Optimisation économique de la prédiction de la note de panification : apport des caractères corrélés et des marqueurs moléculaires

Sarah Ben Sadoun

doctorante, INRA GDEC

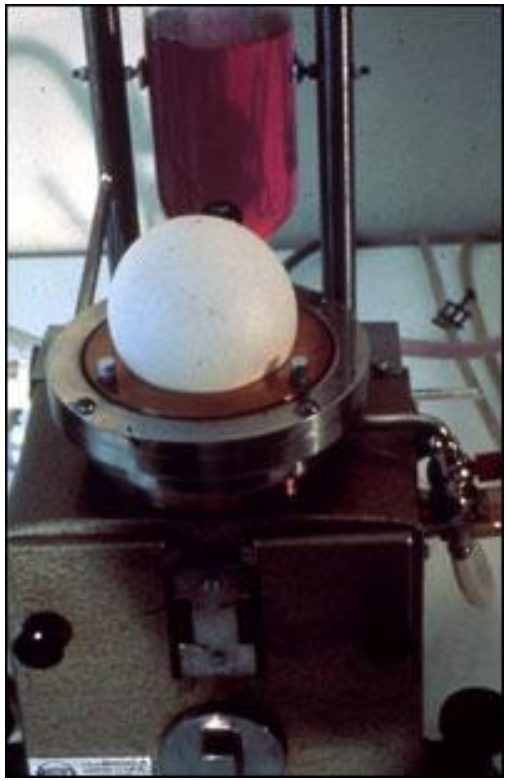
R Rincent, C Ravel, FX Oury, B Rolland, E Heumez,
J Auzanneau, G Charmet, S Bouchet

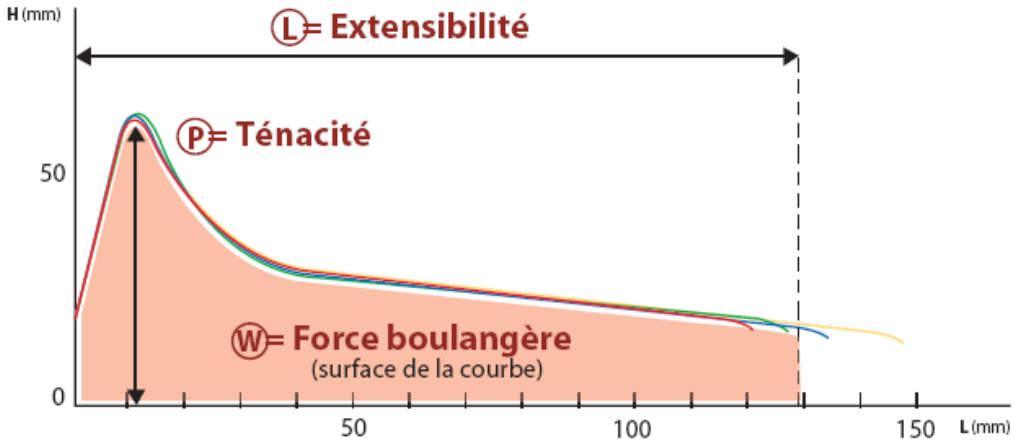
La qualité boulangère : un caractère complexe

- « Bread making score » : BMS
- Intègre plusieurs mesures physiques sur la pâte et le pain (Volume, ténacité, élasticité, force…)
- Nécessite d'avoir plusieurs kg de grains pour chaque lignée pour pouvoir faire les mesures
- Coûte cher à phénotyper

→ Caractère évalué à la fin du programme de sélection

Alvéographe de Chopin



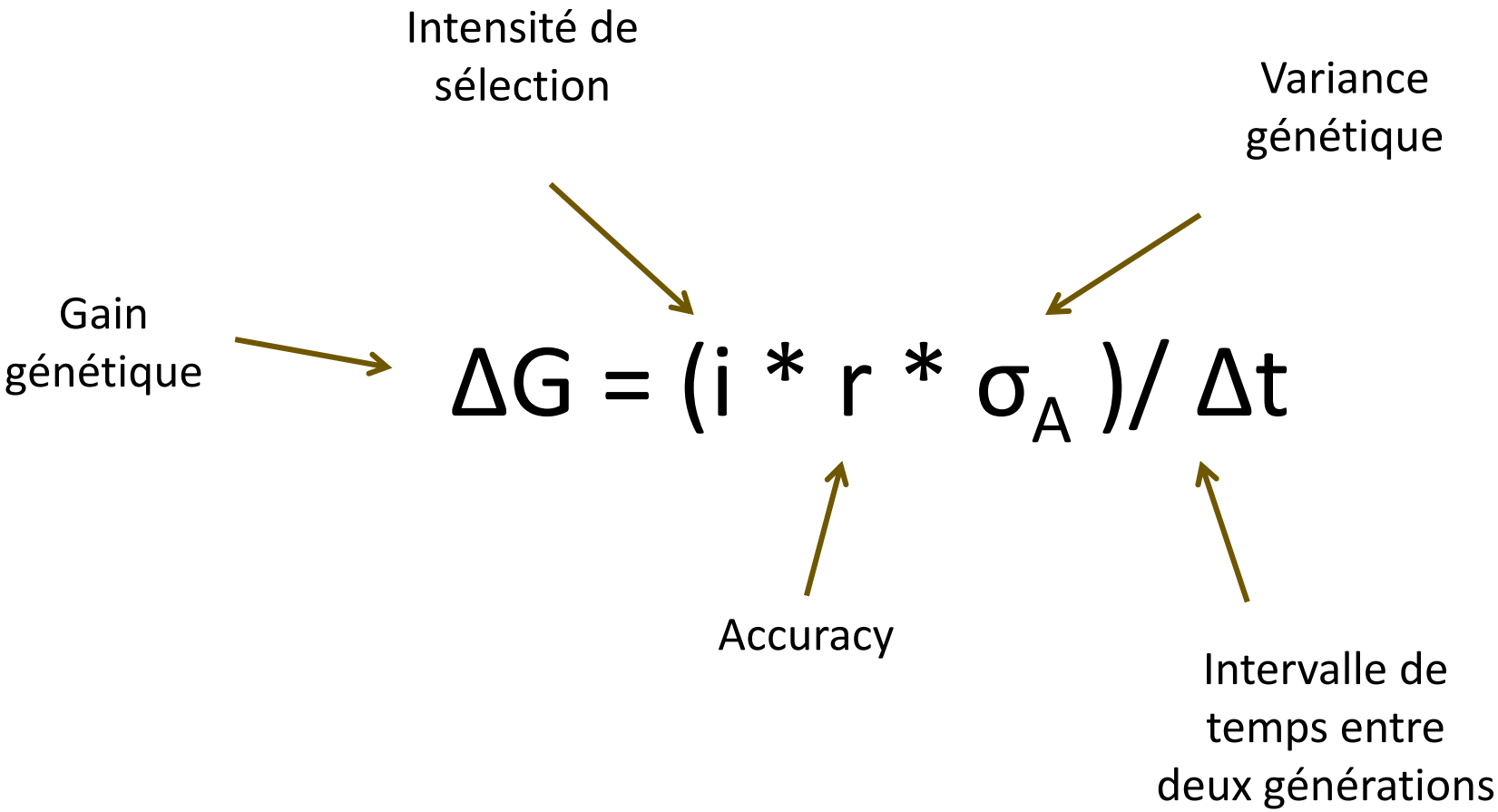


La qualité boulangère à l'inscription

- 7 classes de qualité définies par le CTPS:
  - **A'** : blé ayant des caractéristiques originales dont l'intérêt est attesté par les utilisateurs
  - **A** : blé de force ou améliorant
  - **B.P.S.** : blé panifiable supérieur
  - **B.P.** : blé panifiable
  - **B.B.** : blé biscuitier
  - **B.A.U.** : blé autres usages
  - **B.A.U. impanifiable** : blé autres usages impanifiable

