Genomic selection: from technological development to application to the major challenges of tomorrow

Pascal Croiseau

Habilitation à Diriger des Recherches

La sélection génomique : De l'évolution technologique à son utilisation pour les grands défis de demain

Pascal Croiseau

Membres du Jury

- **Pascale Le Roy**, Directrice de recherches INRAE - Rapporteur
- **Frédéric Farnir**, Professeur à la Faculté de Médecine Vétérinaire de Liège - Rapporteur
- **Xavier Rognon**, Professeur à AgroParisTech, Université Paris Saclay - Rapporteur
- **Chirstine Dillmann**, Professeur à l'Université Paris Saclay - Examinatrice
- **Carole Moreno**, Directrice de recherche INRAE - Examinatrice

SOMMAIRE

I. Introduction

Depuis une dizaine d'années, la recherche autour des évaluations génétiques des bovins laitiers est portée par une révolution technologique : le séquençage du génome et sa conséquence immédiate, la disponibilité des puces de génotypage à SNP (Single Nucleotide Polymorphim). Les bovins laitiers ont été espèce pionnière pour l'utilisation de ces puces SNP dans les évaluations génétiques. Les raisons ont été explicitées par Schaeffer dès 2006 : dans cette espèce à cycle de sélection long et qui dépendait d'un testage sur descendance, la sélection génomique peut doubler le progrès génétique sans augmentation de coût, voire pour un coût réduit, principalement par réduction de l'intervalle de génération. A cette raison majeure se rajoutent beaucoup d'avantages : dans les races les plus importantes, la disponibilité d'un grand nombre de taureaux testés qu'il suffisait de génotyper pour disposer d'une population de référence de qualité et donc d'évaluations précises et ce, pour tous les caractères évalués indépendamment de leur héritabilité ; des perspectives attractives sur l'orientation de la sélection, la gestion de la variabilité ou la prise en compte de nouveaux caractères. En France, la sélection génomique a été mise en œuvre très précocement, dès 2009. Pour en comprendre les raisons, il faut se souvenir que sur la période 2001-2008, les programmes de sélection des trois grandes races laitières (Normande, Montbéliarde et Holstein) incluaient déjà une phase de présélection sur information moléculaire, dite sélection assistée par marqueurs de première génération (SAM1). Vue du prisme d'aujourd'hui, cette sélection assistée par marqueurs était limitée dans son principe : reposant sur le typage de 45 marqueurs microsatellites, couvrant une quinzaine de QTL, elle ne tirait parti que d'une petite proportion de la variabilité génétique, par ailleurs fortement surestimée. L'absence de déséquilibre de liaison populationnel impliquait d'utiliser cette information intra famille uniquement. Le gain de précision limité permettait une amélioration du choix sur ascendance mais ne permettait pas d'éliminer la phase de testage sur descendance. Toutefois, au niveau organisationnel, les circuits d'information étaient prêts, la profession était motivée pour investir dans les développements et surtout les opérateurs de la sélection étaient habitués à cette étape d'évaluation, ce qui a grandement facilité l'adoption de la sélection génomique lorsqu'elle a pu réellement être mise en œuvre.

C'est dans ce contexte que la puce moyenne densité bovine produite par Illumina et nommée Bovine SNP50BeadChip® est arrivée (nous la nommerons puce 50K par la suite). Comparativement aux microsatellites, les marqueurs SNP, de par leur polymorphisme limité par nature, sont individuellement moins informatifs. Cependant, ce désavantage est plus que compensé par leur grand nombre. En effet, la couverture du génome qu'offre la puce 50K permet d'espérer un déséquilibre de liaison entre les marqueurs de la puce et les mutations causales suffisant pour exploiter leur association génétique au sein des modèles d'évaluation génétique. Ainsi, fin 2008, la SAM de 1ère génération a laissé place à une SAM de 2ème génération (SAM2), reposant sur un plus grand nombre de QTL et des marqueurs en fort déséquilibre de liaison avec eux. Pour alimenter cette SAM2, des études d'association ont été réalisées et ont permis l'identification d'une cinquantaine de QTL sur les caractères déjà présents lors de la SAM1 mais également pour des nouveaux caractères tels que : les cellules, la fertilité chez la génisse, la vitesse de traite et six caractères de morphologie. Ce sont ces QTL qui ont été intégrés dans les évaluations génétiques disponibles pour ces 15 caractères. Bien sûr, la qualité d'estimation de l'effet de ces QTL sur les différents caractères dépend, en grande partie, du nombre d'animaux génotypés et phénotypés disponibles. Dans le cadre du projet CartoFine cofinancé par l'ANR et Apis-Gene, environ 8000 taureaux testés sur descendance ont été génotypés, constituant ainsi les premières vraies populations de référence de sélection génomique. Dans le même temps, les entreprises de sélection ont réalisé le génotypage

environ 20 000 animaux par an, pour le choix des mères à taureaux et des jeunes mâles candidats à la sélection. Cependant, la couverture familiale n'a plus besoin d'être aussi forte que pour la SAM1. Ainsi, l'effort de génotypage a été réalisé sur des noyaux familiaux plus simples (père et grand-père maternel du candidat). Cela a permis la couverture d'un plus grand nombre de familles et en particulier des familles peu représentées qui ne pouvaient pas être couvertes avec la SAM1.

Cependant, pour la SAM1 comme pour la SAM2, l'identification des meilleurs taureaux nécessite un testage sur descendance (Figure 1). Pour évaluer les performances d'un jeune taureau, ce taureau est gardé en station de contrôle individuel jusqu'à sa maturité sexuelle (15 mois). Ensuite, ce taureau est utilisé pour procréer une centaine de filles de testage qui sont élevées jusqu'à l'obtention de performances individuelles. A ce stade, le taureau qui a 5 ans peut être évalué sur descendance et, pour les meilleurs, sélectionné pour procréer la génération suivante de vaches (« pères à vaches ») et éventuellement de taureaux (« pères à taureaux »). Il s'agit donc d'un processus long (entre 5 et 6 ans) et couteux (environ 45000 euros par taureau testé). La précision des évaluations génétiques permise par les SAM1 et SAM2 était suffisante pour sélectionner les jeunes taureaux avec un fort potentiel mais trop faible pour se passer d'un contrôle sur descendance.

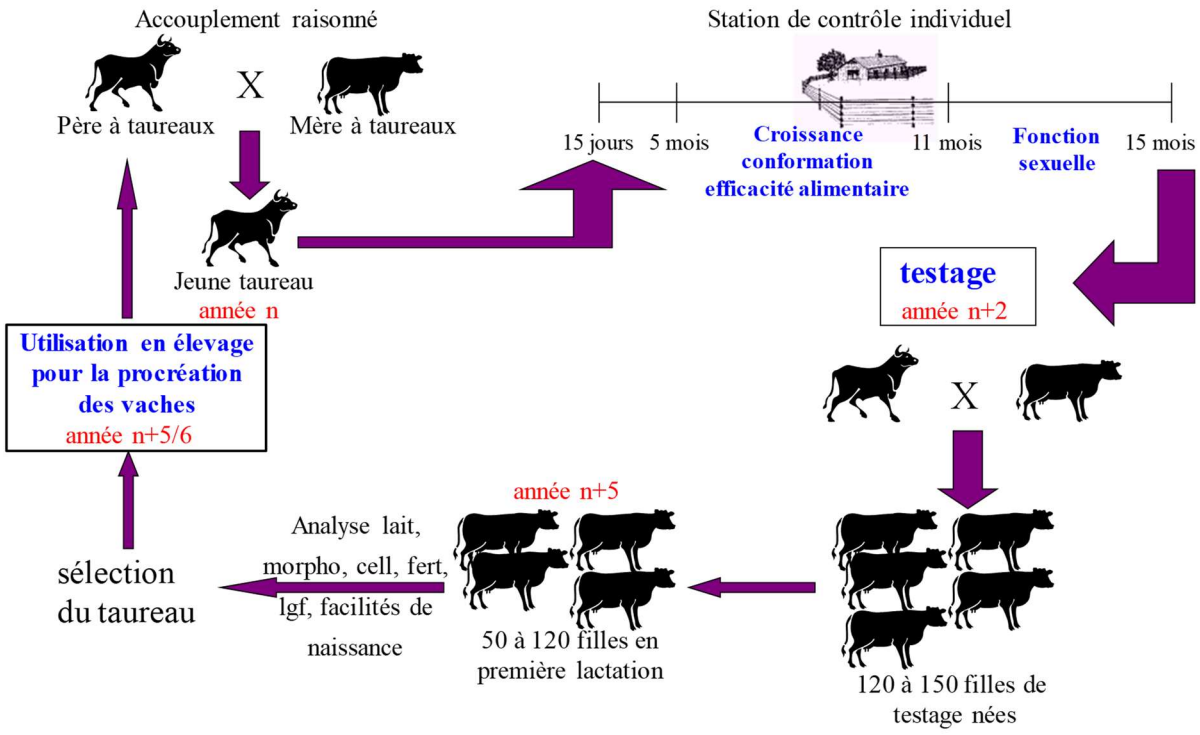


Figure 1: Le testage sur descendance chez les bovins laitiers.

Alors qu'en France, l'approche utilisée initialement reposait sur une sélection assistée par marqueurs utilisant l'information biologique des QTL, les travaux méthodologiques conduits par plusieurs équipes, d'abord par simulation (Meuwissen *et al.* 2001) puis sur de premières applications (Hayes *et al.* 2006), montraient la puissance d'une approche basée sur des marqueurs couvrant tout le génome, sans information QTL a priori, dite sélection génomique. Pour cela, on estime l'effet statistique de chaque marqueur (il s'agit bien d'un effet statistique car les marqueurs n'ont généralement pas d'effet biologique) ou de chaque segment chromosomique. La somme de ces contributions fournit une valeur d'élevage estimée pour un caractère donné dite génomique (GEBV pour Genomic Enhanced Breeding Value en anglais). Et ces GEBV pourraient être estimés à la naissance de l'animal, sans recours au testage sur

descendance ce qui induirait une forte diminution de l'intervalle de génération. Plusieurs autres études ont ensuite été présentées ou publiées (Gianola *et al.* 2006; Hayes *et al.* 2007; Habier *et al.* 2007; Solberg *et al.* 2008). Ces études, majoritairement basées sur des simulations, produisent des résultats moins convainquant que les travaux de Meuwissen et al. Cependant, malgré des conclusions moins optimistes, elles montrent qu'une évaluation génomique a une fiabilité supérieure à une évaluation génétique exploitant l'information pedigree. Les premières études sur données réelles conduites majoritairement par l'Australie, les Pays-Bas et les Etats Unis (Calus and Veerkamp 2007; Long *et al.* 2007; Muir 2007) ont confirmé des premiers résultats prometteurs.

C'est dans ce contexte que je suis arrivée à l'INRA en octobre 2008 avec pour mission (i) (i) de développer une approche d'évaluation génomique intégrant autant que possible l'information QTL disponible et (iii) de comparer les modèles statistiques pour la détection de QTL afin d'isoler le plus précisément possible les régions du génome qui impactent les caractères d'intérêt.

Dans ce manuscrit, je vous présenterai dans une première partie les travaux que j'ai menés au cours de ces 9 ans sur ces deux thématiques étroitement imbriquées que sont : les études de détection de QTL et la méthodologie des évaluations génomiques. Ensuite, dans une seconde partie, je vous présenterai les grands projets dans lesquels je compte m'investir pour les 5 prochaines années. Ces projets visent essentiellement à exploiter les propriétés des méthodes d'évaluation génomique ainsi que de la sélection génomique mise en place dans les entreprises de sélection pour répondre aux grands enjeux de demain tel que le maintien de la diversité génétique ou encore le développement d'une évaluation génomique en croisement.

II. Mise en place d'une première évaluation génomique officielle chez les bovins laitiers

A. Approches méthodologiques appliquées à la sélection génomique (projet ANR AMASGEN)

Ce projet cofinancé par l'ANR et Apis-Gene était piloté par Vincent Ducrocq. Démarré en octobre 2008, il avait pour objectifs de développer une méthode de prédiction des valeurs génétiques des taureaux et des vaches à partir de l'information SNP. Cette méthode de prédiction devait s'adapter à une contrainte forte, à savoir une population de référence (animaux phénotypés et génotypés) relativement petite au regard du nombre de marqueurs disponibles sur la puce 50K. En effet, en 2008, les populations de référence disponibles étaient constituées d'environ 1000 taureaux pour la Montbéliarde et la Normande et d'environ 4000 taureaux pour la Holstein. Plusieurs méthodes avaient déjà été proposées pour réaliser des évaluations génomiques. Conceptuellement, la plus simple d'entre elles est le GBLUP (Genomic Best Linear Unbiased Prediction) (VanRaden 2008). Le GBLUP correspond à la version génomique du BLUP, où la matrice de parenté généalogique du BLUP est remplacée par une matrice de parenté génomique (Habier *et al.* 2007; VanRaden 2008). Pour prendre en compte la disproportion entre le nombre d'animaux et le nombre de SNP, le GBLUP fait l'hypothèse que tous les SNP ont des effets de même variance a priori, reflétant un déterminisme très polygénique avec de très nombreux QTL à petits effets répartis sur tout le génome. D'autres hypothèses a priori peuvent être faites et un large panel de méthodes Bayesiennes ont été développées et que l'on a appelé par la suite l'alphabet Bayesien. Pour les décrire brièvement, le Bayes A (Meuwissen *et al.* 2001) lève l'hypothèse de variance génétique égale pour tous les SNP en estimant une variance pour chaque SNP. Pour rétablir un ratio nombre de SNP utiles/nombre d'animaux plus raisonnable et par là même mimer un modèle biologique avec un nombre modéré de QTL, un modèle Bayesien avec sélection de variables a été développé par Meuwissen *et al.* (2001) et nommé Bayes B. Dans ce modèle, l'utilisateur définit un pourcentage π de SNP dont l'effet sur le caractère est nul. Ainsi, à chaque itération, seule une proportion $(1- \pi)$ de SNP ont un effet non nul avec une variance génétique estimée propre à chaque SNP. Ce modèle, en terme de prédiction de la valeur génétique des animaux, a longtemps été une référence. Toutefois, les temps de calcul restaient incompatibles avec des évaluations de routine. Une méthode plus simple a été proposée par (Habier *et al.* 2011), similaire au BayesB mais imposant une variance commune à tous les effets de SNP non nuls. Cette approche présente l'avantage de n'estimer qu'une seule variance, elle est donc moins sensible à la taille de la population. A la suite de ces travaux, toute une série méthodes dérivées ont été proposées. Notons par exemple la méthode BayesR (Erbe *et al.* 2012), comparable au BayesC mais considérant trois classes de variance d'effet au lieu d'une, autrement dit des petits, moyens et gros QTL.

