Évaluation de la précision des modèles de prédiction de la qualité boulangère du blé tendre (Triticum aestivum L.)

Pierre Colin

AGROCAMPUS OUEST

☐ CFR Angers

☑ CFR Rennes

Année universitaire : 2015 -2016

Spécialité : Agronome

Spécialisation :

Science et productions végétales, option Amélioration des Plantes

Mémoire de Fin d'Études

☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ d'un autre établissement (étudiant arrivé en M2)

Evaluation de la précision des modèles de prédiction de la qualité boulangère du blé tendre (Triticum aestivum L.)

Par : Pierre COLIN

Soutenu à Rennes le 13/09/2016

Devant le jury composé de :

Président : Maria Manzanares-Dauleux

Maître de stage : Sophie Bouchet

Enseignant référent : Maria Manzanares-Dauleux

Autres membres du jury (Nom, Qualité)

Rapporteur : Sophie Allais

Examinateur : Anne Laperche

Remerciements :

Je tiens à remercier l'ensemble des personnes qui m'ont permis de mener à bien ce mémoire de fin d'étude. Je souhaite avant tout remercier ma maître de stage Sophie Bouchet pour son suivi tout au long de mon stage et pour son aide précieuse dans la rédaction de ce mémoire. Je remercie également Gilles Charmet et l'ensemble de l'équipe DGS, de m'avoir accueilli et permis de travailler dans de bonnes conditions. Merci également à Bruno Poupard pour son suivi et ses conseils avisés.

Et enfin, merci à Séverine Rougeol et Nathalie Bernard de m'avoir accueilli dans leur bureau et à l'ensemble du personnel, post-doc, doctorants, stagiaires, de m'avoir occasionnellement apporté leur aide et leurs conseils.

Table des matières

LISTE DES ABREVIATIONS / GLOSSAIRE

ACP : Analyse en Composantes Principales

ANMF : Association nationale de la meunerie française

ANOVA : ANalysis Of VAriance

AFNOR : Association Française de Normalisation

BPS : Blé panifiable supérieur

BP : Blé panifiable

BAU : Blé autre usage

BB : Blé biscuitier

BA: Blé améliorant

BLUE: Best Linear Unbiased Estimator

BLUP: Best Linear Unbiased Predictor

BWGS : BreadWheat Genomic Selection (Pipeline)

DL : Déséquilibre de liaison

DHS : Distinction, Homogénéité, Stabilité

cM : CentiMorgan

CNERNA: Centre National d'Etudes et de recommandations sur la Nutrition et l'Alimentation

FSOV : Fond de soutien à obtention variétal

G : Gonflement

GEBV: Genetic Estimated Breeding Value

G-BLUP: Genomic Best Linear Uniased Prediction

He : Hétérozygotie attendue

Ho : Hétérozygotie observée

IBS: Identity By State

INRA-AO : Institut national de recherche agronomique –Agri-obtention

IWGSC: International Wheat Genome Sequencing Consortium

L : Longueur

MAF : Minor allele frequency

MLMM : Multi Locus Mixed Model

NPATE: Note de pâte

NPANI: Note de panification totale

NPAIN: Note de pain

VOL: Volume

VOLG: Volume/masse

P : Pression maximale

Pb : Paires de Base

QTL : Quantitative Trait Loci

RR-BLUP: Ridge Regression Best Linear Uniased Prediction

SEVEN: Structure et Evolution du Génome du Blé

SNP : Single Nucleotid Polymorphism

TBV: True Breeding Values

VATE : Valeur agronomique technologique et environnementale

W : Force boulangère

RR-BLUP: Ridge Regression Best Linear Unbiased Prediction

SEVEN: Structure et Evolution du Génome du Blé

SNP : Single Nucleotid Polymorphism

TBV: True Breeding Values

VATE : Valeur agronomique technologique et environnementale

W : Force boulangère

LISTE DES TABLES SUPPLEMENTAIRES

LISTE DES FIGURES SUPPLEMENTAIRES

LISTE DES ILLUSTRATIONS

LISTE DES TABLES

Table 1 : Production (en Mt) des cinq premières production végétales au monde en 2013 (FAOSTAT)

Produit	**Production (Mt)**
Canne à Sucre	1898,2
Maïs	1017,53
Riz	738,06
Blé	711,14
Pomme-de-terre	374,46

T**able 2** : Production (Mt) nationale en blé des six premiers producteurs mondiaux en 2013 (FAOSTAT)

Pays	**Production (Mt)**
Chine	121,92
Inde	93,5
USA	57,96
Russie	52,09
France	38,61
Canada	37,52

Introduction

Le blé est la quatrième production végétale et troisième céréale mondiale (Table 1). La France a produit en 2015 40.8 millions de tonnes de blé tendre (France Agrimer), ce qui en fait le cinquième producteur mondial (Table 2). Au total, 58 % de la production annuelle est destinée à l'alimentation humaine, majoritairement sous forme panifiée. La qualité boulangère est donc un critère important à prendre en compte dans le processus de création variétale afin de répondre aux attentes du marché. Les composantes de la qualité boulangère sont des caractères quantitatifs, contrôlé par plusieurs dizaines de régions sur le génome et potentiellement avec interactions, dont l'évaluation est chère et nécessite une quantité de grain importante. Pour cette raison, son évaluation n'intervient que tardivement dans le processus de sélection. La qualité boulangère étant déterminante pour l'inscription en catégorie (Blé Panifiable Supérieur, Blé Panifiable, Blé Autres Usages) (Table S1), le matériel de départ des programmes de sélection est, en majorité, constitué de lignées BPS et BP, d'où une diversité génétique limitée. L'objectif de la filière est de pouvoir prédire dans les premiers cycles de sélection la qualité boulangère des lignées afin de limiter le nombre de lignées et les coûts de phénotypage des générations suivantes. L'objectif du stage est de choisir le modèle de sélection génomique qui optimise la précision des prédictions de la qualité boulangère et de ses composantes.

I. Contexte

A. Contexte économique

Le blé est la quatrième production mondiale, avec une production de 711,14 Mt derrière la canne à sucre, le maïs et le riz en 2013. (FAOSTAT) (Table 1)

Le blé tendre représente 95% des blés cultivés au monde (Shewry, 2009). La France compte 5,16 millions d'hectares de blé tendre en 2013, elle est le cinquième producteur mondial (Table 2) et le premier producteur de l'Union Européenne avec 37,869 million de tonne produite, devant l'Allemagne avec 26,531 millions de tonne. La production Française de 2015 avec 40.8 est en hausse de 6% par rapport à la période 2010-2014, hausse attribués à la fois à l'accroissement de la sole et à celle des rendements stagnants (hausse de 0.9% par rapport à la moyenne 2010-2014) (AGREST). La récolte 2016 ne confirmera pas cette progression avec un recul de près de 30%.

Au total, 58% du blé tendre produit en France est destiné à l'alimentation humaine alors que 34% est destiné à l'alimentation animale, les 8% restant représentent les utilisations industrielles (amidonnerie, production d'éthanol) (www.passiocereales.fr).

B. Contexte scientifique

Ce projet est financé par le Fond de soutien à l'obtention végétale et a pour but l'établissement d'un modèle de sélection génomique pour la qualité boulangère des blés. Ce projet a une durée de 3 ans et a commencé le 1er septembre 2014. A terme, 192 lignées de blé différentes seront implantées sur 4 lieux sur les 3 années du projet (2304 lignées, non

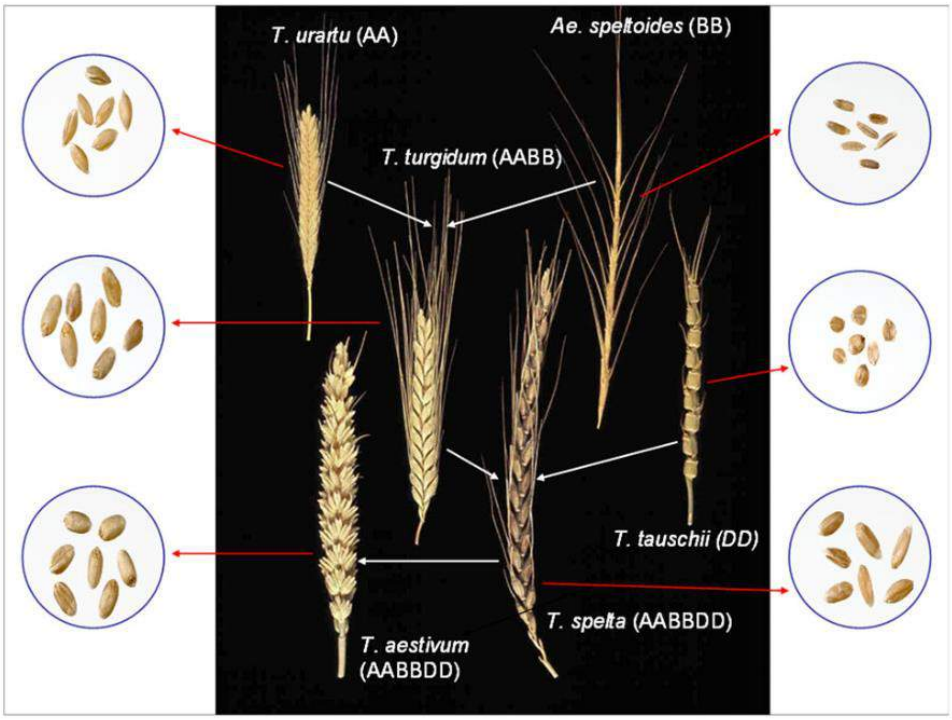


Figure 1 : Evolution et hybridation du blé cultivé (Snape and Pankova (2006), modifié par Shewry (2009))

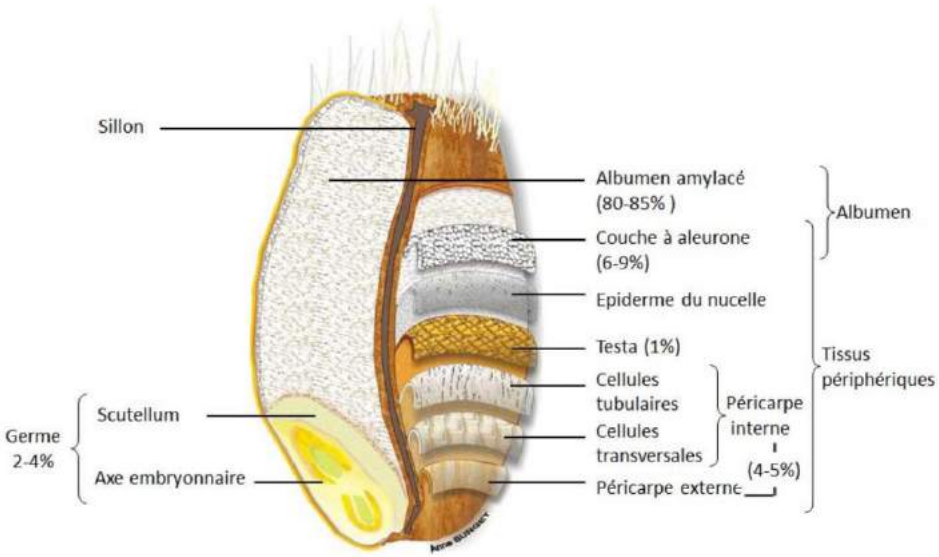


Figure 2 : Anatomie schématique du grain de blé et de ses différents tissus (Surget et Barron, 2005 adapté par Barron *et al.*, 2012)

répétées, analysées pour la qualité au totale) ainsi que 4 témoins (Apache, Arezzo, Symoisson, Solehio) répétés chaque année sur tous les lieux. Le projet est coordonné par LIMAGRAIN EUROPE et est en partenariat avec l'INRA, Arvalis et l'association nationale de la meunerie française (ANMF). En plus du jeu de données, assemblées pour ce projet, nous avons pu utiliser les données issues du projet BREEDWHEAT.

1) Fonds de soutien à l'obtention végétale

Le fonds de soutien à l'obtention végétale est destiné à financer des programmes de recherche allant dans le sens d'une agriculture durable. Cela regroupe une meilleure valorisation de l'azote et de l'eau, l'amélioration des résistances aux maladies, mais également des qualités technologiques et sanitaires. Ce fonds est financé par une cotisation volontaire obligatoire, de 0.7 € par tonne de céréales produites (blé tendre, blé dur, orge, avoine, seigle, triticale, riz, épeautre). Chaque année un appel à projet est lancé, et en 2014 l'un des projets retenus concerne l'établissement d'un modèle de sélection génomique pour la qualité boulangère des blés (fsov.org).

2) Projet BREEDWHEAT

Il s'agit d'un projet d'investissement d'avenir porté par l'UMR GDEC de l'INRA de Clermont-Ferrand et destiné à renforcer la compétitivité de la filière française de sélection de blé. Ce projet est prévu sur 9 ans, dispose d'un budget de 34 millions d'euros et regroupe 26 partenaires publiques comme privés. Les thèmes de recherche de ce programme vont de l'identification de nouvelles sources de résistance, à l'efficacité de l'utilisation de l'azote et de l'eau en passant par les critères de qualité

C. Caractéristiques du blé tendre

1) L'origine génétique du blé tendre

Le blé tendre est une plante allohexaploïde (génome A, B, D) possédant 21 chromosomes (2n=42), il résulte de deux croisements interspécifiques successifs (McFadden and Sears, 1946). Une première hybridation a eu lieu au niveau du croissant fertile il y a environ 500 000 ans entre deux blés diploïdes, T. uratu (2n=14, AA) et Aegilops speltoides (2n=14, BB). Ce croisement donna un blé tétraploïde (T. turgidum, 2n=28, AABB), le blé dur, qui se croisa à son tour avec un blé diploïde sauvage (2n=14, DD) Aegilops tauschii, (Aegilops squarrosa) (Figure 1), il y a environ 10000 ans en Asie centrale continentale. Ces deux événements de dérive suivis de périodes de migration et de différentiation ont eu des impacts différents en terme de diversité génétique au niveau des génomes A, B et D (Choulet 2014).

2) Caractéristiques du grain de blé

a) Caractéristiques histologiques

Le grain de blé est principalement constitué d'amidon (70% de la matière sèche), de protéines (10 à 18% selon les variétés et les milieux), de fibres (7 à 10%) et d'autres éléments mineurs comme les lipides et les minéraux. (Feillet, 2000). L'albumen amylacé qui représente 82 à 85% du grain est la partie la plus valorisée du fait de son aptitude à la panification (Figure 2). Le germe qui représente 3% du grain de blé a une teneur élevée en lipides,

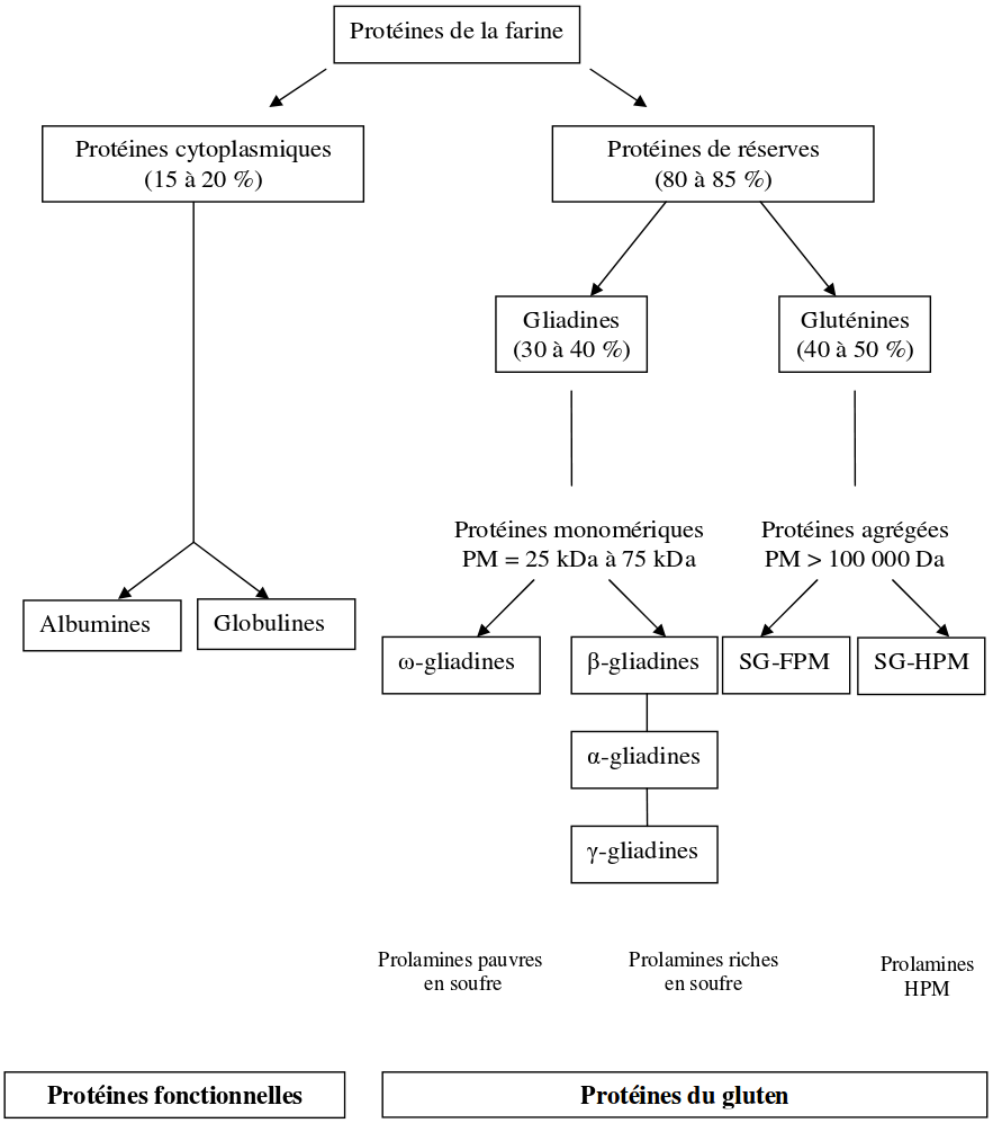


Figure 3 : Composition protéique de la farine de blé (Debiton, 2010 d'après Osborne 1924 et Shewry 1986)

protéines, vitamines et éléments minéraux. Mais cette partie est éliminée des farines par les techniques actuelles de mouture sur cylindre.

Les enveloppes qui représentent 13 à 15% du grain sont riches en cellulose et en protéines. Après mouture, elles forment le « son » (Roussel et Chiron, 2002).

b) Caractéristiques physiques

Les variétés de blé tendre sont classées en différentes catégories : hard, médium hard, soft. Cet indice correspond à la proportion d'amidon endommagé par la mouture. Lorsque l'indice de dureté augmente (blé « hard »), la friabilité diminue et la granulométrie devient plus grossière. Au contraire, pour les «blés « soft », la fragmentation est plus facile et la quantité d'amidon endommagée moindre. La qualité de l'amidon influe sur la rétention d'eau, ce qui a notamment pour conséquence l'augmentation de la viscosité, et la diminution de l'extensibilité.

c) Caractéristiques bio-chimiques

La teneur mais surtout la qualité protéique des farines sont déterminants pour la qualité boulangère. On distingue les protéines métaboliques (albumines et globulines) qui représentent 15 à 20 % des protéines présentes dans la farine de blé, les protéines amphiphiles membranaires qui représentent 5 à 10% des protéines présentes dans la farine de blé connues pour jouer sur la qualité technologique de la pâte et les protéines de réserve qui sont déterminantes pour la qualité boulangère. Parmi elles, les gliadines et les gluténines constituent respectivement 30 à 40 % et 40 à 50% des protéines de la farine. Elles sont regroupées sous le nom de prolamines et sont les principales constituantes du gluten (Figure 3). Les plus importantes pour la qualité boulangère, les gluténines sont codées par des gènes appartenant à une famille multigénique portée par les chromosomes 1A, 1B et 1D Les gliadines se situent sur les chromosomes 6A, 6B et 6D. Deux QTLs ont été trouvés sur le chromosome 5B, (Groos, 2001), et sont impliqués dans le contrôle de la note totale et l'un de ces deux QTL apparaît également significatif pour la note de pâte et de mie. Concernant la teneur en protéine, deux QTLs ont été trouvé sur les chromosomes 1B et 6A et trois QTL associé à la force boulangère W sur les chromosomes 1A, 5D et 3B. (Perretant *et al*., 1999)

La quantité de protéine minimale pour obtenir des résultats de panification satisfaisant est de 10%. Un excès de protéines engendrera un défaut de fabrication du pain. La teneur en protéine ne donne ainsi qu'une indication sur la qualité du blé (Roussel et Chiron, 2002).