Classe technologique	BMS	rendement
BPS	note > 250	> 102% témoins
BP, BB	220 < note < 250	> 104%
BAU	note < 220	> 107%
BAU impanifiable		>109%

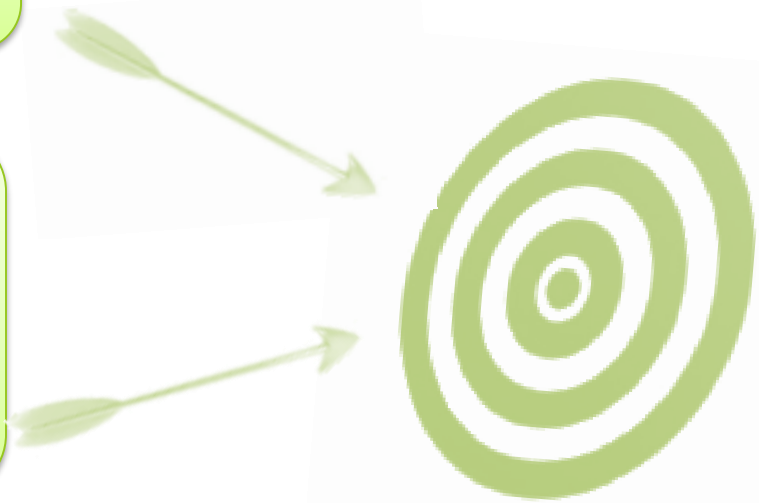
Equation du sélectionneur



Objectif du sélectionneur

Améliorer la qualité des prédictions de la qualité boulangère

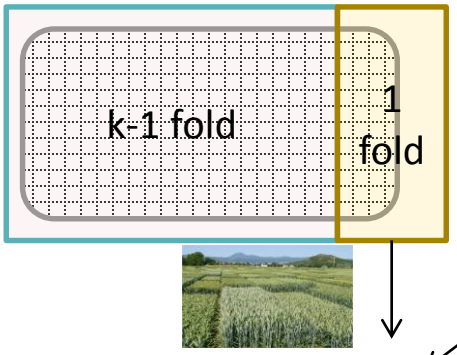
Optimiser l'utilisation du budget de phénotypage pour augmenter le nombre de candidats phénotypés



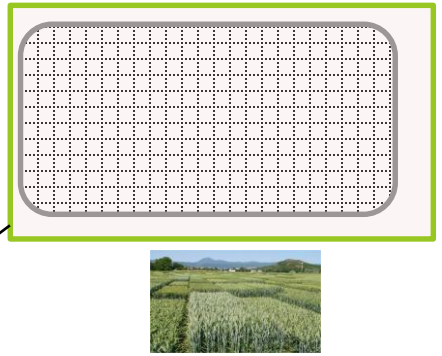
- Phénotyper tous les candidats dans un maximum d'environnement pour BMS (150 euros)
- Phénotyper un caractère corrélé moins cher (20 euros)
- Génotyper des marqueurs moléculaires diagnostiques (10 euros)
- Génotyper des marqueurs moléculaires neutres sur l'ensemble du génome (35 euros): prédictions génomiques
- Une combinaison judicieuse de ces informations

Prédictions génomiques

Cross validation du modèle



Validation du modèle Sur une population déconnectée

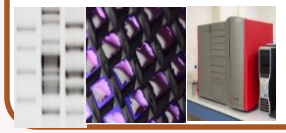


Calibration du modèle

Population de calibration

Phenotypes

Genotypes



Modèle

Calibration du modèle

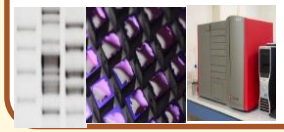
Calibration du modèle

$$y = \mu + Z\beta + e$$

estimation du vecteur β des effets

Population de validation

Genotypes



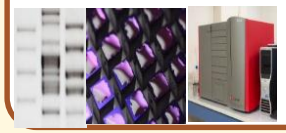
GEBV validation

Predictive ability = cor (y, GEBV)

Sélection

Population de candidats

Genotypes



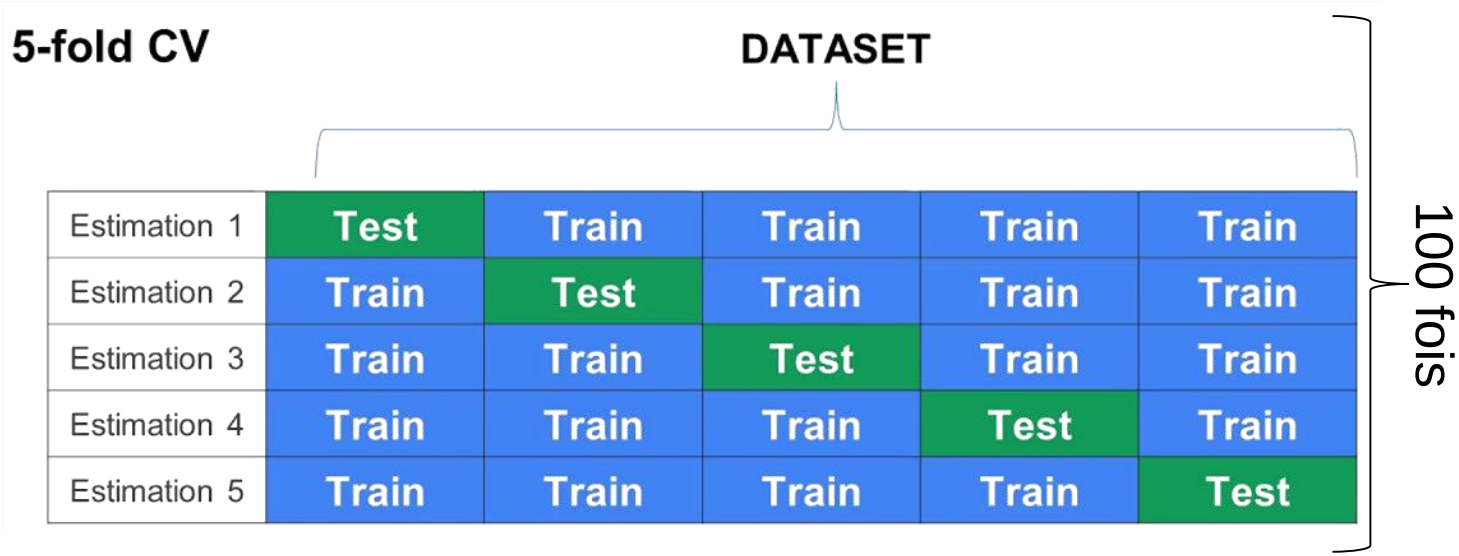
GEBV candidats

Tri des candidats

$$GEBV = Z\hat{\beta}$$

Qualité des prédictions

Cross-validation



Qualité de prédiction

Predictive ability = cor (Prédit, BLUEs)