Fort de notre expertise en Sélection Assistée par Marqueurs, nous avons souhaité comparer la performance des méthodes d'évaluation génomique disponibles (et en particulier le GBLUP) à celle obtenue par une extension de notre méthode. L'idée sous-jacente était d'exploiter la puce 50K pour sélectionner, pour chaque caractère, la liste des marqueurs d'intérêt et de les inclure dans un modèle de Sélection Assistée par Marqueurs. Pour sélectionner les SNP les plus pertinents pour chaque caractère, nous nous sommes intéressés à une méthode de régression pénalisée, l'Elastic Net (EN) (Zou and Hastie 2005). Cette méthode correspond à une combinaison linéaire de deux modèles de régression pénalisée qui sont la Ridge Regression (RR) et le LASSO. En présence de variables corrélées, la RR a la propriété de retenir tous les prédicteurs tandis que le LASSO ne retient que le plus significatif et impose un

effet nul aux autres variables (Zou and Hastie 2003, 2005). Les SNP sont en plus ou moins fort déséquilibre de liaison par conséquent, l'EN, qui est un intermédiaire entre la RR et le LASSO, permet une pénalisation des SNP plus souple en fonction du niveau de déséquilibre de liaison.

Voici comment l'estimation des effets des SNP est calculée :

$$\hat{\beta} = \arg min \left\{ \sum_{i=1}^{Nani} (y_i - x_i\beta)^2 + \lambda \left(\alpha \sum_{j=1}^{Nsnp} \beta_j^2 + (1-\alpha) \sum_{j=1}^{Nsnp} |\beta_j| \right) \right\}$$

où β correspond au vecteur des effets de SNP, y_i correspond au phénotype de l'animal i, x_i correspond au vecteur de génotypes de l'animal i. En ce qui concerne la pénalisation des SNP, λ correspond à l'intensité de la pénalisation tandis que α permet répartir la pénalisation sur de la RR ou sur un LASSO. Ce paramètre, propre à l'EN, prend des valeurs comprises entre 0 et 1. Ainsi, avec un α=0, un modèle LASSO pur est utilisé tandis qu'avec un α=1, un modèle RR pur est appliqué.

Cette approche a été comparée à un BLUP pedigree et à un GBLUP à travers des travaux de validation. Pour cela, chacune des populations de référence (animaux génotypés et phénotypés) des 3 races disponibles a été divisée en une population d'apprentissage (à partir de laquelle l'équation de prédiction est définie) et une population de validation (constituée des 25% plus jeunes animaux et à partir de laquelle les prédictions sont comparées aux phénotypes observés). Les effectifs sont décrits dans le Tableau 1. Les analyses ont été réalisées sur 6 caractères (les 5 caractères de production laitière : quantité de lait, matière grasse, matière protéique, taux butyreux et taux protéique ainsi que pour la fertilité des vaches). Les phénotypes utilisés sont des DYD (Daughter Yield Deviation), qui correspondent aux performances moyennes des filles d'un taureau, corrigées des effets d'environnement et de la valeur génétique de leurs mères (VanRaden and Wiggans 1991; Mrode and Swanson 2004). Pour prendre en compte la variabilité de la précision de ces DYD (due en grande partie au nombre variable de filles d'un taureau), ces DYD sont pondérés par l'EDC (Equivalent Daughter Contribution) tel que défini par Fikse and Banos (2001).

	Montbéliarde	Normande	Holstein
Population d'apprentissage	950	970	2976
Population de validation	222	248	964
Total	1172	1218	3940

Tableau 1: Nombre de taureaux génotypés qui intègrent les populations d'apprentissage et de validation pour les trois races étudiées.

Les résultats de ces travaux de validation sont présentés sur le Tableau 2. Il s'agit de corrélations pondérées par les EDC entre les valeurs génomiques prédites des animaux (GEBV) et les DYD. Tout d'abord, les corrélations moyennes obtenues par le GBLUP et l'EN en intégrant l'ensemble des marqueurs disponibles (50K) sont systématiquement supérieures à celles obtenues par un BLUP pedigree (gain compris entre 7,8 et 20,5 points de corrélation en

fonction de la race). Si on compare les résultats des approches génomiques, l'EN améliore les corrélations avec un gain compris entre 2,2 et 5,5 points en fonction de la race.

	BLUP sur ascendance	GBLUP		EN		SAMg
		50K	PS	50K	PS	
Montbéliarde	0,338	0,463	0,467	0,485	0,482	0,484
Normande	0,325	0,403	0,428	0,458	0,470	0,468
Holstein	0,403	0,583	0,582	0,608	0,603	0,627

Tableau 2: Corrélations pondérées moyennes pour les 6 caractères (5 caractères de production et fertilité vache) et pour les 3 races étudiées en utilisant un BLUP sur ascendance, un GBLUP, l'EN sur l'ensemble des marqueurs disponibles (50K), ou après une présélection des SNP (EN PS) pour les races Montbéliarde, Normande et Holstein.

Cependant, comme le montre le Tableau 3, le nombre de SNP retenus par l'EN reste relativement important (entre 723 et 24037 pour la Montbéliarde, entre 403 et 2879 pour la Normande et enfin entre 1271 et 20904 pour la Holstein). Nous avons testé un modèle ou seuls les SNP présents dans des régions d'intérêt sont inclus dans l'EN. Pour définir les régions d'intérêt, nous avons choisi de réaliser une analyse d'association (GWAS pour Genome Wide Association Study) en réalisant une étude LDLA (Linkage Disequilibrium and Linkage Analysis) qui a la propriété de tester conjointement la liaison et l'association génétique au travers de tests haplotypiques (Meuwissen et al. 2001; Druet et al. 2008). Dans un premier temps, l'existence d'un QTL a été testée sur la population d'apprentissage sur l'ensemble des positions du génome à travers des haplotypes de 6 SNP avec une fenêtre chevauchante de 2 SNP. Un pic LDLA a été défini pour toutes les positions pour lesquelles le test statistique (Likelihood Ratio Test ou LRT) obtenu correspond au plus fort LRT dans une fenêtre de 50 SNP avec une valeur minimale pour les LRT de 3 ou 5. Une fois ces pics LDLA identifiés, les 50 SNP autour de chaque pic sont retenus pour définir une liste de SNP présélectionnés utilisée soit dans un GBLUP soit en utilisant l'EN. Les résultats de cette stratégie sont présentés sur le Tableau 2 et montrent des corrélations très proches de celles obtenues sans présélection (gain de corrélation compris entre -0,1 et 2,5 points pour le GBLUP et entre -0,5 et 1,2 point pour l'EN en fonction de la race). Par contre, cette étape de présélection a permis une réduction considérable du nombre de SNP retenus (Croiseau et al. 2011).

	Montbéliarde		Normande		Holstein	
	EN 50K	EN PS	EN 50K	EN PS	EN 50K	EN PS
Lait	24037	4768	1529	3788	1355	2882
Matière grasse	5444	4693	1474	1755	1271	1113
Matière protéique	23044	14905	861	541	5648	2392
taux butyreux	723	684	403	365	1351	1019
Taux de protéine	1776	3544	737	736	2297	3707
Fertilité vache	8215	3824	2879	3830	20904	12302

Tableau 3: Nombre de SNP retenus par l'EN lorsque l'ensemble des SNP de la puce 50K ont été inclus dans le modèle (EN 50K) ou après présélection des SNP (EN PS) pour les races Montbéliarde, Normande et Holstein.

Fort de notre expérience en Sélection Assistée par Marqueurs, nous avons souhaité tester un modèle SAM exploitant à la fois les résultats de cartographie de QTL (étude LDLA) et les listes de SNP retenus par l'EN. Nous avons donc étendu notre modèle SAM de 2ème génération à l'intégration de toutes les régions génomiques mises en évidence par l'EN. Le modèle choisi, nommé SAMg, est le suivant :

$$y_i = \sum_{j=1}^{nbQTL} \left(h_{ij}^{pat} + h_{ij}^{mat} \right) + u_i + e_i \qquad 7$$

Où y_i correspond au DYD de l'animal i, h_{ij} correspond à l'effet gamétique paternel ou maternel de l'individu i au QTL j, u_i correspond à l'effet polygénique de l'individu i et e_i correspond à l'effet résiduel de l'animal i.

Cette méthodologie permet d'évaluer conjointement l'effet de la composante génomique et l'effet de la composante polygénique restante. L'ensemble des QTL sélectionnés par la cartographie et par l'EN sont intégrés dans ce modèle. Les effets de chaque QTL sont estimés à l'aide d'haplotypes de 3 à 5 SNP sur la population de référence de la race. Les QTL retenus par la cartographie expliquent entre 15 et 25% de la variance génétique totale et, chaque QTL se voit attribuer sa part de variance estimée. Les QTL retenus par l'algorithme EN expliquent, eux, entre 25 et 35% de part de variance. Ces nombreux QTL expliquant individuellement une part de variance faible et difficile à estimer, une part de variance constante est attribuée à chacun d'entre eux. Entre 300 et 700 QTL sont utilisés pour l'évaluation génomique d'un caractère donné (Boichard et al. 2012).

Cette approche, qui est compatible avec des évaluations génomiques de routine, montre un gain de corrélation de 2,4 points en race Holstein par rapport aux résultats de l'EN et des corrélations similaires à celles de l'EN dans les deux autres races, Tableau 2).

Dès 2010, cette approche a été utilisée pour les évaluations nationales officielles en France pour les 3 grandes races laitières.

B. Génomique multi-race des bovins allaitants et laitiers (projet ANR GEMBAL)

En 2010, Illumina a commercialisé une puce SNP haute densité contenant 777 000 marqueurs, dénommée puce HD. La très grande densité de SNP nous offrait la possibilité d'améliorer considérablement les études des variations génétiques présentes chez les bovins tant au niveau individuel que populationnel. Ces études peuvent aider à comprendre l'histoire évolutive de la population et les processus de différenciation mais également à déterminer les gènes responsables de la variation des caractères d'intérêt et à effectuer une sélection assistée par marqueurs à partir de ces gènes d'intérêt. Par ailleurs, cette nouvelle puce SNP ouvre de nouvelles perspectives en sélection génomique pour les bovins allaitants et pour les races régionales aux effectifs limités. En effet, la stratégie envisagée pour obtenir des prédictions génomiques avec une précision élevée, était de réaliser des analyses multiraces lorsque la taille des populations de référence intra-race est trop petite. Avec la puce 50K, la distance génétique entre deux SNP est trop élevée pour que le déséquilibre de liaison observé dans une race soit conservé et donc exploitable pour une autre race. Avec cette nouvelle puce HD, nous espérons avoir accès à un déséquilibre de liaison populationnel qui rendrait des évaluations génomiques multiraciales efficaces.

C'est dans ce cadre que le projet GEMBAL, coordonné par F Phocas et cofinancé par l'ANR, Apis-Gene et Races de France, a démarré en 2011. Les objectifs de ce projet étaient :

- De définir une population d'imputation (pour imputer l'ensemble des animaux génotypés sur la puce 50K vers la puce HD) et d'organiser le génotypage des animaux.

- De mieux comprendre la structuration du génome bovin
- De mettre en place l'imputation des animaux génotypés sur la puce 50K au niveau HD. En effet, il était essentiel d'utiliser les typages 50k et d'éviter de regénotyper les animaux qui l'étaient déjà.
- De mettre en place une méthodologie de prédiction génomique multiraciale

A titre personnel, j'ai été fortement impliqué dans les volets « imputation » et « méthodologie des prédictions génomiques multi-races ».

1. Divergence des structures haplotypiques entre races

En amont du démarrage de ces deux volets (pendant la phase d'accumulation des animaux génotypés sur la puce HD), j'ai encadré le stade de master de Ana Maria Perez O'Brian intitulé : « Niveau de déséquilibre de liaison et divergence des structures haplotypiques entre les races Brune, Holstein et Blonde d'Aquitaine observés sur la puce HD ». Cette étude a été conçue afin de mesurer l'impact de l'augmentation de la densité de marqueurs apportée par la puce HD sur la structure des haplotypes des différentes races bovines. L'objectif était de valider la pertinence d'exploiter la puce HD au sein d'une évaluation génomique multiraciale.

Le projet GEMBAL a permis le génotypage de plus de 5000 animaux issus de 16 races différentes mais au démarrage de ce stage, seuls 270 animaux issus de 3 races (105 Holstein, 67 Brunes et 98 Blondes d'Aquitaine) étaient disponibles. A partir de ces données, nous avons comparé l'étendue du déséquilibre de liaison (LD) et sa décroissance à différentes distances et densités.

Pour cela, la mesure du déséquilibre de liaison (r^2) a été réalisée pour toutes les combinaisons de SNP le long du génome. Le LD moyen pour des SNP séparés par une même distance (jusqu'à 100kb ou 1Mb) a ensuite été calculé et mis sous forme de graphique (Figure 2). Deux densités de marqueurs ont été comparées : la puce HD et la puce 50K.

Pour la Brune et la Holstein, le LD démarre avec un $r^2 \sim 0,7$ et diminue fortement pour atteindre un $r^2 \sim 0,2$ à 100Kb (Figure 2-b), puis se stabilise après 200Kb à un $r^2 \sim 0,1$ (Erreur ! Source du renvoi introuvable.-d). Pour la Blonde d'Aquitaine, le LD à courte distance se situe à un niveau plus faible que pour les deux autres races et décroit plus rapidement pour atteindre un $r^2 \sim 0,2$ à une distance de 40kb et un $r^2 \sim 0,1$ autour des 100kb (Erreur ! Source du renvoi introuvable.-b). La stabilisation ne s'observe qu'après 200kb avec un r^2 proche de 0,05 (Erreur ! Source du renvoi introuvable.-d). Ces résultats sont en accord avec un effectif génétique de la Blonde d'Aquitaine plus élevé (environ 250) que pour les deux autres races (environ 40 pour la Holstein et 50 pour la Brune) (Boichard *et al.* 1996; Bouquet *et al.* 2009).

Si les patterns de déséquilibre de liaison sont cohérents entre la 50K et la HD, on s'aperçoit que l'augmentation de la densité de marqueurs, qui augmente drastiquement le nombre de couples de marqueurs entrant en jeu, permet une meilleure précision des résultats. En effet, les résultats obtenus par la puce 50K apparaissent plus dispersés que ceux de la HD, en particulier sur des distances inférieures à 40Kb (Erreur ! Source du renvoi introuvable.-a,c).