De nombreuses études ont montré l'influence de l'environnement, la fertilisation azotée et la densité de semis notamment, sur la quantité et la qualité des protéines de blé mais aussi sur l'aptitude à la panification. (Jia *et al.*, 1996; Johansson *et al.*, 2001, Geleta *et al*.,2002).

d) Classification des variétés selon leur profil qualité

Les variétés candidates à l'inscription au catalogue passent des tests de DHS (Distinction, Homogénéité, Stabilité) et VATE (Valeur agronomique, technologique, environnementale). Les tests VATE panification qui déterminent la classe d'utilisation de la variété : blé améliorant (BA), blé panifiable supérieur (BPS), blé panifiable (BP), blé biscuitier (BB) et blé autres usages (BAU). Au dépôt à l'inscription, l'obtenteur choisit la

Figure 4 : Gonflement de la pâte à l'alvéographe (source : http://www.agroscope.admin.ch)

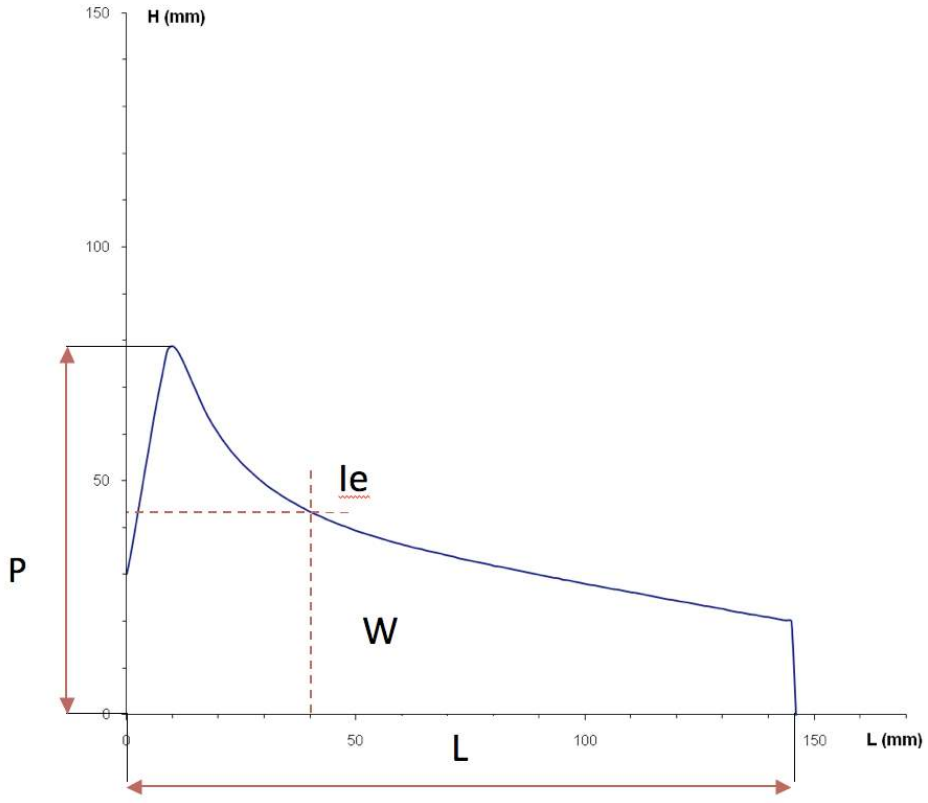


Figure 5: Courbe alvéographique (source : https://ssl10.ovh.net)

classe dans laquelle il envisage d'inscrire sa variété. En fonction des résultats, celle-ci peut-être « sous-classée » ou « sur-classée » (Table S1). Ce classement de variétés en fonction de leur qualité boulangère est typiquement français. Depuis 2000, les surfaces cultivées en blé panifiable et panifiable supérieur, ne cessent d'augmenter, et représentaient en 2015, 94 % de la sole de blé nationale.

e) Caractéristiques génétiques

Le blé tendre est un allo-hexaploïde, possédant 3 génomes A, B et D de 7 chromosomes. La taille physique du génome est de 14.5 Gb et génétique de 3500 cM. (Singh and Singh *et al.* 2015). La séquence complète de référence est disponible depuis juin 2016 (wheatgenome.org). Elle est disponible sur le site de URGI-INRA-Versailles, France. Elle a été assemblée par le logiciel israélien NRGene's DeNovoMAGICTM. Le contrôle qualité a été effectué par l'équipe SEVEN du GDEC. L'annotation du génome est en cours au sein de l'équipe informatique du GDEC. Ces travaux ont été coordonnés par l'IWGSC (*International Wheat Genome Sequencing Consortium*).

D. Qualité boulangère du blé tendre

La qualité boulangère d'un blé peut être définie selon des critères réglementaires, hygiéniques, nutritionnels, organoleptiques ou technologiques. Nous nous intéressons ici à la qualité technologique des farines boulangères, ce qui correspond à son utilisation en panification française.

1) La méthode BIPEA

À partir de 1960, des experts français réunis dans le cadre du CNERNA (Centre national d'études et de recommandations sur la nutrition et l'alimentation) définissent un protocole d'essai de panification. Le protocole utilisé depuis 1988 est la méthode BIPEA (Roussel, 1989), normalisée en 2002 (AFNOR V03-716). La notation se fait grâce à une grille d'évaluation (Figure S1) composée de 31 caractères. Chaque caractère peut être noté comme étant satisfaisant, en excès ou insuffisant. Un critère satisfaisant obtient la note de 10. Un critère non satisfaisant obtiendra la note de 7, 4 ou 1, en excès (positif) ou en insuffisance (note négative). A chaque étape (pétrissage, pointage, façonnage, apprêt, mise au four, aspect du pain et de la mie), un ensemble d'observations est effectué (Table S4). Les notes de pâte, de mie et de pain sont des index (notes intégratives) correspondant à la combinaison linéaire de notes de caractères affectées d'un coefficient qui peut être fixe ou variable. Le volume est le caractère de base le plus influençant (Figure S2). Chacune de ces notes est sur 100. La note de panification est la somme des trois.

2) Les mesures d'alvéographe

L'alvéographe de Chopin permet une mesure indirecte de la valeur boulangère qui présente plusieurs avantages par rapport à un test de panification (Figure 4). Cette méthode est plus rapide, moins chère et plus reproductible. Cet appareil teste la variation de résistance de la pâte pendant son gonflement à débit d'air constant. Il mesure la pression maximale ou ténacité (P), l'indice d'élasticité (le), l'indice de gonflement (G), le rapport P/L et la force boulangère (W) (Figure 5). P dépend de la viscosité de la pâte. Une bonne valeur de P se situe entre 60 et 80 mm. Une bonne valeur d'élasticité (Ie) se situe entre 45 et 55%. Une

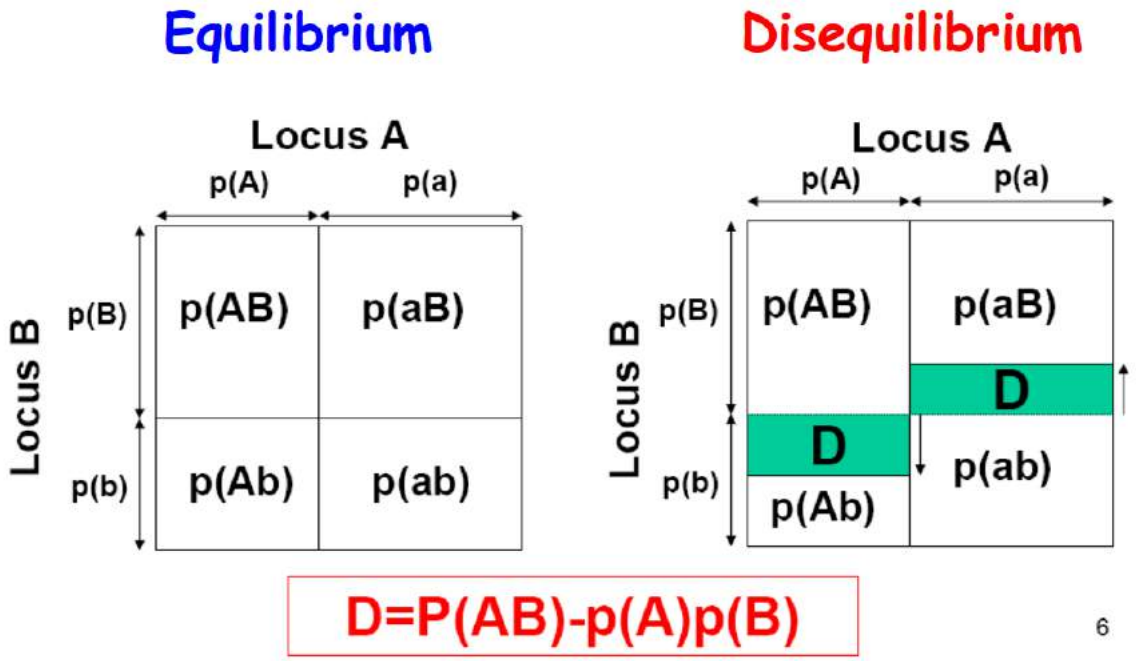


Figure 6 : Représentation schématique et évaluation du déséquilibre de liaison

bonne valeur de G (volume de la bulle) se situe entre 22 et 24. Une bonne valeur de W se situe entre 180 et 220. W est corrélé à la quantité de gluten (Roussel et Chiron, 2002).

E. Statistiques de base pour la génétique

1) Déséquilibre de liaison

Le déséquilibre de liaison (DL) est par définition la différence entre les fréquences gamétiques observées et celles attendues à l'équilibre. Il mesure le degré de dépendance statistique entre les allèles à deux loci différents. Des combinaisons alléliques peuvent être plus ou moins fréquentes que ne le prédit une association aléatoire des loci. Le DL peut être causé par la dérive génétique, la sélection, la mutation ou la migration (Figure 6). Dans une population de blé élite le déséquilibre de liaison est de l'ordre de 10 à 20 cM (Chao, 2007). L'estimation de cette valeur peut permettre de prédire la densité de marqueurs à utiliser en sélection génomique. En effet pour des caractères fortement héritables un R^2 d e0.15 entre marqueurs adjacents était suffisant alors que pour des caractères faiblement héritables, un R^2 de 0.2, améliorait les prédictions en sélection génomique (Calus et Veerkamp, 2007).

2) Moyennes ajustées / modèle linéaire

La moyenne ajustée représente la moyenne obtenue si le dispositif était équilibré. Soit un modèle d'analyse de la variance à deux facteurs avec interactions :

$$Y_{ijk} = \mu + \propto_i + \beta_j + \gamma_{ij} + E_{ijk}$$

Où i est le nombre de niveaux du premier facteur, j est le nombre de niveaux du second facteur, μ la moyenne générale, α l'effet du facteur 1, β l'effet du facteur 2 et γ l'interaction entre le facteur 1 et 2.

La moyenne ajustée pour le niveau i du premier facteur est :

$$\tilde{\eta}_i = \mu + \propto_i + \frac{1}{J}\sum_{j=1}^{J} \beta_j + \frac{1}{J}\sum_{j=1}^{J} \gamma_{ij}$$

3) Diversité génétique

L'indice de diversité génétique peut s'estimer en fonction du taux d'hétérozygotie attendu si on était à l'équilibre de Hardy-Weinberg.

$$He = 1 - \sum_{i=1}^{k} p_i^2$$

Cet indice varie de 0 (peu de diversité) à 1 (beaucoup de diversité).

Cette notion ne doit pas être confondue à l'hétérozygotie observée (Ho) qui est calculée à partir de la fréquence réelle des hétérozygotes (nombre d'individus hétérozygotes/ nombre d'individus total d'individus). Les sélectionneurs qui travaillent sur des lignées fixées, ont pour habitude de remplacer les loci hétérozygotes résiduels par des données manquantes. Dans ce cas, Ho = 0.

Les valeurs d'hétérozygotie attendue sont faibles pour une espèce autogame comme le blé, en effet, chez le riz l'hétérozygotie attendue est de l'ordre de 0.067±0.02 (Gao, Hong, 2000),

contrairement à une espèce allogame comme le pois d'Espagne (P. coccineus) dans laquelle l'hétérozygotie attendue est de l'ordre de 0,314 ± 0,077 (Escalante *et al*., 1994).

4) Génétique d'association / modèle linéaire mixte

L'analyse de QTL par association exploite le déséquilibre de liaison existant au sein du panel étudié entre les marqueurs et les QTLs impliqués dans la variation du caractère considéré. Différents modèles statistiques peuvent être utilisés :

a) Le modèle de régression (modèle linéaire à un facteur : le marqueur, modèle naïf)

Si le panel n'est pas structuré et qu'il est à l'équilibre de Hardy-Weinberg, alors le DL entre deux loci est uniquement dû à leur proximité physique. On peut alors expliquer le phénotype observé avec le modèle de régression linéaire suivant :

$$Y = \mu + \beta x + \varepsilon$$

Le phénotype Y d'un individu est la somme de la moyenne du caractère dans la population, l'effet β des allèles au marqueur x et la résiduelle non expliquée par le modèle. Si l'effet β du marqueur est significatif alors le marqueur x testé est en déséquilibre de liaison avec un QTL.

b) Le modèle linéaire mixte

Si le panel est structuré, tous les individus ne sont pas apparentés de la même manière, le modèle précédent peut donner de nombreux marqueurs associés si le caractère a évolué en parallèle avec la structuration de la population (Pritchard *et al*. 2000). C'est le cas des caractères dits adaptatifs. Ces fausses associations sont appelées faux positifs. Yu *et al*. (2006) ont proposé un modèle linéaire mixte qui prend en compte la structure et l'apparentement au sein du panel.

$$Y = \mu + \beta X + Qv + Zu + e$$

Où X est le vecteur de génotypes, β le vecteur des effets fixes au marqueur, v le vecteur d'effets fixes des (n-1) différents groupes génétiques constituant le panel, Q la matrice d'assignation de chaque lignées à chaque groupe génétique, u le vecteur des effets aléatoires du fond génétique de chaque lignée sachant que $Var(u) = 2KVg$, où Vg est la variance génétique et e la résiduelle.

La matrice de Kinship (K) correspond à la covariance génétique entre les individus, c'est à dire leur apparentement. Elle peut être mesurée grâce au pedigree s'il est connu ou grâce au génotypage de marqueurs moléculaires neutres répartis sur l'ensemble du génome. La matrice de structure (Q) nécessite au préalable de définir des groupes génétiques ayant des fréquences alléliques différentes suite à des événements de dérive et différentiation. Pour chaque individu, sachant ses fréquences alléliques observées, on calcule la probabilité qu'il appartienne à chacun de ces groupes génétiques. En pratique, lorsque le nombre de marqueurs pour calculer K est important, il est inutile d'ajouter la structure Q dans le modèle qui devient :

$$y = X\beta + Zu + \varepsilon \quad (1)$$

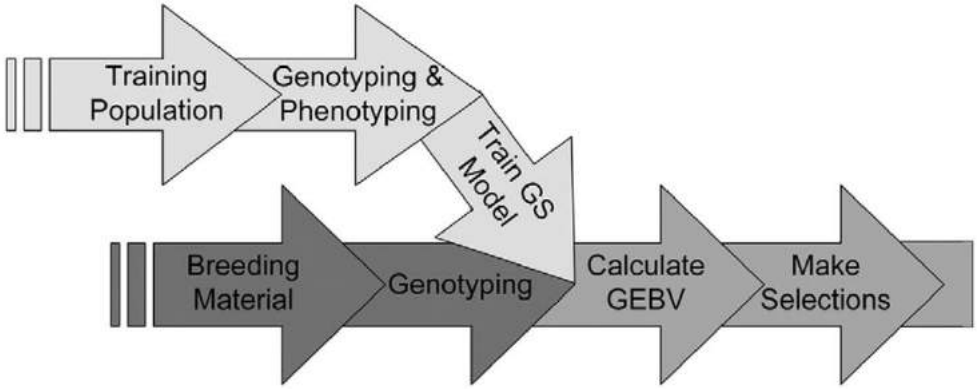


Figure 7 : Les différentes étapes constitutives de la sélection génomique et sa place dans le processus de sélection (Heffner *et al.*, 2009).

A partir, d'une population de calibration, génotypée et phénotypée, un modèle de sélection génomique est mis au point et appliqué à du matériel de sélection précédemment génotypé, leur GEBV (G*enomic Estimated Breeding Value*) peut ainsi être estimée, permettant de sélectionner à partir de ces résultats.

Où y est le vecteur des phénotypes, β le vecteur des effets fixes, u le vecteur des valeurs génétiques, et ε le vecteur des résidus.

5) Sélection génomique

La sélection assistée par marqueur est une technique bien adaptée pour l'introgression d'allèles favorables contrôlant des caractères monogéniques. Mais lorsque le nombre de QTLs est trop important et leurs effets faibles, pour des caractères complexes, le pyramidage de nombreux allèles dans un même fond génétique peut être fastidieux, voire impossible (Bernardo, 2008 et Xu et Crouch, 2008). Contrairement aux méthodes de détection de QTLs, la sélection génomique estime simultanément les effets de tous les marqueurs (Meuwissen *et al*. 2001) pour prédire la valeur génétique des individus appelée GEBV (*Genetic estimated breeding values*). Dans la pratique, deux populations sont définies, une population de calibration et une population de cible. La population de calibration est génotypée et phénotypée et sert à établir le modèle de prédiction. Ce modèle est ensuite appliqué à la population de validation qui a été génotypée mais pas phénotypée. Il permet de prédire les phénotypes à partir des génotypes. (Figure 7). La capacité de prédiction correspond à la corrélation entre les prédictions (GEBV) et les vraies valeurs génétiques (TBV : *True Breeding Value*). Les TBV étant généralement inconnues, les modèles sont testés avec de la cross-validation : les GEBV de plusieurs échantillons (souvent un dixième) des phénotypes observés sont estimés à partir des 9 dixièmes phénotypes restant. La précision de prédiction (*accuracy*) est estimée en divisant la moyenne des corrélations entre phénotypes et GEBV par la racine carrée de l'héritabilité.

