Evaluation multi-environnementale de blé tendre « exotique »

Prédictions génomiques GxE et génétique d'association

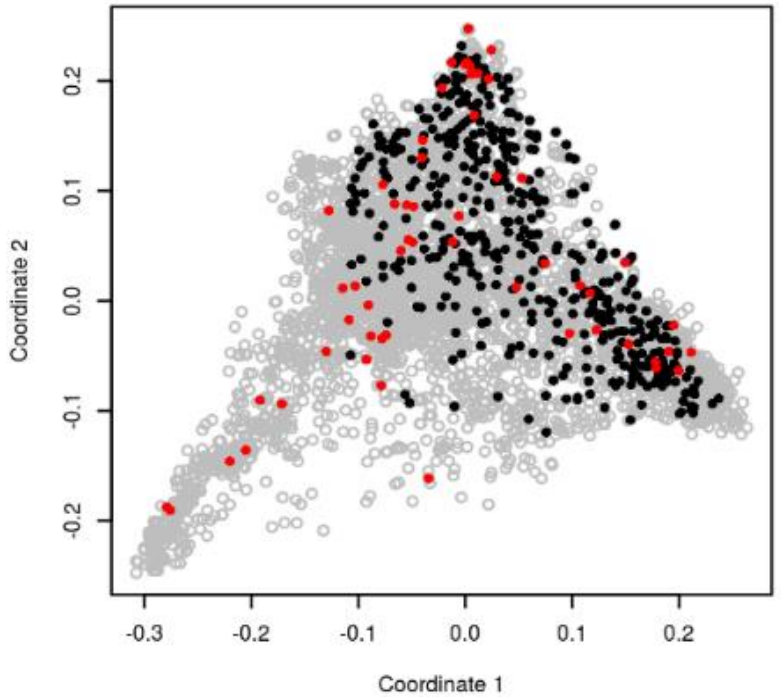
Justin Blancon / Sophie Bouchet

Prédictions génomiques GxE

Blés tendres « exotiques » sous contraintes abiotiques

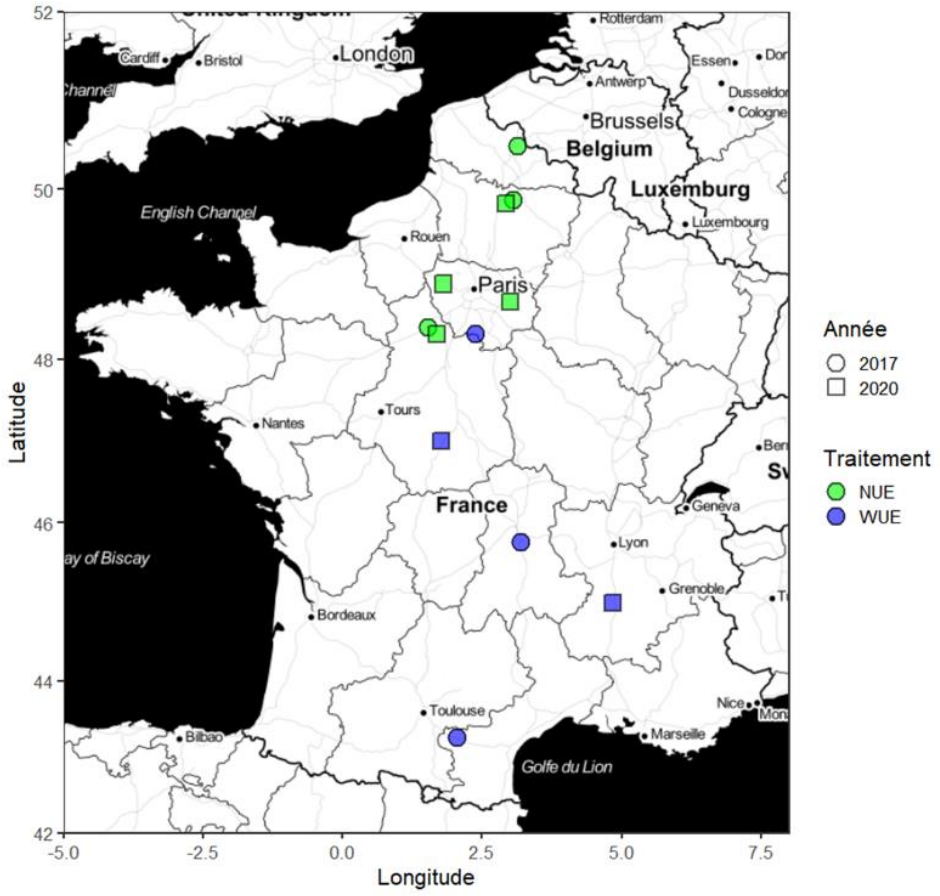
- BWP3ext : Panel de diversité mondiale – 485 individus
- Caractérisation phénotypique du panel dans 12 essais en 2019-2020:
  - N+ et N-
  - Irrigué et pluvial
- Design p-rep avec témoins répétés
- Données historiques du PIA BreedWheat disponibles pour 12 autres essais (N et H_2O) en 2016-2017

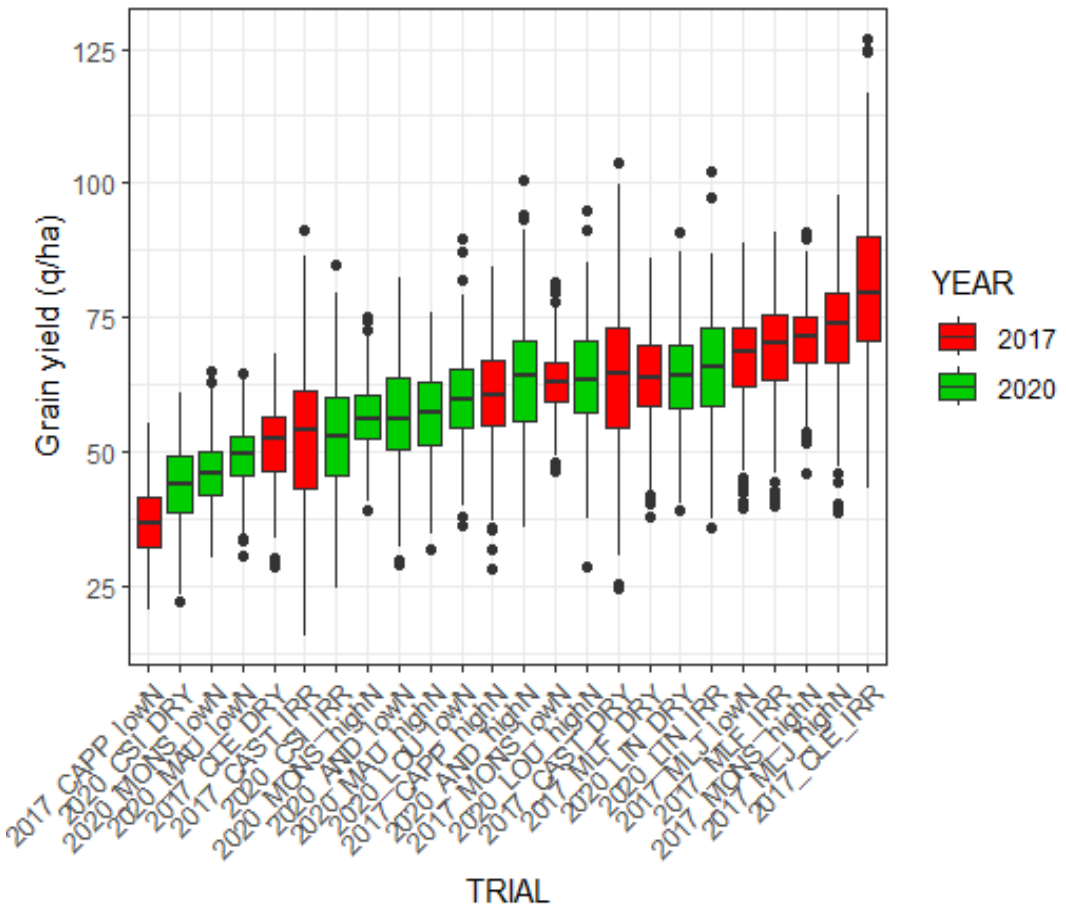
PCoA de la diversité mondiale du blé tendre



En gris: 4500 individus représentatif de la diversité mondiale. En noir et rouge: le panel BWP3ext.