Modèle de prédiction mono-caractère

GBLUP

$$\boldsymbol{y = \mu + Za + e}$$

n = nombre d'individus
y = vecteur (de longueur n x 1) des données phénotypiques (BLUEs)
μ = moyenne du caractère
Z = matrice d'incidence
a = vecteur (de longueur n x 1) des effets des individus, avec $a \sim N(0, K\,\sigma_a^2)$
K = matrice de Kinship (de taille n x n)
σ_a^2 = variance génétique additive
e = vecteur (de longueur n x 1) des résidus avec $e \sim N(0, I\,\sigma_e^2)$
σ_e^2 = variance résiduelle

Modèle de prédiction multi-caractère

GBLUP

$$y_t = X_t\mu_t + Z_t a_t + e$$

n = nombre d'individus

t = nombre de caractères

y = vecteur (de longueur n x t) des données phénotypiques (BLUEs)

μ_t = moyennes des caractères

X_t = matrice des effets fixes

Z_t = matrice d'incidence (effets aléatoires)

a = vecteur (de longueur n x t) des effets des individus, avec $a \sim MVN(0, K \otimes \Sigma_a)$

K = matrice de Kinship (de taille n x n)

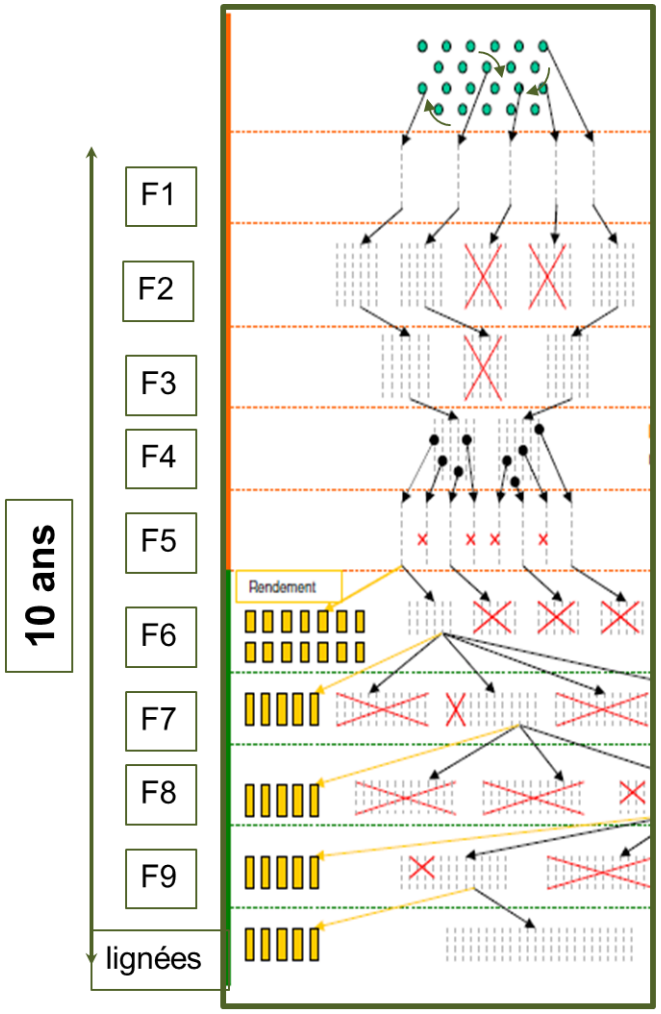
Σ_a = matrice de variance covariance de la forme suivante $\begin{pmatrix} \sigma_{a1}^2 & \sigma_{a12} \\ \sigma_{a12} & \sigma_{a2}^2 \end{pmatrix}$

σ_{a1}^2 = variance génétique additive pour le caractère 1

e = vecteur (de longueur n xt) des résidus avec $e \sim MVN(0, I_N \otimes \Sigma_e)$

Données phénotypiques et génotypiques Du programme de sélection INRA-AO

Schéma de sélection



Croisements : 10^2

Génération F2 : bulks

F3 : bulks

10^5

10^4

10^3

10^2

10^1

10^0

Inscription des nouvelles variétés

Jeu de données: lignées F8 F9

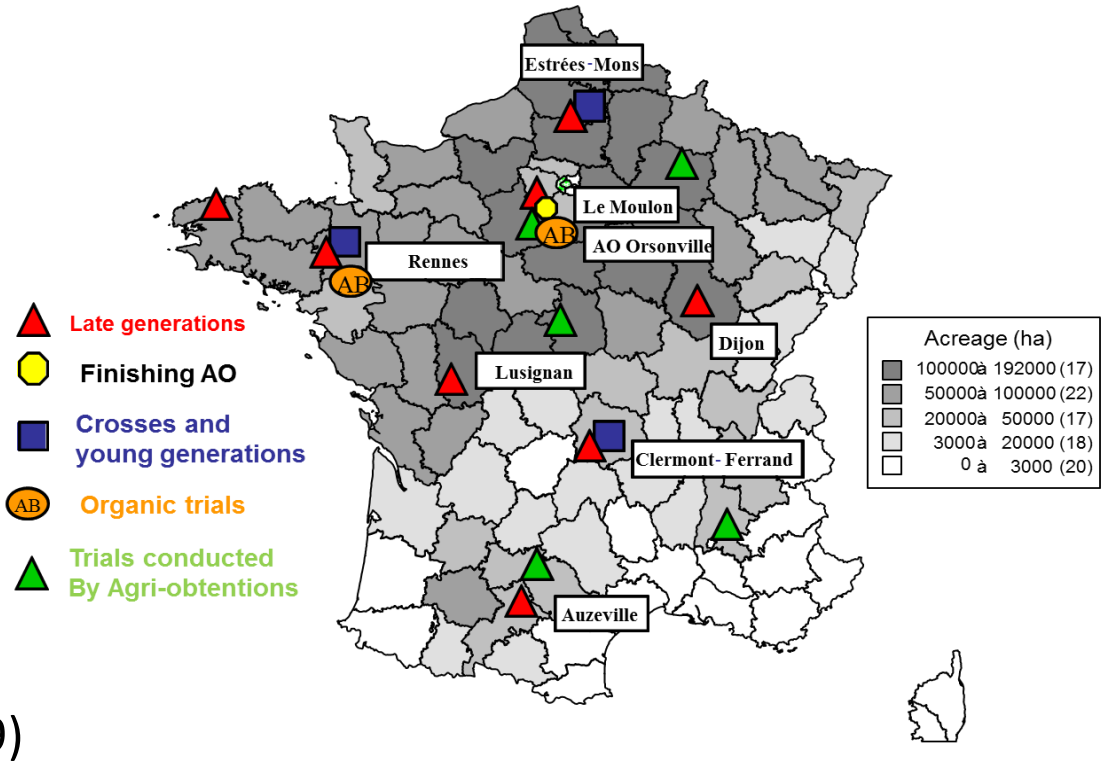
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2000	15																	
2001	13	28																
2002		13	22															
2003			10	37														
2004				8	32													
2005					9	24												
2006						12	38											
2007							17	58										
2008								21	47									
2009									18	44								
2010										16	55							
2011											17	48						
2012												19	55					
2013													29	75				
2014														31	79			
2015															30	83		
2016																32	82	
2017																	31	81

Réseau d'essais

Mesures effectuées :

- De 2000 à 2017

- Sur 11 sites en France :
  - Clermont-Ferrand
  - Colmar (jusqu'en 2007)
  - Dijon
  - Estrées-Mons
  - Le Moulon
  - Lusignan
  - Orsonville (à partir de 2009)
  - Rennes
  - Champagne-Céréales
  - Epi Centre (près de Bourges)



Données phénotypiques

- 13 caractères « agronomiques »

  Dont : le rendement, la précocité, la teneur en protéines, les notes de maladies, etc.