Il existe plusieurs modèles de prédiction :

- Avant la sélection génomique, on détectait les QTLs, on estimait leurs effets. La valeur génétique des individus était calculée en faisant la somme des effets de leurs allèles aux QTLs considérés comme significatifs :

1) Analyse en régression simple de chaque marqueur à l'aide du modèle

$$y = \mu 1n + X\beta + e \quad (2)$$

Où y est le vecteur des phénotypes, μ la moyenne général, β le vecteur des effets des marqueurs et e l'erreur résiduelle

2) Sélection des m marqueurs les plus significatifs

La valeur génétique (GEBV : *Genomic Estimated Breeding Value*) d'un individu se calcule comme la combinaison linéaire des effets des allèles aux QTLs détectés:

$$GEBV = X_m \hat{\beta}_m$$

Cette approche pose plusieurs problèmes. Il faut choisir un seuil de significativité pour les QTLs. De plus, la régression simple peut surestimer les effets des marqueurs. La solution est d'estimer tous les effets simultanément, ou de ne pas les estimer du tout (G-BLUP) grâce aux différents modèles de sélection génomique :

Table 3 : Récapitulatif des variables de panification analysées et éliminées (données Limagrain)

Variable analysée	Variable éliminé
Rapidité de lissage	Collant de la pâte (pétrissage, façonnage, mise au four, de la mie)
Extensibilité (façonnage, pétrissage)	Consistance
Elasticité	Relâchement
Déchirement (façonnage, coup de lame)	Détente/relâchement
Section du pain	Elasticité (façonnage, de la mie)
Pourcentage d'hydratation	Déchirement (apprêt)
Volume	Activité fermentaire
Volume massique	Tenue de la pâte
Note de pain	Couleur (pain, mie)
Note de pâte	Epaisseur (pain et alvéolage)
Note de panification	Croustillant du pain
	Développement
Teneur en protéine	Souplesse
	Régularité de l'alvéolage
	Saveur et arôme
	Note de mie

- La méthode de ridge regression (RR-BLUP : *Random Regression Best Linear Unbiased Prediction*)

Whittaker *et al.* (2000) a développé la méthode de « *ridge regression* » qui se base sur le modèle des moindres carrés (2) mais où les βi sont estimés simultanément en ajoutant un terme λI dans l'équation :

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'y$$

Cette « pénalité » permet de rendre inversible la matrice, ce qui n'est pas le cas quand le nombre de paramètres à estimer (marqueurs) dépasse le nombre d'observation. Cette régression « pénalisée » a tendance à écraser tous les effets, qu'il ne faut donc pas interpréter individuellement. Sous l'hypothèse que les effets des marqueurs suivent une distribution normale centrée sur 0, $\lambda = \frac{\sigma_e^2}{\sigma_\beta^2}$. C'est alors équivalent à la méthode G-BLUP qui utilise une matrice de similarité / apparentement pour estimer directement les valeurs génétiques des individus, sans passer par l'estimation des effets des marqueurs. Elle consiste en la résolution de l'équation du modèle linéaire mixte (1) (Henderson 1975).

-Les méthodes Bayesiennes

L'hypothèse sous-jacente au RR-BLUP est que les effets génétiques sont faibles, tirés d'une même distribution normale et distribués sur l'ensemble du génome, ce qui n'est pas forcément vérifié en pratique. Meuwissen *et al* (2001) ont proposé une méthode plus souple vis-à-vis de la distribution des effets génétiques en utilisant un algorithme bayésien. Les méthodes Bayes B et C supposent que la majorité des marqueurs (π) ont des effets nuls. Elles diffèrent par les lois de distribution des effets (inverse du Chi2 ou gaussienne respectivement). Si π = 0, Bayes A et Bayes B sont équivalents. La proportion π, est souvent fixée à 0.95.Ce taux étant variable suivant les caractères, Habier *et al*, (2011) ont proposé la méthode Bayes Cπ qui estime elle-même à partir des données la part de marqueurs à effet nul (π) et la variance expliquée par les marqueurs inclus dans le modèle. On peut noter qu'il s'agit indirectement de méthodes de sélection de variable.

II. Matériel et méthode

A. Données Limagrain

1) Phénotypes

Le jeu de données du programme de sélection Limagrain est constitué de 785 lignées obtenues par haplo-diploïdisation ou par sélection généalogique. La qualité boulangère a été mesurée avec la méthode BIPEA par l'entreprise QUALTECH. Au total, 31 caractères (notes de base) participent au calcul de trois notes agrégées (note de pâte, note de mie et note de pain) dont la somme correspond à la note de panification finale. La teneur en protéine a été mesurée par spectrométrie proche infrarouge. Chaque caractère a été mesuré une fois pour chaque lignée. Les lignées ont été réparties dans deux lieux en 2015 (397 lignées dans le lieu 1 et 388 dans le lieu 2). Quatre témoins ont été répétés 2 fois dans chaque lieu. Les notes de base ont été linéarisées sur une échelle de 1 à 10. Au total, 24 variables monomorphes ont été

éliminées et 14 caractères ont été analysées (Table 3). Les composantes de la variance de ces caractères ont été calculées avec un modèle linéaire (ANOVA : analyse de la variance) :

$$y_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ij} \quad (3)$$

Où yij est le phénotype du génotype i dans l'essai j, μ est la moyenne, αi est l'effet du fond génétique de l'individu i, βj est l'effet de l'essai j, $\alpha\beta_{ij}$ est l'effet d'interaction entre génotype i et environnement j, e_{ij} est l'erreur résiduelle.

L'effet lieu étant significatif, les moyennes ajustées ont été calculées avec un modèle prenant en compte les interactions Génotype x environnement, en considérant tous les effets comme fixes. Les moyennes ajustées calculées dans R ont été utilisées pour les analyses de diversité, d'association et de sélection génomique. Les caractères ont été décrits par leurs corrélations et une Analyse en Composantes Principales (ACP) avec le package R Factominer (Lê *et al*).

2)Génotypes

Les lignées ont été génotypées pour 14592 marqueurs SNP cartographiés génétiquement à partir de plusieurs populations de cartographie. La carte consensus a été produite avec le logiciel BioMercator (Arcade *et al* 2004) par Biogemma. Les marqueurs ayant une MAF (*Minor Allele frequencie*) inférieure à 1 % ont été supprimés. Les individus et les marqueurs ayant plus de 5 % de valeurs manquantes sont également éliminés. Au total, 716 lignées et 10124 marqueurs ont été utilisés pour les analyses.

B.Données INRA-AGRI-OBTENTIONS (INRA-AO)

1) Phénotypes

Le jeu de données du programme de sélection INRA-AO est constitué de 738 lignées phénotypées pour la teneur en protéine, 370 pour les variables de panification et 357 pour les variables d'alvéographe. Ces mesures ont été recueillies sur 14 ans (entre 2000 et 2013) avec des répétitions dans différents lieux (Clermont-Ferrand, Dijon, Le Moulon, Lusignan, Estrées-Mons, Rennes, Orsonville) et entre années successives (Table S3). En moyenne pour une lignée donnée, les notes de panification et d'alvéographe ont été mesurées six fois (médiane de 4) et la teneur en protéine onze fois (médiane de 4). En moyenne, 41 lignées sont testées chaque année (médiane de 41). De plus, sur ces 14 années, chaque lieu reçoit en moyenne 220 lignées (médiane de 260). Le nombre de lignées répétées entre deux années est inversement proportionnel au nombre d'années qui les séparent. En effet, il s'agit des lignées élites du programme qui sont évaluées pendant deux ou trois années successives avant d'être proposées à l'inscription. Seuls 3 témoins ont été évalués toutes les années (Table S2).

Les moyennes ajustées ont été calculées en considérant le génotype comme effet fixe (BLUE : Best Linear Unbiased Estimator) ou un effet aléatoire (BLUP : *Best Linear Unbiased Predictor*) sont similaires (corrélation supérieure à 0.9) (Table S8). Pour la suite de l'analyse nous avons utilisés les moyennes ajustées BLUE. Les composantes de la variance ont été calculée avec le modèle (3).

La répétabilité a été calculée à partir de ces composantes de la variance :

$$h^2 = \frac{\text{Var}\alpha}{(\text{Var}\alpha + \text{Var}e)}$$

2)Génotypes

Les lignées ont été génotypées avec une puce 420K construite par l'équipe SEVEN du GDEC dans le cadre du projet BREEDWHEAT. Les lignées et les marqueurs avec plus de 5% de données manquantes ont été éliminés ainsi que les marqueurs avec une MAF inférieure à 1%. Au total, 1884 lignées et 172 074 marqueurs ont été analysés.

Pour estimer la part de variance de la qualité boulangère expliquée par les gluténines et leur potentiel de prédiction, un génotypage de marqueurs KASPar (Ravel *et al*, in prep) est en cours. Pour le stage, seul le génotypage des gluténines A pour 180 lignées a pu être utilisé. Les haplotypes obtenus ont été inclus avec les autres caractères de panification et d'alvéographe dans les ACP.

C. Diversité phénotypique, diversité génétique et déséquilibre de liaison

Le paramètre de diversité génétique He a été calculé avec le package R hierfstat (Goudet, 2005). L'étendue du déséquilibre de liaison a été calculée comme la corrélation au carré des doses alléliques entre deux loci (R^2) avec le logiciel Plink (Purcell *et al*, 2007).
Les moyennes et variances des différents caractères ont été calculées dans R.

La structuration phénotypique et génotypique de ces jeux de données a été évaluée avec des Analyses en Composante Principale et une classification hiérarchique ascendante implémentés dans le package R FactoMineR.

La structure génétique des jeux de données a également été évaluée avec le logiciel Admixture (Alexander *et al*., 2009). Le nombre de groupes génétiques a été estimé en mesurant la variation d'assignation par cross-validation.

D. Etudes d'association

Les panels ne présentent pas de structure forte d'après les résultats de classification d'Admixture. La structure génétique a été prise en compte dans les modèles d'association avec une matrice de kinship mais sans matrice de structure d'après le modèle :

$$y = X\beta + Zu + \varepsilon \quad (1)$$

Où y est le vecteur des phénotypes, β le vecteur des effets fixes, u le vecteur des valeurs génétiques, et ε le vecteur des résidus.

Table 4 : Moyenne, variances et héritabilité des principaux caractères de qualités boulangères

Trait	Moy LG	sd LG	Moyenne INRA-AO	sd INRA-AO	Var G INRA-AO	Var E INRA-AO	Var R INRA-AO	h2
P	-	-	85.8	-	458.84	185.02	107.48	0.81
W	-	-	206.7	-	2453.45	705.26	1027.35	0.7
L	-	-	40.1	-	515.43	305.8	295.54	0.64
G	-	-	21.6	-	7.37	4.42	4.11	0.64
PsurL	-	-	7.23	-	0.37	0.16	0.22	0.63
Prot	11.22	-	11.37	-	0.36	1.2	0.35	0.51
VOL	1589.9	200.14	1514.3	211.79	17490.59	8154.79	22268.64	0.44
Npanif	251.1	19.53	220.2	34.25	458.45	163.57	600.03	0.43
Npate	80.81	9.89	77.3	14	64.11	46.17	99.28	0.39
Npain	70.89	12.88	65.9	33.41	138.55	1157.41	234.9	0.37
Nmie	99.02	3.61	93.7	7.73	15.65	14.71	41.22	0.28

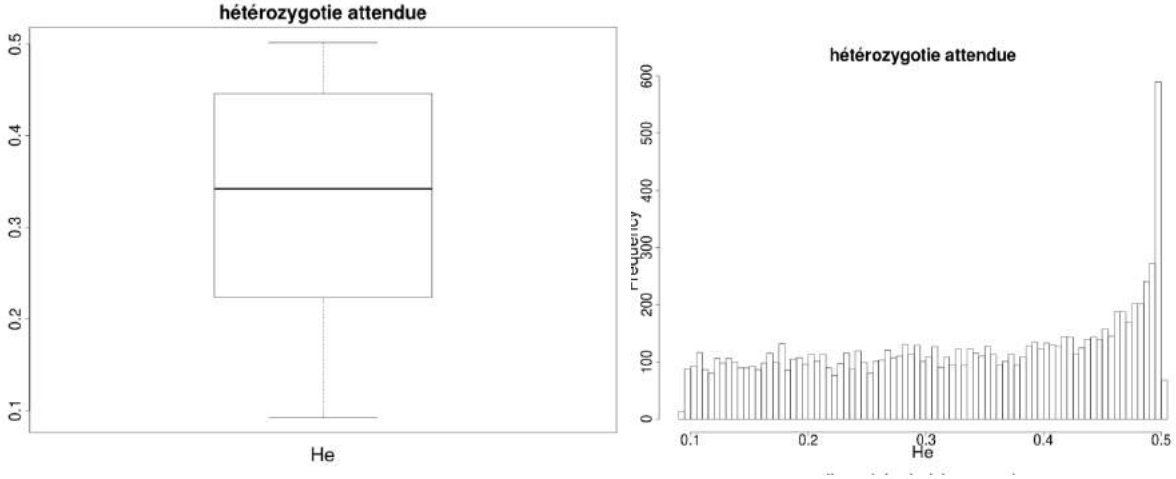


Figure 8 : Boite de dispersion et Histogramme de l'hétérozygotie attendue (données Limagrain)

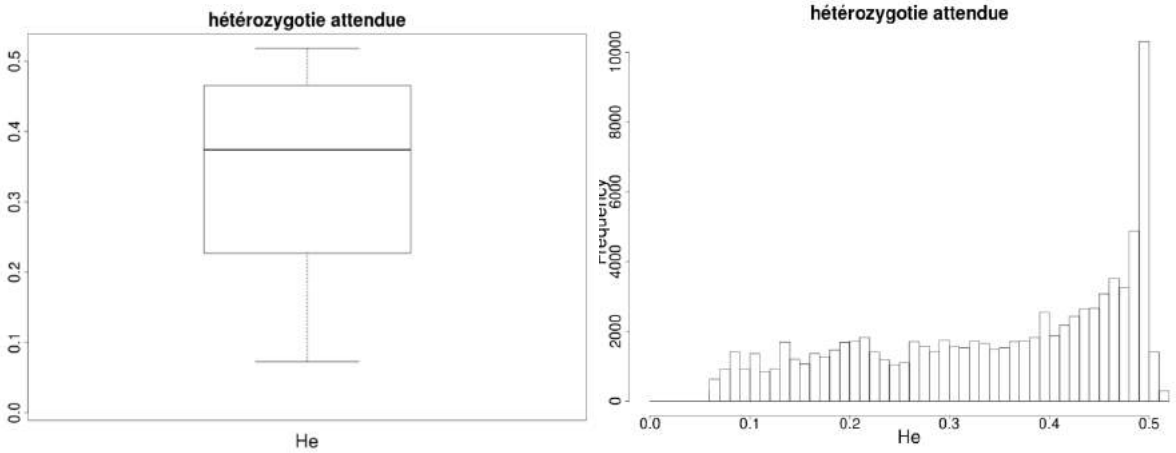


Figure 9 : Boite de dispersion et Histogramme de l'hétérozygotie attendue (données INRA)

Nous avons tout d'abord utilisé le modèle linéaire mixte implémenté dans le package R rrblup (Endelman *et al* 2011).

Les matrices de Kinship K ont été calculées avec la méthode IBS (*Identity By State*) implémentée dans le logiciel Plink (Purcell *et al.,* 2007). L'IBS est le pourcentage d'allèles partagés entre les individus.
Dans un deuxième temps, la méthode bayésienne *Multiple loci mixed model* (MLMM) (Segura *et al*. 2012) a été utilisée. Au niveau de chaque QTL, une régression de type *« forward backward »* est effectuée. Les marqueurs les plus associés sont ajoutés un à un en co-facteur dans le modèle. Les composantes de la variance ($\widehat{\sigma_g}^2$ et $\widehat{\sigma_e}^2$) et la p-value des marqueurs en cofacteurs dans le modèle sont ré-estimées à chaque étape. La régression *forward* s'arrête lorsque la variance génétique expliquée par le dernier marqueur inclus dans le modèle ($\frac{\widehat{\sigma_g}^2}{Var(\boldsymbol{y})}$) est proche de zéro. Une régression *backward* est ensuite effectuée. Elle élimine un par un les marqueurs les moins associés. Le modèle retenu est celui qui maximise la part de variance expliquée par les marqueurs.

Dans un second temps, les études d'association single locus et haplotypiques ont été réalisées en utilisant des programmes en cours de développement (Servin *et al*., Xu *et al*., *personal communication*). Le programme modélise le déséquilibre de liaison local pour définir de manière dynamique le long du génome la taille de la fenêtre de l'haplotype local. Les haplotypes locaux sont centrés sur chaque SNP. Il y a donc autant de tests réalisés que de SNPs dans le jeu de données.

Le nombre de tests / marqueurs indépendants a été estimé selon Li & Ji (2005). Le seuil de Bonferroni 5% correspondant a été utilisé pour définir le seuil de significativité des QTLs.

E. Sélection génomique

1) Modèle G-BLUP

Nous avons fait tourner les modèles G-BLUP et BAYES- Cπ implémentés dans le pipeline BWGS développé dans le cadre de BREEDWHEAT. Nous avons fait varier le nombre de marqueurs de 100 à 10 000 pour le panel Limagrain et de 100 à 40 000 pour le panel INRA-AO. Pour chaque nombre de marqueurs, 30 cross-validations ont été effectuées : la population est divisée en 10, la population de calibration correspond aux 9/10 du jeu de données et la population de validation correspond au 1/10 restant.

Afin de trouver la taille de la population de calibration minimale pour obtenir une précision des prédictions optimales, nous avons également fait varier la taille de la population en utilisant 10 000 marqueurs (données Limagrain) ou 15 000 marqueurs (données INRA-AO).

2) Modèle G-BLUP avec sélection de variable

Nous avons dans un second temps regardé si l'utilisation d'un sous-ensemble de marqueurs associé aux caractères plutôt que l'ensemble des marqueurs permettait d'améliorer les prédictions génomiques. Nous avons fait varier le seuil de significativité de l'ANOVA à un facteur (le marqueur) pour choisir les marqueurs à inclure dans le modèle. Dans un second

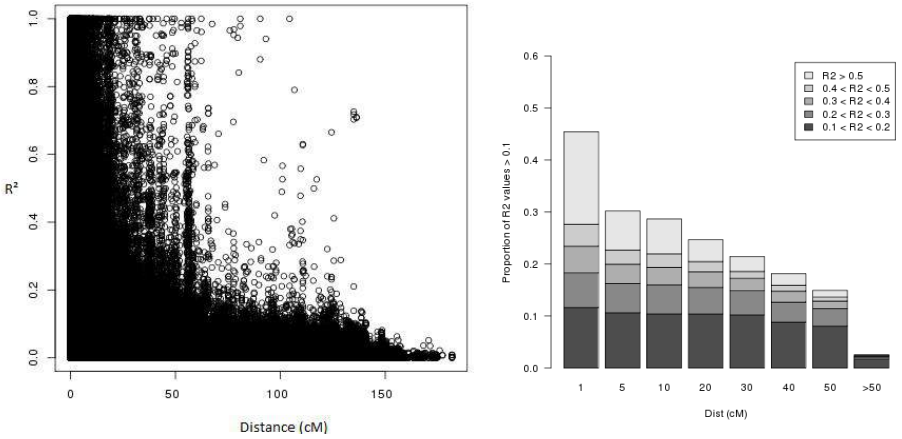


Figure 10 : Représentation graphique du déséquilibre de liaison en fonction de la distance entre marqueurs (données Limagrain)

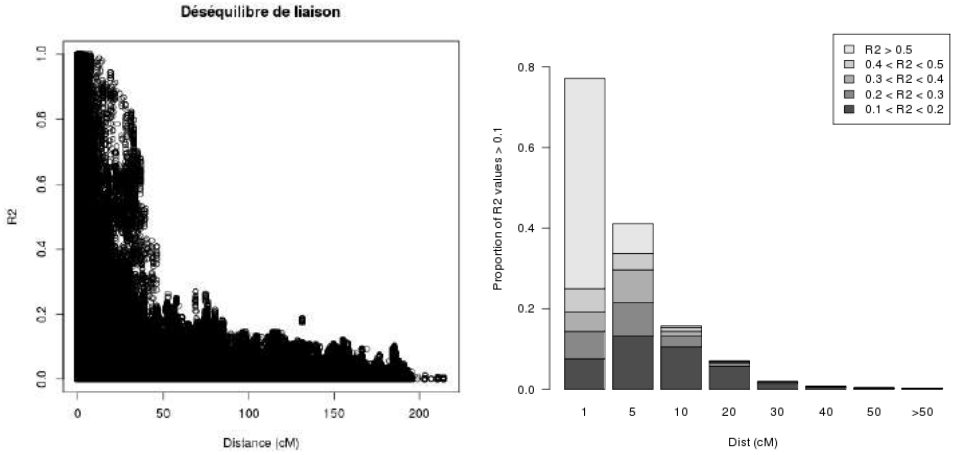
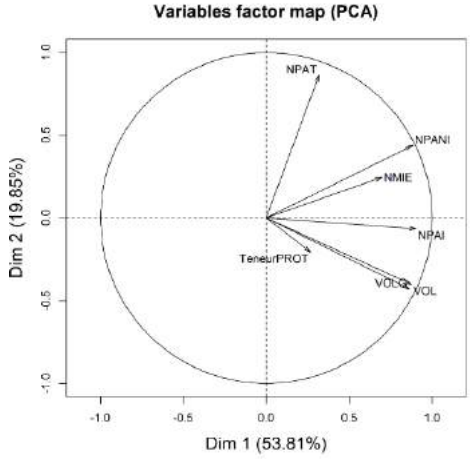


Figure 11 : Représentation graphique du déséquilibre de liaison en fonction de la distance entre les marqueurs (données INRA)

a)



b)

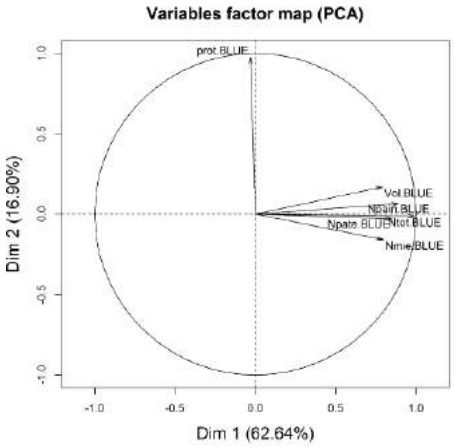


Figure 12 : Analyse en composante principales (a : donnée Limagrain, b : donnée INRA-AO)

temps, nous avons utilisé les p-values d'association utilisées avec le modèle MLMM qui fait de la sélection de variable décrit plus haut. L'objectif est de donner un poids équivalent à tous les QTLs et de limiter la redondance des marqueurs en déséquilibre de liaison.