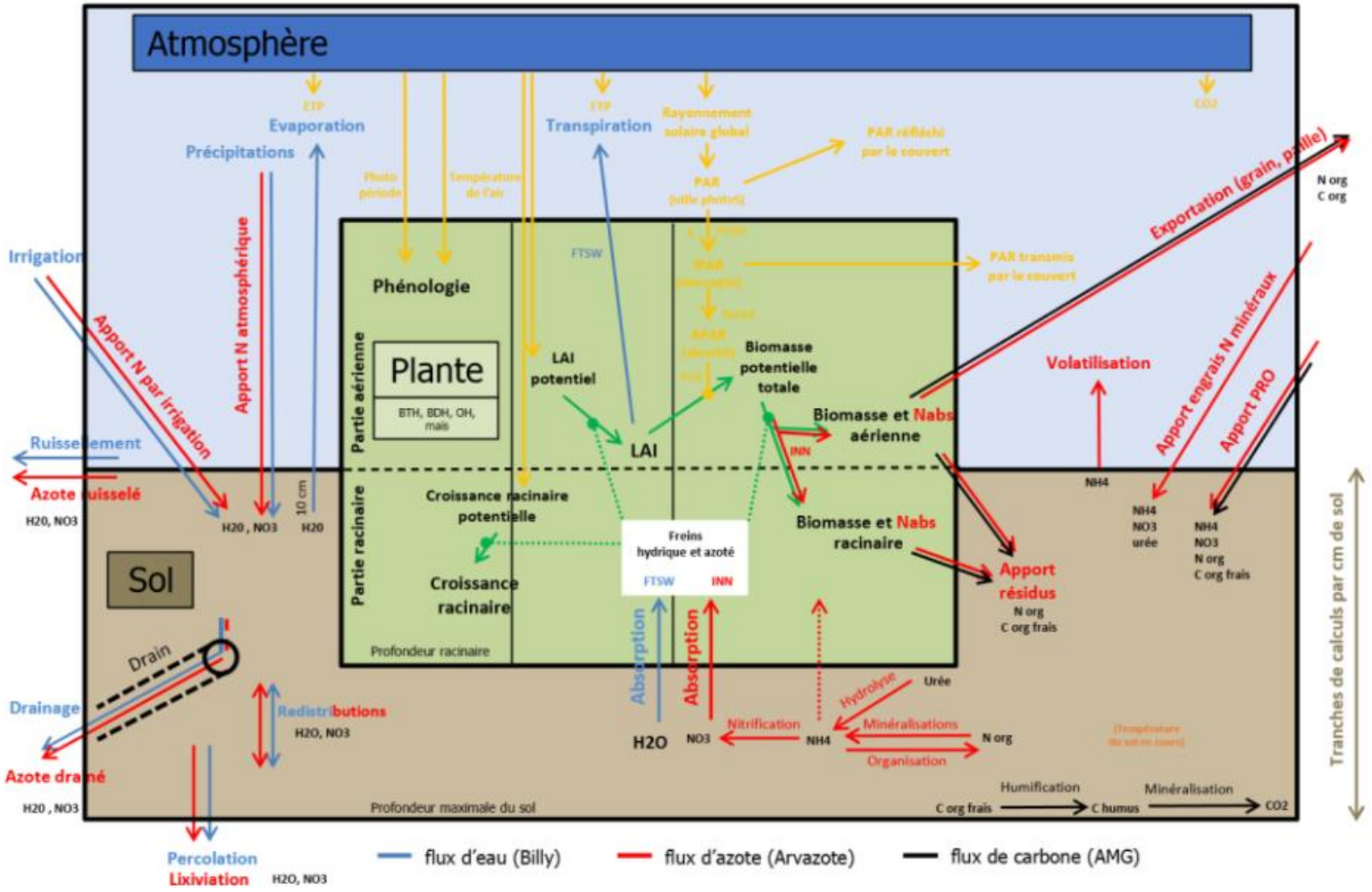
Présentation du réseau d'essai





Analyse des conditions environnementales

Le modèle écophysiologique CHN



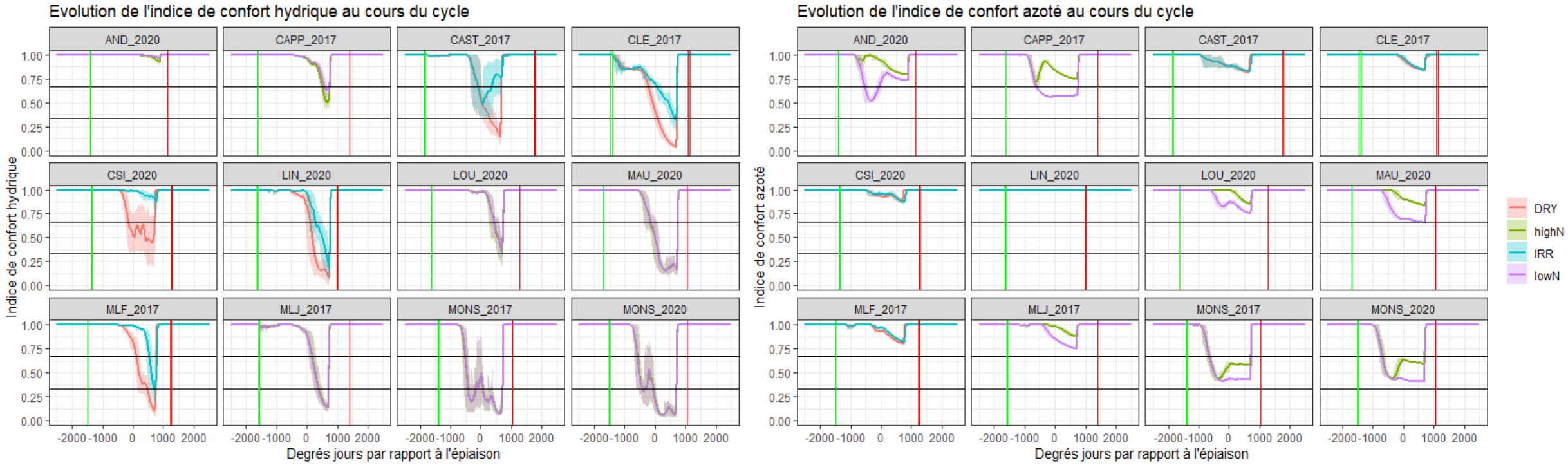
Simulations CHN

192 indices environnementaux (facteur environnemental x stade physiologique)

- Indices climatiques
- Indices intégrés

Deux indices intégrés journaliers par variété dans chaque essai

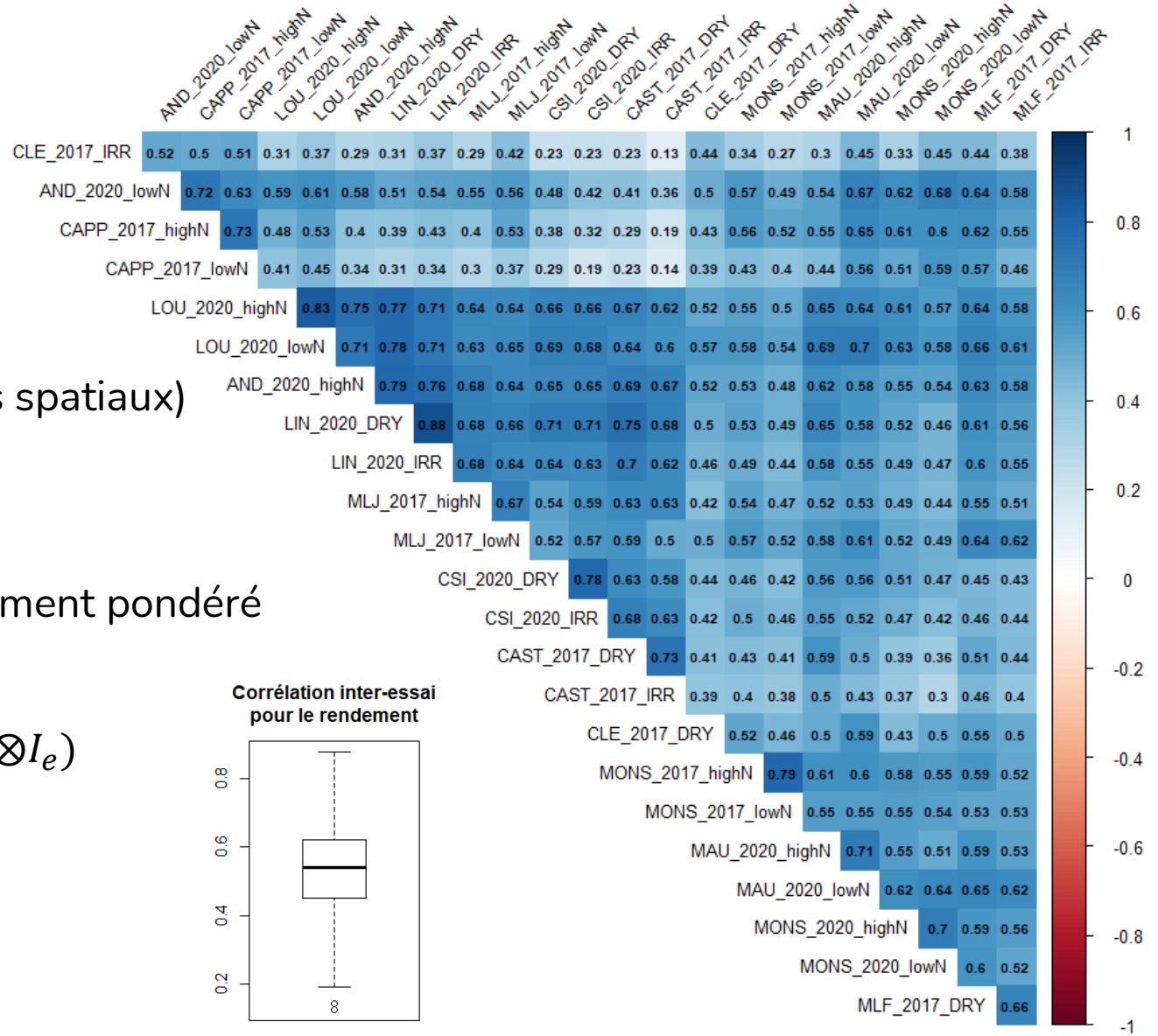
- Indice de confort hydrique
- Indice de confort azoté