- 19 caractères « technologiques » pour la moitié des lignées

  Dont : la dureté du grain, les paramètres de l'alvéographe de Chopin, les tests de panification, etc.

Certains caractères ne sont pas mesurés tous les ans sur chaque site

Données génotypiques

1. 180K SNP (816 lignées) et 35K SNP (pour les 91 lignées les plus récentes)
   - ➔ 35K SNP gardés pour toutes les lignées
   - ➔ matrice de Kinship calculée à partir de ces données
   - ➔ 398 lignées génotypées et phénotypées pour W et la qualité boulangère (Bread making score=BMS)

2. 12 marqueurs développés par Catherine Ravel (INRA GDEC) liés aux loci Glu-A1, Glu-B1, et Glu-D1 (sous-unités des gluténines de haut poids moléculaire)
   - ➔ 200 lignées phénotypées pour W et pour la qualité boulangère et génotypées avec ces marqueurs et avec la puce

Two-step GBLUP

Données phénotypiques obtenues sur plusieurs années et dans plusieurs locations

BLUEs pour les n individus

Prédictions des performances des candidates

Première étape : corriger les effets environnementaux

Deuxième étape : prédire les performances des candidats

Résultats

Héritabilité des caractères

$$\text{Héritabilité} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GxE}^2 + \sigma_\varepsilon^2}$$

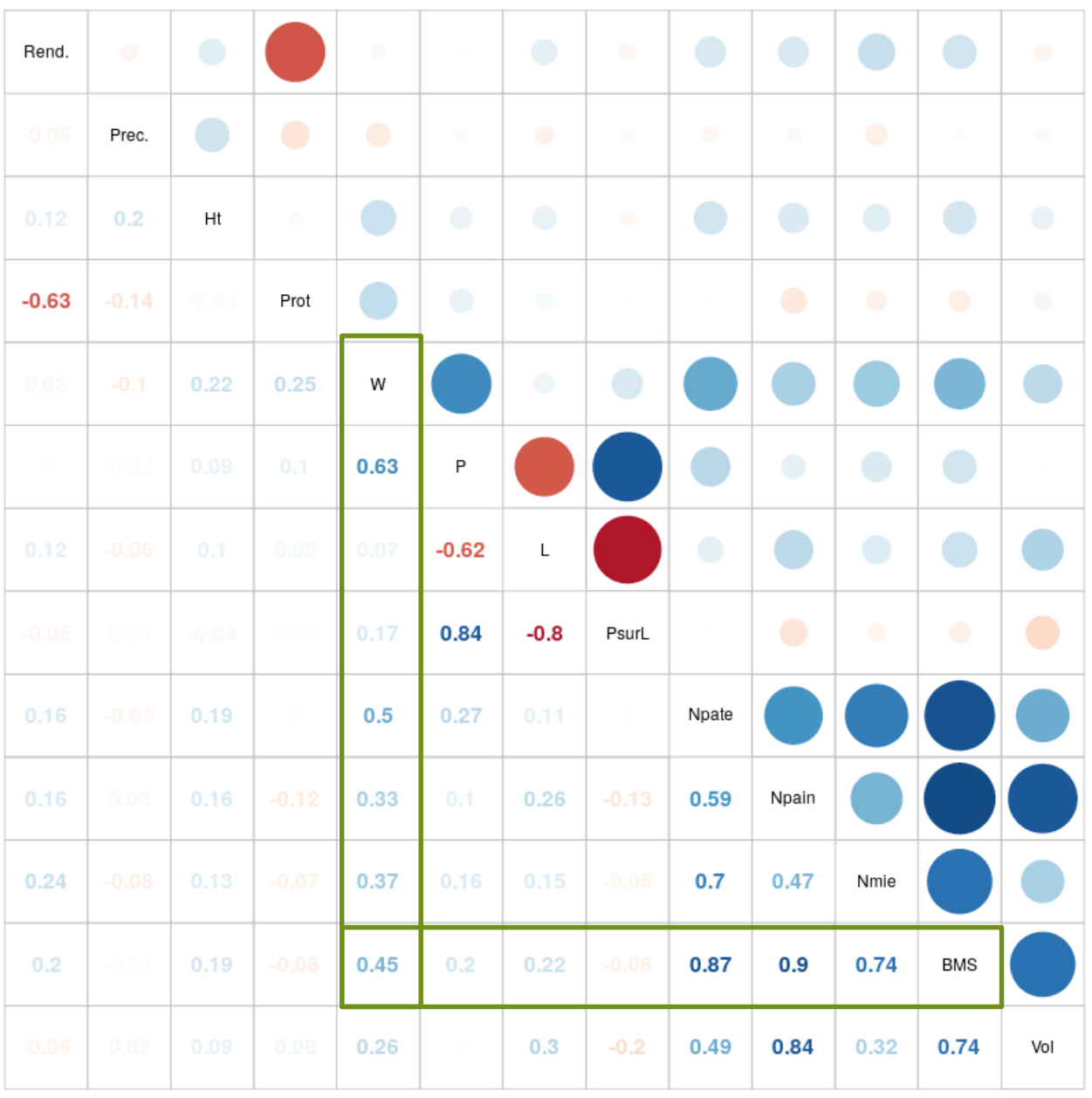
Qualité

Caractères rhéologiques (alveographe de Chopin)

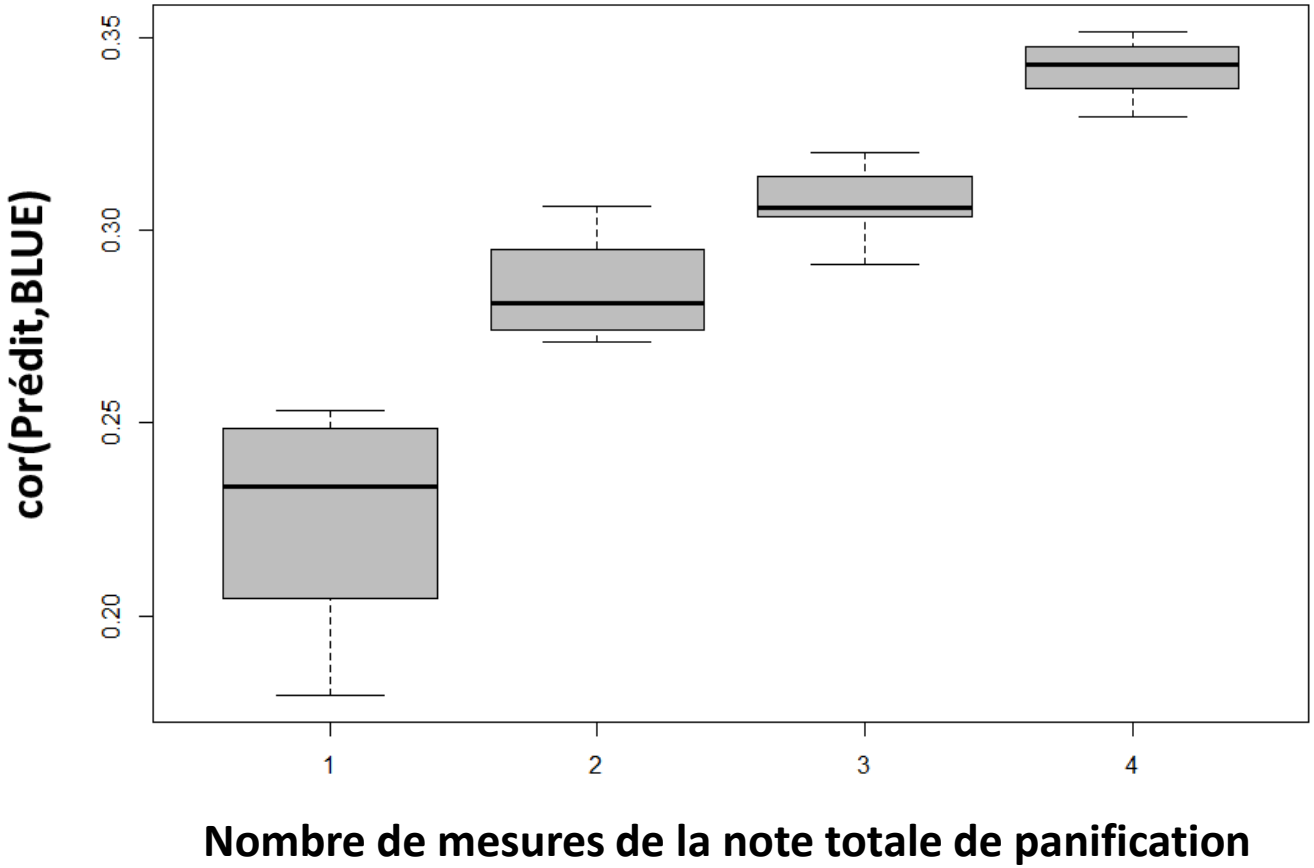
Note de panification: utilisée en France pour évaluer la qualité boulangère du blé