Nous avons regardé pour chaque méthode et chaque caractère le nombre de marqueurs optimal pour les prédictions génomiques.

3) Modèle G-BLUP avec covariables (QTLs majeurs) en effet fixe

Nous avons ensuite regardé si l'ajout de marqueurs à gros effet en effet fixe dans le modèle RR-BLUP améliorait les prédictions. Nous avons ajouté les trois marqueurs les plus associés à chaque caractère. Nous avons également ajouté un haplotype associé aux gluténines A pour un panel INRA-AO de 180 lignées. Ce marqueur est issu de la combinaison de deux SNP définissant 3 formes différentes pour la protéine GLU-1A. Nous avons utilisé le package rrBLUP (Endelman *et al.*, 2011) pour ces analyses.

III Résultats

A. Description des variables et héritabilité

Les composantes de la variance et les héritabilités des caractères (Table 4) ont été calculées pour le panel INRA-AO pour lequel nous disposions de répétitions. Les notes d'alvéographe possèdent les héritabilités les plus élevées situées entre 0.64 et 0.81. Les héritabilités des notes de panification se situent entre 0,28 et 0,44. La variance génétique est 2 à 4 fois supérieure à la variance environnementale pour les variables d'alvéographe. Cette tendance s'inverse pour les notes de panification, où la variance environnementale est 1.5 à 2.5 fois supérieure à la variance génétique.

Les valeurs de teneur en protéine sont similaires dans les deux jeux de données. Les lignées Limagrain ont des notes de panification plus élevées que les lignées du panel INRA-AO (Note de panification (NPANI) : INRA-AO : 220 ; Limagrain : 250) (Table 4). La variance phénotypique est également plus élevée pour les caractères du panel INRA-AO, que dans le panel Limagrain.

B. Corrélation entre variables

Nous observons que la teneur en protéine est peu corrélée aux caractères liés à la qualité boulangère. (Table S5)

La teneur en protéine est faiblement corrélée aux valeurs d'alvéographe INRA-AO (r^2 W = 0.34, r^2 G = 0.15, r^2 P = 0.14, r^2 L = 0.12) et aux notes de panification Limagrain (r^2 VOL = 0.24, r^2 NPAIN = 0.19, r^2 NPANIF = 0.16). Notons que les corrélations entre la teneur en protéine et les notes de panification sont nulles pour INRA-AO.

La note de panification sont relativement corrélées aux valeurs d'alvéographe (r^2 W = 0.44, r^2 G = 0.30, r^2 P = 0.13, r^2 L = 0.32). Par contre, chacun de ces groupes de caractères sont très corrélés entre eux (Table S7), avec des valeurs comprises entre 0.72 et 0.88 pour l'INRA-AO, 0.58 et 0.83 pour Limagrain (Table S8).

Table 6 : Correspondances des marqueurs significatifs (Données INRA-AO) avec les gènes candidats

Traits	Marqueur	Chromosome	pos physique (10^6 pb)	gène candidat	Position physique gène candidat
G	cfn2171682	2A	727,698,502	GS2	729,292,149
L	cfn0712783	2A	727,704,636	GS2	729,292,149
Protéine	cfn0761212	3D	596,565,773	ASR1	600,631,622
Protéine	P	4D	87,653,038	ASR5	88,700,367
L	cfn0353395	6D	335,302,434	Sad	334,226,078

Table 7 : Nombre de marqueurs associé aux caractères communs entre les jeux de données INRA-AO et Limagrain (MLMM)

Caractères	Nb QTL Limagrain	Nb QTL INRA-AO	Zone commune	Chromosome commun
Protéine	7	31	5	3A, 4A, 5B, 4D, 6D
NPATE	4	2	1	2B
NPAI	3	4	-	-
VOL	7	18	3	4A, 1D, 2D
NPANIF	-	2	-	-

Table 8 : Nombre de marqueurs et QTLs (régions de 1cM) associés (modèle MLMM)

Caractères	**Limagrain**	**INRA-AO**
Protéine	7	31
NPATE	4	2
VOL	7	17
EXTF	5	
DECF	2	
DECL	3	
ELAP	-	
EXTP	6	
LISP	-	
NPAIN	3	4
NPANIF	-	2
% Hydra	6	
SECT	-	
VOLG	7	
P		29
G		16
L		15
PsurL		3
W		14
NMIE		7

C. Analyse en composantes principales (ACP)

Pour les données Limagrain, le premier axe explique 54% de la variance et est corrélé aux caractères liés à la note de pain et au volume. Le deuxième axe de l'ACP explique 19% de la variance et est corrélé à la note de pâte. Le troisième axe est corrélé à la teneur en protéine et explique 14% de la variance (Figure 12, a).

Pour les données INRA-AO, l'ACP construite avec les mêmes caractères ne donne pas exactement les mêmes résultats. Le premier axe est corrélé aux notes de panification (note de pâte et pain) et explique 62.64% de la variance, le deuxième axe à la teneur en protéine et explique 17 % de la variance (Figure 12 b). Pour les ACP INRA-AO augmentées des notes d'alvéographe et des haplotypes de gluténines A, les caractères liés aux notes de panification (note de panification, note de pain, note de pâte, note de mie et volume) et à la force boulangère W sont corrélés au premier axe de l'ACP qui explique 41.1 % de la variance phénotypique. Le deuxième axe est fortement corrélé aux variables d'alvéographe (Pression maximale P, Longueur L et Gonflement G, rapport P sur L) et explique 26.8% de la variance. Le troisième axe est corrélé à la teneur en protéine et explique 12.1% de la variance. Les haplotypes de gluténines A sont corrélés au quatrième axe de l'ACP qui explique 7.9% de la variance. Notons que pour les deux panels, la teneur en protéine est mal projetée sur les axes principaux qui expliquent la variation des notes de panification, ce n'est pas la variable la plus importante pour prédire la qualité boulangère.

Les classifications ascendantes hiérarchiques permettent de distinguer 3 groupes principaux de qualité boulangère différente dans les deux panels. Dans le cas du jeu de données INRA-AO, (Figure S3, Table S 9), le groupe 1 (noir) a une mauvaise qualité boulangère. Il a des notes de panification, une force boulangère et des notes d'alvéographe faibles. La majorité des variétés inscrites dont Folklor, un BPS sélectionné par Agri-obtention, sont dans le groupe 2 qui correspond à des notes de panification variable avec W et P élevés. Le groupe 3 (vert), correspond à des blés qui ont une bonne note de panification avec W et L élevés. Il contient les variétés Galibier et Apache. Apache est une variété BPS phare de Limagrain inscrite en en 1998.

Concernant les données Limagrain, 3 groupes ressortent également de l'analyse. Le groupe 1 correspond à des blés qui ont de mauvaises notes de panification mais une bonne note de pâte. Le groupe 2 correspond à des blés qui ont de bonnes notes de panification malgré une faible note de pâte. Le groupe 3 correspond à des blés qui ont de bonnes notes de panification et de pâte grâce à une bonne élasticité (Table S10).

D. Diversité génétique et déséquilibre de liaison des panels

Nous avons comparé les panels Limagrain et INRA-AO en matière de diversité génétique et déséquilibre de liaison. En moyenne la diversité génétique (hétérozygotie attendue : He) est de 0.34 pour le panel INRA-AO et 0.33 pour le panel Limagrain. (Figure 8 et 9).

Concernant le déséquilibre de liaison moyen, pour le panel Limagrain, 95% des valeurs de DL (R^2) sont inférieures à 0.12. Parmi les valeurs supérieures à 0.2, 80% concernent des couples

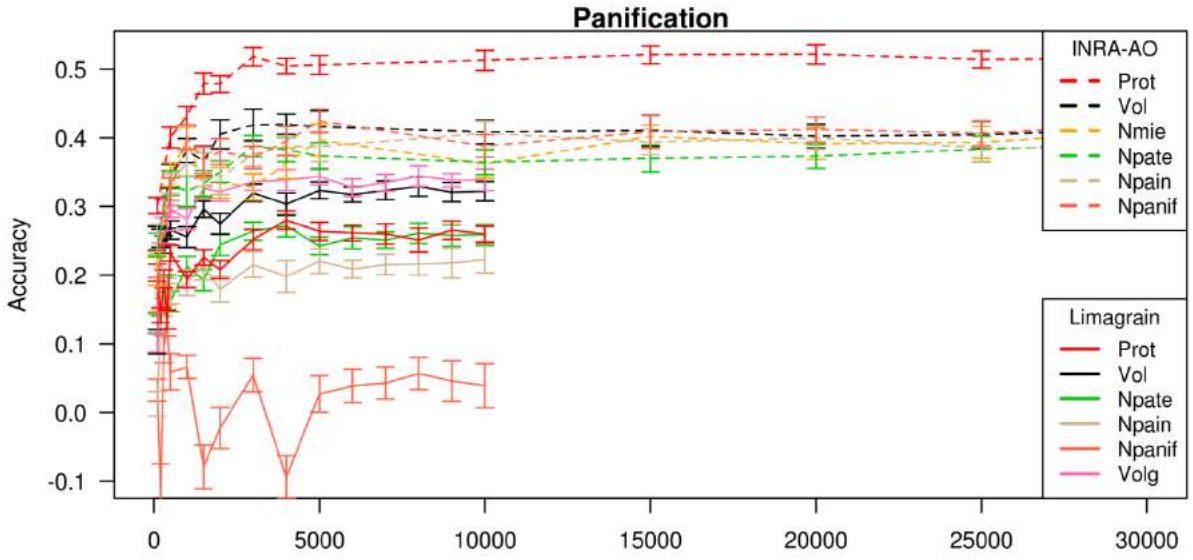


Figure 13: Représentation de la précision de prédiction en fonction du nombre de marqueurs choisi aléatoirement pour les caractères de panification

(Données INRA-AO/Limagrain, modèle GBLUP, population de calibration de 313 lignées (panif), 654 lignées (protéine), 288 lignées (Nmie), 300 lignées (alveo), Limagrain : 644 lignées)

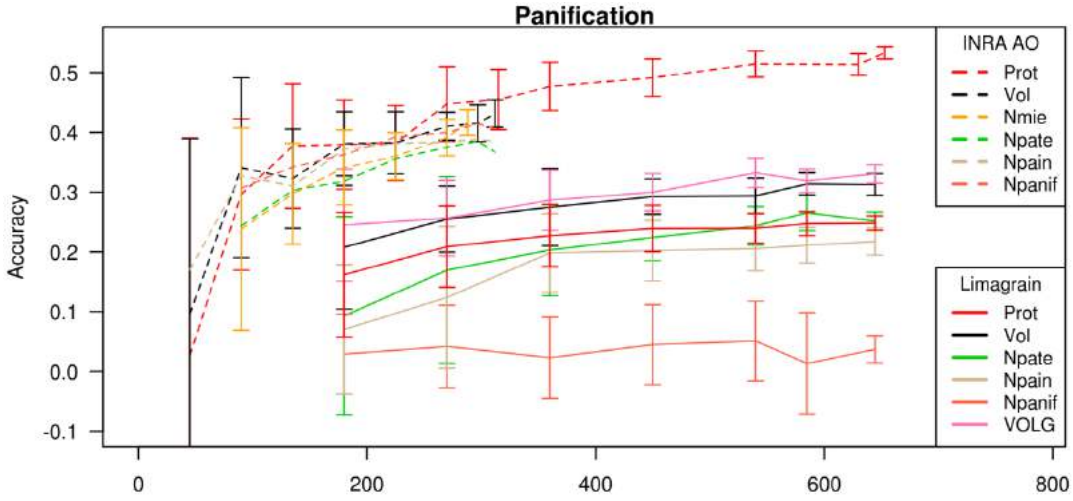


Figure 14 : Représentation de la précision de prédiction en fonction du nombre de lignées choisies aléatoirement pour les caractères de panification

(modèle G-BLUP, données INRA-AO/Limagrain : nombre de marqueurs : 8 000, données INRA-AO : 15 000)

de marqueurs situés à moins de 6 cM l'un de l'autre. Pour les données INRA-AO, 90% des valeurs sont inférieurs à 0.22 et 80% des valeurs supérieures à 0.2 concernent des marqueurs situés à moins de 4.34 cM l'un de l'autre. Le déséquilibre de liaison est moins étendu dans le panel INRA-AO que Limagrain (Figure 10 et 11)

E. Analyses d'association

1) Modèle linéaire mixte single locus

Nous reportons les QTLs détectés avec le modèle MLMM. Pour la note de panification, 2 QTLs chromosomiques ont été détectées sur les chromosomes 1A et 1D dans le panel INRA-AO et aucun dans le panel Limagrain. Pour le volume 17 QTLs ont été détectés dans le panel INRA-AO et 7 dans le panel Limagrain. Pour W, 14 QTLs ont été détectés dans le panel INRA-AO et aucun dans le panel Limagrain (Table 8).

Pour les caractères communs aux deux jeux de données (Protéine, NPATE, NPAIN, volume), nous retrouvons 5 QTLs sur des chromosomes communs pour le caractère de teneur en protéine, 1 pour la note de pâte, 3 pour le volume et aucune pour la note de pain (Table 7).

2) Modèle linéaire mixte avec prise en compte des haplotypes locaux

Pour le jeu de données Limagrain, la pvalue minimale obtenue en utilisant des haplotypes locaux est toujours inférieure ou égale à la pvalue obtenue avec un modèle single locus, sauf pour le caractère déchirement au façonnage. Le nombre de QTLs détecté est également systématiquement supérieur. Pour la note de panification par exemple, 7 QTLs sont détectés avec l'approche haplotypique alors qu'aucun ne sont détectés avec l'approche single locus (Figure S4, Table S 14).

F. Sélection génomique

1) Précision des prédictions en fonction du nombre de marqueurs

Pour analyser l'effet de l'augmentation du nombre de marqueurs sur les prédictions génomiques, toutes les lignées disponibles ont été utilisées, soit en moyenne 644 lignées pour Limagrain et 300 lignées pour INRA-AO, sauf pour la teneur en protéine disponible pour 735 lignées).

Pour les deux jeux de données la précision de prédiction atteint un plateau lorsque le nombre de marqueurs tirés aléatoirement est supérieur à 5000. L'écart-type entre cross-validations est faible (Figure 13 et Figure S7 a). Pour un même nombre de marqueurs, la précision de prédiction est plus élevée pour les caractères d'alvéographe (entre 0.3 pour PsurL et 0.62 pour la pression maximale, P) que pour les caractères de panification (entre 0.35 pour la note de pâte et 0.4 pour la note de panification d'après les données INRA-AO.

D'après les données Limagrain, la précision de prédiction est supérieure pour les notes de panification (entre0.2 pour la note de pain et 0.25 pour la de pâte) que pour les caractères de base (entre 0.02 pour la note de section et 0.22 pour le pourcentage d'hydratation).

Notons que la prédiction de la note de panification Limagrain est faible (0.05) par rapport à la prédiction INRA-AO (0.4).

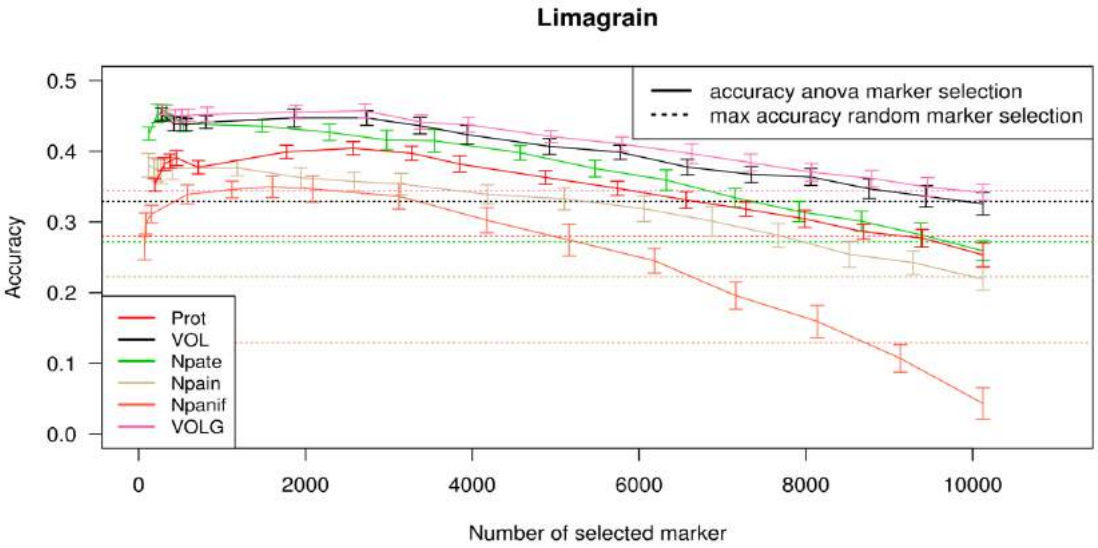


Figure 15 : Gain de précision de prédiction entre un modèle G-BLUP avec et sans sélection de marqueurs
Données Limagrain avec une population de calibration de 644 lignées
Pointillés : précision de prédiction maximale obtenue sans sélection
Ligne pleine : précision de prédiction maximale obtenue avec sélection des marqueurs
Barre d'erreur : écart-types associé à chaque valeur

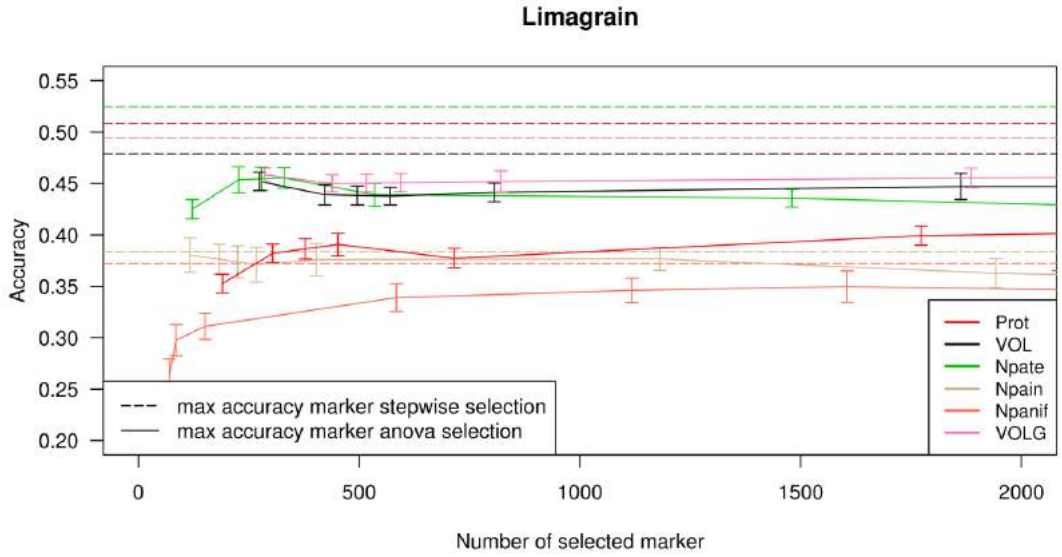


Figure 16 : Gain de précision de prédiction entre un modèle G-BLUP avec sélection de type ANOVA ou MLMM

Population de calibration : 644 lignées (population cible de 71)

Pointillés : précision de prédiction maximale obtenue avec sélection de type MLMM

Lignes pleines : précision de prédiction maximale obtenue avec sélection par ANOVA

Barre d'erreur : écart-type associé à chaque valeur

2) Précision des prédictions en fonction de la taille de la population de calibration

Pour les deux jeux de données, la précision des prédictions augmente avec le nombre de lignées dans la population de calibration et n'atteint pas un plateau. Les écart-types diminuent avec l'augmentation de la taille de la population de calibration. (Figure 14 et S7 b).

