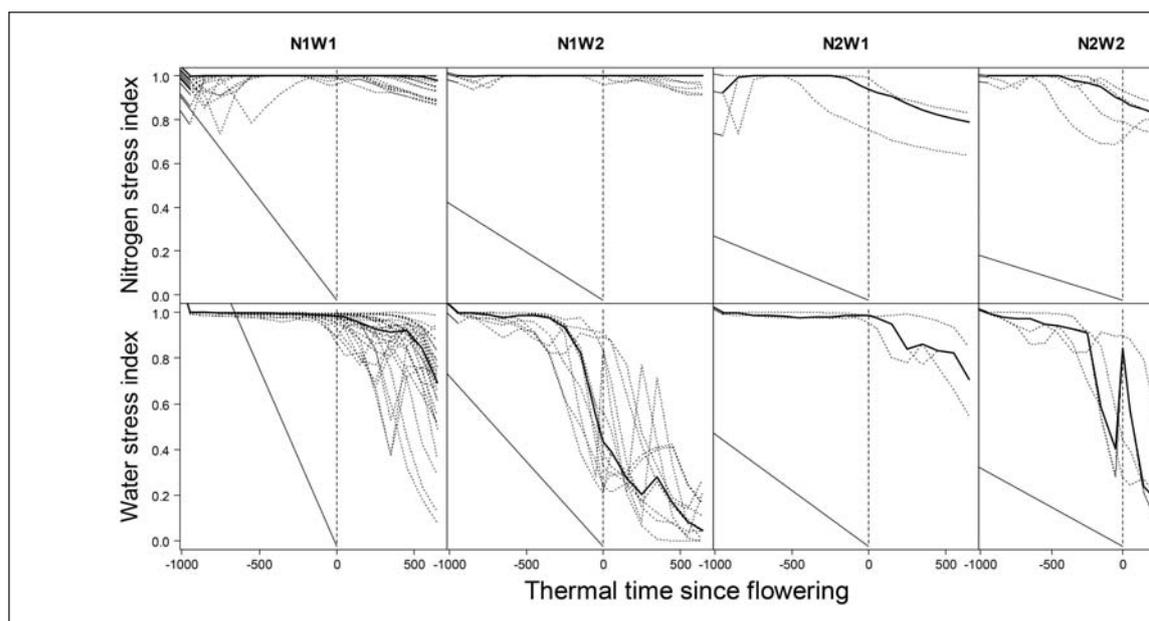
Ce qui semble normal même si on aurait pu imaginer que l'ajout du GxE dans le modèle entraîne une non-orthogonalité des

effets et donc une précision réduite vis-à-vis de la valeur génétique totale. A noter qu'une prédiction d'une lignée non phénotypée mais génotypée avec le meilleur modèle, MAI, donne une précision équivalente à un phénotypage seul dans le dispositif actuel, soit l'héritabilité (le résultat de MAI est de 0.3, qui divisée par H^2 donne $0.3 / 0.55 = 0.55$, équivalent à H^2).

En ce qui concerne les prédictions des valeurs génétiques dans un lieu donné, cas du b), il semble que l'introduction des co-variables environnementales dans le modèle MAIE aient été bénéfiques comme le montre la figure ci-dessous.



Ce résultat signifie que la prédiction d'une lignée génotypée mais pas phénotypée avec MAIE dans un lieu donné est supérieure à une évaluation purement phénotypique dans ce même lieu. Cela vient du fait que la corrélation au carré entre la valeur génétique prédite dans un lieu donné et la vraie valeur dans ce même lieu est égale à (0.31^2) divisé par la répétabilité dans un lieu ($r_{env} = 0.25$, qui est la précision de la valeur génétique dans un lieu donné), ce qui donne $\sim 0.1/0.25 = 0.4 > r_{env}$. En d'autres termes, une prédiction génomique est plus précise qu'une observation phénotypique seule dans un lieu donné.



Dynamique des indices de stress hydrique et azoté (0 = stress maximal, 1 = pas de stress) calculés à l'aide d'un modèle de culture pour la variété Soissons dans 42 essais regroupés en 4 groupes (N1W1, absence de stress azoté ou hydrique ; N2W1, essais stress azoté ; N1W2, essais stress hydrique ; N2W2, essais combinant stress hydrique et stress azoté). Les lignes tiretées représentent la dynamique des indices de stress pour les différents groupes d'essai. Les lignes pleines représentent les médianes des indices de stress pour chaque groupe d'essai.