Analyse des interactions GxE

Corrélations entre environnements

- Pas de corrélation négative
- r^2 autour de 0,25 en moyenne (de 0,02 à 0,77)



Analyse de variance en deux étapes

- Etape 1 : Ajustement d'un modèle par essai (effets spatiaux)

$$Y = X\beta + Zu + e \longrightarrow \text{BLUE et sem}$$

- Etape 2 : Ajustement d'un modèle multi environnement pondéré

$$\widehat{Y_{ij}} = \mu + E_j + G_i + GxE_{ij} + \varepsilon_{ij}$$

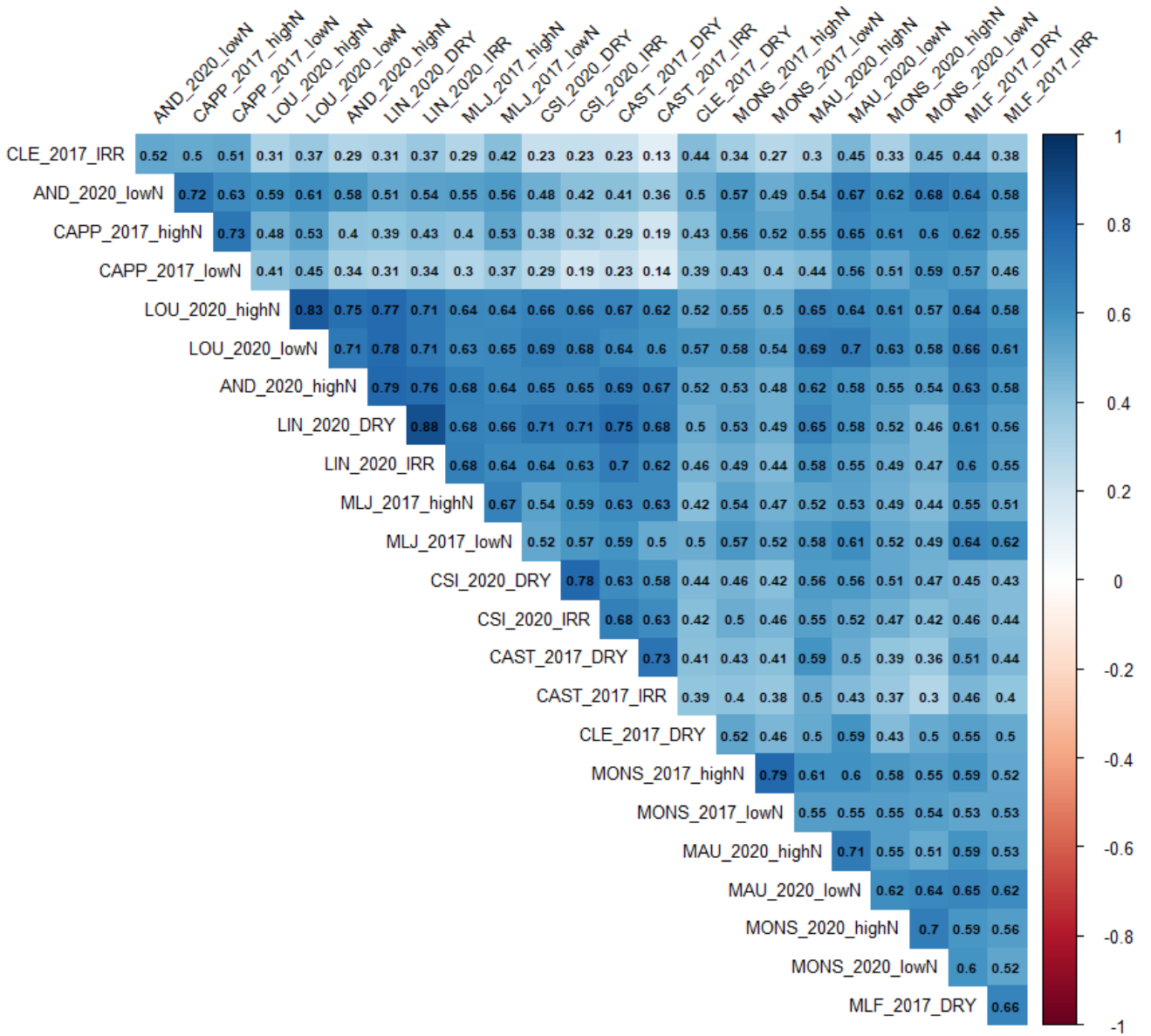
$$E_j \sim N(0, \sigma_e I_e),\ G_i \sim N(0, \sigma_g K),\ GxE_{ij} \sim N(0, \sigma_{ge} K \otimes I_e)$$

$$\varepsilon_{ij} \sim N(0, V) \quad \text{avec } V = \bigoplus_{j=1}^{e} V_j \text{ et } V_j = diag(sem_{ij}^2)$$

Analyse des interactions GxE

Corrélations entre environnements

- Pas de corrélation négative
- r^2 autour de 0,25 en moyenne (de 0,02 à 0,77)



Analyse de variance en deux étapes

Estimation des variances G, E et GxE

	Variance	Erreur standard
Génotype	136.52	10.22068
Environnement	321.3613	140.98888
GxE	300.9044	4.73078

AMMI – Additive Main effects and Multiplicative Interaction

$$Y_{ij} = \underbrace{\mu + g_i + e_j}_{\text{effets principaux}} + \underbrace{\sum_{k=1}^{K} \lambda_k u_{ik} v_{jk}}_{\text{interactions}} + \varepsilon_{ij}$$

$\rightarrow$ $U\Lambda V^t$ avec

- U la matrice $(g \text{ x } K)$ des scores génotypiques
- Λ la matrice diagonale $(\text{K x } K)$ des valeurs singulières
- V la matrice $(e \text{ x } K)$ des scores environnementaux

$U\Lambda V^t = \Phi_{AMMI}$

- Distance euclidienne sur les lignes = D_g
- Distance euclidienne sur les colonnes = D_e

Matrice de covariance
de l'interactivité des génotypes

$$K_{AMMI} = 1_g - \frac{D_g}{\max(D_g)}$$

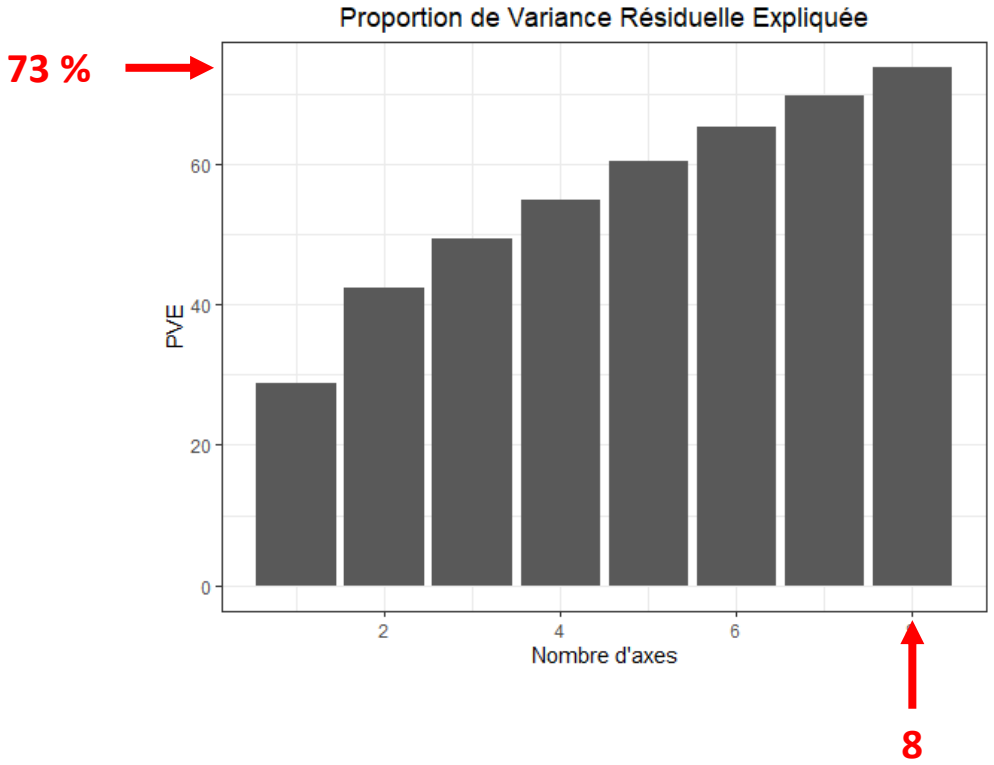
Matrice de covariance
de l'interactivité des environnements

$$W_{AMMI} = 1_e - \frac{D_e}{\max(D_e)}$$

AMMI – Additive Main effects and Multiplicative Interaction

$$Y_{ij} = \mu + g_i + e_j + \sum_{k=1}^{K} \lambda_k u_{ik} v_{jk} + \varepsilon_{ij}$$

?