3) Comparaison des prédictions entre modèles G-BLUP et BAYES- Cπ

Un modèle BAYES- Cπ a été testé sur le jeu de données Limagrain avec 10 000 SNP. Les résultats n'ont pas montré de différence significative entre les précisions de prédiction obtenues avec les deux modèles (Table S15).

4) Comparaison des prédictions avec sélection de variables associées au caractère

Pour tester le niveau d'information apporté par les marqueurs non associés aux caractères pour la prédiction génomique, nous avons utilisé un nombre croissant de marqueurs triés selon leur significativité dans le test d'association avec le caractère. Dans la continuité de cette idée, le DL étant très étendu (> 6cM) et le nombre de tests indépendants étant très inférieur au nombre de marqueurs utilisés (10 000 pour les données Limagrain et 170 000 pour les données INRA-AO), nous avons testé si la sélection de marqueurs associés indépendants améliorait les prédictions en utilisant les résultats du modèle d'association MLLM qui fait de la sélection de variable stepwise forward backward (Segura *et al*, 2012). Après un tri des marqueurs selon ces deux méthodes, un modèle G-BLUP a été utilisé avec un nombre de marqueurs croissant ayant des pvalues croissantes.

a) Sélection de marqueurs par ordre de significativité décroissant

Pour les données Limagrain, la précision de prédiction en fonction du nombre de marqueurs sélectionnés par ANOVA est représentée sur la Figure 16. Pour tous les caractères, nous observons des courbes en cloche qui présentent un maximum au-delà duquel la précision des prédictions diminue. Le seuil de pvalue pour lequel la précision est maximale est de 0.05 en général, ce qui correspond en moyenne à 1200 marqueurs. Le gain est d'autant plus important que la précision de prédiction de base était faible. Pour les caractères liés à la panification (EXTP, ELAP, DECF, DECL, SECT, % Hydratation), la précision de prédiction maximale est multipliée par 3.65 par rapport à une sélection de marqueurs aléatoire. Le gain est de 1.54 pour les caractères ayant de bonnes précisions de prédiction (Prot, VOL, VOLG, NPATE et NPAIN).

b) Sélection de marqueurs indépendants par ordre de significativité décroissant

Pour les données Limagrain, la précision de prédiction en fonction du nombre de marqueurs indépendants associés au caractère est représentée sur la Figure 16. Nous observons une augmentation de la précision de prédiction par rapport au modèle précédent qui ne prenait pas en compte la non-indépendance entre marqueurs. Le maximum des précisions

de prédiction est obtenu pour un seuil de p-value de 0.1, ce qui correspond à 1023 marqueurs (Table S16) en moyenne pour les différents caractères. Des résultats similaires sont observés sur les données INRA-AO, avec un seuil de pvalue compris entre 0.05 (correspondant à 1430 marqueurs indépendants) et 0.001 (correspondant à 585 marqueurs indépendants) suivant les caractères (Figure S 8).

5) Comparaison des prédictions avec QTLs majeurs en effet fixe

Nous n'avons pas trouvé de QTLs majeurs qui améliore la qualité des prédictions lorsqu'ils sont mis en covariable dans les modèles de sélection génomique.

IV. Discussion et perspectives

1. Gènes candidats pour les caractères de qualité boulangère

La disponibilité de l'annotation dans les prochains mois permettra d'affiner l'interprétation fonctionnelle des QTLs trouvés et de les comparer avec d'autres QTLs publiés détectés dans des panels différents.

Cinq QTLs détectés sont à proximité de gènes candidats. La teneur en protéine est associée à des marqueurs proches (<4 et <1 Mb) des gènes ASR1 et ASR5. La famille de gènes ASR (Abscissic Acid Stress Ripening), et ASR1 et 5 en particulier, est connue pour être impliquée dans la tolérance au stress hydrique chez le blé (Hu *et al*., 2013), le riz (Arenhart *et al*., 2016) et d'autres espèces. Les caractères de gonflement et longueur sont associés à des marqueurs proches (<2Mb) du gène GS2 qui code pour une enzyme de la voie de bio-synthèse des acides aminés (Gadaleta *et al*. 2011). Le caractère de longueur est également associé à un marqueur proche (<1Mb) du gène SAD impliqué dans la biosynthèse des acides gras (Table 6). La Table 5 résume les marqueurs les plus associés aux caractères principaux.

Notre analyse ne nous a pas permis de retrouver les gènes candidats majeurs de gluténines liés à la qualité boulangère. D'après Oury *et al* (2010), les gluténines seules expliquent 37% de la force boulangère W. Le gène Glu- B1 est le plus informatif. Mais plusieurs combinaisons alléliques entre les gènes homologues des 3 génomes sont favorables à la qualité boulangère, ce qui explique que les analyses mono-locus ne les détectent pas. Ces gènes sont en cours de génotypage avec des marqueurs KASPar (Ravel *et al*, in prep) sur le matériel de sélection pour comparer les prédictions en utilisant les gluténines avec les prédictions génomiques.

Notons que la teneur en protéine n'est pas suffisamment corrélée aux variables de qualité boulangère pour la prédire. C'est cependant un critère très important pour l'exportation qui nécessite une teneur minimale de 11.5%. Nous confirmons également que les mesures d'alvéographe sont faiblement corrélées aux index de panification (Oury *et al.*, 1999), ce qui n'en font pas de bons prédicteurs de la panification prises une par une.

2. Perspectives d'amélioration des prédictions

2.1 Améliorer la taille et la qualité de la population de calibration

La note de panification est composée de critères subjectifs du boulanger, comme le comportement de déchirement de la pâte après un coût de couteau, qui sont déterminants pour l'inscription de la variété, mais risquent d'être difficilement prédictibles. Malgré tout, la précision de prédiction de la note de panification est encourageante pour le jeu de données INRA-AO (0.4). De plus, de nombreuses pistes d'amélioration des équations de prédiction sont envisageables.

L'augmentation du nombre de lignées permet une augmentation continue des précisions de prédiction et une diminution des écarts-types entre échantillons pour tous les caractères. Donc 700 lignées ne sont pas suffisantes pour une prédiction de la qualité boulangère. Le panel Limagrain sera étendu à 2304 et le panel INRA sera incrémenté de 200 lignées chaque année. La population de calibration peut également être optimisée afin d'améliorer les précisions de prédiction. Cette optimisation peut se faire sur la base du Coefficient de Détermination (CD), correspondant au carré de l'espérance de la précision de prédiction (Rincent *et al*., 2012). Cette méthode d'optimisation est plus performante lorsque le panel est peu structuré comme chez le blé et que les caractères concernés impliquent peu de gènes majeurs (Isidro *et al*., 2014).

Pour un même nombre de lignées utilisées, les précisions de prédiction restent plus élevées pour les variables du panel INRA-AO que Limagrain. Cette différence de précision de prédiction pourrait venir d'une différence de niveaux d'information des jeux de marqueurs. Cependant la puce 15K Limagrain est un sous ensemble de marqueurs issus de la puce INRA 420K. La diversité phénotypique des panels peut aussi expliquer une part de cette différence, une plus grande diversité dans la population de calibration permettant de mieux prédire des phénotypes variés. Cette différence peut également venir de la qualité des données phénotypiques, notamment d'un nombre de répétions important pour INRA-AO qui permet de calculer des moyennes ajustées et ainsi de s'affranchir des interactions génotype x environnement. Il serait intéressant de tester de manière empirique ou par simulation l'impact du nombre de répétitions sur les qualités de prédiction.

2.2. Tester de nouvelles covariables ou calculer de nouveaux index

Les variables d'alvéographe sont mieux prédites que les index de panification. Cette différence s'explique par des valeurs d'héritabilité plus élevées pour les caractères d'alvéographe. L'héritabilité et la précision de prédiction sont en effet fortement corrélés (r^2 = 0.67, pvalue = 0.023) (Figure S6). Mais nous confirmons (Groos, 2001) que les variables d'alvéographe ne sont pas suffisamment corrélées aux notes de panification (<0.3) pour espérer les utiliser comme prédicteurs. Les informations fournies par ces deux types de tests semblent donc complémentaires. Il serait intéressant de tester si un index construit avec ces variables peut améliorer les qualités de prédiction. Il a été par exemple montré qu'il est possible de prédire certaines propriétés de la pâte à biscuits à partir des variables d'alvéographe et de la teneur en protéine. La prédiction du volume à partir de L, W et de la teneur en protéine, permettrait d'obtenir des corrélations de l'ordre de 0.95 (Bettge *et al*. 1988).

On pourrait de même envisager de définir une combinaison de caractères importants (volume, force boulangère W, haplotypes de gluténines) moins chers à mesurer que les tests de panification, qui permettent de prédire l'appartenance d'une lignée à un groupe de qualité donné (BPS, BP, BAU). On peut imaginer une approche de type ACP, où on prédirait les coordonnées des lignées sur les deux premiers axes, correspondant à une combinaison linéaire de caractères qui distingue les groupes de qualité boulangère. Il pourrait être intéressant de combiner l'utilisation de tests indirects comme l'alvéographe avec les marqueurs pour la prédiction de la valeur boulangère (BIPEA), comme Oury *et al* l'avaient fait avec les marqueurs HMW-Glu.

Bernardo *et al*. 2013, montre qu'il est plus intéressant de rajouter deux à trois gènes majeurs, expliquant plus de 10% de la variance génétique en effets fixes dans un modèle de sélection génomique, plutôt que de les laisser en effet aléatoire. Nous n'avons pas trouvé de gènes majeurs avec les marqueurs dont nous disposions (mais d'autres sont en cours de génotypage). Le blé étant polyploïde, plusieurs combinaisons d'allèles sont possibles entre les 3 génomes pour donner le même phénotype. Ainsi, l'ajout de gluténines de type A dans un modèle G-BLUP ne présente pas de gain particulier, quel que soit le caractère étudié. C'est pourquoi l'ajout d'haplotypes de gluténine A B et D sera testé en covariable.

2.3. Utiliser des haplotypes locaux plutôt que des SNPs

L'approche haplotypique est plus efficace que l'approche single marqueur pour les études d'association en termes de pvalue et de nombre de QTLs détectés. Il faudrait prouver par simulation qu'il s'agit bien d'un gain de puissance et pas d'une augmentation du nombre de faux positifs. Cette méthode permettra d'avoir une meilleure interprétation biologique et évolutive des QTLs grâce à l'identification de plusieurs haplotypes à effets différents appelés séries alléliques (Lin *et al.*2012). Il serait intéressant de tester l'efficacité de cette approche en sélection génomique, dans le cadre d'un futur projet SELGEN (Servin, Bouchet *et al*.).

3. Nécessité de tester la portabilité des équations de prédiction

Les différents échantillonnages de marqueurs montrent que 5000 marqueurs sont suffisants pour faire des prédictions génomiques avec des panels élite. La sélection de marqueurs associés indépendants avec la méthode MLLM pour calibrer un modèle de prédiction améliore très significations les précisions, en les doublant pour certains caractères. Donner un poids identique à chaque QTL apparaît important. Utiliser l'ensemble des marqueurs semble bruiter les prédictions de la qualité boulangère. N'étant pas contrôlée par un très grand nombre de gènes à effet faible, un modèle infinitésimal ne serait donc pas adapté. Par contre, le risque de choisir un sous-ensemble de marqueurs est de sur-ajuster les modèles et de diminuer leur portabilité. Bien que l'on ne dispose pas des noms des marqueurs Limagrain et qu'on ne puisse pas comparer finement les QTLs obtenus avec le panel INRA-AO, peu de QTLs semblent communs. Ceci peut être un problème pour la portabilité des prédictions, notamment si on préconise une sélection de marqueurs associés aux caractères pour construire l'équation. Dans ce projet, la population de calibration et la population de cible sont issues du même jeu de données. Afin de valider les modèles de sélection génomique, il faudra regarder

la portabilité des équations de prédictions sur des jeux de données différents (core-collection/INRA-AO, INRA-AO/Limagrain, INRA-AO/Arvalis).

BIBLIOGRAPHIE :

Alexander D. H., Novembre J., & Lange K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9), 1655-1664.

Arcade A., Labourdette A., Falque M., Mangin B., Chardon F., Charcosset A., & Joets J. (2004). BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. Bioinformatics, 20(14), 2324-2326.

Arenhart R. A., Schunemann M., Bucker Neto L., Margis R., Wang Z.-Y., and Margis-Pinheiro M. (2016) Rice *ASR1* and *ASR5* are complementary transcription factors regulating aluminium responsive genes. Plant, Cell and Environment, 39: 645–651.

Bettge A., Rubenthaler G. L., & Pomeranz Y. (1989). Alveograph algorithms to predict functional properties of wheat in bread and cookie baking. Cereal Chem, 66(2), 81-86.

Bernardo R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Science, 48(5), 1649-1664.

Bernardo R. (2013). Genomewide selection when major genes are known. Crop Science, 54(1), 68-75.

Bonjean A. P., Angus W. J., (2001). The World Wheat Book: A History of Wheat Breeding, Lavoisier

Calus M. P. L., & Veerkamp R. F. (2007). Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics, 124(6), 362-368.

Chao S., Zhang W., Dubcovsky J., & Sorrells M. (2007). Evaluation of Genetic Diversity and Genome-wide Linkage Disequilibrium among US Wheat (L.) Germplasm Representing Different Market Classes. Crop Science, 47(3), 1018-1030.

Choulet F., Alberti A., Theil S., Glover N., Barbe V., Daron J., ... & Leroy P. (2014). Structural and functional partitioning of bread wheat chromosome 3B. Science, 345(6194), 1249721.

Debiton, C. (2010). Identification des critères du grain de blé (Triticum aestivum L.) favorables à la production de bioéthanol par l'étude d'un ensemble de cultivars et par l'analyse

protéomique de lignées isogéniques waxy (Doctoral dissertation, Université Blaise Pascal-Clermont-Ferrand II).

Endelman J.B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255.

Escalante A. M., Coello G., Eguiarte L. E., & Pinero D. (1994). Genetic structure and mating systems in wild and cultivated populations of Phaseolus coccineus and P. vulgaris (Fabaceae). American Journal of Botany, 1096-1103.

Feillet P. (2000). In Lipides du grain de blé, INRA éditions: pp 114-121.

Gadaleta A., Nigro D., Giancaspro A., & Blanco A. (2011). The glutamine synthetase (GS2) genes in relation to grain protein content of durum wheat. *Functional & integrative genomics, 11*(4), 665-670.

Gao L. Z., Ge S., & Hong D. Y. (2000). Low levels of genetic diversity within populations and high differentiation among populations of a wild rice, Oryza granulata Nees et Arn. ex Watt., from China. International Journal of Plant Sciences, 161(4), 691-697.

Geleta B., Atak M., Baenziger P. S., Nelson L. A., Baltenesperger D. D., Eskridge K. M., ... & Shelton D. R. (2002). Seeding rate and genotype effect on agronomic performance and end-use quality of winter wheat. Crop Science, 42(3), 827-832.

Goudet J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes, 5(1), 184-186.

Groos Cyril., *Analyse génétique de critères de qualtié en panification française chez le blé tendre*, thèse de doctorat, Ressources génétiques et amélioration des plantes, Clermont-Ferrand, 2011

Habier D., Fernando R. L., Kizilkaya K., & Garrick D. J. (2011). Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics, 12(1), 1.

Henderson C. R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics, 423-447.

Hu W., Huang C., Deng X., Zhou S., Chen L., Li Y., Wang C., Ma Z., Yuan Q., Wang Y., Cai R., Liang X., Yang G. and He G. (2013), *TaASR1*, a transcription factor gene in wheat, confers drought stress tolerance in transgenic tobacco. Plant Cell Environ, 36: 1449–1464.

Isidro J., Jannink J. L., Akdemir D., Poland J., Heslot N., & Sorrells M. E. (2015). Training set optimization under population structure in genomic selection. Theoretical and applied genetics, 128(1), 145-158.

Jia Y. Q., Masbou V., Aussenac T., Fabre J. L., & Debaeke P. (1996). Effects of nitrogen fertilization and maturation conditions on protein aggregates and on the breadmaking quality of Soissons, a common wheat cultivar. Cereal chemistry, 73(1), 123-130.

Johansson E., Prieto-Linde M. L., & Jönsson J. Ö. (2001). Effects of wheat cultivar and nitrogen application on storage protein composition and breadmaking quality. Cereal Chemistry, 78(1), 19-25.

Lê S., Josse J. & Husson F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. 25(1). pp. 1-18.

Li J., & Ji L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity, 95(3), 221-227.

Lin Z., Li X., Shannon L. M., Yeh C. T., Wang M. L., Bai G., ... & Doebley, J. (2012). Parallel domestication of the Shattering1 genes in cereals. Nature genetics, 44(6), 720-724.

McFadden E.S. et Sears E.R. (1946) The origin of Triticum spelta and its free-threshing hexaploid relatives. Journal of Heredity 37, 81–91

Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001; 157:1819–29.

Oury F. X., Chiron H., Pichon M., Giraud A., Bérard P., Faye A., ... & Rousset M. (1999). Reliability of indirect selection in determining the quality of bread wheat for French bread-baking. Agronomie, 19(7), 621-634.

Oury F. X., Chiron H., Faye A., Gardet O., Giraud A., Heumez E., ... & Branlard G. (2010). The prediction of bread wheat quality: joint use of the phenotypic information brought by technological tests and the genetic information brought by HMW and LMW glutenin subunits. Euphytica, 171(1), 87-109.

Perretant M. R., Cadalen T., Charmet G., Sourdille P., Nicolas P., Boeuf C., ... & Bernard S. (2000). QTL analysis of bread-making quality in wheat using a doubled haploid population. Theoretical and Applied Genetics, 100(8), 1167-1175.

Pritchard J. K., Stephens M., Rosenberg N. A., & Donnelly P. (2000). Association mapping in structured populations. The American Journal of Human Genetics, 67(1), 170-181.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A., Bende, D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3), 559-575.

Rincent R., Laloë D., Nicolas S., Altmann, T., Brunel D., Revilla P. & Schön, C. C. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics, 192(2), 715-728.

Roussel P. et Chiron H. (2002). Les pains français. – MAÉ-ERTI Editeurs. 433pp.