Caractère	Héritabilité
Rendement	0.28
Précocité	0.81
Hauteur	0.74
Protéines	0.46
Dureté	0.86
W	**0.66**
P	0.78
L	0.62
P/L	0.59
Note de pâte	0.37
Note de mie	0.31
Note de pain	0.37
Note totale (BMS)	**0.40**
Volume	0.41

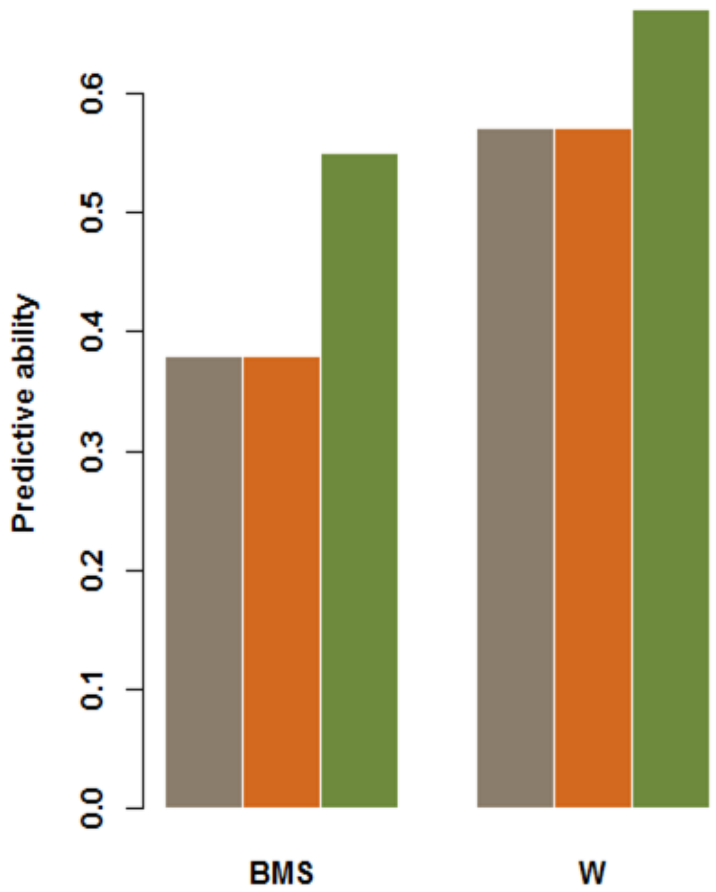
Corrélation des caractères



Impact du nombre de mesures de la note totale sur la qualité des prédictions

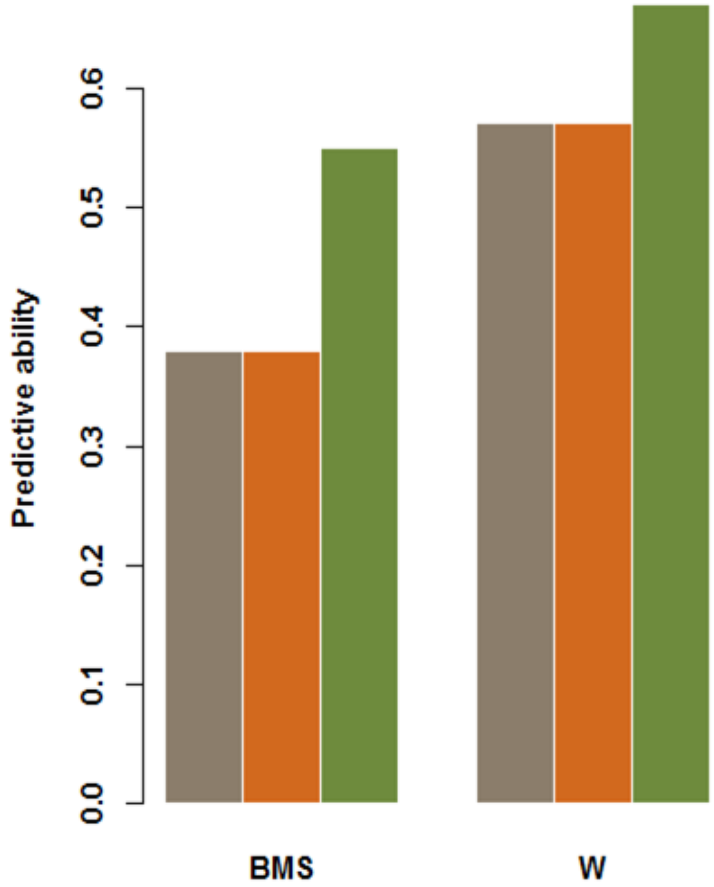


Mono-caractère vs. multi-caractère



- **Single-trait (ST)** = prédictions génomiques **mono-caractères**
- **Multi-trait (MT)** = prédictions génomiques multi-caractères,
  Train: BMS et **W phénotypés**
  Test: pas de phénotypage
- **Trait-assisted (TA)** = prédictions génomiques multi-caractères
  Train: BMS et **W phénotypés**
  Test: W phénotypé

Mono-caractère vs. multi-caractère



➔ **Pas d'amélioration** de la qualité de prédiction **avec le modèle MT** comparé au modèle ST

➔ **Amélioration** de la qualité des prédictions avec le **modèle TA**

➔ Gain plus important pour BMS que pour W (dont l'héritabilité est supérieure à celle de BMS)