AMMI – Additive Main effects and Multiplicative Interaction

$$Y_{ij} = \underbrace{\mu + g_i + e_j}_{\text{effets principaux}} + \underbrace{\sum_{k=1}^{K} \lambda_k u_{ik} v_{jk}}_{\text{interactions}} + \varepsilon_{ij}$$

$\rightarrow U\Lambda V^t$ avec

- U la matrice $(g \text{ x } K)$ des scores génotypiques
- Λ la matrice diagonale $(\text{K x } K)$ des valeurs singulières
- V la matrice $(e \text{ x } K)$ des scores environnementaux

$U\Lambda V^t = \Phi_{AMMI}$

- Distance euclidienne sur les lignes = D_g
- Distance euclidienne sur les colonnes = D_e

Matrice de covariance de l'interactivité des génotypes

$$K_{AMMI} = 1_g - \frac{D_g}{\max(D_g)}$$

Matrice de covariance de l'interactivité des environnements

$$W_{AMMI} = 1_e - \frac{D_e}{\max(D_e)}$$

W_{AMMI} : Matrice de covariance de l'interactivité des environnements

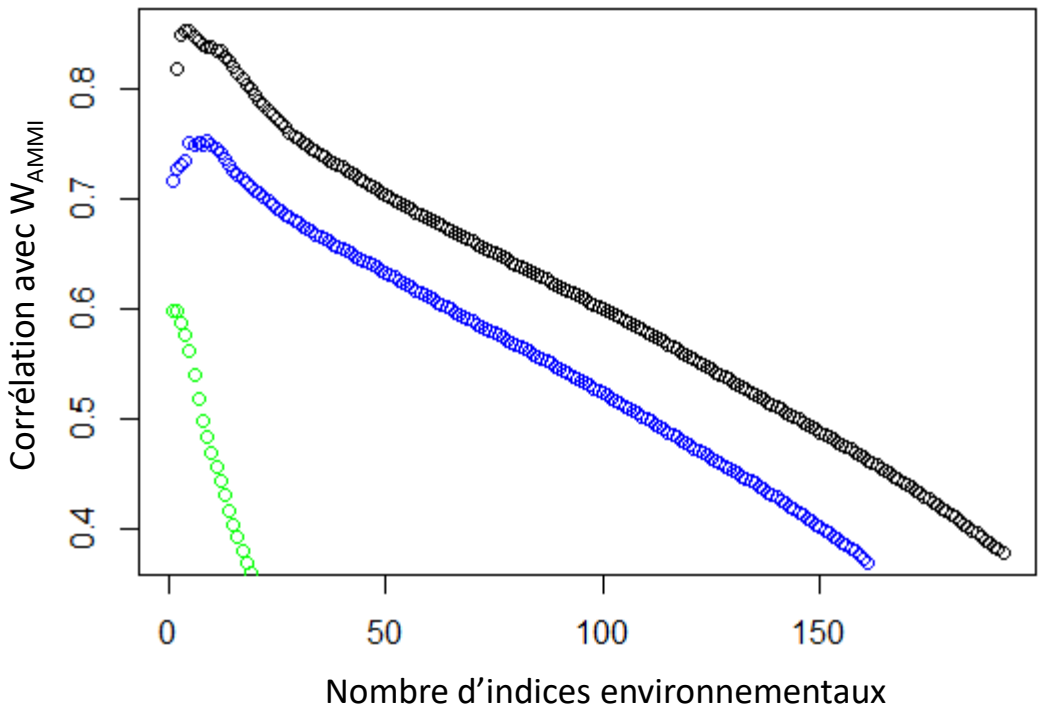
Estimation d'une matrice de covariance environnementale $W_{selpred}$ à partir des indices CHN

- Calcul d'une matrice de distance D_ω similaire à D_e à partir de la matrice des indices environnementaux
- Prise en compte séquentielle des indices pour le calcul de D_ω basé sur la corrélation entre $W_{selpred}$ et W_{AMMI}

Un gain du sous-échantillonnage

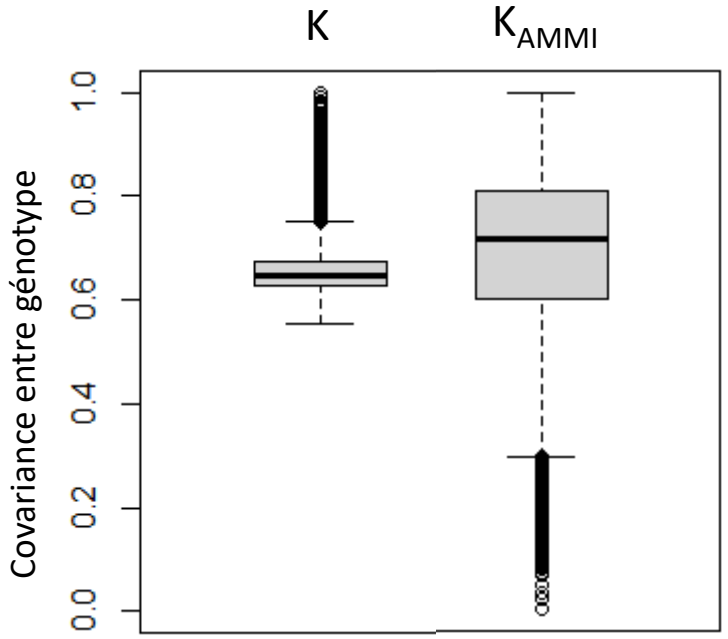
- $W_{selpred}$ est davantage corrélée à W_{AMMI} que W_{all}
- Les indices intégrés apportent une information complémentaire à celle des indices climatiques
- Six indices retenus

Evolution de la corrélation entre $W_{selpred}$ et W_{AMMI}

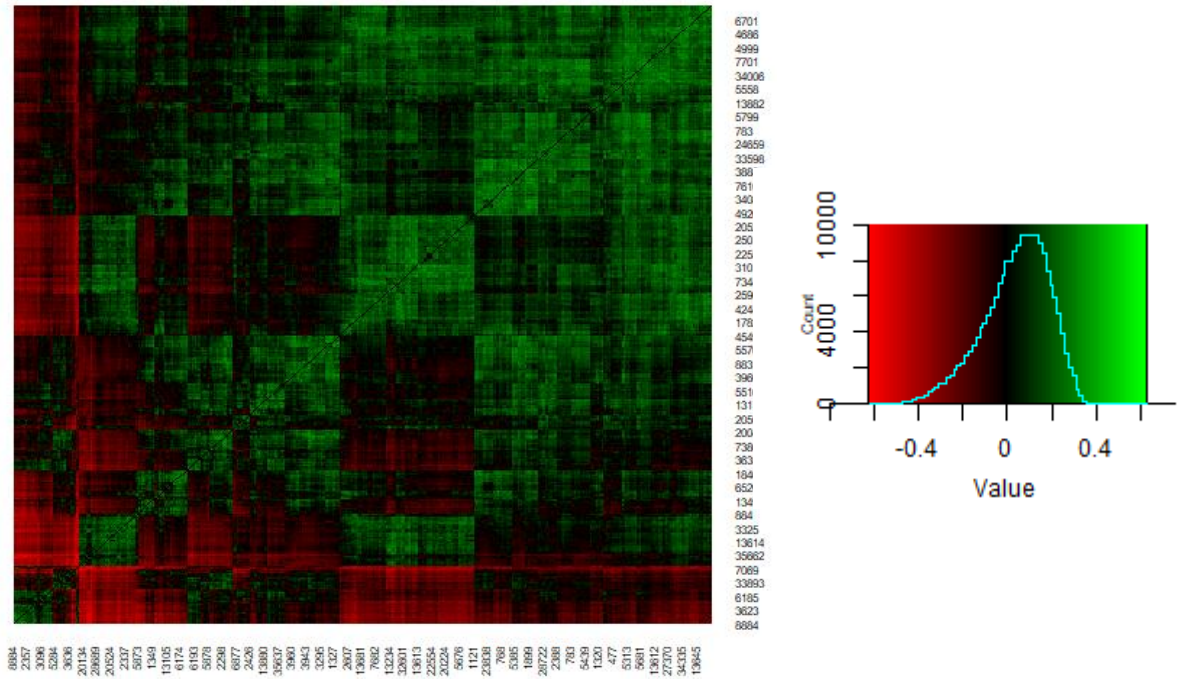


K_{AMMI} : Matrice de covariance de l'interactivité des génotypes

K ou K_{AMMI}, quelle différence ?