Roussel P., 1989, Contribution à la normalisation et à la codification des essais de panification et des critères d'appréciation de la valeur boulangère. Mémoire d'ingénieur DPE, ENSMIC Paris.

Segura V., Vilhjálmsson B. J., Platt A., Korte A., Seren Ü., Long Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics, 44(7), 825-830.

Shewry, P. (2009). Wheat. Journal of experimental botany, 60(6):1537.

Singh B. D., & Singh A. K. (2015). Marker-assisted plant breeding: principles and practices. Springer.

Surget A., & Barron C. (2005). Histologie du grain de blé. Industries des céréales, (145), 3-7.

Whittaker J. C., Thompson R., & Denham M. C. (2000). Marker-assisted selection using ridge regression. Genetical research, 75(02), 249-252.

Xu Y., & Crouch, J. H. (2008). Marker-assisted selection in plant breeding: from publications to practice. Crop Science, 48(2), 391-407.

Yu J., Pressoir G., Briggs W. H., Bi I. V., Yamasaki M., Doebley J. F. & Kresovich, S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38(2), 203-208.

SITOGRAPHIE:

AGRESTE, (2015) AGRESTE Conjoncture : Grande cultures et fourrages http://www.agreste.agriculture.gouv.fr/IMG/pdf/conjinforap201505cult.pdf (consulté le 01/07/2016)

Arvalis, (2015), Qualité boulangère des blés français : à l'entrée des silos de collecte (consulté le 01/03/2016)

FAOSTATS, (2015), Food and Agricultural commodities production, http://faostat3.fao.org/browse/rankings/commodities_by_regions/E (consulté le 01/07/2016)

France AgriMer, (2016), Céréales, http://www.franceagrimer.fr/filiere-grandes-cultures/Cereales (consulté le 01/09/2016)

FSOV, (2014), Etablissement d'un modèle de sélection génomique pour la qualité boulangère des blés, http://www.fsov.org/mod-select-genom-qual-boulang.html (consulté le 01/07/2016)

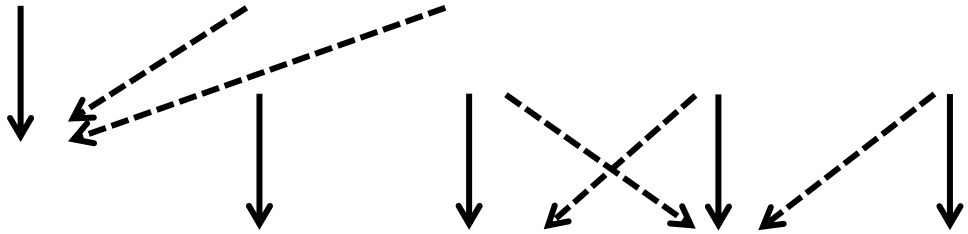
Pasion ceréales (2015), Le blé tendre, http://www.passioncereales.fr/dossier-thematique/le-bl%C3%A9-tendre (consulté le 01/07/2016)

URGI-INRA-Versailles, (2010), URGI, https://wheat-urgi.versailles.inra.fr/ (consulté le 01/07/2016)

Wheat genome, (2010), International Wheat genome sequencing Consortium, http://www.wheatgenome.org/ (consulté le 01/07/2016)

Annexes :

Table S 1 : Classes et critère d'inscription des blés selon leurs utilisations (Bonjean *et al.* 2001)

Classe de blés	BAU	BB	BP	BPS	BA
	Teneur en Protéine – Indice de Zélény				
	Alvéographe - dureté du grain				
Test effectué	Test européen de machinabilité	Test biscuitier	Teste de panification (BIPEA)		Test de panification (en tant que correcteur)
Conclusion					
Classe d'inscription	BAU	BB	BP	BPS	BA
Rendement exigé (% des témoins)	106	103	103	100	90

BA : Blés Améliorants

BP(S) : Blés Panifiables (Supérieurs)

BB : Blés Biscuitiers

BAU : Blés Autres Usages

Table S2 : Distribution des variétés communes d'une année à l'autre (données INRA-AO)

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
2000	**15**													
2001	13	**28**												
2002	7	13	**22**											
2003	2	3	10	**37**										
2004	1	2	4	8	**32**									
2005	1	1	2	3	9	**24**								
2006	1	1	1	1	6	12	**38**							
2007	1	1	2	2	5	6	17	**58**						
2008	1	1	1	1	3	2	6	21	**47**					
2009	1	1	1	1	2	1	3	5	18	**44**				
2010	1	1	1	1	3	2	3	4	8	16	**55**			
2011	0	0	0	0	1	0	1	2	4	5	17	**48**		
2012	0	0	0	0	1	0	0	1	2	3	6	19	**55**	
2013	0	0	0	0	1	0	0	1	2	3	3	11	29	**75**

Nombre de lignées communes tous lieux confondus (Rouge : 10 ou plus / Orange : entre 5 et 10 / Jaune : entre 2 et 5 / vert : 1 / bleu : 0)

Table S3: Distribution des lignées communes d'un site à l'autre (données INRA-AO)

	CF	DI	EM	LM	LU	OR	RE
CF	**371**	262	355	324	28	145	60
DI	262	**262**	256	217	26	106	24
EM	355	256	**355**	311	26	145	55
LM	324	217	311	**324**	27	101	60
LU	28	26	26	27	**28**	13	20
OR	145	106	145	101	13	**145**	14
RE	60	24	55	60	20	14	**60**

Nombre de lignées communes d'un site à l'autre toutes années confondues

Observations laboratoire		Insuffisance 1	7	4	10	Excès 7	4	1	Coef.	Note x Coef.
Evaluations « note pâte »										
Pétrissage	Rapidité Lissage								0,5	
	P- Collant Pâte								0,5	
	Extensibilité								0,5	
	Elasticité								0,5	
	P- Relâchement								0,5	
		Note sur 25 =								*
Pointage	Détente / relâchement								1	
		Note sur 10 =								
Façonnage	**P-** Allongement								1	
	Elasticité								0,5	
	P- Collant Pâte								1	
		Note sur 25 =								*
Apprêt	Action fermentaire								0,5	
	Pate Déchirement								0,5	
		Note sur 10 =								
Mise au four	**P-** Collant Pâte									
	P- Tenue Pâte									
		Note sur 30 =								*
Evaluations « note pain »										
Expansion	Volume	(cm^3)				Note sur 30 =				
Aspect du pain	Section								1	
	Couleur								2	
	Epaisseur								0,5	
	Crosutillant								0,5	
	P- Développement CL								1	
	Régularité CL								1	
	Déchirement CL								1	
		Note sur 70 =								*
Evaluations « note mie »										
Aspect de la mie	Couleur									
	Souplesse texture									
	Elasticité texture									
	Collant texture									
	Régularité A									
	Epaisseur A									
	Flaveur									
		Note sur 100 =								

* Les notes globales pour les domaines d'évaluation (« Note sur 25 » etc.) sont chacune constituée de la somme des valeurs (note x coefficient) multipliée par un éventuel coefficient d'abattement déterminé par les critères prépondérants (indiqués par un suffixe P-)

Figure S1 : Feuille de notation pour le test de panification (méthode BIPEA)

Table S4 : Les différents aspects évalués en panification

La consistance	l'état de fermeté de la pâte, elle s'apprécie par enfoncement progressif des doigts dans la pâte, la pâte se déforme en s'écoulant, on peut alors évaluer ses caractéristiques visqueuses. Cette caractéristique est jugée en fin de pétrissage.
Le collant	il s'apprécie à plusieurs étapes de la fabrication : en fin de pétrissage (par contact avec le dos de la main), au façonnage (par adhérence sur la surface de repos) et à la mise au four (adhérence sur la toile de deuxième fermentation), il est toujours jugé en excès (l'absence d'adhérence étant un caractère normal). Ce phénomène a pour origine un excès d'hydratation, une mauvaise qualité des protéines, une prise de force insuffisante et/ou une humidité relative excessive du local.
Le relachement	il s'agit de l'écoulement de la pâte sous son propre poids (tenue insuffisante de la pâte). Il s'apprécie en fin de pétrissage, en fin de première fermentation et à la mise au four. Il est toujours noté en excès.
L'élasticité	capacité d'un corps à reprendre totalement ou partiellement sa forme initiale. Elle s'apprécie en fin de pétrissage et en fin de façonnage.
Le développement	il permet de juger de l'activité fermentaire, mais également de l'aptitude à la déformation et la rétention gazeuse.
L'extensibilité	il s'agit d'apprécier les capacités d'allongements ou de déformations de la pâte généralement jusqu'à un stade de rupture. Elle s'apprécie à différents stades : - le lissage au stade pétrissage : il est en lien avec la formation de la structure gluténique et ses capacités d'extensions, l'aspect lisse est un caractère normal. - l'extensibilité au stade pétrissage : le caractère normal est défini pour une rupture, après un étirement de 20 à 30 cm. - l'allongement et déchirement au façonnage : l'allongement d'un pâton définit l'extensibilité au stade façonnage, il est jugé normal si en sortie de la façonneuse la longueur du pâton atteint 32 cm. On note également les déchirures de surface. - déchirement à l'apprêt : s'apprécie par la présence de craquelures ou de petits trous, l'absence est jugée normal, et ce défaut est jugé uniquement en excès

Autres caractères	Volume, couleur, flaveur, texture

$$NPAT = (0.5\ x\ LISP + \boldsymbol{COLP}x0.5 + EXTPx0.5 + ELAPx0.5 + \boldsymbol{RELP}x0.5)\ x\ coef(*) + REPO + (\boldsymbol{EXTF}x0.5 + DECFx0.5 + ELAFx0.5 + \boldsymbol{COLF})\ x\ coef(*) + (ACTPx0.5 + DECAx0.5) + (\boldsymbol{COMF} + \boldsymbol{TENU}x2)\ x\ coef(*)$$

$$NPAI = (SECT + COUPx2 + EPAPx0.5 + CROPx0.5 + \mathbf{DVCL} + \mathrm{RECL} + \mathrm{DECL})\ x\ coef(*) + VOL$$

$$NMIE = (SOUMx2 + ELAM + COLM) + (REAL + EPALx2 + FLAVx3)$$

$$NPANI = NPAT + NPAI + NMIE$$

Note du critère prépondérant	10	7	4	1
Coef (*)	1	0.75	0.5	0.25

Figure S 2 : Combinaison linéaire des différentes notes de base en panification aboutissant à la note de pâte, pain, mie et de panification (Roussel et Chrion, 2002)

Table S 5 : Corrélations positives (>0.5) et négatives (<-0.5) entre les variables issues du jeu de données Limagrain et INRA-AO

Variable 1	Variable 2	Corrélation (Limagrain)	p-value (Limagrain)	Corrélation (INRA-AO)	p-value (INRA-AO)
Volume	Volume/masse	0,99	0		
Note de pain	Note de panification	0,83	2.30 e-200	0,85	1.33 e-108
Note de pain	Volume	0,74	2.03 e-138	0,87	6.04 e-116
Note de pain	Volume/masse	0,73	9.29 e-134		
Note de panification	Note de mie	0,66	1.48 e-98	0,78	5.63 e-72
Extensibilité pétrissage	Extensibilité façonnage	0,66	7.94 e-99		
Longueur de la courbe (L)	Gonflement (G)			0,74	1.06 e-64
Note de pâte	Note de mie			0,73	1.88 e-59
Note de panification	Note de pâte	0,65	1.85 e-94	0,88	7.94 e-122
Note de panification	Volume	0,57	1.24 e-69	0,71	1.98 e-59
Force boulangère (W)	Pression max (P)			0,65	5.84 e-45
Gonflement (G)	Ténacité (PsurL)			0,61	3.02 e-39
Note de pâte	Note de pain			0,54	1.40 e-30
Note de panification	Volume/masse	0,55	5.51 e-63		
Note de pain	Note de mie	0,53	1.33 e-57		
Note de pâte	Force boulangère (W)			0,5	6.29 e-25
Volume	Déchirement-façonnage	-0,51	1.90 e-52		
Volume/masse	Déchirement-façonnage	-0,52	4.73 e-56		
Pression max (P)	Longueur de la courbe (L)			-0,56	5.20 e-32
Note de pâte	Extensibilité-pétrissage	-0,64	6.59 e-92		
Note de pâte	Extensibilité-façonnage	-0,76	3.35 e-149		

Table S6 : Corrélation variables phénotypiques (Limagrain)

	PROT	% Hydratation	Vol	VOLG	NPATE	NPAIN	NMIE	NPANI	LISP	EXTP	ELAP	EXTF	DECF	SECT
PROT	1													
% Hydratation	0,27	1												
Vol	0,24	0,29	1											
VOLG	0,23	0,3	**0,99**	1										
NPATE	0,07	-0,1	0	-0,03	1									
NPAIN	0,19	0,21	**0,75**	**0,74**	0,15	1								
NMIE	0	0,07	0,46	0,45	0,25	**0,53**	1							
NPANI	0,16	0,1	**0,58**	**0,55**	**0,65**	**0,83**	**0,66**	1						
LISP	-0,03	0,02	-0,03	**-0,02**	-0,06	-0,12	-0,08	-0,12	1					
EXTP	-0,07	0,11	0,08	0,1	-0,64	-0,04	-0,1	-0,37	0,04	1				
ELAP	0,01	-0,1	-0,07	-0,07	0,39	0,01	0,04	0,21	0,02	-0,38	1			
EXTF	-0,04	0,14	0,06	0,09	**-0,76**	-0,04	-0,09	-0,43	0,04	**0,66**	-0,36	1		
DECF	-0,15	-0,27	**-0,51**	**-0,53**	0,16	-0,44	-0,24	-0,25	0,02	-0,26	0,12	-0,26	1	
SECT	0,07	0,04	0,03	0,03	0,04	-0,1	-0,04	-0,05	-0,06	0,03	-0,03	-0,02	0,03	1

Table S7 : Corrélation phénotypique (INRA-AO)

	prot	NPATE	NPAIN	NPANI	Vol	NMIE	W	P	L	G	PsurL
Prot	1,00										
NAPTE	-0,01	1,00									
NPAI	-0,08	**0,55**	1,00								

NNPANI	-0,07	**0,88**	**0,86**	1,00							
				Groupe 1		**Groupe 2**			**Groupe 3**		
Vol	0,00	0,44	**0,87**	**0,72**	1,00						
NMIE	-0,05	**0,74**	0,49	**0,78**	0,35	1,00					
W	0,34	**0,51**	0,29	0,44	0,27	0,37	1,00				
P	0,14	0,19	0,08	0,13	-0,02	0,09	**0,65**	1,00			
L	0,12	0,25	0,29	0,32	0,40	0,25	0,06	**-0,57**	1,00		
G	0,15	0,30	0,21	0,30	0,31	0,28	0,25	-0,19	**0,75**	1,00	
PsurL	0,12	0,09	-0,02	0,03	0,00	-0,20	0,31	0,44	-0,03	**0,62**	1,00

Table S 8 : Corrélation entre les moyennes ajsutées issu d'un modèle BLUE et d'un modèle BLUP (données INRA-AO)

Trait	Corrélation entre les deux modèles
P	0,99
W	0,99
L	0,98
G	0,98
PsurL	0,99
Vol	0,97
NPanif	0,94
Npate	0,90
Npain	0,93
Nmie	0,88
Prot	0,95

Table S 9 : Test de chi2 faisant ressortir les variables caractérisant le mieux la partition et les groupes d'ACP (données INRA-AO)

Traits	Eta2	Pvalue	v.test	p.value	v.test	p.value	v.test	p.value
Ntot	**0,577**	**2,31 e-61**	**-13,49**	**1,75e-41**	4,57	4,78e-06	**8,79**	**1,47e-18**
P	**0,477**	**2,09 e-46**	-5,6	2,13e-08	**12,45**	**1,28e-35**	-6,97	2,98e-12
L	**0,457**	**1,08 e-43**	-2,87	4,03e-03	**-8,83**	**9,83e-19**	**11,75**	**6,99e-32**
G	**0,456**	**1,18 e-43**	-3,15	1,61e-03	**8,63**	**5,92e-18**	**11,82**	**3,01e-32**
Npate	**0,446**	**2,15 e-42**	**-11,82**	**2,94e-32**	3,8	1,40e-04	7,9	2,63e-15
Npain	**0,437**	**2,91 e-41**	**-11,84**	**2,39e-32**	4,53	5,76e-06	7,19	6,38e-13
Nmie	0,382	1,13 e-34	-10,84	2,07e-27	3,1	1,88e-03	7,64	2,16e-14
PsurL	0,350	4,57 e-31	NA	NA	**9,32**	**1,13e-20**	**-9,25**	**2,24e-20**
Volume	0,318	9,74 e-28	-9,89	4,45e-23	2,81	4,90e-03	6,99	2,72e-12
W	0,275	2,37 e-23	-13,49	5,14e-20	6,65	2,90e-11	2,4	1,63e-02

Table S 10 : Test de chi2 faisant ressortir les variables caractérisant le mieux la partition et les groupes d'ACP (données Limagrain)

			Groupe 1		**Groupe 2**		**Groupe 3**	
Traits	**Eta2**	**P-value**	**v.test**	**p.value**	**v.test**	**p.value**	**v.test**	**p.value**
NPAIN	**0,575**	**5,58 e-144**	**-21,07**	**1,40 e-98**	3,22	1,27 e-03	9,38	6,54 e-21
NPANI	**0,519**	**3,17 e-123**	**-17,43**	**4,21 e-68**	-6,68	2,37 e-11	**16,34**	**4,76 e-60**
EXTP	**0,487**	**1,24 e-112**	-	-	**19,02**	**1,09 e-80**	**-17,75**	**1,49 e-70**
EXTF	**0,484**	**1,36 e-111**	-	-	**18,96**	**3,62 e-80**	**-17,7**	**4,09 e-70**
NPAT	**0,454**	**4,42 e-102**	-	-	**-18,12**	**1,86 e-73**	**17,53**	**8,35 e-69**
VOLG	0,39	1,68 e-83	**-17,01**	**6,19 e-65**	6,33	2,40 e-10	4,14	3,32 e-05
VOL	0,388	4,43 e-83	**-17,09**	**1,75 e-65**	5,76	8,03 e-09	7,41	2,45 e-06
ELAP	0,335	4,27 e-69	-	-	**-15,89**	**7,05 e-57**	**14,45**	**2,35 e-47**
DECF	0,302	5,80 e-61	**13,66**	**1,57 e-42**	-9,09	9,64 e-20	-	-
NMIE	0,299	3,29 e-60	**-15,17**	**5,02 e-52**	-	-	7,26	3,65 e-13
RECL	0,291	1,86 e-58	**15**	**6,52 e-51**	-2,19	2,81 e-02	-6,77	1,26 e-11
Pourcentage hydra	0,044	2,02 e-08	-4,15	3,23 e-05	4,83	1,35 e-06	-2.00	4,52 e-02
Prot	0,016	1,88 e-03	-3,25	1,14 e-03	NA	NA	2,63	8,41 e-03
SECT	0,013	5,11 e-03	2,19	2,79 e-02	1,96	4,91 e-02	-3,09	1,98 e-03