3.3 - Calibration de modèles de SG spécifiques à différents scénarios environnementaux

► Regroupement des essais

La méthode appliquée a permis de regrouper les essais en quatre groupes correspondant à quatre scénarios environnementaux :

- N1W1 : absence de stress hydrique ou azoté
- N1W2 : absence de stress azoté / stress hydrique de fin de cycle
- N2W1 : stress azoté au cours du cycle / absence de stress hydrique
- N2W2 : présence de stress hydrique et azoté

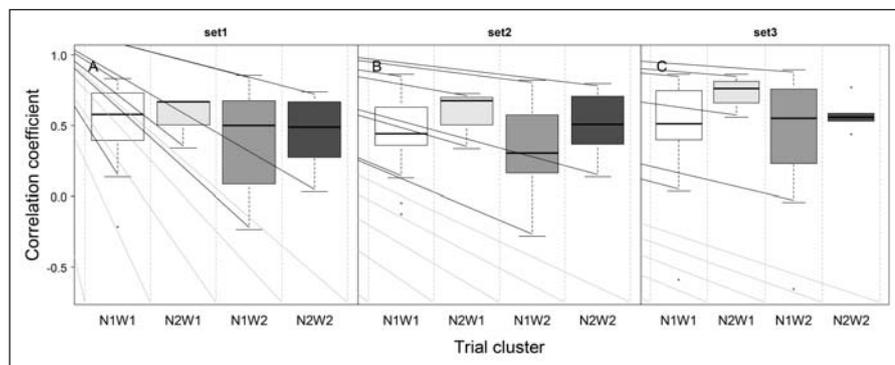
Le groupe N1W1 comprenait 28 essais, dont 17 essais en condition de fertilisation optimale, deux essais en condition de fertilisation limitante, quatre essais en condition pluviale et cinq essais « date de semis tardive ». Le groupe N1W2 comprenait 11 essais, dont six essais en condition de fertilisation optimale, un essai en condition de fertilisation limitante, un essai en condition pluviale, deux essais « date de semis tardive » et un essai « densité de semis réduite ». Le groupe N2W1 comprenait trois essais en condition de fertilisation limitante. Enfin, le groupe N2W2 comprenait quatre essais en condition de fertilisation limitante et un essai « date de semis tardive ».

► Comparaison des capacités prédictives

L'analyse des capacités prédictives des modèles calibrés sur différents jeux de données montre que l'utilisation de modèles spécifiquement calibrés pour un scénario environnemental donné semble permettre un léger gain en termes de prédiction. C'est le

cas pour trois des quatre groupes d'environnements testés avec N1W2 (r médian égal à 0.55 avec le set 3 contre 0.50 avec le set 1), N2W1 (r médian égal à 0.76 avec le set 3 contre 0.67 avec le set 1) et N2W2 (r médian égal à 0.56 avec le set 3 contre 0.49 avec le set 1). En revanche, pour le groupe N1W1, la calibration sur des données spécifiques à ce groupe conduit à une diminution des capacités prédictives (r médian égal à 0.51 avec le set 3 contre 0.57 avec le set 1). Il est à noter que ce groupe est celui qui contient le plus d'essais et compte tenu de la nature quantitative des dynamiques de stress, la classification des essais dans ce groupe (comme dans les autres) reste relativement sujette à imprécision. Comme attendu, les capacités prédictives des modèles calibrés sur le set 2 sont systématiquement plus faibles que celles obtenues avec les modèles calibrés sur les sets 1 et 3 (r médian égal à 0.44 et 0.30 pour les groupes N1W1 et N1W2 avec le set 2 contre 0.57 et 0.50 pour le set 1). Néanmoins les différences sont parfois relativement faibles (r médian égal à 0.67 et 0.51 pour les groupes N2W1 et N2W2 avec le set 2 contre 0.67 et 0.49 pour le set 1).

Même si les différences entre le set 1 et le set 3 sont relativement faibles, il est notable que leurs capacités prédictives ont été obtenues avec un nombre d'essais de calibration très différent (45 pour le set 1 et deux à 27 pour le set 3 suivant le scénario environnemental considéré). Il semble donc qu'il soit possible d'atteindre des capacités prédictives similaires pour un scénario environnemental donné avec un nombre d'essais restreint, à condition de caractériser les essais et d'inclure dans la base de données de prédiction des essais dans lesquels les conditions environnementales correspondent au scénario ciblé.



Boxplots des corrélations intra-essais après validation croisée de type « leave-one-out » entre valeurs prédites et observées pour le rendement en grains. Les prédictions ont été obtenues à l'aide d'un modèle intégrant un effet génotype (basé sur les données de génotypage), un effet environnement (basé sur l'utilisation de co-variables agroclimatiques) et un effet d'interaction GxE (interactions marqueurs × co-variables agroclimatiques). Ce modèle a été calibré en utilisant trois jeux de données différents : **set1** l'ensemble des essais sauf l'essai j (appartenant au groupe d'essais k) pour lequel les prédictions étaient réalisées, **set2** l'ensemble des essais sauf les essais appartenant au groupe k, **set3** l'ensemble des essais appartenant au groupe k sauf l'essai j.