Différence de covariance entre K et K_{AMMI}



La covariance génomique (K) ne retranscrit pas
la covariance de l'intéractivité des génotypes

Prédictions génomiques GxE – Scénarios et modèles

L'information contenue dans W_{AMMI} et K_{AMMI} peut elle nous permettre de prédire les interactions GxE ?

- Comparaison de **8 modèles** dans **4 scénarios** : **oGoE** (sparse), **nGoE** (new G), **oGnE** (new E), **nGnE** (new GE)

	$Y_{ij} = \mu + e_j + G_i + (GxE_{ij}) + \varepsilon_{ij}$
Modèle E+G [1]	-
Modèle E+G+GxE [2]	$GxE_{ij} \sim N(0, \sigma^2_{ge} \cdot K \otimes I_e)$
Modèle E+G+GxW$_{all}$ [3]	$GxE_{ij} \sim N(0, \sigma^2_{ge} \cdot K \otimes W_{all})$
Modèle E+G+GxW$_{selpred}$ [4]	$GxE_{ij} \sim N(0, \sigma^2_{ge} \cdot K \otimes W_{selpred})$
Modèle E+G+G$_{AMMI}$xW$_{all}$ [5]	$GxE_{ij} \sim N(0, \sigma^2_{ge} \cdot K_{AMMI} \otimes W_{all})$
Modèle E+G+GxW$_{AMMI}$ [6]	$GxE_{ij} \sim N(0, \sigma^2_{ge} \cdot K \otimes W_{AMMI})$
Modèle E+G+(GxW)$_{AMMI}$ [7]	$GxE_{ij} \sim N(0, \sigma^2_{ge} \cdot K_{AMMI} \otimes W_{AMMI})$
Modèle E+G+ (GxW)$_{AMMIpred}$ [8]	$GxE_{ij} \sim N(0, \sigma^2_{ge} \cdot K_{AMMIpred} \otimes W_{AMMIpred})$

Prédiction des scores génétiques par GBLUP pour **nGoE** et **nGnE**

Prédiction des scores environnementaux par PLS à partir des indices CHN pour **oGnE** et **nGnE**

$U\Lambda V^t = \Phi_{AMMIpred}$

Distance euclidienne sur les lignes = $D_{g,pred} \longrightarrow K_{AMMIpred}$

Distance euclidienne sur les colonnes = $D_{e,pred} \longrightarrow W_{AMMIpred}$

Prédictions génomiques GxE – Predictive ability

Modèles	Scénarios de Cross-Validation			
	oGoE	nGoE	oGnE	nGnE
E+G [1]	0.671	0.461	0.656	0.439
E+G+GxE [2]	0.726	0.522	-	-
$E+G+GxW_{all}$ [3]	0.753	0.532	0.650	0.435
$E+G+GxW_{selpred}$ [4]	0.750	0.514	0.626	0.396
$E+G+G_{AMMI}xW_{all}$ [5]	0.871	0.740	0.642	0.481
$E+G+GxW_{AMMI}$ [6]	0.762	0.501	0.732	0.487
$E+G+(GxW)_{AMMI}$ [7]	0.871	0.740	0.788	0.662
$E+G+(GxW)_{AMMIpred}$ [8]	0.753	0.530	0.644	0.432

- **Amélioration des prédictions dans tous les scénarios**
- Modèle [2] ≈ Modèle [3] > Modèle [4]
- Modèle [7] permet un gain de r de +0.13 à +0.28
- **Apports respectifs des matrices K_{AMMI} et W_{AMMI} pour les scénarios nG et nE**
- Modèle [8] ≈ Modèle [3] => $K_{AMMIpred}$ et $W_{AMMIpred}$ ne sont pas plus informatives que K et W_{all}

- **Plus difficile de prédire des nG que des nE mais gains importants du modèle [7] pour nGoE et nGnE**
  => Intérêt majeur d'améliorer l'estimation de $K_{AMMIpred}$ (BayesA, BayesB, … ?)

- **Difficulté à prédire $W_{AMMIpred}$**
  - Structuration faible du réseau
  - Qualité de la modélisation écophysiologique (données d'entrée, modèle choisi)
  - Taille et diversité limité du réseau (24 essais - 2 années et 11 lieux - majoritairement au Nord de la France)

Conclusion

- La modélisation des interactions GxE en prédiction génomique par l'approche AMMI est prometteuse
- La précision de prédiction des matrices d'interactivité génétique et environnementale n'est pas suffisante pour permettre un gain de predictive ability
- Des hypothèses moins fortes sur les déterminismes de l'interactivité génétique pourrait permettre d'améliorer sa prédiction
- ↗ de la taille et de la diversité du jeu de calibration environnementale et de la qualité de la modélisation pour améliorer la prédiction de l'interactivité environnementale
- La multiplicité des étapes de prédictions entraine une perte d'information
  => utilisation d'une approche single-step ?

Génétique d'association

Caractères agronomiques

Données phénotypiques et génotypiques

- **Données phénotypiques collectées au cours des 24 essais du projet PIA BreedWheat et du projet FSOV ExIGE**
  - Précocité (HD)
  - Hauteur (PH)
  - Rendement à 15% d'humidité (GY15)
  - Nombre d'épis par m² (SA)
  - Nombre de grains par épi (GPS)
  - Nombre de grains par m² (GPA)
  - Poids de mille grains (TKW15)
  - Surface, longueur et largeur du grain (GA, GL, GW)
  - Teneur en protéine (GPC)
  - Grain Protein Deviation (GPD)

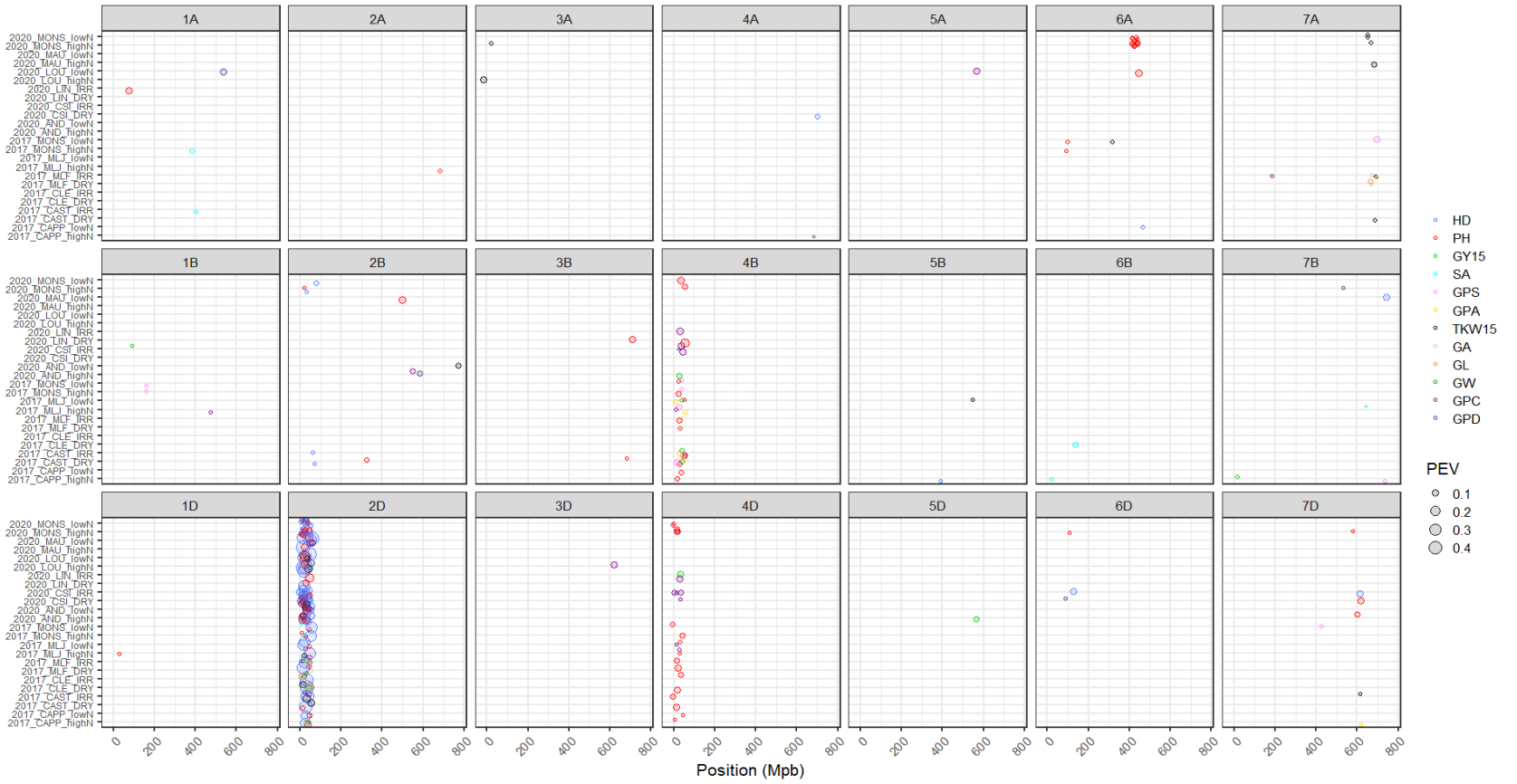
- **Données de génotypage**
  - Puce 420K BreedWheat ≈ 185K SNP polymorphes