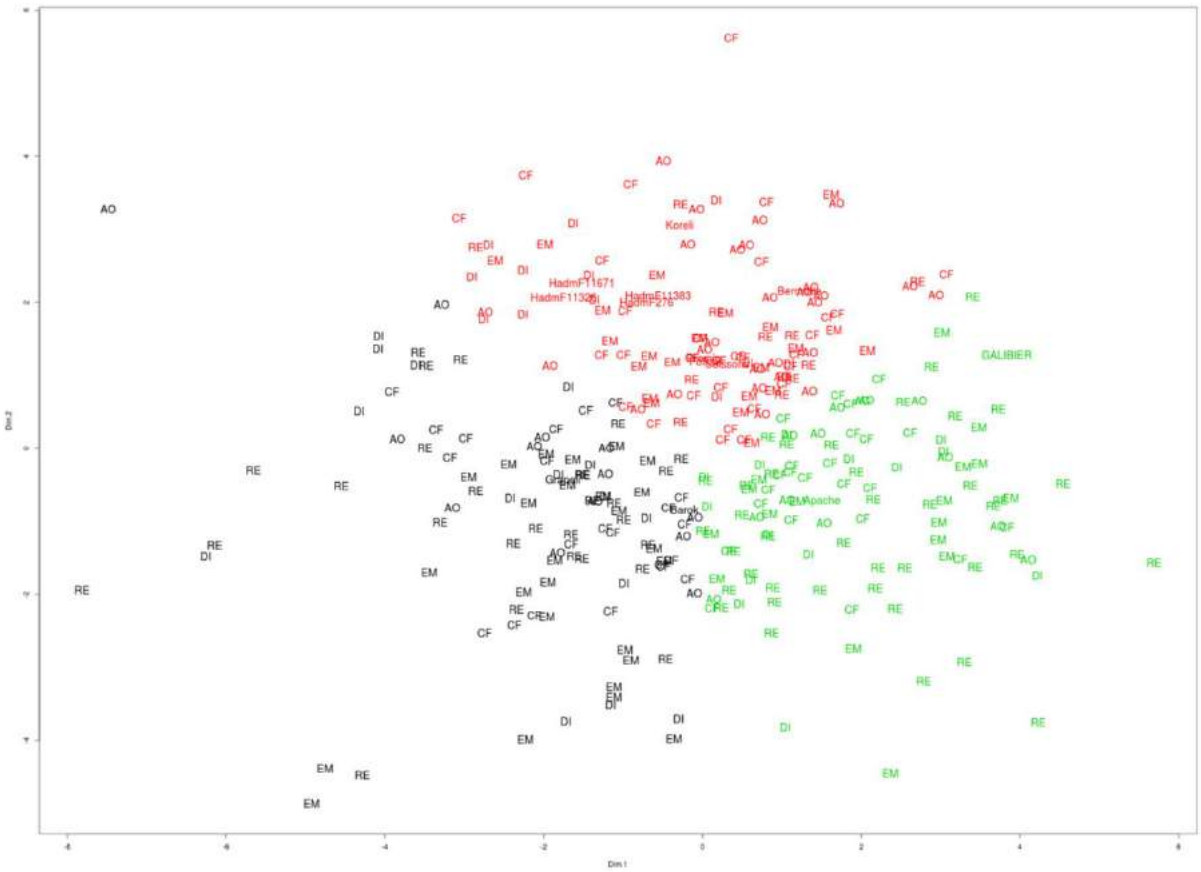


Figure S3: Classification ascendante hiérarchique (données INRA-AO)

Lieu d'obtention : RE : Rennes ; CF : Clermont-Ferrand ; AO : Agri-Obtention ; D : Dijon ; EM : Estrées-Mons

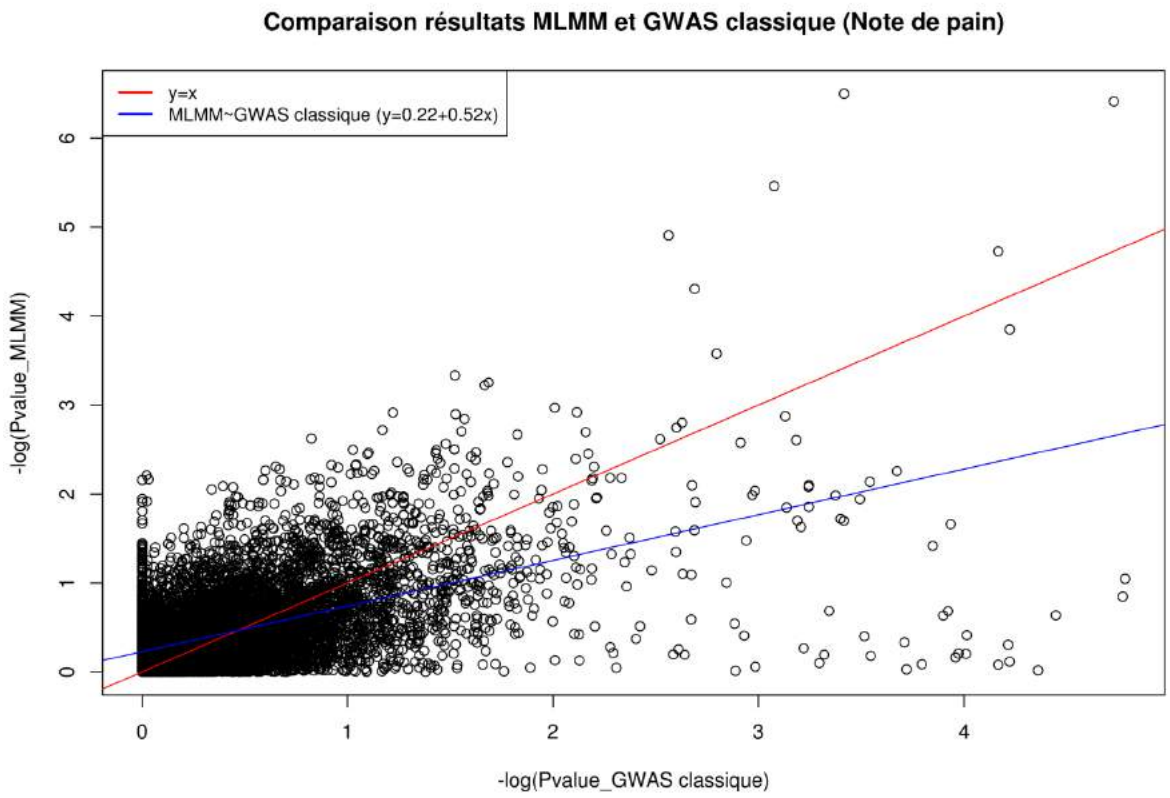


Figure S 4 : Comparaison des résultats après une GWAS classique et après analyse MLMM (résultat note de pain, données Limagrain)

Les valeurs sont corrélées à hauteur de 0.53. L'analyse MLMM, permet de contrôler le nombre de marqueurs apparaissant comme significatifs, et augmente leur pvalue.

Table S 11 : Marqueurs associés les plus significatifs des principaux caractères de qualité boulangère (Données INRA-AO)

	Maqueurs	Chromosome	Position génétique	Pvalue
W	cfn1313476_G	7A	166,86	6,79E-17
	cfn2823590_G	4B	77,75	2,08E-24
	Q	6D	146,48	6,09E-16
VOL	R	6B	59,61	4,35E-35
	cfn1821727_G	1D	59	2,31E-24
	cfn0336967_G	3D	49,17	1,94E-22
NPAIN	cfn3126111_C	5A	82,52	1,17E-10
	cfn0255654_C	6A	19,85	1,91E-12
	cfn0771948_C	3D	29,48	9,98E-12
Prot	cfn0707514_C	2A	61,67	2,61E-18
	cfn0592506_C	5B	131,57	8,36E-16
	cfn0531717_G	6D	7,16	2,44E-15
	cfn0535184_C	6D	22,39	9,38E-19
	cfn1575732_G	7D	11,88	3,29E-32

	cfn0628857_C	7D	84,05	6,74E-26
P	cfn0747681_G	3A	40,58	3,29E-16
	cfn0933547_C	6A	110,16	3,20E-19
	cfn1309728_G	7A	165,54	3,14E-30
	BS00022768_51_G	7A	166,86	9,89E-57
	cfn2823590_G	4B	77,75	2,45E-53
	cfn1689793_C	1D	54,66	7,07E-44
	cfn0658419_T	1D	48,35	6,05E-25
	cfn0763934_G	3D	55,63	4,15E-23
	cfn0172945_C	4D	114,38	3,81E-18
	cfn0892630_G	5D	25,73	1,06E-32
	cfn0397291_C	5D	70,25	4,14E-43
	cfn0540010_G	6D	134,58	1,83E-21
	cfn0532062_C	6D	18,31	1,59E-59
	S	7D	151,95	3,43E-29
Npanif	cfn0658353_T	1D	19	2,04E-09
	cfn1003563_C	1A	55,69	2,92E-11

Table S12 : Marqueurs associés significativement aux variables de qualité boulangère, avec un seuil Bonferroni à 5%, (données Limagrain) (MLMM)

Traits	**Marqueur**	**Chromosome**	**Position (cM)**	**p-value**
Protéine	WC.0220141	3	11,7	2,5876e-06
	WC.0225880	3	52,2	4,3308e-08
	WC.0226485	4	136,4	2,5003e-06
	WC.0217572	11	59,0	4,3771e-08
	WC.0227558	12	78,9	1,9827e-06
	WC.0213051	18	25,5	2,0700e-12
	WC.0228558	20	139,9	1,3108e-06
NPAT	WC.0220881	4	136,4	1,7912e-06
	WC.0223685	7	91,3	2,4061e-07
	WC.0213263	8	101,1	4,5105e-07
	WC.0224844	15	66,2	3,7445e-19
VOL	WC.0221203	4	44,9	4,9380e-07
	WC.0213425	7	95,1	2,5642e-08
	WC.0228921	7	70,5	5,0777e-07
	WC.0214627	10	22,4	2,8600e-06
	WC.0215966	10	48,2	2,0377e-06
	WC.0216017	15	63,6	9,5723e-08
	WC.0225618	16	58,5	1,6096e-07
EXTF	WC.0223685	7	91,3	1,2969e-06
	WC.0226828	11	23,5	3,3716e-06
	WC.0224844	15	66,2	1,8209e-28

	WC.0227835	19	160,2	1,3406e-07
	WC.0229391	21	88,0	2,7593e-09
	WC.0229411	21	88,0	3,7162e-06
DECF	WC.0218362	1	4,4	7,8799e-06
	WC.0224844	15	66,2	5,5328e-13
DECL	WC.0216825	5	65,6	4,8002e-06
	WC.0215903	13	33,9	8,6130e-08
	WC.0218809	15	65,7	2,6104e-08
ELAP	-	-	-	-
EXTP	WC.0213187	1	65,1	3,3377e-08
	WC.0218743	8	30,8	4,2346e-08
	WC.0224652	8	92,7	5,1469e-10
	WC.0213104	9	63,7	2,9320e-09
	WC.0229301	14	18,7	9,5468e-07
	WC.0224844	15	66,2	1,5145e-14
LISP	-	-	-	-
NPAIN	WC.0220011	3	118,6	3,1636e-07
	WC.0228864	7	34,9	3,4559e-06
	WC.0219821	16	58,3	3,8463e-07
NPANIF	-	-	-	-
%Hydratation	WC.0220902	4	114,0	2,0835e-08
	WC.0223366	7	146,5	9,2562e-10
	WC.0214190	9	71,4	4,0808e-06
	WC.0213051	18	25,5	1,5476e-10
	WC.0221497	18	37,3	4,9415e-07
	WC.0228528	20	139,9	4,3587e-07
SECT	-	-	-	-
VOLG	WC.0222509	6	54,9	7,9239e-07
	WC.0213425	7	95,1	1,0572e-07
	WC.0228921	7	70,5	2,1126e-06
	WC.0214627	10	22,4	1,7000e-07
	WC.0215966	10	48,2	4,3569e-07
	WC.0216017	15	63,6	2,1106e-10
	WC.0225618	16	58,5	2,2960e-08

Table S13: Marqueurs associés significativement aux variables de qualité boulangère, avec un seuil Bonferroni à 5%, données INRA-AO, analyse MLMM

Traits	Marqueur	Chromosome	Position (cM)	p-value
P	A	21	151,95	3,43E-29
	B	3	81,34	1,06E-07
	cfn0509127_G	1	140,25	1,25E-12
	cfn1309728_G	7	165,54	3,14E-30
	cfn3573083_C	6	56,02	3,91E-08

	cfn0547572_C	20	169,13	1,25E-08
	C	15	144,43	5,75E-09
	cfn0763934_G	17	55,63	4,15E-23
	cfn0933547_C	6	110,16	3,20E-19
	cfn0747681_G	3	40,58	3,29E-16
	cfn0540010_G	20	134,58	1,83E-21
	cfn2455213_C	10	68,03	3,02E-12
	D	7	166,86	9,89E-57
	cfn0172945_C	18	114,38	3,81E-18
	cfn0658419_T	15	48,35	6,05E-25
	cfn0412199_C	19	18,26	5,47E-12
	cfn2456725_C	10	123,72	3,21E-13
	cfn0892630_G	19	25,73	1,06E-32
	cfn3601364_G	14	49,66	2,00E-14
	cfn0281186_G	17	115,66	2,42E-28
	cfn2823590_G	11	77,75	2,45E-53
	cfn0532062_C	20	18,31	1,59E-59
	cfn0829022_G	18	144,68	1,05E-08
	cfn0512892_G	1	155,93	2,48E-07
	cfn1689793_C	15	54,66	7,07E-44
	cfn2969413_C	5	170,25	4,43E-11
	cfn0397291_C	19	70,25	4,14E-43
	cfn2773486_G	4	119,16	2,11E-14
	cfn0163144_C	17	59,36	1,27E-37
	cfn0327481_C	2	148,94	2,23E-06
	cfn0210917_C	15	49,28	1,25E-10
	cfn0420678_G	19	66,99	2,32E-29
	cfn2822606_C	11	107,02	1,60E-20
G	cfn0931108_G	6	94,82	9,27E-07
	cfn2458536_C	17	131,11	7,41E-25
	E	7	165,54	4,58E-12
	cfn0637522_C	21	160,35	2,37E-21
	cfn2171682_G	2	11,52	1,54E-15
	cfn0580479_C	12	147,78	4,74E-09
	cfn0883924_C	13	166,62	3,49E-14
	cfn1905775_G	15	78,11	3,70E-28
	cfn0854482_G	5	155,4	5,30E-08
	cfn2511902_G	17	63,15	2,20E-07
	cfn0927352_C	6	39,69	2,19E-11
	cfn2377732_G	3	138,04	1,07E-09
	cfn0580525_G	21	20,74	6,01E-10
	cfn1705569_C	21	159	1,75E-12
	cfn1207818_C	20	136,29	2,73E-08
	cfn1321258_C	7	96,56	2,55E-13
	cfn0926243_G	6	41,14	8,36E-22
L	F	5	83,98	4,49E-09
	cfn0540583_G	20	117,81	1,23E-18
	cfn0539506_C	20	18,48	1,10E-09
	cfn0902732_G	19	130,98	2,83E-07
	cfn2517148_G	17	131,11	6,62E-15
	cfn0712783_G	2	11,62	6,75E-11

	G	17	69,94	1,15E-09
	cfn1300788_G	7	165,71	8,57E-19
	cfn0985664_G	1	164,85	1,99E-08
	cfn3171720_G	13	90,29	2,47E-07
	cfn0691526_G	2	144,05	2,74E-19
	cfn0220902_G	7	114,21	1,07E-10
	cfn2418915_G	3	40,58	4,45E-13
	cfn0353395_T	20	120,08	6,84E-11
	cfn2215741_G	9	88,76	2,08E-06
PsurL	cfn3168128_C	13	165,83	1,06E-08
	cfn0244750_C	19	103,01	6,56E-10
	H	18	115,82	1,44E-08
W	I	20	146,48	6,09E-16
	cfn2904423_C	18	107,99	3,48E-11
	cfn3521312_G	6	89,5	1,26E-09
	cfn1313476_G	7	166,86	6,79E-17
	cfn2237415_G	16	46,99	2,64E-13
	cfn2823590_G	11	77,75	2,08E-24
	cfn0892467_T	19	65,83	1,07E-14
	cfn0754009_G	3	75,3	6,16E-08
	contig78617_211_BS00047058_G	19	131,1	3,57E-13
	contig74618_275_BS00010625_G	7	165,53	1,30E-08
	cfn0171510_C	2	125,27	7,13E-11
	cfn3305843_G	19	73,92	3,45E-09
Vol.BLUE	cfn0274514_C	15	153,92	7,44E-09
	cfn0371846_G	12	88,67	1,62E-08
	cfn1894898_G	15	85,71	9,81E-09
	cfn0843779_C	18	99,82	1,70E-11
	cfn0419879_G	17	49,5	3,77E-12
	cfn2361372_G	16	95,34	2,30E-12
	K	21	145	2,50E-09
	BS00036102_51_G	13	59,61	4,35E-35
	cfn0748720_T	3	139,43	2,78E-14
	cfn1821727_G	15	59	2,31E-24
	cfn0254435_C	1	29,29	1,46E-07
	cfn2226436_G	16	63,76	2,43E-07
	cfn0807059_G	4	50,89	4,21E-12
	cfn3175004_C	13	171,24	1,42E-09
	cfn0336967_G	17	49,17	1,94E-22
	cfn0765916_C	17	61,05	2,90E-19
	cfn2699375_C	4	55,09	2,91E-12
	cfn2826088_A	11	81,54	4,24E-10
Npain.BLUE	cfn3163730_G	13	59,61	6,19E-08
	cfn0771948_C	17	29,48	9,98E-12
	cfn3126111_C	5	82,52	1,17E-10
	cfn0255654_C	6	19,85	1,91E-12
Npate.BLUE	cfn1037354_C	1	51,04	6,73E-08

	cfn1765308_C	8	54,19	6,73E-08
Npanif.BLUE	cfn0658353_T	15	19	2,04E-09
	cfn1003563_C	1	55,69	2,92E-11
Prot.BLUE	cfn2008789_G	2	140,9	7,94E-14
	cfn0535184_C	20	22,39	9,38E-19
	cfn1131180_G	20	128,86	1,70E-13
	L	18	108,33	5,23E-10
	cfn1575732_G	21	11,88	3,29E-32
	cfn3613803_G	14	25,43	8,05E-14
	cfn0515261_G	1	156,44	1,82E-12
	cfn0398129_C	21	12,51	1,37E-12
	cfn0199242_C	21	171,89	3,09E-08
	cfn0531717_G	20	7,16	2,44E-15
	cfn2975911_C	5	15,33	1,15E-07
	M	5	138,6	3,80E-12
	cfn0592575_G	12	97,77	4,65E-09
	cfn0794677_C	4	63,35	4,13E-12
	cfn1188253_G	20	142,33	1,95E-08
	cfn0774917_C	17	34,35	2,85E-09
	cfn1622646_G	21	15,24	2,79E-07
	cfn0913315_T	19	69,98	2,79E-10
	cfn0441638_C	2	172,17	2,29E-10
	cfn0363209_C	12	140,71	6,42E-12
	cfn0592506_C	12	131,57	8,36E-16
	N	8	80	3,06E-09
	cfn0628857_C	21	84,05	6,74E-26
	cfn2381663_C	3	122,32	7,73E-11
	O	15	141,6	4,81E-16
	cfn0761212_G	17	36,36	7,67E-10
	cfn0146523_C	1	10,26	1,14E-09
	cfn0891517_C	19	41,47	4,43E-11
	cfn0723933_C	16	46,5	6,93E-10
	cfn0750740_C	3	136,41	3,38E-16
	cfn2333065_G	16	15,77	5,68E-08
	cfn0239634_C	17	158,23	4,05E-17
	cfn0707514_C	2	61,67	2,61E-18
	cfn1784696_G	8	153,38	5,05E-08
	cfn0859340_C	5	142,02	6,21E-17
	cfn0358702_C	21	45,21	6,24E-08
	cfn3593778_T	14	40,69	5,07E-13
	cfn2134004_C	2	132,81	6,94E-10
	cfn0501802_T	1	152,9	2,27E-11
	cfn2398522_G	3	138,84	2,72E-13
	cfn3289608_G	19	27,36	9,95E-15
Nmie.BLUE	cfn0686288_C	2	45,58	2,54E-10
	cfn2808309_C	11	114,07	9,83E-10
	cfn0539960_G	20	158,97	3,10E-09
	cfn0559113_G	20	60,6	2,79E-09
	cfn1390470_G	12	34,54	1,12E-07
	cfn0372913_A	8	64,71	1,78E-26

cfn0613793_C	21	23,58	1,71E-07

Table S14 : Comparaison du nombre de QTLs significatifs entre une analyse d'association sur marqueurs seuls ou haplotypes, données Limagrain