L'utilisation de la puce HD nous permet d'améliorer nettement la description du déséquilibre de liaison à courte distance (entre 0 et 20Kb). Dans un second temps, nous avons recherché un éventuel déséquilibre de liaison populationnel qui ouvrirait la porte à une évaluation génomique multiraciale. Pour avoir une idée plus précise de cette information, Anna-Maria a utilisé le logiciel PLINK (Chang et al. 2015) pour identifier les blocs haplotypiques à partir des deux densités de puces pour les trois races. A partir de cette information, trois catégories de

SNP ont été mises en évidence : les SNP appartenant à un bloc haplotypique au sein des 3 races, les SNP n'appartenant pas à un bloc haplotypique au sein des 3 races et les SNP dont le statut est différent en fonction de la race. Cette étude a montré qu'avec la puce 50K, seuls 2% des SNP appartiennent à un bloc haplotypique présent au sein des trois races. Cela signifie que la majorité des SNP ne sont pas regroupés au sein de blocs haplotypiques et que les blocs haplotypiques identifiés diffèrent en fonction de la race. Avec la puce HD, 51% des SNP sont inclus dans des blocs haplotypiques présents dans les 3 races. Cette statistique est intéressante car elle montre que la densité de marqueurs de la puce HD pourrait permettre un travail multiracial qui n'était pas possible avec la puce 50K.

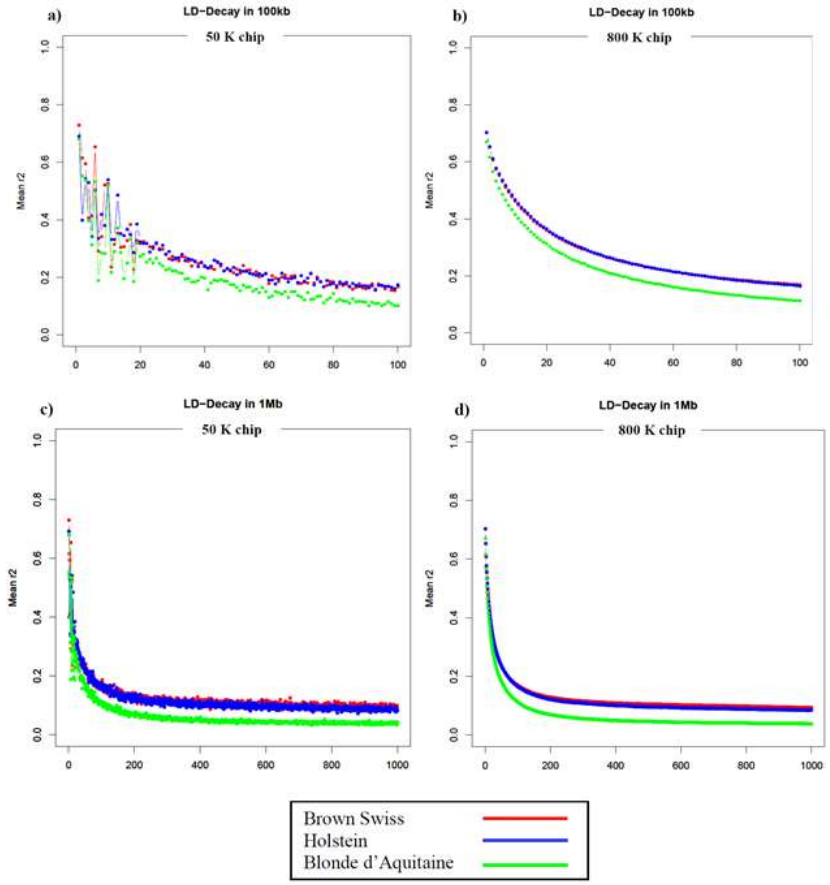


Figure 2: Décroissance du Déséquilibre de Liaison (r^2) chez les races Brune, Holstein et Blonde d'Aquitaine sur une distance de 100kb avec la puce 50K (a) ou HD (b) et, à une distance de 1Mb avec la puce 50K (c) ou la puce HD (d).

2. Imputation des animaux génotypés à l'échelle de la puce HD

Pour réaliser une évaluation génomique exploitant la puce HD, une étape d'imputation (prédiction des génotypes manquant sur une puce grâce à l'information apportée par les animaux génotypés sur la puce de densité supérieure) est nécessaire pour amener l'intégralité des animaux génotypés (quelle que soit la densité) à un génotypage HD. Une fois l'ensemble des génotypages HD disponible, les travaux d'imputations ont débuté et Chris Hozé, doctorante que j'ai co-encadrée avec Vincent Ducrocq entre 2011 et 2014, s'est beaucoup investie dans cette tâche pour évaluer l'efficacité de l'imputation en fonction de la population de référence disponible.

Pour mener cette étude, nous disposions alors de 5153 animaux issus de 16 races différentes (entre 89 et 788 animaux génotypés en HD en fonction de la race) (Tableau 4). Ces animaux ont été choisis sur la base de leur contribution marginale à la population (Boichard et al. 1997),

estimée sur à partir du pedigree de chaque race. Dans la mesure du possible, deux ou trois descendants de chaque fondateur ont également été génotypés pour améliorer la qualité du phasage.

	Nb of genotyped animals	Nb of genotyped families (sire + progeny)	Mean nb of genotyped progeny per sire	Nb of effective ancestors
Dairy breeds				
Abondance (ABO)	209	54	3.69	15
Brown Swiss (BSW)	99	52	1.90	28
Holstein (HOL)	788	204	2.30	21
Montbéliarde (MON)	530	139	3.77	18
Normande (HOR)	551	138	3.82	23
Simmental (SIM)	125	55	2.24	39
Tarentaise (TAR)	185	65	2.77	15
Beef breeds				
Aubrac (AUB)	254	116	2.17	112
Bazadaise (BAZ)	89	60	1.45	46
Blonde d'Aquitaine (BLA)	327	187	1.74	78
Charolais (CHA)	672	310	2.14	249
Gasconne (GAS)	163	76	2.12	197
Limousine (LIM)	462	235	1.96	185
Parthenaise (PAR)	304	97	3.02	89
Rouge des Prés (RDP)	149	80	1.83	99
Salers (SAL)	246	186	1.31	99

Tableau 4: Nombre d'animaux génotypés sur la puce HD et structure de population génotypée pour chaque race.

Pour mesurer l'efficacité de l'imputation de la puce 50K vers la puce HD, les données de chaque race ont été divisées en deux parties : les animaux les plus vieux ont été utilisés comme population d'apprentissage pour mimer une population de référence génotypée sur la puce HD. Les 20% plus jeunes animaux forment quant à eux la population de validation. Pour la population de validation, les marqueurs qui étaient présent sur la puce HD ont été masqués pour mimer une population cible génotypée sur la puce 50K. L'ensemble des imputations ont été réalisées avec le logiciel Beagle 3.3.0 identifié, en 2013, comme étant le logiciel le plus performant (Browning and Browning 2009). Cette étude a conclu à un taux d'erreur d'imputation très faible pour la plupart des races, avec un taux d'erreur moyen de 1,36% (Tableau 5).

	Training population size	Validation population size	Allelic imputation error rate (%)	LD level at 70 kb	Average $R_{T/V}$
Dairy breeds					
Abondance (ABO)	169	40	0.75	0.217	0.146
Brown Swiss (BSW)	79	20	1.92	0.255	0.074
Holstein (HOL)	634	154	0.73	0.255	0.078
Montbéliarde (MON)	424	106	0.51	0.196	0.116
Normande (HOR)	444	107	0.33	0.233	0.104
Simmental (SIM)	100	25	2.55	0.209	0.050
Beef breeds					
Aubrac (AUB)	204	50	2.03	0.177	0.028
Bazadaise (BAZ)	72	17	2.07	0.239	0.038
Blonde d'Aquitaine (BLA)	262	65	1.80	0.175	0.038
Charolais (CHA)	539	133	0.68	0.176	0.018
Gasconne (GAS)	131	32	2.26	0.174	0.026
Limousine (LIM)	370	92	1.09	0.164	0.014
Parthenaise (PAR)	245	59	1.88	0.161	0.024
Rouge des Prés (RDP)	119	30	2.39	0.206	0.028
Salers (SAL)	197	49	1.27	0.213	0.024

Tableau 5:Taux d'erreur d'imputation intra-race et autres paramètres affectant le taux d'erreur d'imputation.

Cependant, une certaine variabilité existe entre la race Normande qui affiche un taux d'erreur de 0,31% et la race Simmental qui affiche un taux d'erreur de 2,41%. Un des facteurs clés pour réduire drastiquement le taux d'erreur d'imputation est la taille de la population de référence. Ainsi, nous avons constaté que lorsque la population de référence contient plus de 500 animaux, le taux d'erreur est inférieur à 0,7% (Figure 3).

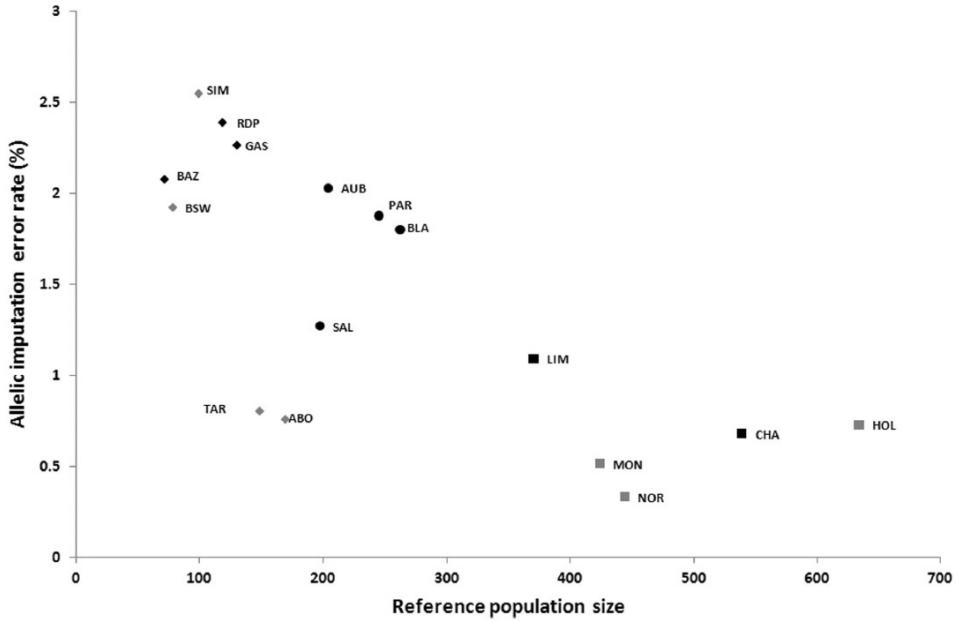


Figure 3: Relation entre le taux d'erreur d'imputation et la taille de la population de référence chez les bovins allaitants (en noir) et laitiers (en gris). Les races avec plus de 300 animaux dans leur population de référence sont représentées par un carré, celles avec plus de 200 animaux par un cercle et celles avec 200 animaux ou moins par un diamant.

Par ailleurs, aucune relation n'a pu être détecté entre la MAF (fréquence de l'allèle mineur) et le taux d'erreur d'imputation pour les SNP ayant un taux d'erreur inférieur à 0,1%. Par contre,

le taux d'erreur augmente avec la MAF pour les SNP avec un taux d'erreur élevé. En ce qui concerne le niveau de déséquilibre de liaison à faible distance, celui-ci ne semble avoir d'influence sur le taux d'erreur d'imputation, tandis que la taille efficace de la population a un effet modéré (Hozé et al. 2013).

3. Evaluations Génomiques multiraciales

Les taux d'erreur d'imputation étant relativement faibles, ce facteur ne représente pas un frein pour la mise en place d'une sélection génomique multiraciale à partir de génotypes 50K imputés HD. Cependant, les effectifs disponibles pour les races régionales rendent difficile l'estimation de l'efficacité de la sélection génomique. Nous avons donc réalisé un premier test utilisant les grandes races laitières (Holstein, Normande et Montbéliarde). Plusieurs évaluations utilisant des génotypes 50K ou HD et une population de référence mono ou multiraciale ont été réalisées. Arbitrairement, la race Normande a été considérée comme race à évaluer à partir de sa population de référence et/ou des populations de référence des deux autres races. Afin de simuler différentes tailles de population de référence, la population d'apprentissage Normande a été soit considérée dans son intégralité (1597 individus), soit réduite à 198 ou 404 individus (Tableau 6Erreur ! Source du renvoi introuvable.).

Feature	Training A	Training B	Training C	Training D
BLUP	1,597 NO	404 NO	198 NO	
Single-breed GS[1]	1,597 NO	404 NO	198 NO	
Multi-breed GS	1,597 NO	404 NO	198 NO	4,989 HO
	4,989 HO	4,989 HO	4,989 HO	1,788 MO
	1,788 MO	1,788 MO	1,788 MO	
Phenotypes[2]	DYD/sDYD	DYD/sDYD	sDYD	sDYD
SNP panel[3]	50K/HD	50K/HD	HD	HD
Eff_SNP[4]	7,000	700/7,000	7,000	7,000
Polygenic (%)	30	10/20/30/40	30	30
Validation (with sire + mgs)[5]	394 NO (380)	39 NO (326)	394 NO (326)	394 NO (0)

[1]GS = genomic selection.
[2]DYD = daughter yield deviation; sDYD = standardized DYD.
[3]50K = ~50,000 SNP panel; HD = high-density SNP chip.
[4]Eff_SNP = expected number of markers with a nonzero effect.
[5]Number of validation bulls with sire and maternal grandsire (mgs) in the training set.

Tableau 6 :Description des différents scenarios testés pour les races Montbéliarde (MO), Normande (NO) et Holstein (HO).

La précision des différents types d'évaluations a été comparée en utilisant la corrélation entre les phénotypes et les valeurs génétiques estimées de 394 taureaux normands. La méthode d'évaluation génomique utilisée est un BayesC (Habier *et al.* 2011) implémentée dans le logiciel GS3 (Legarra et al. 2013). Cette méthode est sensée sélectionner les marqueurs fortement associés aux QTL dans les trois races tout en éliminant les variants distants, dont les effets ne sont pas conservés entre races. Six caractères ont été étudiés : les quantités de lait, de protéine et de matière grasse, les taux butyreux et protéique ainsi que le nombre de cellules somatiques (Tableau 7Erreur ! Source du renvoi introuvable.).

Training set	Model	Trait[2]						Average corr.[3]	Average slope[4]
		Milk	FatY	ProY	Fat%	Pro%	SCS		
A	Multi-breed GS	0.505	0.518	0.514	0.647	0.644	0.568	0.566	0.875
	Single-breed GS	0.484	0.489	0.495	0.638	0.632	0.561	0.550	0.863
	BLUP	0.316	0.346	0.305	0.350	0.397	0.430	0.357	0.760
B	Multi-breed GS	0.345	0.401	0.353	0.478	0.494	0.427	0.416	0.726
	Single-breed GS	0.312	0.389	0.333	0.398	0.470	0.418	0.387	0.710
	BLUP	0.234	0.299	0.244	0.305	0.382	0.399	0.311	0.663
C	Multi-breed GS	0.347	0.401	0.368	0.434	0.444	0.389	0.397	0.717
	Single-breed GS	0.319	0.396	0.340	0.368	0.420	0.368	0.368	0.729
	BLUP	0.243	0.306	0.249	0.305	0.373	0.393	0.311	0.685

[1]Genomic evaluations were based on a BayesC approach. Phenotypes used were standardized within breed; the proportion of residual polygenic component was set to 30% and the number of expected SNP with a nonzero effect to 7,000.