Apport des marqueurs liés aux gluténines

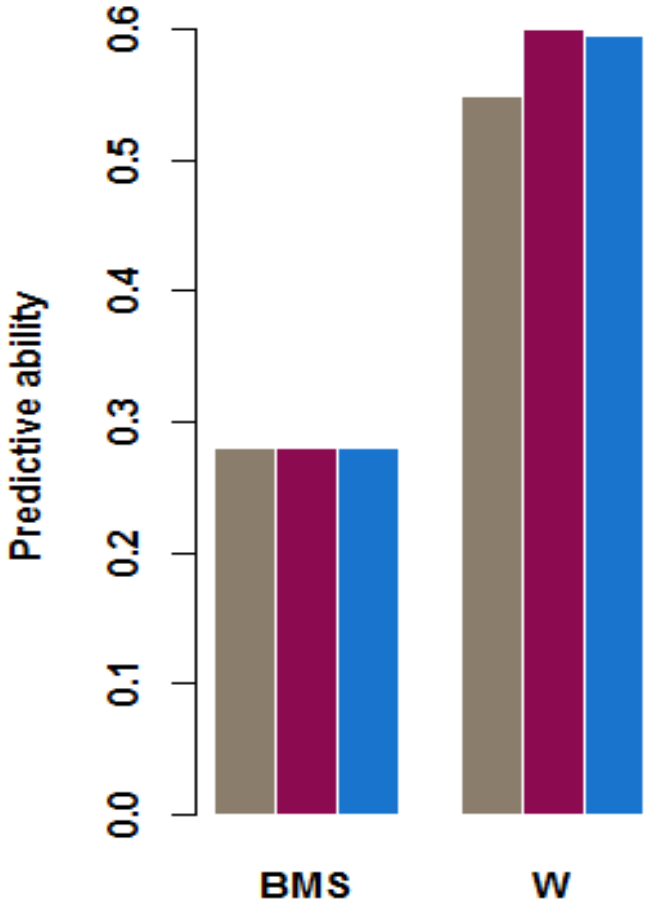
- 200 lignées phénotypées, génotypées 35K et marqueurs gluténines

- Sélection des marqueurs (modèle stepwise) qui expliquent la variation du phénotype

1. 3 marqueurs expliquent environ 31% de W

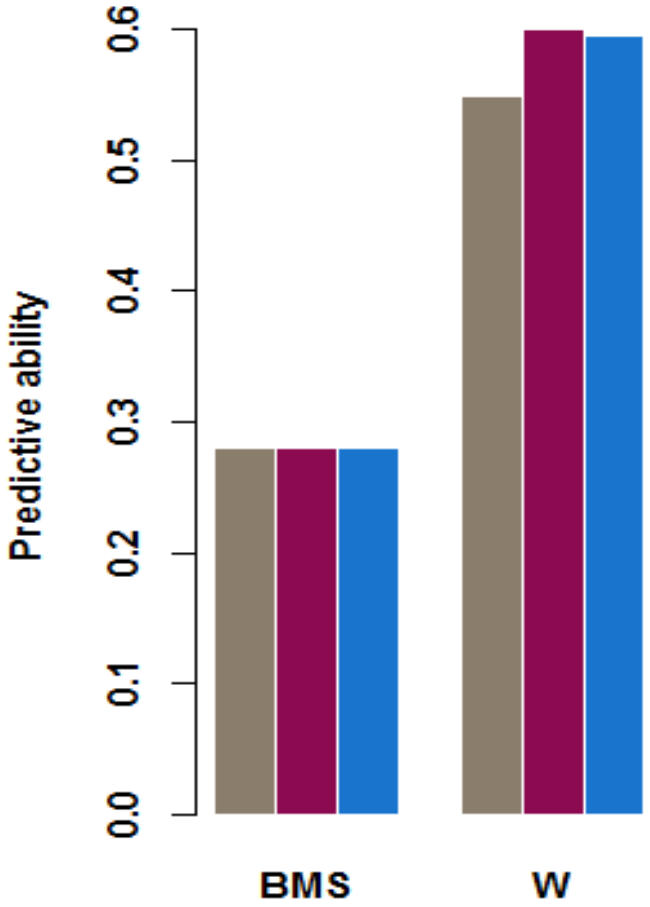
2. 4 marqueurs expliquent 7% de BMS

Apport des marqueurs liés aux gluténines



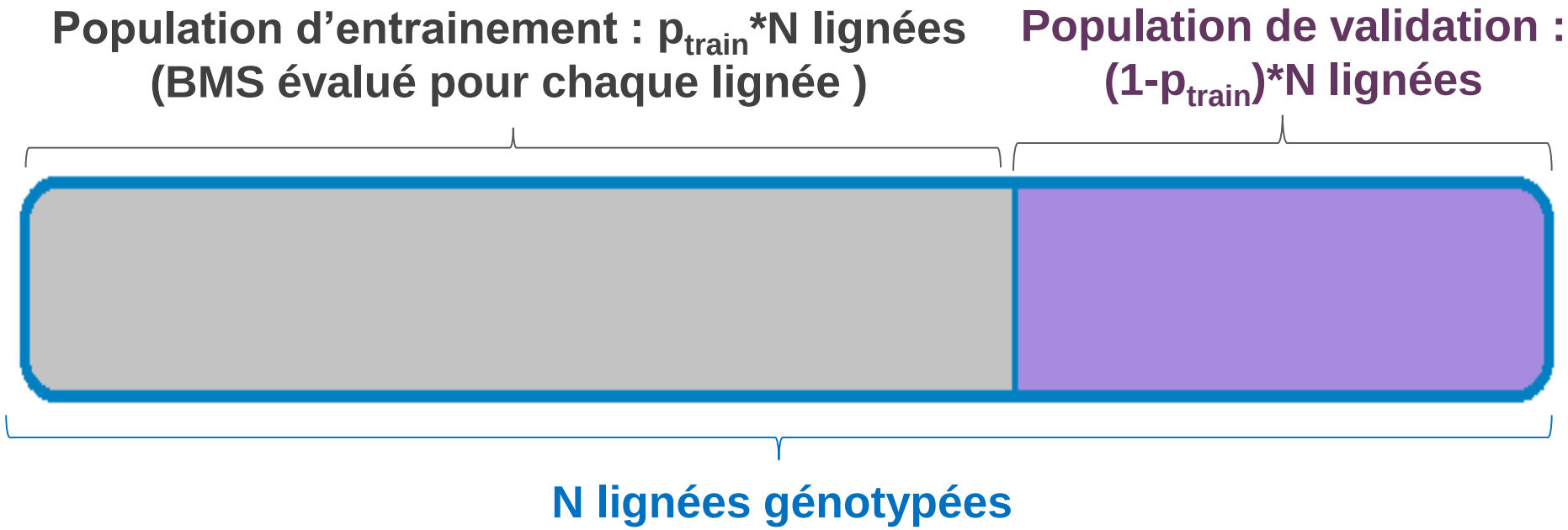
- **Single-trait (ST)** = prédictions génomiques **mono-caractères**
- **ST-Glu** = prédictions génomiques **mono-caractères**, avec **12 marqueurs** gluténines en **effet fixe**
- **ST-Glu_stepwise** = prédictions génomiques **mono-caractères**, avec 4 ou 3 **marqueurs** gluténines en **effet fixe**

Apport des marqueurs liés aux gluténines



- Pas de gain de qualité de prédiction pour la note totale de panification (BMS)
- Meilleure qualité de prédiction du W avec les modèle ST-Glu et ST-Glu_stepwise où les marqueurs gluténines sont en effet fixe