Caractère	-log10 (Pv.single.min)	-log10 (Pv.hap.min)	Nb.single 5_1cM	Nb.hap 5_1cM	Nb.single 6_1cM	Nb.hap 6_1cM	Nb.single 5_5cM	Nb.hap 5_5cM	Nb.single 6_5cM	Nb.hap 6_5cM
decf	12.202	10.734	4	19	3	15	2	8	2	6
elap	4.334	5.067	0	1	0	0	0	1	0	0
extf	10.789	10.674	6	25	3	15	4	12	2	6
extp	5.199	6.145	1	11	0	1	1	7	0	1
lisp	3.488	3.98	0	0	0	0	0	0	0	0
nmie	5.391	5.061	2	2	0	0	2	2	0	0
npai	5.496	6.185	2	10	0	1	2	8	0	1
npani	4.933	6.719	0	7	0	1	0	7	0	1
npat	8.185	7.796	9	53	2	19	6	35	2	12
hydra	5.318	7.2	2	16	0	6	2	13	0	5
prot	11.461	12.164	71	533	30	334	48	247	21	176
recl	5.199	6.666	2	7	0	3	2	7	0	3
sect	4.145	5.579	0	2	0	0	0	2	0	0
teneur	8.661	11.113	49	343	24	186	42	172	21	101
volg	6.068	8.992	11	102	1	41	10	70	1	26
vol	5.484	8.405	9	91	0	22	7	66	0	19

single: modèle single locus

hap: modèle avec haplotype local

5 ou 6: seuil de significativité E-05 ou E-06

1cM ou 5 cM: nombre de QTLs dans une fenêtre de 1 ou 5 cM

a)

b)

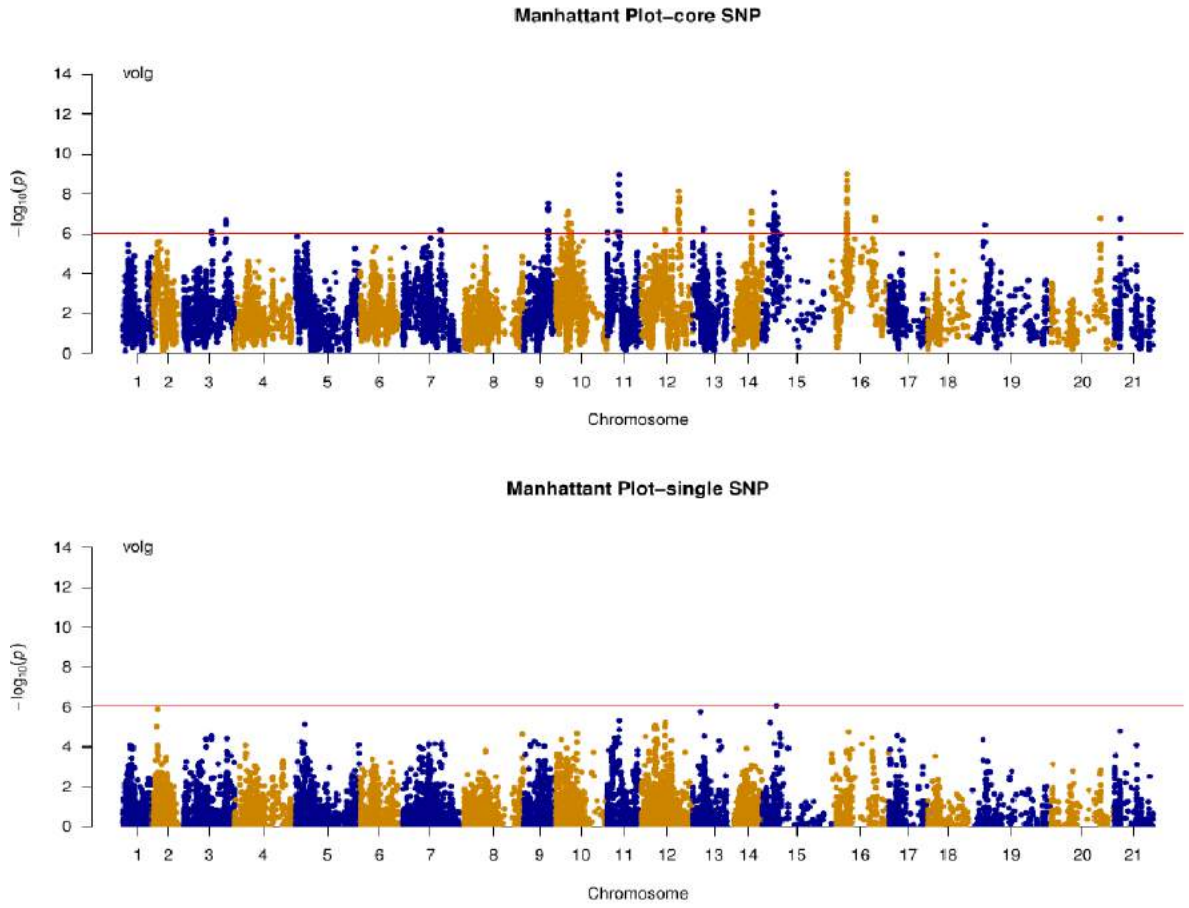


Figure S 5 : Manhattan plot approche haplotypique (a) et marqueur seul (b) (Données Limagrain, caractère VOLG)

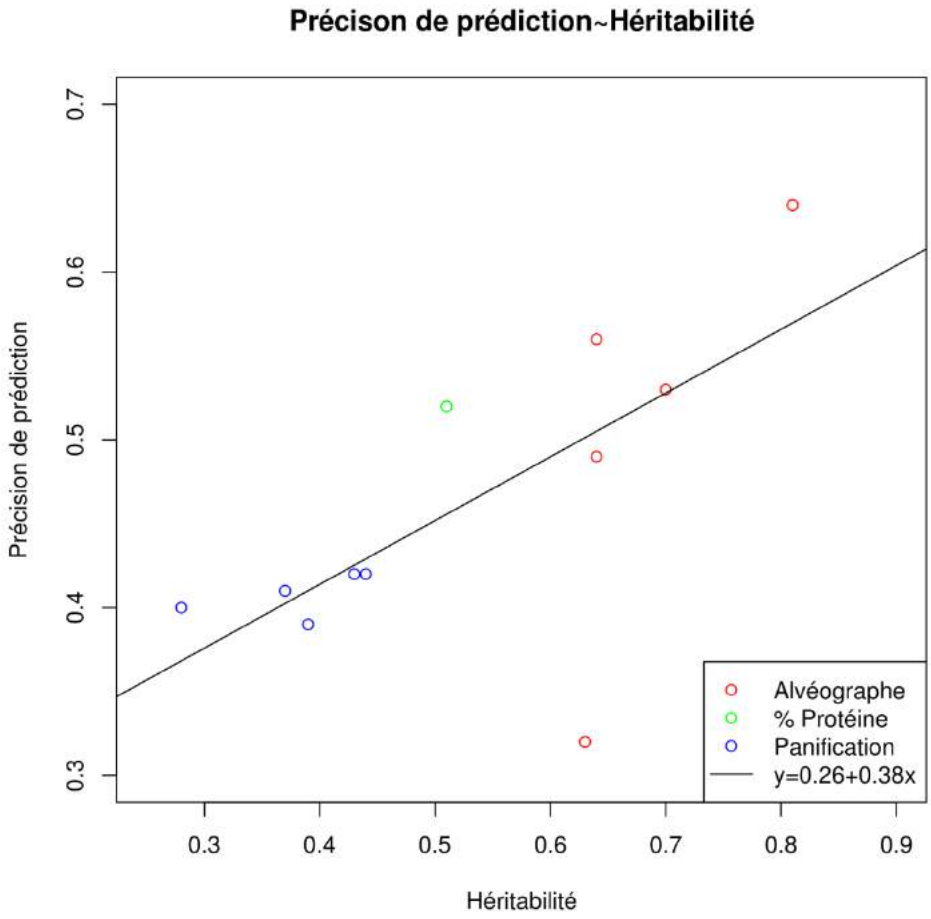


Figure S6 : Précision de prédiction en fonction de l'héritabilité (donnée INRA-AO)

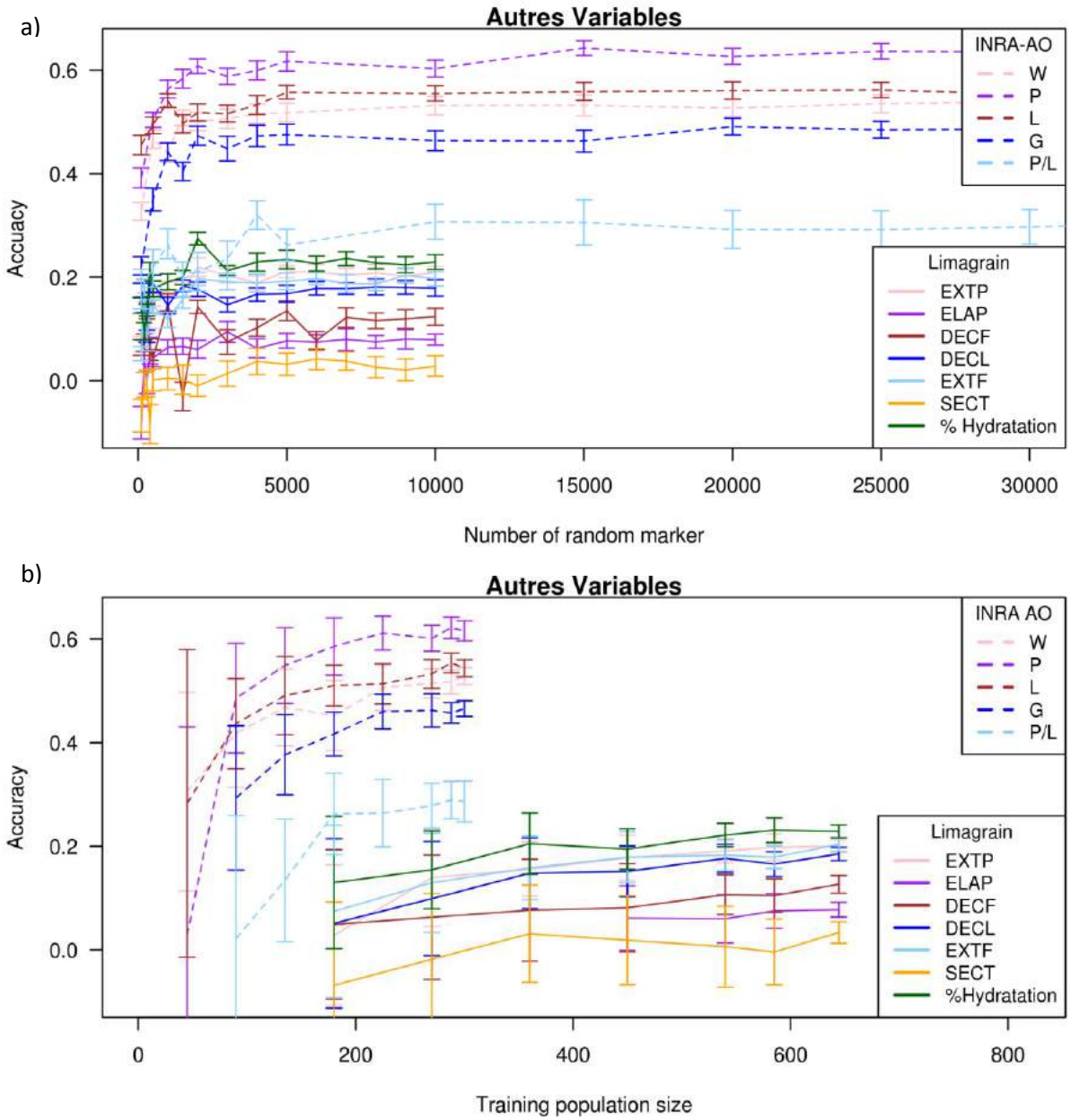


Figure S7 : Représentation de la précision de prédiction en fonction du nombre de marqueurs choisies aléatoirement (a) et du nombre de lignées choisies aléatoirement (b) pour les caractères d'alvéographes et les variables de base de panification

(modèle G-BLUP, données INRA-AO/Limagrain)

(a) : population de calibration de 313 lignées (panif), 654 lignées (protéine), 288 lignées (Nmie), 300 lignées (alveo), Limagrain : 644 lignées

(b) : nombre de marqueurs : 8 000, données INRA-AO : 15 000

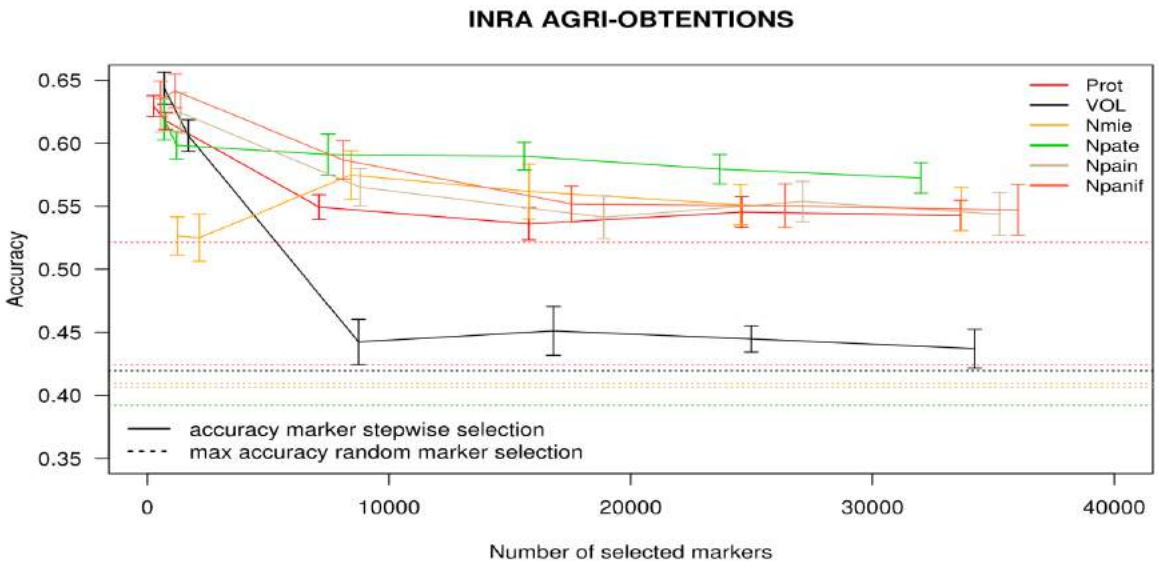


Figure S8: Gain de précision de prédiction entre un modèle G-BLUP avec ou sans sélection « MLMM »

Population de calibration : 313 lignées (panif), 654 lignées (protéine), 288 lignées (Nmie), 300 lignées (alvéo),

Pointillés : précision de prédiction maximale obtenue sans sélection de marqueurs

Lignes pleines : précision de prédiction maximale obtenue avec sélection de type MLMM

Barre d'erreur : écart-type associé à chaque valeur

Table S15 : Comparaison des précisions de prédiction entre un modèle de sélection génomique GBLUP et Bayes-C

(Données Limagrain, 10 000 marqueurs et une population de calibration de 644 lignées et population cible de 71 lignées)

	GBLUP	**Bayes C**
protéine	0,254	0,267
note de pâte	0,244	0,278
note de pain	0,214	0,201
Volume	0,306	0,314
Volume/masse	0,347	0,315
note de panification	-0,009	0,06
EXTP	0,2	0,218
ELAP	0,067	0,066
DECF	0,1	0,133
DECL	0,17	0,169
SECT	-0,002	-0,006
EXTF	0,186	0,214
Pourcentage Hydratation	0,228	0,245

Table S16 : Gain de précision en sélection génomique avec sélection de variable par MLMM

	Précision INRA-AO	GAIN / SG classique	Précision Limagrain	GAIN / SG classique
NPAT	0,61	**x1,6**	0,52	**x1,9**
NPAI	0,624	**x1,5**	0,38	**x1,7**
NPANI	0,641	**x1,5**	0,37	**x2,8**
VOL	0,643	**x1,5**	0,48	**x1,4**
Proteine	0,629	**x1,2**	0,51	**x1,8**
Nmie	0,574	**x1,4**		
W	0,643	**x1,2**		
P	0,693	**x1,1**		
L	0,647	**x1,1**		
G	0,544	**x1,1**		
PsurL	0,554	**x1,7**		

	Diplôme : Ingénieur Agronome Spécialité : Agronome Spécialisation / option : Science et productions végétales, Amélioration des plantes Enseignant référent : Maria Manzanares-Dauleux

Auteur(s) : Pierre Colin Date de naissance : 10/05/1993	Organisme d'accueil : INRA Site de Crouël Adresse : 5 Chemin de Beaulieu 63039 CLERMONT-FERRAND Maître de stage : Sophie Bouchet
Nb pages : 22 Annexe(s) : 26 Année de soutenance : 2016	

Titre français : Evaluation de la précision des modèles de prédiction de la qualité boulangère du blé tendre (Triticum aestivum L.).

Titre anglais : Testing accuracy of prediction models for wheat (Triticum aestivum L.) bread-making qualities

Résumé : La qualité boulangère est déterminante pour l'inscription d'une variété de blé au catalogue français. C'est un caractère complexe qui se mesure par des tests de panification ou par des mesures physiques sur la pâte (alvéographe). La sélection génomique pourraît augmenter le gain génétique par unité de temps ou de coût sur ces caractères. L'objectif est de définir la taille de la population de calibration, le nombre de marqueurs moléculaires et le modèle statistique optimaux pour établir l'équation de prédiction. Au total, 716 lignées et 10 124 SNPs provenant des programmes de sélection Limagrain et 370 lignées et 172 074 marqueurs INRA-Agri-Obtentions sont analysés. L'architecture de 14 caractères est décrite grâce aux résultats d'études d'association. Différents modèles de prédiction adaptés à des caractères contrôlés par de nombreux gènes à effets faibles (GBLUP) ou quelques gènes à effets forts (Bayes Cπ) sont testés. Alors que 5000 marqueurs tirés au hasard sont suffisants pour permettre une précision optimale avec un GBLUP classique, la sélection de 1000 marqueurs indépendants associés au caractère améliore les prédictions de tous les caractères, jusqu'à un facteur 3 pour la note de panification Limagrain. Le modèle and Bayes Cπ donne des précisions équivalentes au GBLUP. La taille de la population de calibration devra être augmentée pour stabiliser les précisions de prédiction. Les modèles nécessitent une validation sur des jeux de données indépendants.

Abstract: Bread-making qualities are crucial to register a new variety in France. It is a complex trait that can be measured either by bread-making tests or by indirect physical measurements of the dough (alveograph). Genomic selection could increase genetic gain per time or cost unit for this trait. The objective is to define the size of the training population, the number of molecular markers and the optimal statistical model to build the prediction equation. In total, 716 lines and 10 124 SNPs from Limagrain's breeding program, 370 lines and 172 074 SNPs from INRA-AgriObtentions are analysed. The architecture of 14 traits is determined using association analyses. Different prediction models adapted to traits controlled by many genes with small effect (GBLUP) or a few genes with large effect (Bayes Cπ) are tested. Although random samples of 5000 molecular markers allow optimal predictions for GBLUP, 1000 independent markers associated with the trait give significantly higher accuracy for all traits, up to three times higher for bread-making ability at Limagrain. The Bayes Cπ model gives similar results than GBLUP. The size of the training population needs to be increased to stabilize predictions. These models necessitate validation on independent data sets.

Mots-clés : Blé tendre - Panification – Alvéographe – Association – Haplotypes locaux – Sélection génomique

Key Words: Bread wheat– Bread-making- Alveograph– Association study – local haplotypes – Genomic selection