[2]Milk yield (Milk), fat yield (FatY), protein yield (ProY), fat content (Fat%), protein content (Pro%), and SCS.

[3]Average correlation between DYD and EBV over the 6 traits.

[4]Average regression slope over the 6 traits.

Tableau 7: Corrélations entre DYD observés et prédits pour six caractères (quantité de lait, matière protéique, matière grasse, taux protéique, taux butyreux et comptage des cellules somatiques) en utilisant un BLUP sur ascendance, un BAYES C intra-race ou un BAYES C multi-races

Cette étude a montré que, quelle que soit la taille de la population de référence, l'utilisation de la puce HD ne permet pas d'améliorer significativement l'efficacité d'une sélection génomique intra-race. Ce résultat est d'ailleurs cohérent avec la littérature (Erbe et al. 2012; Su et al. 2012; VanRaden et al. 2013) et avec le fait que les haplotypes conservés sont relativement grands. En revanche, dans le cas d'une évaluation multiraciale, l'utilisation de la puce HD est recommandée puisqu'elle permet d'augmenter la précision des valeurs génétiques estimées (Hoze et al. 2014a). Ce résultat est cohérent avec une meilleure conservation du déséquilibre de liaison ainsi qu'une meilleure cohérence des phases sur la puce HD (Larmer et al. 2014). Toutefois, malheureusement, les résultats acquis ne débouchent pas sur un gain de précision des index suffisant pour justifier la généralisation d'une approche multiraciale qui permettrait une mutualisation des efforts de constitution des populations de référence. Autrement dit, chaque race doit constituer une population de référence suffisante pour disposer d'index précis.

Chris Hozé a également réalisé un test d'évaluation génomique combinant les populations de référence de deux races génétiquement proches : la Simmental et la Montbéliarde. Cette étude a été réalisée en exploitant le puce 50K. Trois tests ont été réalisés (Tableau 8):

- EstMo : Utilisation en race Simmental des effets de SNP estimés chez la Montbéliarde pour prédire les GEBV des animaux Simmental.
- PreselMO : Utilisation de la race Montbéliarde pour présélectionner les SNP. Le fait de diminuer le nombre d'effets à estimer est une manière de réduire la complexité calculatoire et ainsi, d'améliorer la précision des effets estimés. Cette présélection est rendue possible grâce au BayesCπ, méthode d'évaluation génomique basée sur la sélection de variables.
- Multi-races : Analyse multiraciale combinant les populations de référence Montbéliarde et Simmental.

	Quantité de lait	Matière grasse	Matière protéique	Comtage des cellules somatiques	moyenne
EstMO	0.28	0.36	0.28	0.19	0.28
PreselMO	0.54	0.38	0.58	0.16	0.42
multi-race	0.53	0.49	0.61	0.27	0.48

Tableau 8: Corrélation entre performances observées et prédites pour 4 caractères en utilisant un BayesCπ.

Cette étude a montré un fort gain de précision des valeurs génétiques des animaux avec un gain de 9 points de corrélation (ce qui est trois supérieur à celui observé lors de la précédente étude) (Hoze et al. 2014b). Il est donc possible, même à partir de génotypes 50K, d'améliorer les évaluations génomiques grâce à une population de référence multiraces. Cette stratégie nécessite toutefois que les deux races soient suffisamment proches pour que le déséquilibre de liaison soit conservé entre races malgré la relativement faible densité de marqueurs.

4. Utilisation d'haplotypes dans les évaluations génomiques

Ces travaux ont été réalisés au sein du Work Package « méthodologie de la sélection génomique multiraciale » du projet GEMBAL dont je portais l'animation. Nous avons fait le constat que la plupart des méthodes d'évaluation décrites dans la littérature exploitent l'information SNP. Cependant, les différences de précision de ces méthodes d'évaluation génomique sont relativement faibles et l'utilisation de puces SNP à forte densité en intra-race, comme en multi-race a conduit à de faibles améliorations. L'utilisation d'haplotypes à la place des SNP, en augmentant fortement l'informativité des marqueurs, pourrait augmenter le déséquilibre de liaison entre allèles et QTLs. Dans ce contexte, nous avons proposé une extension du BayesCπ à l'utilisation d'haplotypes. Le modèle implémenté est le suivant :

$$y_i = \mu + u_i + \sum_{j=1}^{N} z_{ij}\alpha_j\delta_j + e_i$$

où y_i correspond à la performance de l'individu i, μ correspond à la moyenne, u_i est la composante polygénique de l'animal i, N est le nombre total d'haplotypes, z_{ij} représente le nombre de copies de « l'allèle » à l'haplotype j pour l'animal i, α_i correspond à l'effet de substitution du marqueur j, δ_j est une variable 0/1 indiquant si l'haplotype j est inclus dans le modèle (et a un effet estimé) ou non et enfin e_i correspond à l'erreur résiduelle de l'animal i.

Ce modèle a été implémenté au sein du logiciel GS3 (Legarra *et al.* 2013; Croiseau *et al.* 2014).

En étendant le BayesCπ à l'utilisation d'haplotypes, nous espérions un gain de précision des évaluations conséquent malgré un nombre d'effets à estimer plus important en utilisant ce modèle et une capacité des SNP à capter les effets de QTL même si leur LD individuel n'est pas complet. Malheureusement, les résultats de cette approche n'ont pas validé nos espoirs. En effet, nous avons testé cet algorithme sur un jeu de données Montbéliarde contenant 2235 taureaux soit en exploitant la puce 50K, soit en exploitant la puce HD après imputation. Les tests en intra-race sur puces 50K n'ont montré aucun gain de précision et les tests en multi-races à l'échelle de la HD ne sont pas envisageables car le nombre d'effets à estimer serait gigantesque (pour des haplotypes de 4 SNP, cela représente environ 2,4 millions d'effets à estimer) par rapport au nombre de phénotypes disponibles (Tableau 9).

Trait name[1]	Correlation coefficient				Regression slope			
	HS[2]: 2	HS: 3	HS: 4	HS: 5	HS: 2	HS: 3	HS: 4	HS: 5
MY	0.502	0.497	0.507	0.500	0.863	0.869	0.885	0.895
FY	0.557	0.557	0.563	0.559	0.863	0.871	0.912	0.905
PY	0.490	0.491	0.497	0.491	0.763	0.779	0.799	0.792
FC	0.576	0.572	0.571	0.559	0.868	0.874	0.894	0.894
PC	0.596	0.589	0.593	0.581	1.055	1.052	1.090	1.094
Average[3]	0.544	0.541	0.546	0.538	0.140	0.132	0.120	0.122

1: Trait name abbreviations: MY – milk yield; FY – fat yield; PY – protein yield; FC – fat content; PC – protein content

2: Haplotype size

3: Average deviations from 1 are indicated for regression slopes

Tableau 9: Corrélation et pente de régression entre performances observées et estimées mesurées sur la population de validation Montbéliarde avec le BayesCπ haplotypique.

Cette période correspond au début de la thèse de David Jonas (décembre 2013) que j'ai co-encadrée avec Vincent Ducrocq. Un des premiers objectifs de cette thèse était de rendre exploitable le BayesCπ haplotypique en intra-race. Pour cela, plutôt que d'inclure tout le génome sous forme d'haplotypes dans le modèle, nous avons sélectionné les régions d'intérêt à partir de résultats de détection de QTL. Puis, ces QTL ont été introduits sous forme d'haplotypes en utilisant les marqueurs flanquants au QTL. Cette stratégie a permis de réduire considérablement la complexité du modèle (avec 3000 QTL, environ 40000 effets à estimer) ainsi que les temps de calcul (1-2 heures).

Une réflexion est aussi née sur la meilleure manière de former les haplotypes. L'utilisation des marqueurs flanquants au QTL n'est pas optimale car le polymorphisme de ces haplotypes peut être faible (plus forte probabilité d'avoir un fort déséquilibre de liaison entre deux marqueurs proches) et le nombre d'allèles rares peut être élevé (or l'estimation de l'effet des allèles rares est moins précise). Un travail méthodologique a donc été entrepris pour sélectionner, au sein d'une région, les marqueurs qui permettent de maximiser le polymorphisme de l'haplotype et d'avoir une répartition des allèles la plus homogène possible.

Deux manières de sélectionner les SNP inclus dans un haplotype ont été étudiées. Pour ces deux approches, les SNP constituant les haplotypes sont choisis dans une fenêtre de marqueurs autour d'un SNPQTL. Les SNPQTL correspondent ici aux SNP ayant les probabilités d'inclusion les plus fortes après utilisation d'un BayesCπ. Une fois la taille de la fenêtre définie ainsi que la taille de l'haplotype (noté par la suite HS pour Haplotype Size et qui correspond au nombre de SNP constituant l'haplotype), tous les haplotypes possibles incluant le SNPQTL sont considérés pour les deux approches : critère A et B.

- Critère de choix de sélection des haplotypes A :

  Ce critère repose sur un seuil basé sur la fréquence des allèles (AFT : Allele Frequency Threshold) pour déterminer les allèles pour lesquels l'effet peut être estimé avec une efficacité satisfaisante (dans l'article de David Jonas, on parle de PA : predictable alleles, (Jonas *et al.* 2015)). Dans un premier temps, pour chaque haplotype au sein d'une fenêtre, le nombre de PA est calculé. Ensuite, les haplotypes qui portent le nombre maximal de PA, un score représentant l'écart à une répartition homogène des allèles est calculé de la manière suivante :

$$SD_{hi} = \sum_{k=1}^{N_i} \left(OF_{i,k} - \frac{1}{N_i} \right)^2$$

où h_i est l'haplotype i, N_i est le nombre de PA de l'haplotype i et $OF_{i,k}$ est la fréquence observée de l'allèle k de l'haplotype i.

Sous ce critère, retenir l'haplotype avec le SD le plus faible donne la garantie que cet haplotype possède les fréquences alléliques observées les plus homogènes.

- Critère de choix de sélection des haplotypes B :
  Un des inconvénients du critère A est qu'il est toujours possible d'avoir un fort déséquilibre des fréquences alléliques car les haplotypes qui maximise le nombre de PA sont toujours préférés aux haplotypes ayant un plus faible nombre de PA.
  Le critère B permet d'éviter cette dérive en remplaçant, dans le calcul de SD, $\frac{1}{N_i}$ par $\frac{1}{2^{HS}}$.
  Ainsi, on a la garantie que pour un écart de fréquences alléliques donné, l'haplotype constitué du plus grand nombre de SNP minimisera toujours le critère.
  Par ailleurs, un poids w est ajouté à ce critère sur le nombre de PA. Cela permet de garantir que pour différents haplotypes qui portent le même nombre d'allèles, celui qui porte le plus de PA aura le plus petit score.
  En pratique, ce score s'écrit de la manière suivante :

$$Critère\ B_{hi} = \sum_{k=1}^{N_i} \left(OF_{i,k} - \frac{1}{2^{HS}} \right)^2 - w \times N_i$$

  où *w* est le poids fonction du nombre de PA.

Le Tableau 10 illustre la différence entre le critère A et B. Sous le critère A, seuls les 2 premiers haplotypes sont sélectionnables. Les haplotypes 3 et 4 ne peuvent pas être retenus car il ne maximise pas le nombre d'allèles PA (un allèle avec une fréquence inférieure à 5% pour l'haplotype 3 et seulement 5 allèles présents pour l'haplotype 4). Ainsi, sous ce critère, l'haplotype 1 minimise le score car ses fréquences alléliques sont plus homogènes que celles de l'haplotype 2. Au contraire, pour le critère B, l'haplotype 2 est classé quatrième car, bien qu'il ait 6 allèles, ses fréquences alléliques sont très déséquilibrées. Ce critère préfèrera l'haplotype 4 qui, bien qu'ayant un allèle de moins, a des fréquences alléliques parfaitement équilibrées.

		Allele frequencies					
Criterion-A	Criterion-B	A1	A2	A3	A4	A5	A6
1	1	0.167	0.167	0.167	0.167	0.167	0.165
2	4	0.70	0.06	0.06	0.06	0.06	0.06
—[1]	3	0.2	0.2	0.2	0.19	0.19	0.02
—[1]	2	0.2	0.2	0.2	0.2	0.2	—

[1]As the first 2 haplotypes have 6 predictable alleles (assuming a threshold of allele frequency threshold = 5%), these haplotypes are not considered in the second step of criterion-A.

Tableau 10 : Fréquences alléliques pour 4 haplotypes et classement de ces haplotypes en fonction du critère de sélection utilisé.

Nous nous sommes intéressés à l'impact du critère de sélection des haplotypes sur la distribution des fréquences alléliques. L'utilisation des critères A et B conduit à une plus forte proportion d'allèles avec une fréquence comprise entre 5 et 30% mais aussi à une plus faible proportion des allèles à très forte fréquence (Figure 4). Par ailleurs, le critère B sélectionne des haplotypes avec moins d'allèles rares et surreprésentés que le critère A. Ces propriétés devraient être bénéfiques pour la prédiction génomique des caractères d'intérêt.

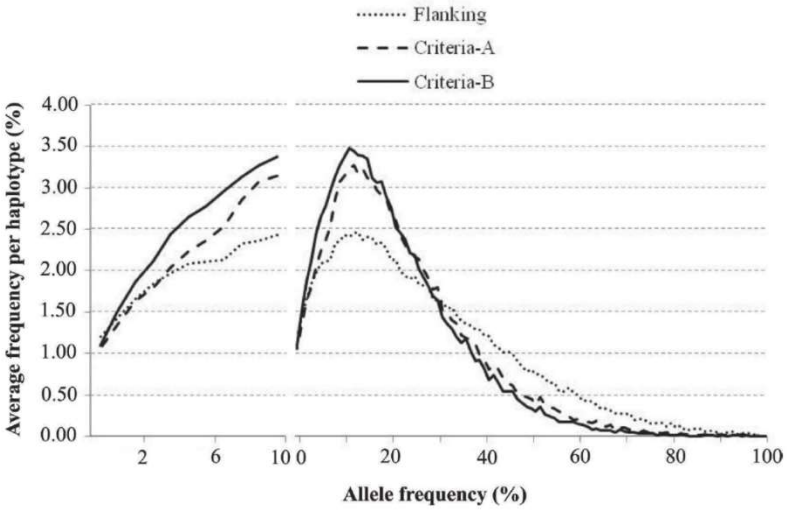


Figure 4 : Distribution des fréquences alléliques des haplotypes retenus en fonction du critère utilisé (taille de l'haplotype: 4 SNP, taille de la fenêtre: 10 SNP, nombre de SNP_{QTL}: 6000).