Pipeline d'analyse

- Pipeline R **tidygwas**
- Détection de **QTL par environnement** pour chaque caractère
- Modèle K avec **approche LOCO**
- Seuil MAF à 5%
- **Seuil de significativité** : $-\log10(\text{p-value}) < 10^{-5}$
- Définition automatique des bornes des QTL par analyse du déséquilibre de liaison

Résultats d'associations

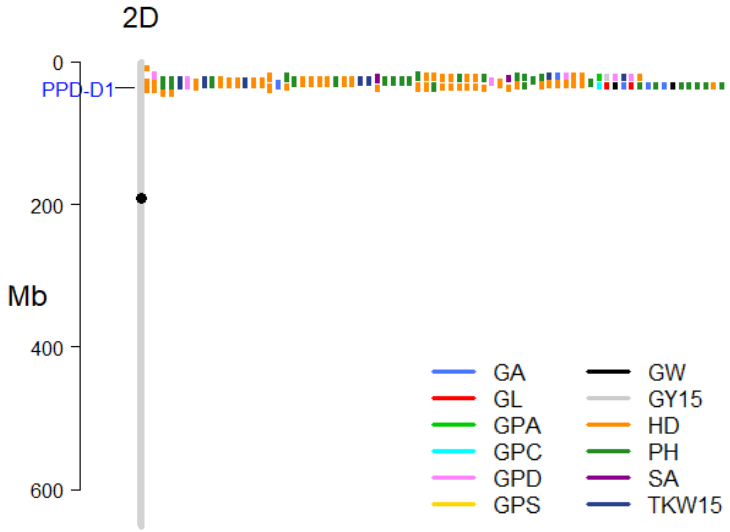
- 272 QTL répartis sur l'ensemble des chromosomes (11 QTL de rendement)

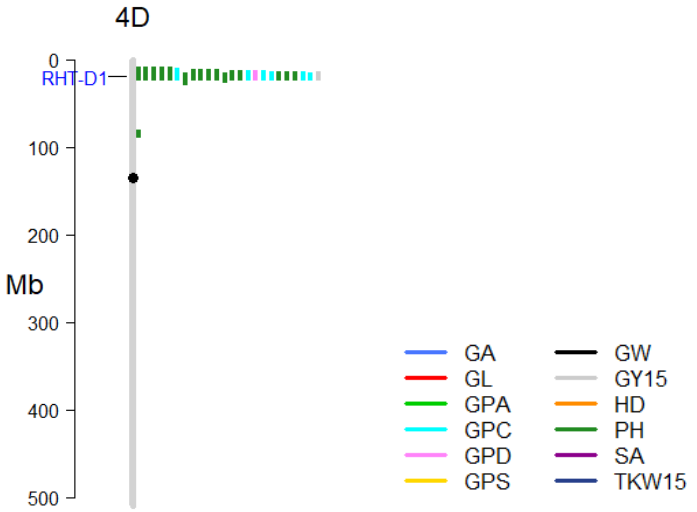


Résultats d'associations

- 272 QTL répartis sur l'ensemble des chromosomes (11 QTL de rendement)

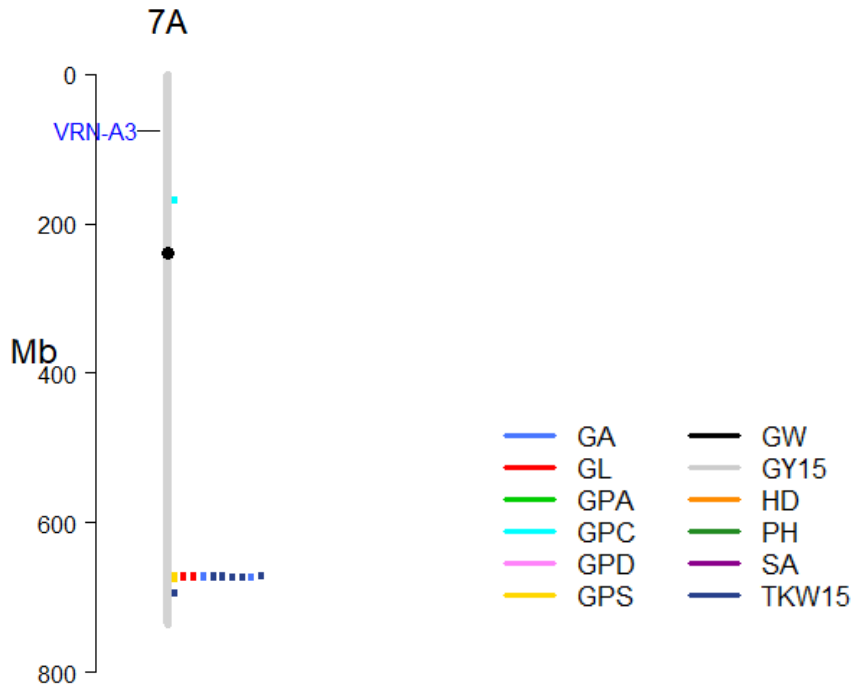
- Importantes colocalisations au niveau de gènes majeurs
  - PPD-D1 : -log10(pvalue)=48, PEV=43%, Effet=-39GDD
  - RHT-D1 : -log10(pvalue)=12, PEV=11%, Effet=-31cm





Résultats d'associations

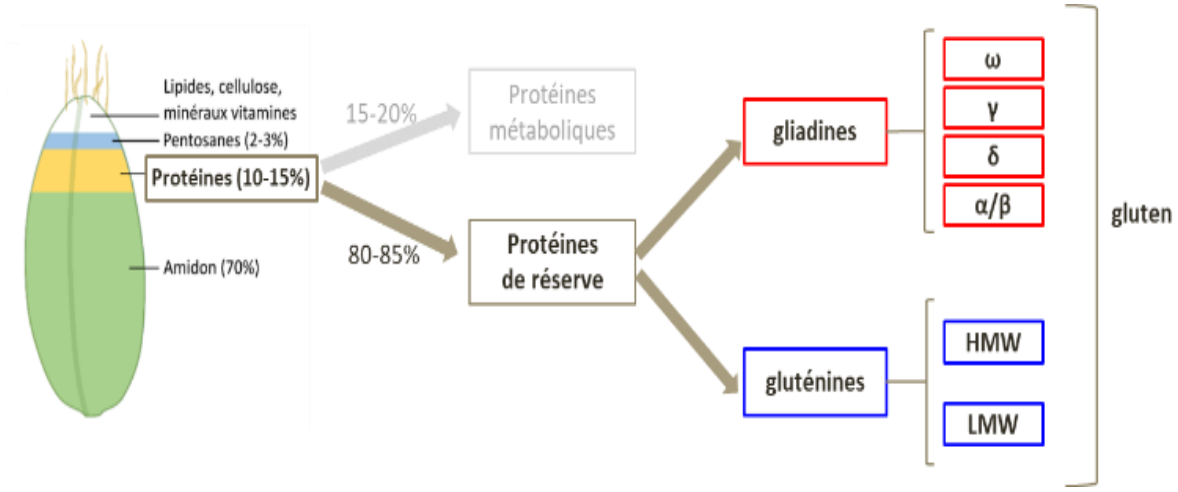
- Colocalisation de QTL de composantes du rendement sur le chromosome 7A
  - Indépendante de la précocité et de la hauteur
  - 5 QTL de TKW15, 2 QTL de longueur de grain, 2 QTL de surface de grain et 1 QTL de nombre de grains par épi
  - En moyenne 6.7% de la variance expliquée
  - L'allèle minoritaire diminue le nombre de grains et augmente leur taille et leur poids