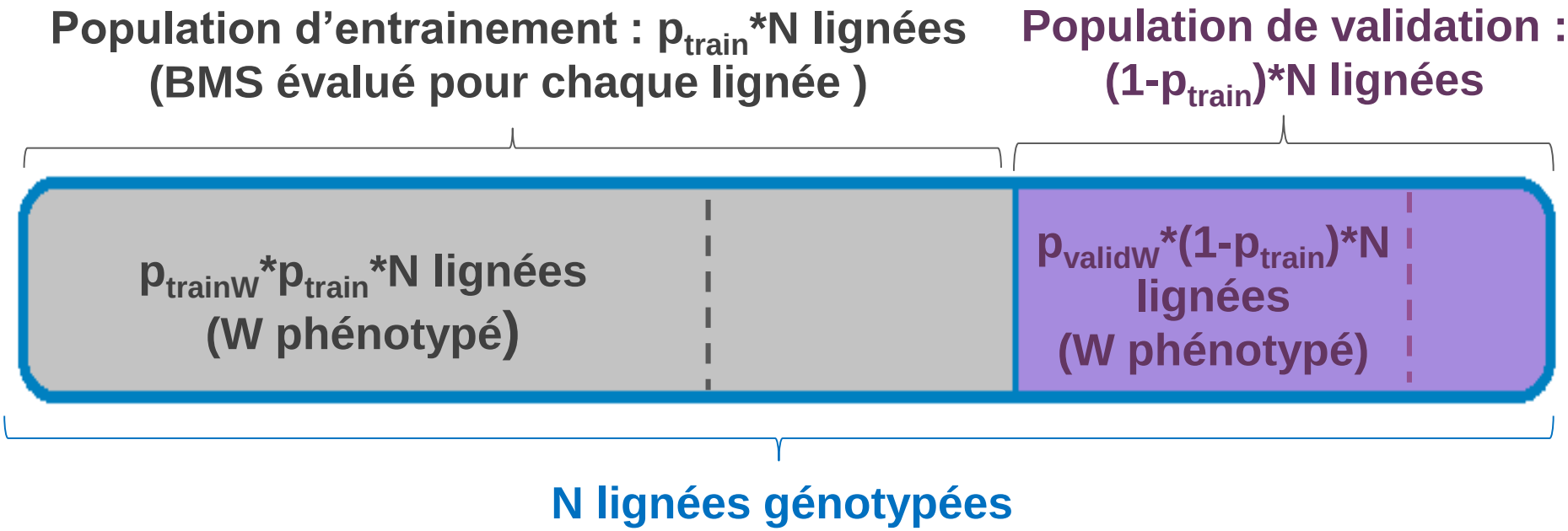
Comparaison des coûts des différents modèles



Avec :

- **p_{train} la proportion de lignées dans la population d'entrainement**

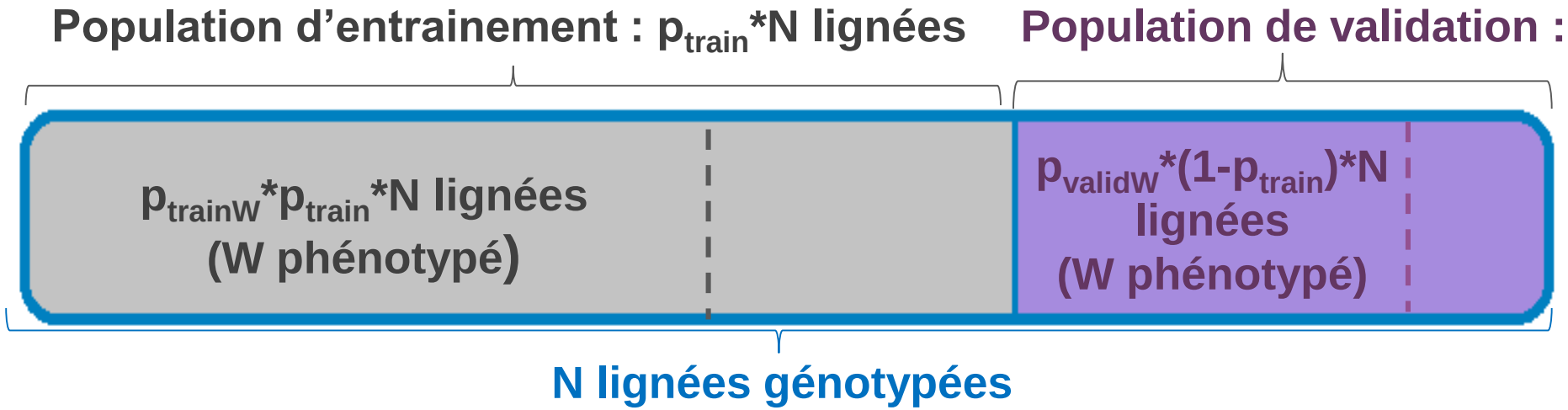
Comparaison des coûts des différents modèles



Avec :

- p_{train} **la proportion de lignées dans la population d'entrainement**
- p_{trainW} **la proportion de lignées de la pop. d'entrainement phénotypées pour W**
- p_{validW} **la proportion de lignées dans la pop. de validation phénotypées pour W**

Comparaison des coûts des différents modèles



$$Ct = (Cg*N) + (Cp_{BMS}*p_{train}*N) + (Cp_{w}*p_{trainW}*p_{train}*N) + (Cp_{w}*p_{validW}*(1-p_{train})*N)$$

Ct : coût total

Cg : coût de génotypage pour une lignée = **35€**

Cp_{BMS} : coût de phénotypage de BMS pour une lignée = **150€**

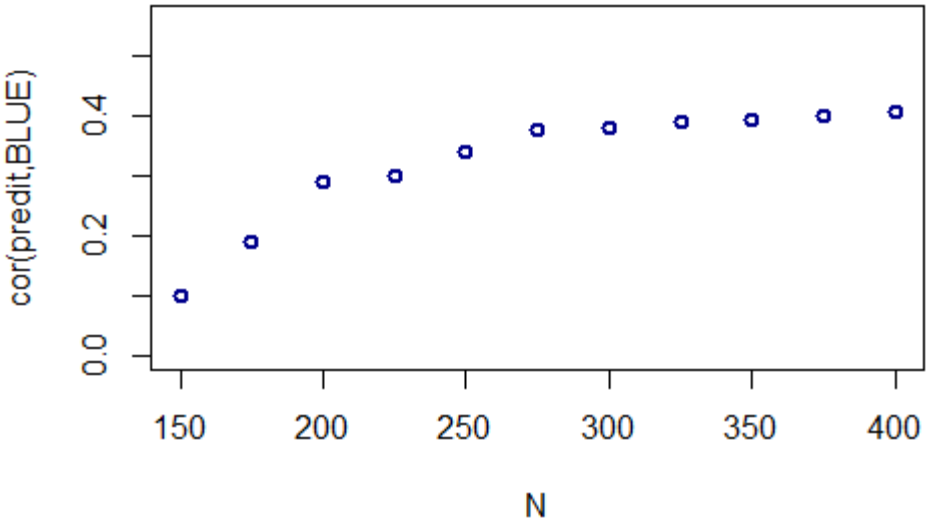
Cp_{W} : coût de phénotypage de W pour une lignée = **20€**