Nous avons voulu vérifier l'impact de ces approches de sélection d'haplotypes sur la prédiction génomique. Cette étude a été menée sur un panel de 2235 taureaux Montbéliards génotypés sur la puce 50K. Les évaluations génomiques ont été réalisées sur les cinq caractères de production laitière en utilisant le BayesCπ haplotypique développé dans le cadre du projet GEMBAL. Les résultats de cette étude sont présentés sur la Figure 5Erreur ! Source du renvoi introuvable.. Tout d'abord, nous avons constaté un gain de corrélation systématique apporté par l'utilisation d'haplotypes, quelle que soit la taille de l'haplotype testée. Si on s'intéresse aux différentes approches de sélection d'haplotypes, les critères A et B permettent un gain de corrélation par rapport aux SNP flanquants. Cependant, ce gain est plus fort pour des haplotypes de 3 SNP que du 4 ou 5 SNP. Dans tous les cas, c'est le critère B qui maximise les gains de corrélation, même si les écarts restent mesurés. En effet, nous observons un gain de corrélation moyen avec le critère A comparativement aux marqueurs flanquant de 1,3 et 0,6% pour des haplotypes de 3 et 4 SNP respectivement. L'utilisation du critère B permet d'améliorer le gain de corrélation du critère A de 0,3% en moyenne (Jonas *et al.* 2015).

En complément de cette étude, nous avons voulu vérifier l'intérêt d'exploiter cette stratégie pour une évaluation génomique sur puce HD. La littérature sur le sujet était très contrastée, certains articles vantant l'intérêt de la puce HD (Brøndum *et al.* 2011) tandis que d'autres n'y voient aucun intérêt (Chen *et al.* 2014; Hoze *et al.* 2014a). Les performances du critère B ont donc été testées sur puce HD, sur données Montbéliarde en intra-race. La Figure 6 confirme que l'utilisation des haplotypes avec la puce HD permet également un gain de corrélation substantiel (+1,8% en utilisant les marqueurs flanquant et +1% en utilisant le critère B).

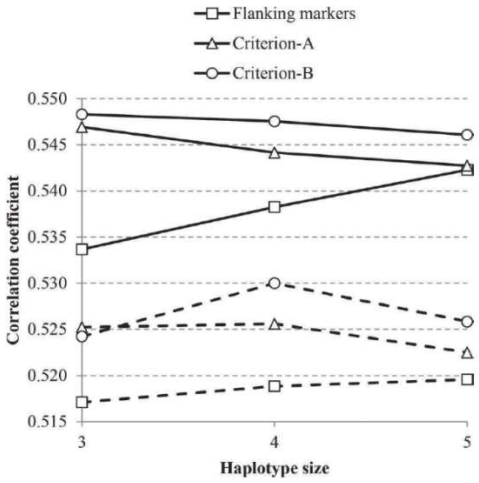


Figure 5: Corrélations entre performances observées et prédites dans la population de validation Montbéliarde en fonction du critère de sélection des haplotypes utilisés et de la taille des haplotypes. Les corrélations moyennes pour les 5 caractères de production sont montrées. Les lignes pleines correspondent aux corrélations des analyses basées sur haplotypes tandis que les lignes pointillées correspondent aux corrélations observées lorsque les mêmes SNP sont utilisés comme simples marqueurs SNP. Une fenêtre de 10 SNP a été utilisée pour les critères A et B.

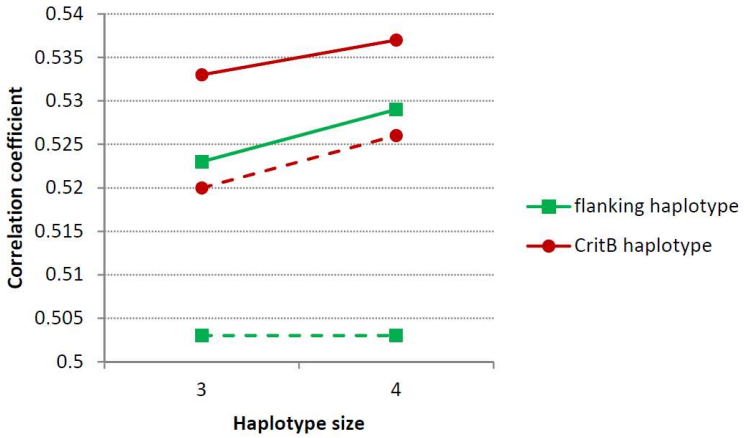


Figure 6: Corrélations moyennes observées entre performances estimées et observées pour les 5 caractères de production avec différentes méthodes de sélection d'haplotypes et de taille d'haplotype. Les lignes pleines indiquent les corrélations pour les tests basés sur haplotypes tandis que les lignées pointillées montrent les corrélations observées quand les mêmes marqueurs sont inclus en tant que simple SNP dans le modèle.

5. Construction des haplotypes à partir de l'information de déséquilibre de liaison

Dans l'étude précédente, les QTL étaient supposés ségréger au sein d'une petite région (±10 SNP) autour du SNP identifié dans l'étape de détection de QTL. Bien que cette fenêtre de ±10 ait été jugée raisonnable, cette approche n'est pas parfaite. Par exemple, une taille de fenêtre qui varie en fonction du taux de recombinaison local pourrait améliorer la recherche de l'haplotype optimal (Coop *et al.* 2008; Beissinger *et al.* 2015). En effet, les taux de recombinaison peuvent varier en fonction de la région, du chromosome ou de la population (Jeffreys *et al.* 2005; Weng *et al.* 2014; Ma *et al.* 2015). Par ailleurs, les SNP ne sont pas équidistants sur les puces SNP, ainsi, pour une fenêtre fixe, les différentes régions génomiques n'auront pas la même longueur. Pour s'affranchir de cette contrainte, les fenêtres ont été définies à partir de l'information de déséquilibre de liaison (blocs haplotypiques ou haploblocs).

Les SNP utilisés pour construire les haplotypes sont ensuite sélectionnés pour représenter au mieux ces « blocs haplotypiques » au lieu de représenter les régions entourant un SNP présélectionné (Jónás *et al.* 2016). L'information QTL n'étant pas exploitée, les mêmes haplotypes sont donc utilisés quel que soit le caractère. Le **Erreur ! Source du renvoi introuvable.** donne quelques statistiques concernant les haplotypes définis selon cette procédure avec comme valeur seuil de déséquilibre de liaison pour définir un haplobloc un D'=0,45.

Parameter	Value
Total number of markers	43,801
Number of haploblocks	8,393
Number of haplotypes built	7,804
Average number of SNP per haploblock	5.2
Average number of alleles per haplotype	13.3

Tableau 11: statistiques descriptives des haploblocs

Ces 7804 haplotypes identifiés ont été testés sur les données Montbéliardes pour les 5 caractères de production. Nous avons pu montrer que la construction d'haplotypes au sein de blocs haplotypiques conduit à une sensible amélioration de la prédiction des performances (+1,5 point de corrélation par rapport aux haplotypes construit autour des QTL) (Tableau 12).

	GBLUP[2]		Preselection method[2]		Haploblock information[3]	
Trait name	Correlation	Slope	Correlation	Slope	Correlation	Slope
Milk quantity	0.490	0.810	0.496	0.789	0.504	0.910
Fat yield	0.551	0.850	0.562	0.806	0.564	0.943
Protein yield	0.478	0.738	0.476	0.697	0.493	0.803
Fat content	0.570	0.785	0.594	0.865	0.637	0.933
Protein content	0.584	0.987	0.609	0.971	0.613	1.071
Average[4]	0.535	0.166	0.547	0.174	0.562	0.096

[1]DYD = daughter yield deviation; GBLUP = genomic BLUP; GEBV = genomic EBV.
[2]Results were taken from Jónás et al. (2016).
[3]Results obtained using haploblock information with a D′ threshold of 45%.
[4]For regression slopes, average absolute deviations from one are shown.

Tableau 12: Corrélations entre performances observes (DYD) et estimées (GEBV) et pentes de régression en utilisant différentes manières de sélectionner les haplotypes.

Cette approche présente un avantage important. Comme nous l'avons déjà mentionné, elle n'est pas basée sur une information QTL. Les mêmes haplotypes peuvent donc être utilisés quel que soit le caractère. Cette propriété est très intéressante car elle permet de s'affranchir d'une détection de QTL sur une population indépendante. Autrement dit, nous pourrions envisager d'identifier les haplotypes les plus pertinents en exploitant l'intégralité des animaux génotypés disponibles.

6. Rénovation de la stratégie d'évaluation génomique Française

a) Le cas des grandes races laitières

Fort de ces travaux, nous avons entrepris de rénover notre modèle de sélection génomique. Ce projet, dans la continuité du projet original AMASGEN, a été nommé AMASGEN2 et a été piloté par Vincent Ducrocq et Didier Boichard entre 2013 et 2015. Un des objectifs était de s'adapter à la massification des données, d'inclure les performances femelles dans les évaluations et de mettre à jour la liste de QTL utilisées pour définir les équations de prédiction des caractères déjà évalués (l'étude sur les haploblocs n'était pas encore achevée). J'ai été fortement impliqué dans ces travaux. Cela nous a conduit à changer la méthode d'identification des QTL : l'Elastic Net ayant été remplacé par le Bayes Cπ. En fonction des

caractères, les 1000 ou 3000 SNP ayant la probabilité d'inclusion la plus forte dans un BayesCπ ont été définis comme QTL. Suivant les travaux de David Jonas, ces QTL ont été modélisés sous la forme d'haplotypes en utilisant le critère B décrit précédemment.
En rénovant notre modèle d'évaluation génomique, nous avions le souhait de ne pas inclure de composante polygénique. Utiliser une composante polygénique permet, pour bon nombre de caractères, d'améliorer les pentes de régression chez les candidats à la sélection (ce qui permet d'avoir un classement non biaisé des jeunes taureaux par rapport à la population de référence). Cependant, en marge des évaluations génomiques officielles, nous fournissons de manière hebdomadaire des index non officiels calculés uniquement à partir d'une information génomique (Direct Genomic Value ou DGV). Ce dernier modèle n'ayant pas de composante polygénique, la cohérence entre les DGV des indexations non officielles et les GEBV des indexations officielles n'est pas garantie. En s'affranchissant d'une composante polygénique, nous pouvions éviter cet écueil.
Afin de s'assurer de la pertinence de ce modèle, nous avons donc comparé différentes approches d'évaluation génomique :

- Un GBLUP et un BayesCπ. Pour ces deux approches, l'inclusion d'une composante polygénique expliquant entre 0 et 90% de la variance génétique a été testée. Pour chaque caractère, la part de composante polygénique qui maximise la précision des GEBV a été retenu.
- Un Marker-Assisted BLUP avec composante polygénique pedigree ou **ped_HAPsel** :

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{n} \mathbf{H}_i\boldsymbol{\beta}_i + \mathbf{Z}u + \mathbf{e}$$

où **y** est le vecteur de performances, **β** correspond au vecteur d'effets alléliques des haplotypes pour les *n* QTL inclus dans le modèle, **u** correspond aux effets de la composante polygénique basée sur une matrice de parenté pedigree, **H** est la matrice indiquant les « allèles » paternel et maternel portés par chaque individu pour chaque haplotype (*i)*, **Z** est la matrice d'incidence associant individu et performance, et **e** correspond à l'erreur résiduelle.

Pour cette stratégie, chaque SNP_{QTL} a été intégré au sein d'un haplotype de 4 marqueurs selon la méthode présentée précédemment (page 17) dans le cadre des travaux de thèse de David Jonas (critère B).

Pour chaque caractère, 1000 ou 3000 SNP_{QTL} ont été inclus dans le modèle. Si pour la plupart des caractères les meilleurs résultats sont obtenus avec 3000 SNP_{QTL}, pour les caractères affectés par un/des gène(s) majeur(s), comme DGAT pour le taux butyreux, les meilleurs résultats sont obtenus avec 1000 SNP_{QTL}. Dans la Figure 7, pour chaque caractère, les résultats présentés ont été obtenus avec le nombre optimal de SNP_{QTL}.

Nous avons testé des parts de variance génétique expliquée par la composante polygénique entre 10 et 90%. Pour chaque caractère, la valeur optimale a été retenue.

- Un Marker-Assisted BLUP sans composante polygénique ou **10k_HAPsel** :

Pour ce modèle, qui s'appuie sur le modèle ped_HAPsel, la composante polygénique pedigree a été remplacée par l'information d'un pool de SNP qui nous permet de construire une matrice de parenté génomique remplaçant la composante polygénique pedigree.

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{n} \mathbf{H_i}\boldsymbol{\beta_i} + \mathbf{S}\,\gamma + \mathbf{e}$$

où ϒ et **S** correspondent au vecteur d'effets de m SNP et à la matrice des génotypes correspondants.

Dans cette étude, les SNP de la puce basse densité (LD) ont été utilisés pour construire cette matrice de parenté génomique car ces SNP assurent une bonne couverture du génome et nous apportent la garantie d'avoir des génotypes observés (et non des génotypes imputés).

Les résultats de cette étude sont présentés sur la Erreur ! Source du renvoi introuvable.. Dans l'ensemble, nous ne constatons pas d'énorme différence d'efficacité de prédiction entre les différentes méthodes d'évaluation génomique. Le GBLUP est un peu moins performant en race Normande que le BayesCπ, ped_HAPsel et 10K_HAPsel. L'approche 10K_HAPsel semble apporter un peu plus de précision pour les caractères morphologiques en race Montbéliarde et pour les caractères de production en race Brune mais fait un peu moins bien que les autres approches pour les caractères fonctionnels en races Montbéliarde et Brune.

Cependant, l'approche 10K_HAPsel nous permet d'avoir des indexations hebdomadaires non officielles en complète cohérence avec les indexations officielles sans détériorer la qualité de précision des index génomiques.