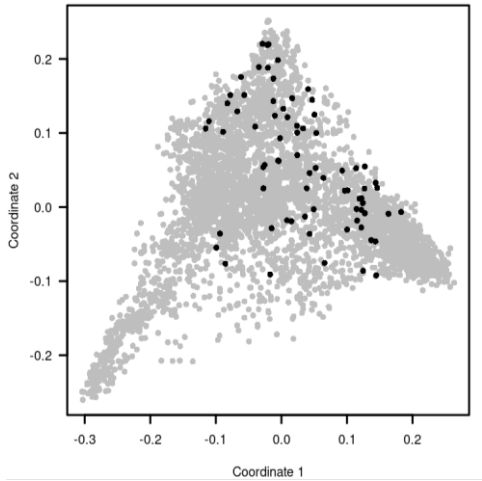
Séquençage gluténines

Catherine Ravel

Matériel



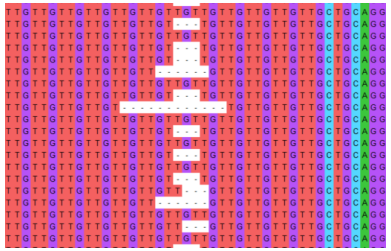
69 séquencées glu



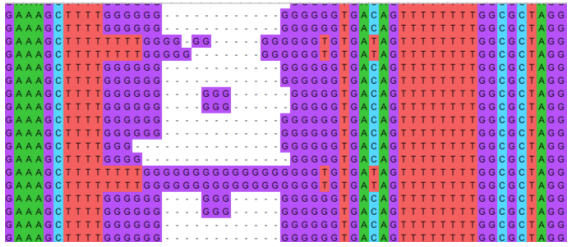
- Annotation des gènes de 44 gliadines et 18 gluténines (6 HMW, 12 LMW) sur Chinese-Spring V2
- Design d'amorces pour PCR long range (4500 < fragment < 8500bp)
- Séquençage avec PACBIO HiFi des 18 gluténines (échec d'amplification des gliadines)
- 2102 SNP, 1309 avec MAF>0.05 et 53 en équilibre de liaison
- Enrichissement gliadines pull catch < 10%, futurs tests par adaptive sampling

Alignements et SNP calling

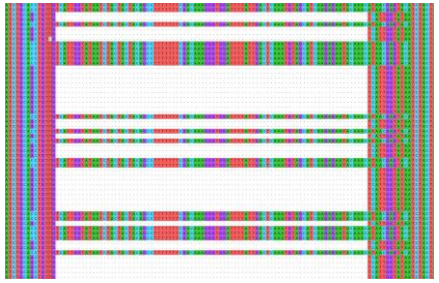
annotation des SNP synonymes et non synonymes, décalage du cadre de lecture avec indels (pas complètement automatisable)



LMW1Ai1

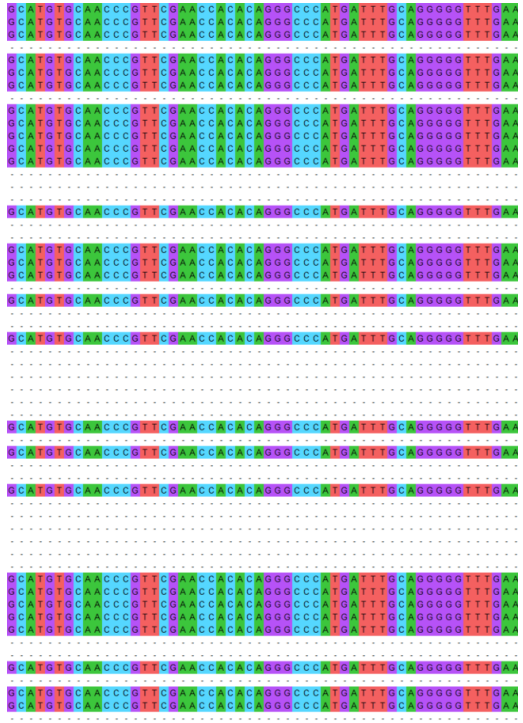


LMW1Ai1



LMW1Am3

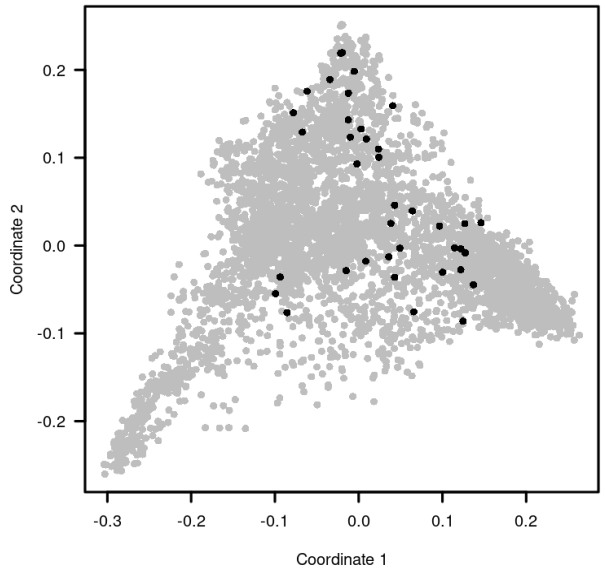
Gène	#SNP	#hap	PAV
HMW1Ax	2	3	0
HMW1Ay	2	4	1
HMW1Bx	9	4	0
HMW1By	5	3	1
HMW1Dx	2	3	0
HMW1Dy	5	5	1
LMW1Ai1	2	4	0
LMW1Am3	7	5	1
LMW1Bm4	7	6	1
LMW1Bm5	2	4	1
LMW1Dm1	1	2	0
LMW1Dm3	4	5	1
LMW1Dm4	3	5	1
LMW1Dm7	1	3	1
LMW1Dm8	2	3	0



HMW1Ay

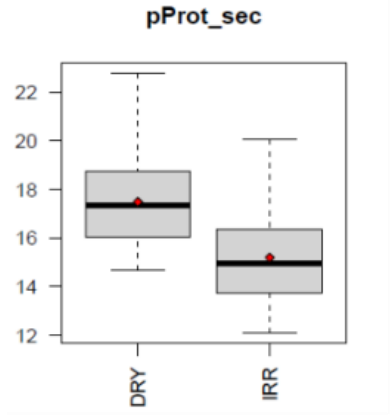
Phénotypage qualité protéique

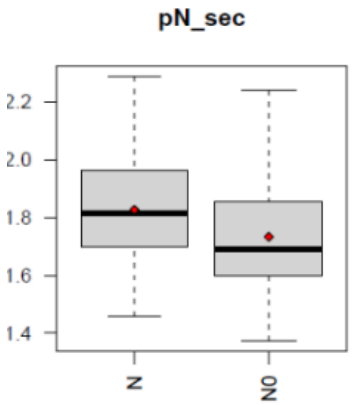
45 alvéo

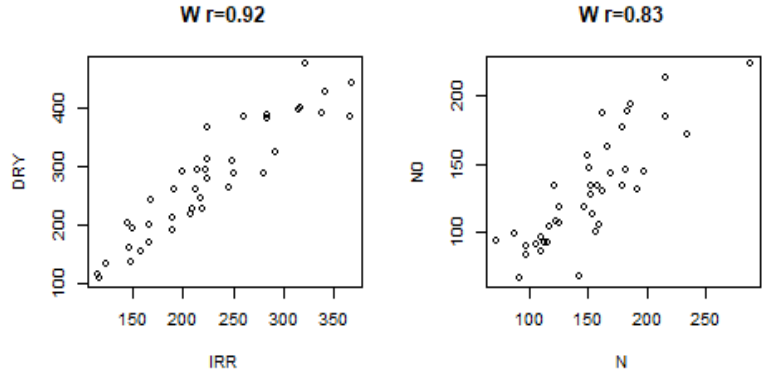


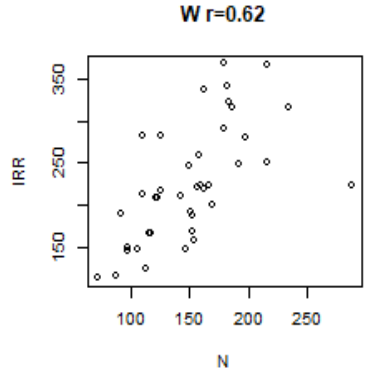
- échantillons de 50g de farine, récoltés dans 4 essais (Mons N-, Mons N+, Alixan irr, Alixan dry)
- mesures HPLC (taux de protéine, taux de gliadine, gluténine dans le grain et la farine, qté HMW, LMW) pour 69 accessions
- mesures alvéographe (W, P, L, P/L, Ie) pour 45 accessions