4. Conclusions

Les résultats de prédiction obtenus varient énormément d'une année sur l'autre ($0 < r^2 < 0.3$) et confirment les difficultés du sélectionneur à anticiper les effets année.

Quoi qu'il en soit, il ressort de ces analyses que le sélectionneur aurait moins de risque de se tromper en appliquant un modèle de sélection génomique rendement sur son matériel plutôt que de se fier à une seule parcelle rendement dans un environnement donné. De même, combiner à la fois les résultats obtenus par sélection génomique et par sélection phénotypique permettrait au sélectionneur d'éliminer sans risque les individus les plus mauvais et de mettre en évidence rapidement les plus prometteurs.

Un modèle classique G-BLUP prenant en compte à la fois les effets d'additivité et d'épistasie ressort comme étant le modèle le plus efficace pour prédire la valeur génétique moyenne d'un individu de blé tendre. Introduire les interactions GxE dans ce modèle après calculs de co-variables environnementales permet d'améliorer considérablement les prédictions génétiques dans un environnement non-testé et est susceptible d'apporter des informations sur la stabilité des lignées et de compléter les données de réseaux d'essais multi-environnementaux qui sont souvent déséquilibrés. Travailler par groupes d'environnements, après caractérisation fine de ces derniers, pourrait donc être une solution afin d'améliorer considérablement les modèles de sélection génomique rendement par prise en compte de des interactions GxE.

Références bibliographiques

- Bernado, R., and J. Yu.** 2007. Prospects for genome-wide selection for quantitative traits in maize. *Crop Sci.* 47:1082-1090. Doi:10.2135/cropsci2006.11.0690
- Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa.** 2012. Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52:707-719.
- Chenu K., Cooper M., Hammer G.L., Mathews K.L., Dreccer M.F., Chapman S.C..** 2011. Environment characterization as an aid to wheat improvement: interpreting genotype environment interactions by modelling water deficit patterns in north eastern Australia. *Journal of Experimental Botany* 62: 1743–1755.
- Comstock, R.E., and R.H. Moll.** 1963. Genotype-environment intersections. p. 164-196. In: W.D. Hanson and H.F. Robinson (ed.) *Statistical genetics and plant breeding.* NAS-NRC Publ. 982.
- Crossa, J., G. de los Campos, P.Pérez, D. Gianola, S. Dreisigacker, J. Burgueño, J.L. Araus, D. Makumbi, J. Yan, R. Singh, V. Arief, M. Banziger, and H.-J. Braun.** 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186:713-724. Doi:10.1534/genetics.110.118521
- Crossa, J., P. Pérez, G. de los Campos, G. Mahuku, S. Dreisigacker, and C. Magorokosho.** 2011. Genomic selection and prediction in plant breeding. *J. Crop. Improv.* 25:239-261. Doi:10.1080/15427528.2011.558767
- Crossa, J., P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, D. Bonnett and K. Mathews.** 2013. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* (2013), 1-13.
- de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel and J.M. Cotes.** 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182:375-385. Doi10.1534/genetics.109.101501
- de los Campos, G., D. Gianola, G.J.M. Rosa, K.A. Wiegeler, and J. Crossa.** 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* 92:295-308. Doi10.1017/S0016672310000285
- Endelman, J.B.** 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4:250-255. Doi:10.3835/plantgenome2011.08.0024
- Jarquín, D., J. Crossa, X. Lacze, P. D. Chevron, J. Daucourt et al.** 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *TAG* 127:595-607.
- Legarra, A., A. Ricard, O. Filangi.** 2013. GS3 software. Available at: <http://snp.toulouse.inra.fr/~alegarra/>. Version 6. INRA, Toulouse.
- Long, N., D. Gianola, G. J. M. Rosa, S. Weigel and S. Avendano.** 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J. Anim. Breed. Genet.* 124:377-389
- Meuwissen, T.H.E., B.J. Hayes, and M. E. Goddard.** 2001. Prediction of total genetic values using genome-wide dense marker maps. *Genetics* 157 :1819-1829.
- Pérez, P., G. de los Campos, J. Crossa and D. Gianola.** 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression package in R. *Plant genome* 3(2):106-116.
- Pérez, P., G. de los Campos, J.** 2014. Genome-wide Regression and Prediction with the BGLR statistical package. *Genetics* 198 (2): 483-495.
- R Core Team.** 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
- Tier, B., J. Cavanagh, R. Crump, M. Khatar, G. Moster,** 2007. Genome wide selection : experiences from the Australian Dairy Industry. The 3rd International Conference on Quantitative Genetics, August 2007, Hangzhou, China.
- Woolaston, A.F., B. Tier and R.D. Murison.** 2007. Principal Component regression of SNP to predict genetic merit. Papers and Abstracts From the Workshop on QTL and Marker-Assisted Selection, March 22-23, 2007, Toulouse, France, edited by A. Legarra.