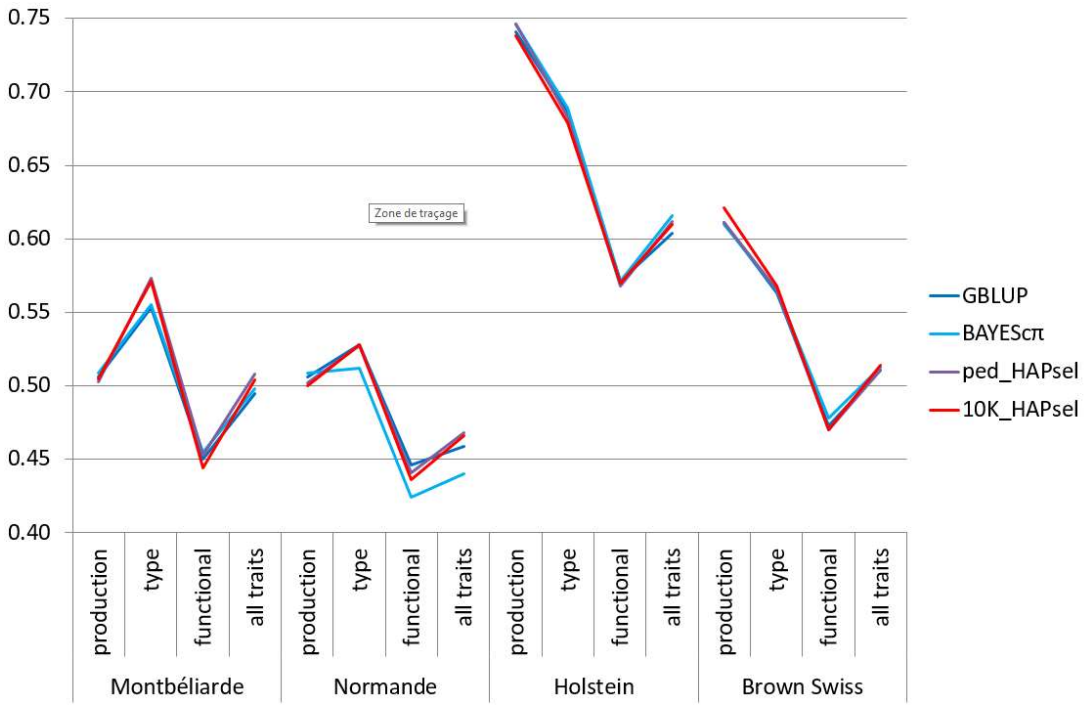


Figure 7: Corrélation moyenne par groupe de caractères et pour l'ensemble des caractères obtenue en utilisant un GBLUP, un BayesCπ, un MABLUP avec composante polygénique (ped_HAPsel) et sans composante polygénique (10K_HAPsel)

Cette méthode est donc devenue la méthode d'évaluation génomique de référence pour les évaluations génomiques Française des 4 grandes races bovines laitières (Montbéliarde, Normande, Holstein et Brune) à partir de juin 2015.

b) Extension de notre modèle d'évaluation pour les races bovines laitières régionales

Dans l'objectif de mettre en place une évaluation génomique intra-race, les professionnels des races Abondance, Tarentaise et Vosgienne ont étendu leurs populations de référence en génotypant avec la puce 50K des vaches représentatives de la diversité raciale et disposant de performances pour les caractères évalués. Afin d'adapter notre modèle d'évaluation génomique aux races régionales, le projet G2R a vu le jour en 2011. La sélection des animaux et leur génotypage ont été réalisés entre 2011 et 2016. En mars 2016, les populations de référence étaient constituées de 389 males et 2769 femelles en race Abondance, 323 males et 1569 femelles en race Tarentaise et 66 males et 1171 femelles en race Vosgienne.

Dans les trois races, l'évaluation a porté sur 40 à 46 caractères disposant d'index polygénique (production, fertilité, résistance aux mammites, morphologie, longévité), ainsi que divers composites. Les phénotypes des taureaux sont les performances moyennes de leurs filles non typées (DYD), tandis que pour les vaches, les phénotypes sont les performances propres (YD) ou leur moyenne en cas de données répétées. Le modèle d'évaluation génomique utilisé est celui des races nationales. Pour un caractère donné, il comprend des QTL suivis par des haplotypes de 4 SNP, et une composante génomique résiduelle estimée avec les 9000 SNP de la puce LD. Compte tenu de la taille des populations de référence, pour bien estimer les effets des QTL, leur nombre a été limité à 500-1000 dans les races régionales (contre jusqu'à 3000 dans les races nationales). Des études de validation croisée ont été réalisées sur les caractères de production et ont montré un gain de corrélation important en Abondance, Tarentaise contre un gain modéré pour la Vosgienne (Tableau 13).

	Abondance	Tarentaise	Vosgienne
pedigree-based BLUP	0.346	0.391	0.418
MABLUP	0.459	0.449	0.43

Tableau 13: Corrélations moyennes sur les 5 caractères de production entre GEBV et YD dans la population de validation.

Le passage à la sélection génomique a pour objectif de mettre en avant des jeunes taureaux au détriment d'animaux plus vieux testés sur descendance. Cependant, pour diffuser un taureau, il doit garantir que la précision apportée par ses phénotypes (niveau de CD) soit suffisante. Les seuils établis sont de 50% pour la production et la morphologie et 35% pour les caractères fonctionnels. Une étude a été réalisée pour s'assurer que les niveaux de CD étaient bien suffisants. Nous avons pu montrer que la majorité des animaux génotypés (sans phénotypes mais avec des parents évalués et leur père génotypé) peuvent avoir leurs index génomiques publiés pour l'ensemble des caractères évalués (Tableau 14).

Race	Nombre	Lait	Cellules	Morphologie	Fertilité vache
Abondance	56	54,1 (59)	50,5 (54)	50,7 (54)	39,5 (45)
Tarentaise	95	52,3 (57)	47,6 (51)	48,7 (51)	34,4 (37)
Vosgienne	24	53,8 (56)	44,8 (48)	49,1 (52)	33,1 (37)

Tableau 14: CD moyens (maximaux) des mâles de moins de 24 mois au traitement de mars 2016

L'évaluation génomique de ces races a été officialisée en mars 2016. Il s'agit d'une première mondiale pour des races de cette taille, qui disposent maintenant des outils les plus modernes pour leur sélection (Sanchez *et al.* 2016).

Compte tenu de la taille limitée de la population de référence et de l'origine génétique relativement proche de ces trois races régionales, une étude multi-race a également été menée dans le cadre de la thèse de David Jonas. Cette étude a conclu à un intérêt de combiner certaines races pour un gain de corrélation. C'est particulièrement vrai pour les races Abondance et Simmental (Figure 8). Cependant, la non réciprocité de ces résultats d'une race à l'autre et la compétition qu'il peut exister entre les races n'ont pas permis d'exporter ces résultats dans les évaluations génomiques officielles.

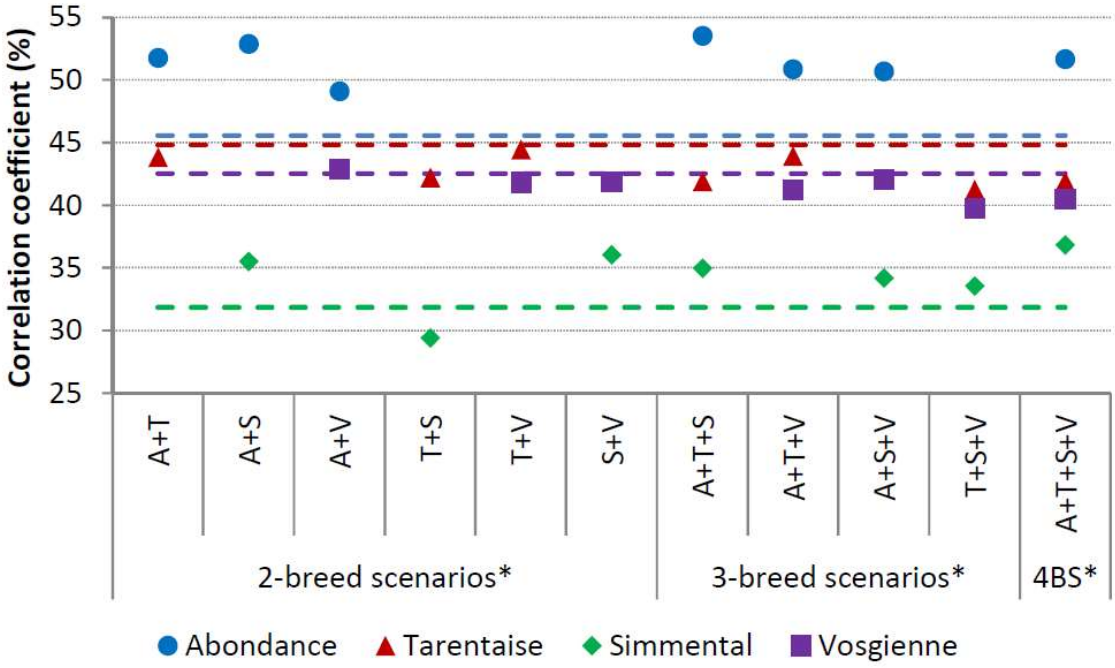


Figure 8: Corrélations observées dans la population de validation après une évaluation génomique multi-raciales pour 4 races différentes. Pour chaque analyse, la population de validation est uni-raciale. Les lignes hachurées correspondent aux scenarios intra-races. Les Abréviations de l'axe des abscisses signifient : A – Abondance ; T – Tarentaise ; S – Simmental ; V – Vosgienne.

III. De l'identification des mutations candidates à leur inclusion dans les modèles d'évaluation génomique

Vu sous le spectre des évaluations génomiques, un bon moyen d'améliorer les prédictions génomiques est d'inclure dans le modèle les marqueurs impliqués dans les caractères d'intérêt. Avec les puces SNP, nous disposons, au mieux, de marqueurs en déséquilibre de liaison (plus ou moins fort) avec les mutations causales mais pas des mutations causales. Leurs effets sont donc potentiellement partiels et surtout ils s'érodent avec l'accumulation de recombinaisons, induisant une dégradation de la précision des prédictions lorsque la distance augmente entre candidat et population de référence. Au cours de ces dernières années, je me suis fortement impliqué dans la comparaison de méthodes d'identification des mutations candidates pour, ensuite, définir des stratégies pour les inclure dans les modèles d'évaluation génomique.

A. Méthodes statistiques pour la dissection de la variabilité des caractères à l'aide de puces SNP

Les méthodologies autour de la détection de QTL sont nombreuses avec une littérature étayée mais ces méthodes ont, pour la plupart, été développées pour répondre aux besoins d'analyse en génétique humaine. Or, les espèces animales d'élevage ont leurs spécificités telles que leur structuration en races chacune d'effectif génétique limité, des effets fondateurs marqués ou une très grande capacité de reproduction.
Identifier, en fonction de la population d'étude, quelle est l'approche de détection de QTL la plus pertinente a été l'objectif du projet ANR « Rules and Tools » piloté par Jean-Michel Elsen de 2010 à 2015. Au sein de ce projet, j'ai été impliqué dans l'identification des méthodes les plus prometteuses pour la détection des SNP associés à un QTL afin de les comparer selon différents critères (Legarra *et al.* 2015): le calcul de l'efficacité de localisation des QTL ; l'estimation de la précision de l'effet des SNP ; la robustesse de ces méthodes. Trois méthodes ont retenu notre attention pour leur aptitude à prendre en compte la structure de nos populations :

- La méthode LDLA (Linkage Disequilibrium and Linkage Analysis) développée par Meuwissen *et al.* 2002 et qui combine deux types d'information, la liaison intra famille et le déséquilibre de liaison au niveau populationnel. Cette méthode qui utilise des haplotypes nécessite d'avoir des données phasées. Des haplotypes de plusieurs SNP consécutifs sont construits. La transmission de ces haplotypes est suivie le long du pedigree et permet d'estimer la probabilité d'identité au QTL entre apparentés. Les composantes de la variance sont estimées pour chaque locus en utilisant un algorithme de maximum de vraisemblance restreinte, soit avec (hypothèse alternative), soit sans (hypothèse nulle) les effets haplotypiques. Le modèle est le suivant :

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{Th} + \mathbf{e}$$

où **y** est le vecteur des phénotypes, **u** et **h** sont les vecteurs d'effets polygéniques et haplotypiques (avec leur matrice d'incidence respective **Z** et **T**) et **e** est le vecteur d'effet résiduel. $Var(\mathbf{h}) = \mathbf{H}\sigma^2_{\mathrm{h}}$ et $Var(\mathbf{u}) = \mathbf{A}\sigma^2_{\mathrm{u}}$ sont les matrices de covariances haplotypiques et polygéniques avec comme composantes de la variance respectives σ^2_h et σ^2_{u}.

Un test de rapport de vraisemblance est utilisé pour comparer les hypothèses nulle et alternative. Le processus est répété pour chaque locus du chromosome d'intérêt.

- La méthode EMMA (Efficient Mixed-Model Association) qui est une méthode de régression pour réaliser des GWAS, étendue pour la prise en compte d'un effet polygénique pour tenir compte de l'apparentement dans la population (Kang *et al.* 2010; Habier *et al.* 2011; Teyssèdre *et al.* 2012). Pour chaque marqueur, le modèle suivant a été utilisé :

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{w}s + \mathbf{e}$$

où **u** est le vecteur d'effets polygéniques, comme pour le modèle LDLA mais avec $Var(\mathbf{u}) = \mathbf{G}\sigma_u^2$ où **G** correspond à la matrice d'apparentement génomique. **w** est un vecteur de covariables codées (0,1,2) pour chaque génotype au SNP étudié et s est l'effet de substitution au marqueur considéré. Un test de Student a été utilisé pour tester la significativité des marqueurs.