Mono-caractère vs. multi-caractère

1ère situation : mono-caractère

$p_{trainW}=0$
$p_{validW}=0$

Ct = (35*N) + (150*0.8*N)
= 155 * N

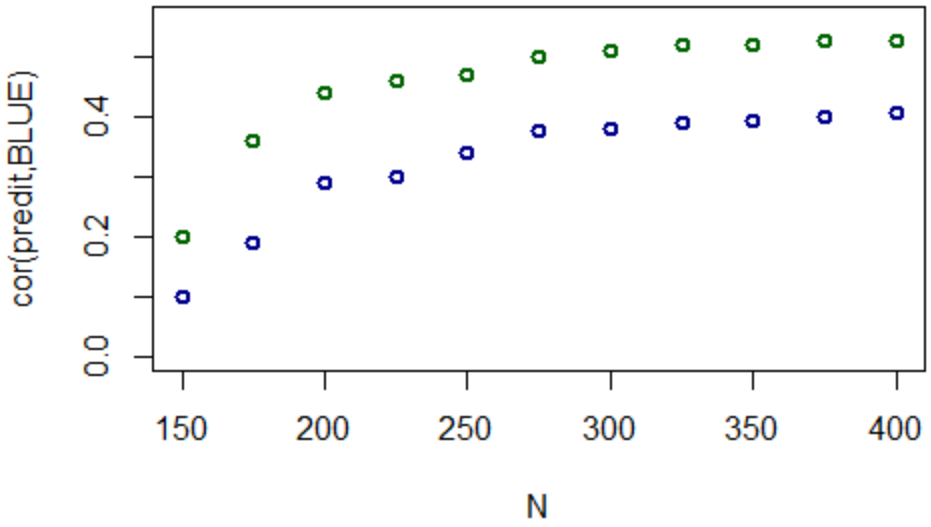


Mono-caractère vs. multi-caractère

2ème situation : multi-caractère « trait assisted » ,
W phénotypé pour tous les individus

p_{trainW}=1
p_{validW}=1

Ct = (35*N) + (150*0.8*N) + (20*N)
= 175 * N



Mono-caractère vs. multi-caractère

1ère situation : mono-caractère

B = 155 * N

➔ cor(prédit,BLUE) = 0.37 quand (N=300, Ntrain=240)

➔ B = 46 000€

2ème situation : multi-caractère, W phénotypé pour tous les candidats

B = 175 * N

➔ Pour un budget de 43 750€ (N=250, Ntrain=200), cor(prédit,BLUE) = 0.48

➔ cor(prédit,BLUE) = 0.36 pour un budget de 30 625€ (N=175, Ntrain=140)

Mono-caractère vs. multi-caractère

1ère situation : mono-caractère
B = 155 * N

➔cor(prédit,BLUE) = 0.37 quand N=300
➔B = 46 000€

2ème situation : multi-caractère, W phénotypé pour tous les individus
B = 175 * N

➔Pour un budget de 43 750€, cor(prédit,BLUE) = 0.48
➔cor(prédit,BLUE) = 0.36 pour un budget de 30 625€

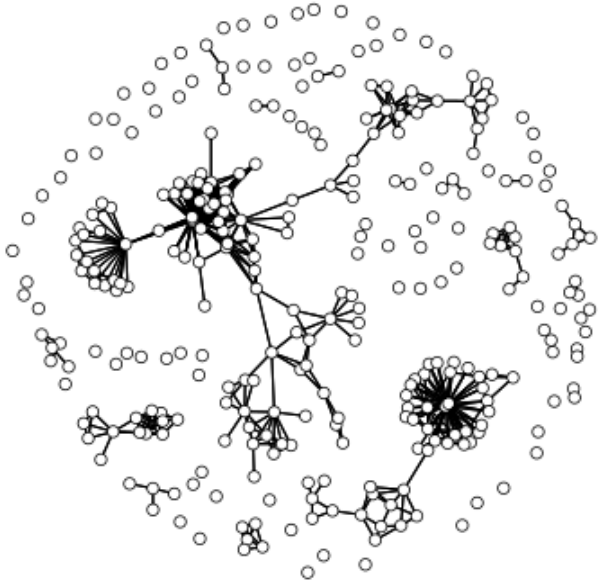
Les économies réalisées peuvent permettre d'augmenter le nombre de candidats testés

Optimisation du phénotypage de BMS et W dans la population candidate

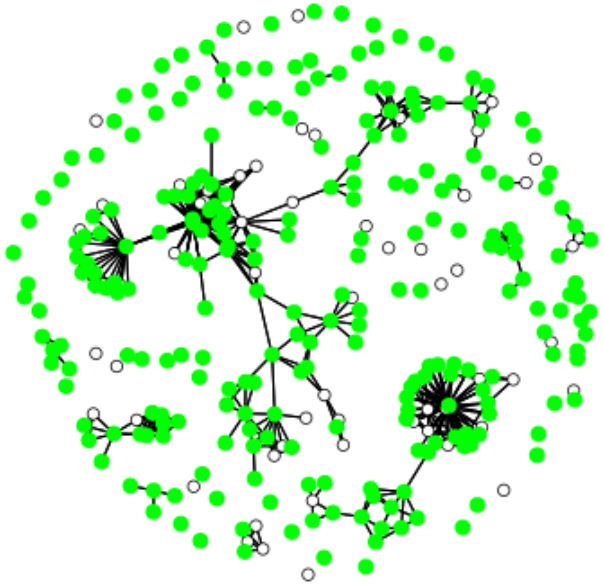
- Maximisation de la diversité allélique?
- PEV: minimisation de la variance d'erreur de prédiction?
- Cdmean: compromis minimisation de la variance d'erreur de prédiction, maximisation de la variance génétique (contrastes entre individus sélectionnés) : minimisation de l'apparentement (Laloe et al 1996, Rincent et al., 2012)
- dérivation d'un CDmean multi-caractère (Laloe et Rincent)

- Optimisation du choix des individus de la population candidate à phénotyper pour W, pour BMS ou pour les 2 caractères

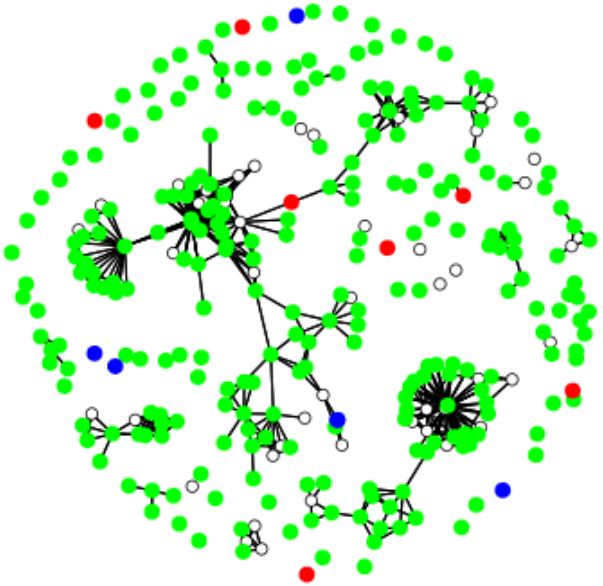
Choix des individus à phénotyper



Tous les individus
(populations d'entrainement et de validation)



Vert = population d'entrainement



Population de validation :
Bleu = individus phénotypés pour W
Rouge = individus phénotypés pour Ntot
Blanc = individus non phénotypés

Conclusions / perspectives

- L'utilisation de modèles de prédiction génomique « trait-assisted » peut
  - améliorer la précision de prédiction de la qualité boulangère
  - diminuer le coût et augmenter le nombre de candidats testés en optimisant le set d'individus phénotypés

- Simulation de schémas de sélection (plans de croisements) pour améliorer simultanément (meilleur compromis en gain génétique) deux caractères négativement corrélés (rendement et qualité)

Merci de votre attention

INRA
SCIENCE & IMPACT