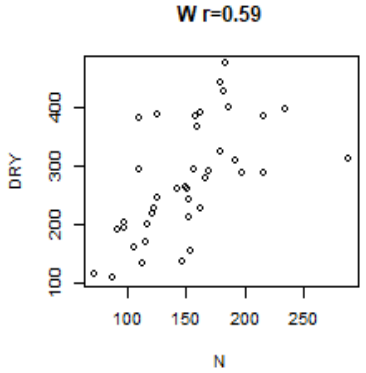
Corrélations entre essais







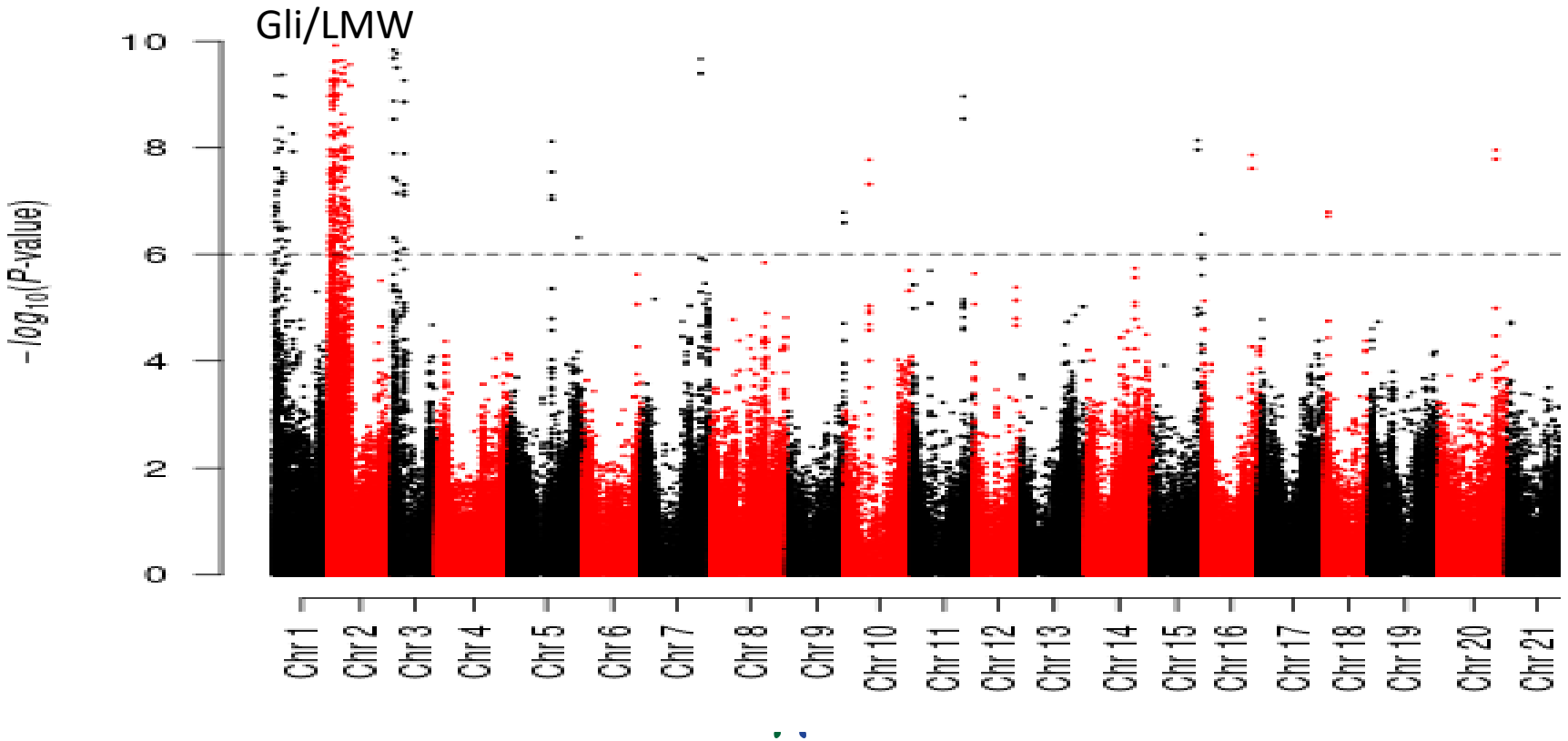




Génétique d'association

Qualité protéique

- 16 QTL principaux, colocalisation de plusieurs caractères, plusieurs environnements
- la majorité sur les chromosomes 1A, 1B et 1D



Génétique d'association

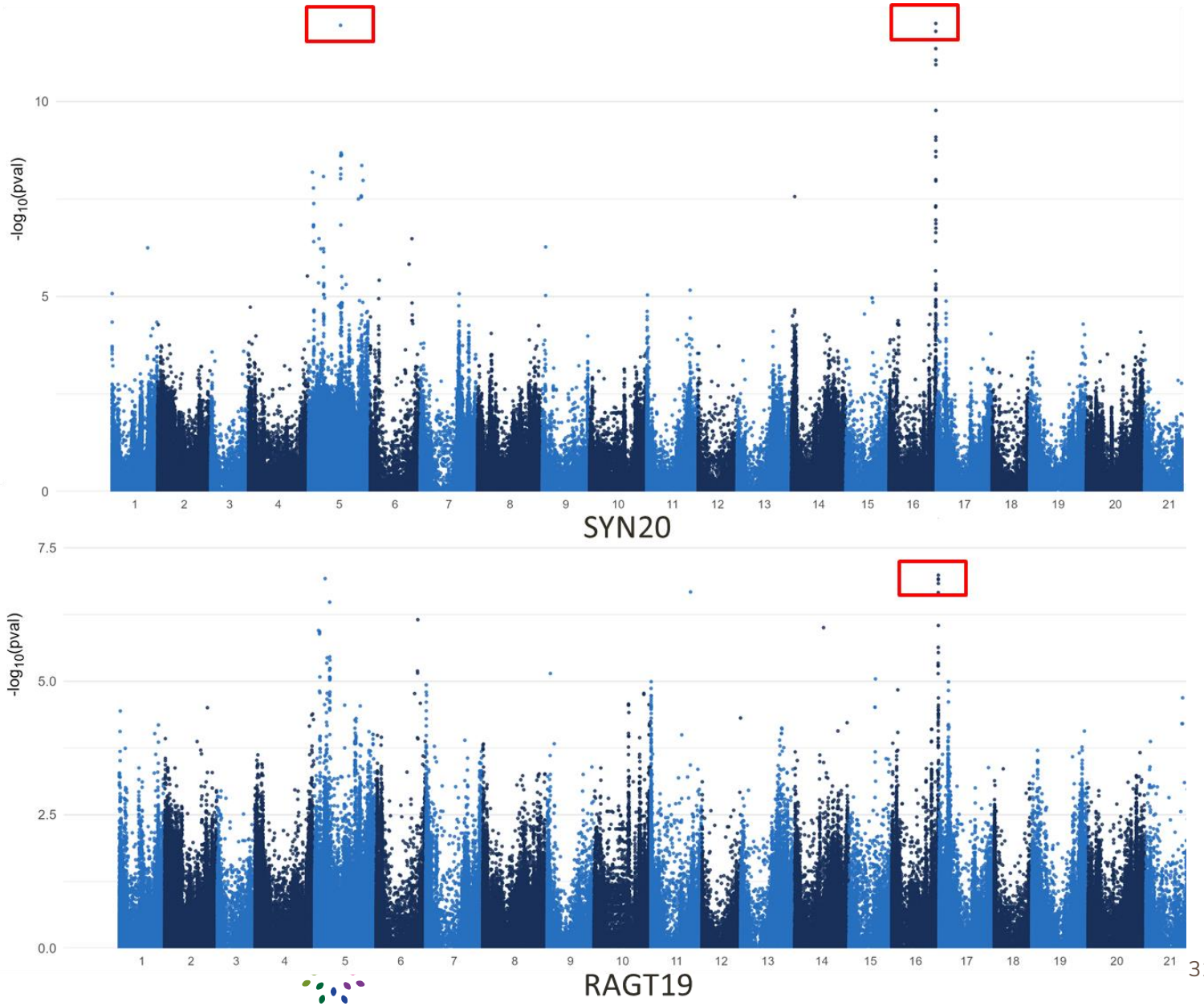
Maladie

Rouille jaune

- 54 QTL maladie
- la plupart spécifiques

- 6A, 613Mb, MAF=0.06
- SYN20,
- INRAE20
- RAGT19
- 5-10% variance expliquée
- l' allèle minoritaire

améliore la résistance



Téléchargement des données

- phénotypes
- génotypes gluténines
- résultats d'études d'association
- https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/VA8HMV

Conclusion / Perspectives

- 12 essais Breedwheat + 12 essais FSOV Ex-IGE + 8 essais en cours avec stress hydrique et azoté
- rendement pour 480 accessions N+, N-, IRR, DRY
  - besoin d'essais très stressés pour prédire le GxE (essais en cours plus au Sud: Montpellier, Saragosse)
  - composantes de rendement
  - variables environnementales

Conclusions / Perspectives

- qualité protéique et séquences protéiques pour 69 accessions
- alvéographe pour 45 accessions

- acquisition de la technologie adaptive sampling: séquencer les gliadines
- typer allèles PAGE sur l'ensemble du panel: set de marqueurs SNPs pour détecter les différents haplotypes (les connus + les nouveaux)
- prédire l'effet des SNP et PAV sur la protéine (codon stop, changement d'acide aminé)
- acquisition d'un alvéographe: mesurer sur l'ensemble du panel
- mesures HPLC sur l'ensemble du panel sous stress hydrique (Espagne) et azoté (Mons)
- estimer les effets des haplotypes qualité

- imputation de tous les marqueurs pour augmenter la puissance de détection des QTLs, optimisation des paramètres en cours
- Proposer des plans de croisements pour le pre-breeding


INRAE


KWS


Fsov
Fsov
Fsov


FLORIMOND
DESPREZ


Lidea


merci


SECOBRA
RECHERCHES


Limagrain
de la terre à la vie


syngenta


R·A·G·T
2n