- La méthode BayesC qui fait partie de la famille des méthodes Bayésiennes conçues initialement pour prédire la valeur génétique des animaux mais qui peut également être utilisée pour la cartographie de QTL (Meuwissen *et al.* 2001; Hoggart *et al.* 2008; Habier *et al.* 2011). Contrairement aux précédentes, cette méthode permet d'analyser simultanément l'intégralité des marqueurs. Pour réduire la contrainte du grand nombre de variables à estimer comparativement au nombre d'animaux disponibles, le modèle BayesC inclut des variables indicatrices d=$\{d_1, \ldots, d_n\}$ qui indiquent si un marqueur est $\{d_i=1\}$ ou n'est pas $\{d_i=0\}$ inclus dans le modèle:

$$\mathbf{y} = \mathbf{1}\mu + \sum \boldsymbol{w}_i\, d_i s_i + \boldsymbol{e}$$

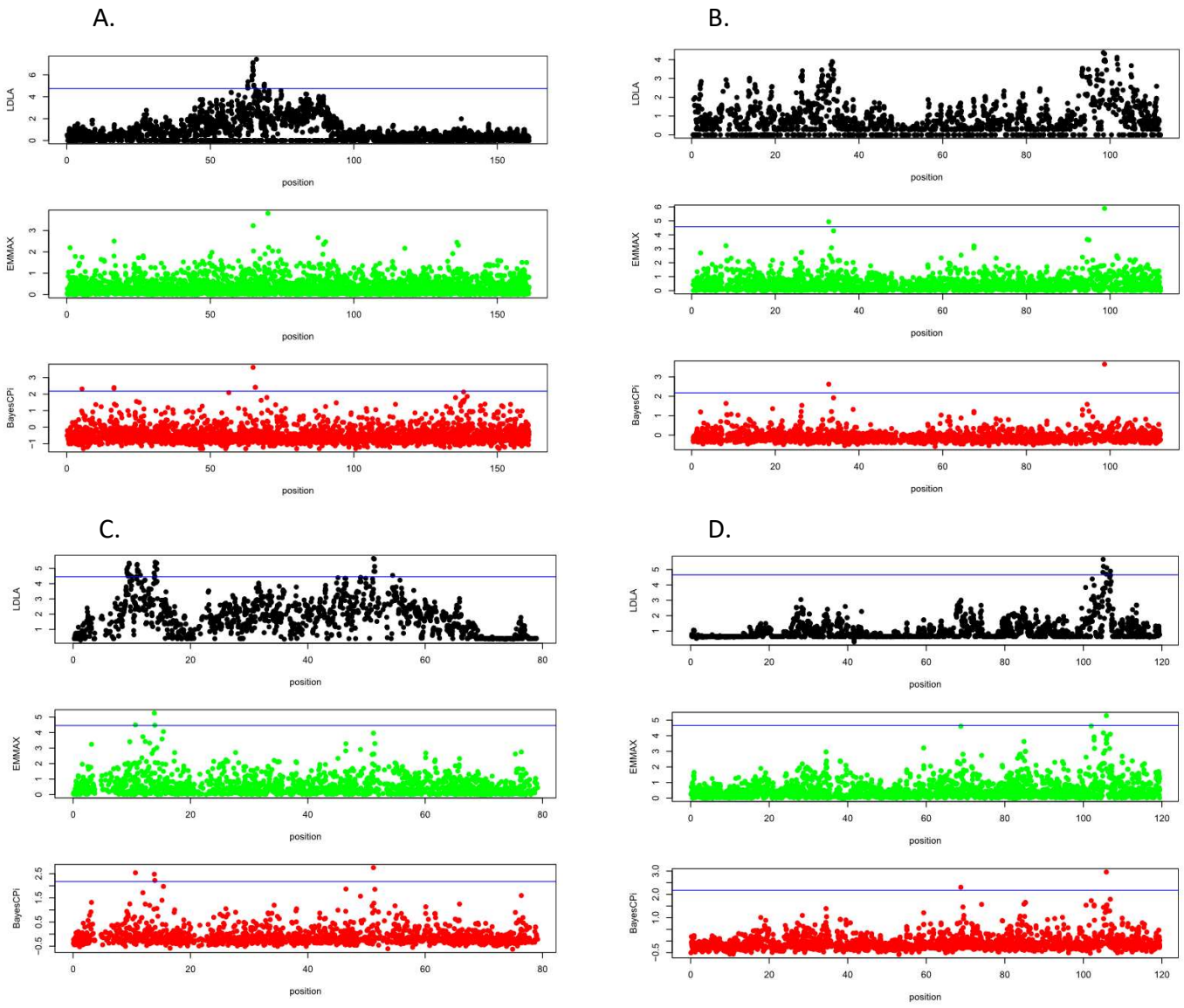
Un paramètre clé du modèle est le nombre de SNP retenus dans le modèle *a priori*. Dans cette étude, ce paramètre a été fixé à $\Pr(d_i = 1) = 1/1000$, sachant que des tests avec 1/100 n'ont pas changé qualitativement les résultats.
La statistique utilisée est un Bayes Factor (BF) qui correspond à l'augmentation entre les probabilités a priori et a posteriori d'un SNP d'être retenu dans le modèle (Sorensen and Gianola 2007; Wakefield 2009).

Pour évaluer la performance de ces approches, nous les avons testées à travers cinq jeux de données différents issus de quatre espèces. Pour chaque jeu de données, un seul chromosome a été analysé, pour lequel une région QTL connue ou candidate avait déjà été identifiée :

- Bovins laitiers : Le chromosome 1 a été testé pour la quantité de lait sur une population de 1221 bovins de race Montbéliarde.
- Bovins allaitants : Le chromosome 7 a été testé pour la tendreté de la viande sur une population de 936 bovins de race Blonde d'Aquitaine.
- Ovins allaitants : Le chromosome 12 a été testé pour la résistance à l'infection aux nématodes sur une population de 1067 moutons issus de trois générations d'un croisement backcross Blackbelly x Romane (F1, backcross et backcross x backcross).
- Cheval : le chromosome 3 a été testé pour l'incidence de l'ostéochondrose du jarret sur une population de 627 trotteurs.

- Porc : Le chromosome 17 a été testé pour la longueur de la carcasse sur une population de 764 porc issus de 3 races (495 Large White, 129 Landrace et 140 Pietrain).

La *Figure 9* présente les GWAS réalisés sur les différents chromosomes et populations décrits ci-dessus. Globalement, les trois approches de détection de QTL détectent les mêmes régions QTL. Par contre, les profils sont assez différents. Les régions identifiées par l'approche LDLA impliquent souvent un nombre de SNP plus important qu'avec les approches EMMA et BayesC (*Figure 9*. A-C-D), principalement du fait de l'impact de la liaison intra famille à l'origine d'une forte corrélation entre tests de marqueurs proches. Par ailleurs, la structure de covariance des haplotypes change peu entre haplotypes adjacents (où seul un SNP diffère dans la composition des haplotypes).



E.

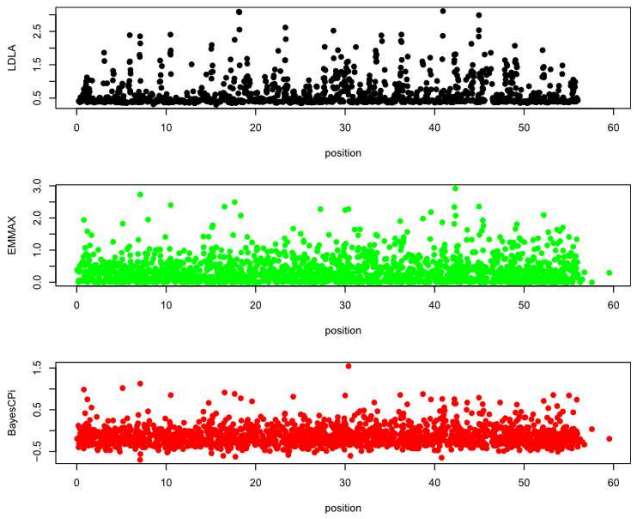


Figure 9: Manhattan plot des différents chromosomes analysés. L'axe des abscisses représente la position sur le chromosome tandis que l'axe des ordonnées représente la statistique de test (-log(1/p-value) pour les méthodes LDLA et EMMAX et log(Bayesian Factor) pour le BayesCpi). Le chromosome 1 des bovins laitiers est présenté en A, le chromosome 7 des bovins allaitants en B, le chromosome 12 des ovins allaitants en C, le chromosome 3 du cheval en D et le chromosome 17 porcin en E.

Si les profils GWAS sont similaires entre méthodes, la significativité des régions diffère selon l'espèce et l'approche étudiées. Ainsi, pour les données porcines, aucune méthode n'a été capable de détecter la région candidate (Figure 9.E). Bien que les méthodes utilisées soient capable de prendre en compte la structure de population, l'utilisation d'une population intégrant trois races différentes est peut-être à l'origine de cet échec. Le cas des bovins allaitants est aussi intéressant car aucun QTL significatif n'a pu être trouvé par l'approche LDLA alors que le QTL est clairement mis en évidence avec EMMA et BayesC (Figure 9.B).
L'analyse multi-marqueurs (BayesC) présente l'avantage de mettre en compétition les SNP entre eux, de sorte que les marqueurs distants ne sont pas retenus alors qu'ils sont significatifs dans les analyses mono-marqueur du fait du LD à grande distance. La cartographie est donc plus fine. Par contre, le nombre d'effets est plus important et la puissance est plus réduite.
Parmi ces 3 approches, le BayesC est tout de même plus complexe à utiliser car il nécessite un certain nombre de paramètres tels que la fréquence des marqueurs dans la population ou la valeur de π (proportion de SNP avec un effet non nul à chaque itération) qui sont subjectifs (nous n'avons accès qu'aux fréquences alléliques du jeu de données et la valeur du π optimale n'est pas connue). Par ailleurs, le seuil de rejet du Bayesian Factor n'a pas de fondement statistique et ne fait donc pas consensus. C'est pourquoi, en conclusion de cette étude, nous recommandions plutôt EMMA (Legarra *et al.* 2015). Toutefois, avec la densification en marqueurs des puces, ces méthodes sont devenues trop chronophages en raison de l'estimation de la composante polygénique. Dès 2007, une approche a été développée pour résoudre ce problème : le modèle GRAMMAR (Genome-wide Rapid Association using Mixed Model And Regression) développé par Aulchenko *et al.* (2007) Cette approche repose sur l'estimation d'une composante polygénique unique qui sera utilisée par la suite pour l'ensemble des SNP analysés. Le modèle sous-jacent est exactement le même que pour EMMA, à savoir :

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{ws} + \mathbf{e}$$

Où **u** est le vecteur d'effets polygéniques, comme pour le modèle EMMA mais avec $Var(\mathbf{u}) = \mathbf{G}\sigma_u^2$ ou **G** correspond à la matrice d'apparentement génomique. **w** est un vecteur de covariables codées (0,1,2) pour chaque génotype au SNP étudié et s est l'effet de substitution au marqueur.

Mais, pour l'approche GRAMMAR, une première étape consiste à estimer **u** sans effet de marqueurs puis, de générer une performance corrigée de ces effets :

$$\boldsymbol{y}^* = \mathbf{y} - (\mathbf{Zu})$$

Ensuite, on applique une régression linéaire simple de y* sur chacun des marqueurs analysés, soit :

$$\boldsymbol{y}^* = \mathbf{ws} + \mathbf{e}$$

Cette approche en deux étapes n'est pas strictement équivalente au modèle EMMA mais il s'agit d'une approximation permettant de réaliser des GWAS en minimisant les temps de calcul. Dans le cadre de nos études sur données de séquence, nous avons beaucoup utilisé le modèle GRAMMAR.

Dans le cadre de ce projet, j'ai également encadré David Jonas durant son stage de M2. Nous avons développé un modèle LDLA multipoint pour réduire l'intervalle de confiance autour d'un QTL. La localisation de ces QTL dépend principalement de l'information de déséquilibre de liaison à longue distance dans une région et cette information de liaison peut être perçue comme un bruit pour la détection de QTL. L'objectif de ce stage était de corriger le test de ce déséquilibre de liaison à grande distance (Jónás *et al.* 2014).

Pour cela, le modèle LDLA de (Meuwissen *et al.* 2002) a été étendu en ajoutant des haplotypes en cofacteurs à gauche et à droite de la position testée, dans le but de masquer le déséquilibre de liaison à longue distance dans l'analyse. Dans cette étude, la distance qui sépare les haplotypes a été définie à 3cM. Ce nouveau modèle, nommé ***lrLD***, prend la forme suivante :

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{w_1}\boldsymbol{h_1} + \mathbf{w_2}\boldsymbol{h_2} + \mathbf{w_3}\boldsymbol{h_3} + \mathbf{e}$$

Où $\mathbf{h_1}$-$\mathbf{h_3}$ sont les vecteurs d'effets aléatoires d'haplotype et $\mathbf{w_1}$-$\mathbf{w_3}$ sont les matrices d'incidences des haplotypes h_1-h_3.

Cette équation est utilisée en tant qu'hypothèse alternative dans un test de rapport de vraisemblance (Likelihood Ratio Test ou LRT). Le modèle sous l'hypothèse nulle est similaire mais ne contient pas l'haplotype h_2 (qui est défini comme étant l'haplotype testé) :

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{w_1}\boldsymbol{h_1} + \mathbf{w_3}\boldsymbol{h_3} + \mathbf{e}$$

Cette approche a été comparée au modèle original de Meuwissen et al. (que nous avons nommé ***MG*** par la suite) sur un jeu de données constitué de 3940 taureaux Holstein génotypés sur la puce 50K. Le caractère étudié est le taux de matière protéique.

Dans un premier temps, nous nous sommes intéressés au chromosome 20 pour lequel un gène majeur pour le taux de matière protéique a été précédemment identifié dans la population Holstein (Blott *et al.* 2003). Ce chromosome est également un bon exemple pour tester si les différents modèles sont capables de distinguer deux QTL étroitement liés car un second QTL est localisé à 6,9Mb en aval du gène GHR : le gène du récepteur à la prolactine ou

PRLR. Les résultats de l'analyse d'association réalisé sous les deux modèles sont présentés sur la Figure 10.

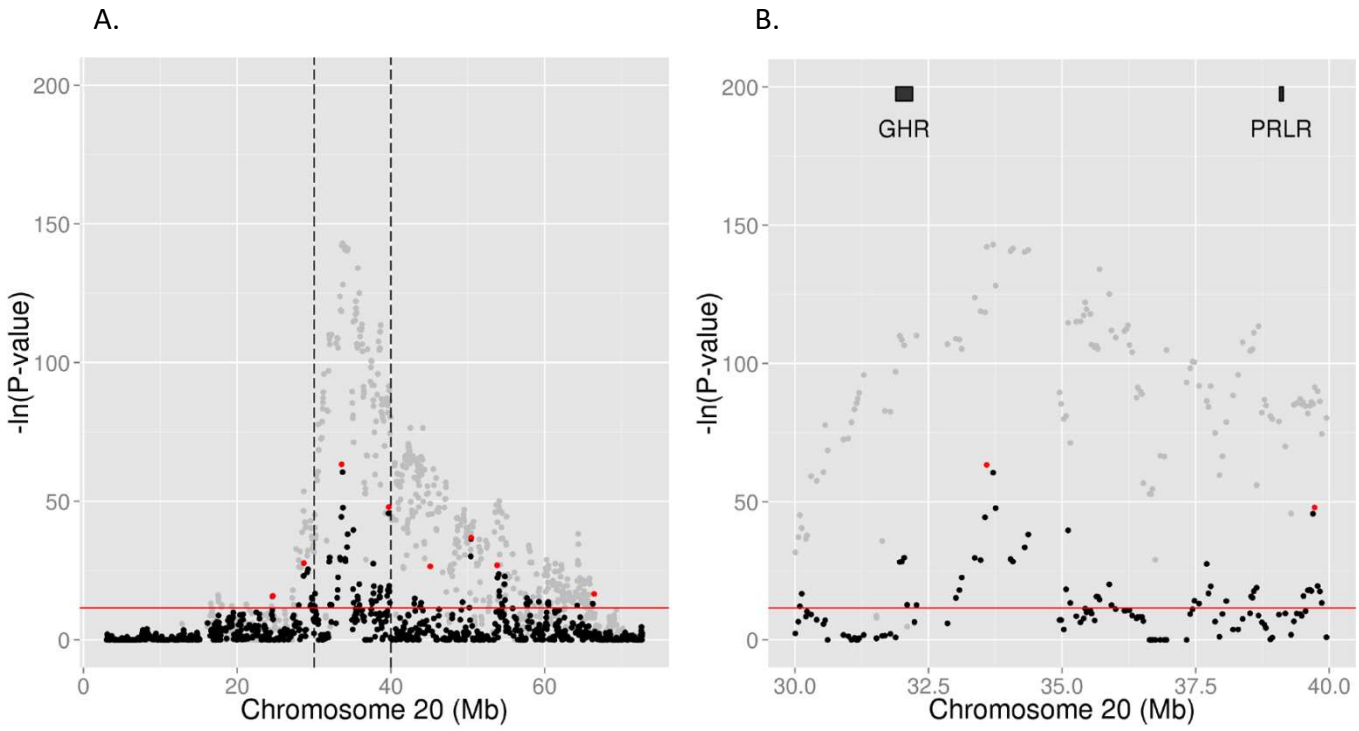


Figure 10: Etude d'association réalisée sur le chromosome 20 (A) ou sur une portion du chromosome 20 (B) pour le caractère taux protéique. Le seuil de rejet de l'hypothèse nulle après correction de Bonferroni est représenté par la droite rouge (α=1%). Les points gris représentent les résultats obtenus par le modèle MG tandis que les points noirs représentent ceux obtenus par le modèle lrLD. Les points rouges indiquent les QTL détectés par le modèle lrLD. Les lignes verticales pointillées représentent la région des gènes GRH-PRLR qui a été conservée sur le graphique B. Les deux rectangles en haut du graphique B indiquent la position des gènes GHR et PRLR.

Le premier résultat apparent est la diminution drastique des –ln(p-value) sous le modèle lrLD. Si 535 des 1045 positions testées sont significatives avec le modèle MG, seules 93 positions restent significatives avec le modèle lrLD. La finesse des pics est aussi sensiblement améliorée ce qui permet l'identification de 5 QTL majeurs et plusieurs QTL mineurs dans les 30cM autour du pic détecté par le modèle MG. Parmi ces 5 QTL majeurs, 2 sont situés dans la région d'intérêt. Ces deux pics sont situés à 1,39 et 0,59Mb des gènes GHR et PRLR respectivement. Notons qu'un des petits QTL détectés par le modèle lrLD se trouve exactement sous le gène GHR.

En plus du chromosome BTAU20, ces deux modèles ont été testés sur trois autres chromosomes bovins, BTAU3, BTAU5, BTAU15. Le nombre de QTL identifiés a été calculé ainsi qu'un score correspondant au ratio entre le –ln(p-value) observé au pic et la valeur moyenne des –ln(p-value) des SNP dans une fenêtre de +/- 3Mb autour du pic. Ce score décrit la forme du pic, ainsi, les scores les plus élevés sont attribués aux pics les plus étroits. Ces statistiques sont présentés sur le Tableau 15.

Chromosome ID	Number of predicted QTLs		Average peak score	
	MG	lrLD	MG	lrLD
BTA3	8	5	4.61	5.98
BTA5	9	2	3.46	6.37
BTA15	9	3	5.03	8.98
BTA20	4	8	2.66	5.81

Tableau 15 : Nombre de QTL identifiés par chromosome avec le modèle publié par Meuwissen et al. (modèle MG) et avec le modèle intégrant des haplotypes en cofacteur (modèle lrLD). Les scores moyens des pics sous ces deux modèles sont aussi présentés.

De manière générale, le modèle lrLD identifie moins de QTL que le modèle MG. Analysant trois QTL liés, il est inévitablement moins puissant. La seule exception est le chromosome 20 mais cela est dû à une région en fort déséquilibre de liaison autour du gène GHR, où la détection de QTL proches n'était pas possible avec le modèle MG (Figure 10). Cependant, après avoir masqué le déséquilibre de liaison à longue distance avec le modèle lrLD, plusieurs QTL ont pu être détectés dans cette région. Par ailleurs, notre modèle montre des scores moyens de pic plus élevés que le modèle MG. Ce constat est particulièrement visible sur le chromosome 20 qui possède un gène majeur. Le modèle lrLD permet donc de réduire substantiellement la taille des régions QTL détectées.
Ce travail a été présenté au congrès mondial WCGALP mais n'a finalement pas été publié, les approches LDLA étant passées de mode, considérées comme trop complexes, et remplacées par les analyses d'association.

B. Exploitation des données de séquences bovines

1. L'apport du projet 1000 génomes bovins

Même si leur densité peut être élevée, les puces commerciales disponibles ne nous donnent pas accès à l'intégralité des variations du génome et donc pas à tous les variants causaux. Initié en 2012, le projet international « 1000 génomes bovins » qui rassemble aujourd'hui 36 partenaires a pour objectif le partage des données de séquences complètes (incluant toutes les variations du génome) d'animaux de l'espèce bovine (Daetwyler et al., 2014, Bouwman et al., 2018). Depuis le début du projet, plusieurs « runs » se sont succédés et lors de son sixième run en 2017, les données de séquences de 2333 bovins de près de 70 races ou types génétiques étaient disponibles. Lorsque nous avons démarré nos travaux sur données de séquence, nous avons exploité les données du run4 qui intégrait 1147 animaux séquencés provenant de 27 races bovines différentes, dont une majorité d'Holstein (288 animaux) et quelques dizaines d'animaux Normands et Montbéliards. Cette population de référence a rendu possible l'imputation de la séquence complète des taureaux et des vaches génotypés pour les puces commerciales. Nous avons suivi les recommandations de (Van Binsbergen *et al.* 2014) qui ont montré qu'en présence de différentes densités de puces, il était préférable d'imputer en deux étapes, de la puce 50k vers la puce haute densité en intrarace, puis de la puce haute densité vers la séquence en multirace. Ce travail en deux étapes permet, au cours de la 1ère étape, de tirer parti des nombreux génotypages HD disponibles intra race, puis, au cours de la 2ème étape, de tirer parti du déséquilibre de liaison entre races, partiel mais déjà fort, observé à l'échelle HD, expliquant le gain de précision apporté par la population de référence multiraciale séquencée. Les imputations ont été

réalisées par le logiciel FImpute (Sargolzaei *et al.* 2014) dans un premier temps, puis, plus récemment par Minimac (Fuchsberger *et al.* 2015).

A la fin de cette étape, 6382 taureaux Holstein, 2616 taureaux Montbéliards et 2344 taureaux Normands avec séquence imputée étaient disponibles pour des études d'association sur une liste de variants (SNP ou insertions-délétions de quelques nucléotides (indel)) proche de l'exhaustivité et qui contient donc, en théorie, les variants causaux qui ont des effets sur les caractères. Notons cependant que cette phrase doit être modulée dans la mesure où il existe d'autres variants structuraux (insertions, délétions, duplications, inversions, ...), souvent négligés dans une première approche, et que certains variants, surtout ceux de MAF faible, sont imputés avec une précision réduite.

a) Méta-analyse de la stature chez les bovins

Au niveau international, une méta-analyse des études d'association réalisées au sein de chaque pays et pour chaque race a été menée sur la stature des bovins. Notre équipe a été sollicitée pour réaliser les études d'association des animaux génotypés en France et j'ai pris part à cette étude. Les effets de chaque SNP ont ensuite été mutualisés pour réaliser une méta-analyse. Pour chaque SNP, la méta-analyse calcule une statistique z (et une p-value associée) à partir de la somme pondérée des statistiques z des études d'association réalisées dans chacune des 17 populations bovines qui ont contribué à l'étude, avec des poids proportionnels à la racine carrée du nombre d'animaux utilisé pour chaque étude d'association indépendante (Figure 11) :

$$\boldsymbol{Z} = \frac{\sum_{i=1}^{k} Z_i * \delta_i}{\sqrt{\sum_{i=1}^{k} \delta_i^2}}$$

Avec $Z_{i=}\, \phi^{-1}\left(\frac{p_i}{2}\right) * \Delta_i$ ou ϕ^{-1} est l'inverse de la loi normale cumulative et $\Delta_i$correspond à la direction de l'effet du SNP dans la race i ; $\delta_{i=}\sqrt{n_i}$ et n_i le nombre d'animaux dans la race *i*. Cette approche permet de combiner des analyses réalisées dans des unités différentes.

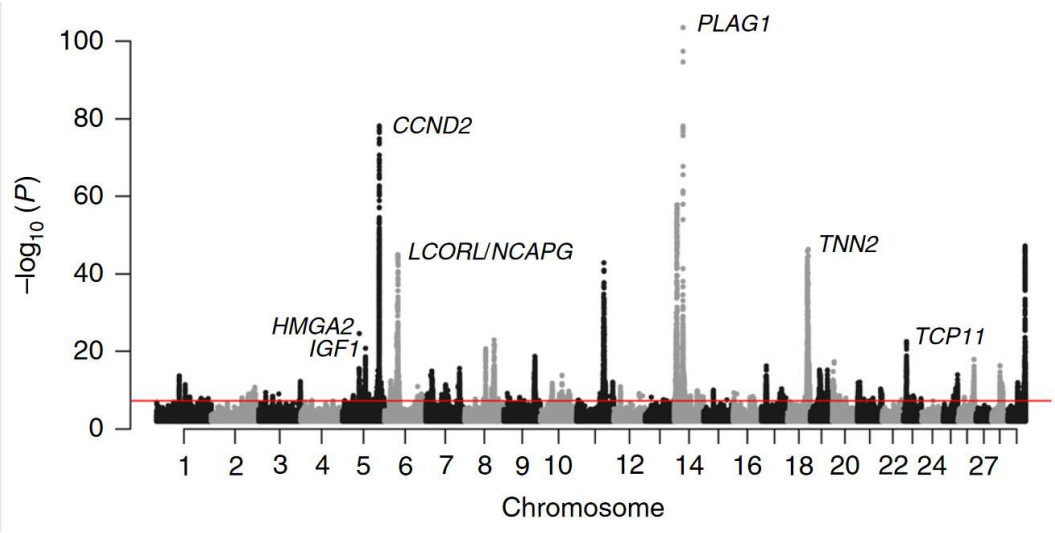


Figure 11 : Manhattan plot de la méta-analyse de la stature des bovins avec 58265 animaux. La ligne rouge correspond à un seuil de significativité de $5x10^{-8}$. Les gènes candidats les plus probables au sein des régions candidates les plus significatives ont été annotés.

Cette étude a permis de mettre en évidence 163 régions du génome bovin impliquées dans la variabilité de la stature des animaux. Par ailleurs, comme le montre la **Erreur ! Source du renvoi introuvable.**, la plupart des gènes en cause ont été identifiés. Une fraction significative des gènes impliqués dans la variabilité de la stature des bovins le sont également chez d'autres mammifères tels que l'homme ou encore le cheval. Parmi ces gènes, plusieurs, comme PLAG1 (Pleomorphic adenom gene 1) ou LCOR (ligand corepressor gene) ainsi que des régions génomiques qui les entourent, ne présentent plus de variabilité dans certaines races, ce qui témoigne de l'importante pression de sélection qui s'est exercée au fil du temps à leur égard et donc sur la taille des animaux.
Cette étude a également montré que la complexité du déterminisme génétique de la stature des bovins s'apparente à celle qui a été observée chez l'homme où les gènes mis en évidence n'expliquent que 10 à 20% de la variabilité (Bouwman *et al.* 2018).

b) Design de la partie custom d'une puce basse densité

Entre 2015 et 2017, un travail collaboratif à grande échelle impliquant toute l'équipe G2B a permis l'étude du déterminisme génétique des caractères à l'échelle de la séquence afin de mettre en évidence les polymorphismes causaux des QTL et permettre leur utilisation en sélection. Ces travaux ont été réalisés sur les trois grandes races laitières (Holstein, Normande et Montbéliarde) sur l'intégralité des caractères disponibles. Les analyses GWAS ont été réalisées avec le logiciel GCTA, qui utilise un modèle GRAMMAR. L'exploitation de ces analyses a permis d'identifier un grand nombre de QTL significatifs. La Figure 12 montre un exemple de GWAS sur le caractère « implantation des trayons » en race Holstein sur le chromosome 26.

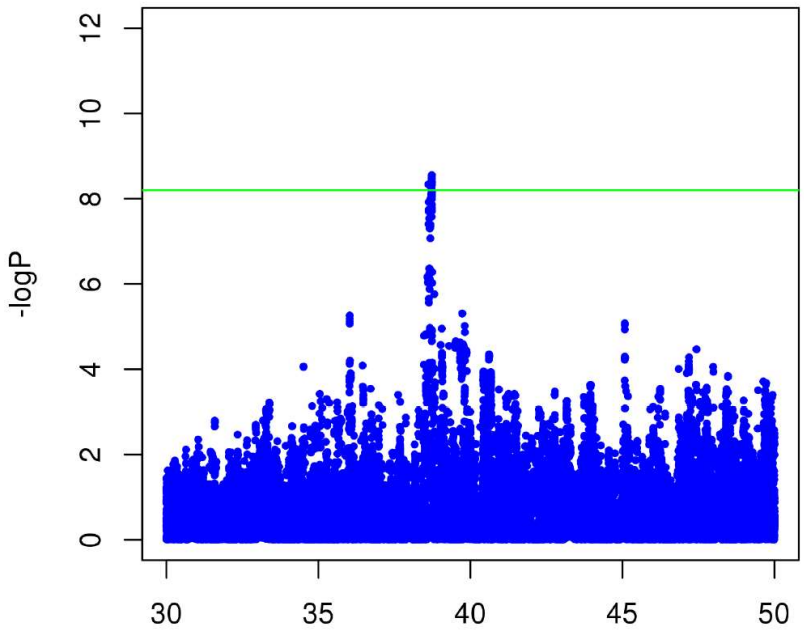


Figure 12: GWAS sur séquence des taureaux pour l'implantation des trayons en race Holstein sur une région du chromosome 26 (Tribout et al, séminaire du département 2018). La ligne verte correspond au seuil de rejet de l'hypothèse nulle après correction de Bonferroni pour les tests multiples.

A partir de ce résultat, le gène candidat du pic observé a été recherché en exploitant les données d'annotation du génome bovin. Pour cet exemple, c'est le gène RAB11FIP2 qui correspond au meilleur candidat (Tribout et al., 2017). Ce type d'analyse a été réalisé pour l'ensemble des caractères et pour les trois races.

Depuis 2014, nous participons à l'évolution de la puce EuroG10k d'Illumina qui, en plus des marqueurs présents sur la puce basse densité, inclut de nouveaux marqueurs (Boichard *et al.*):

- Des marqueurs de la puce 50K ou HD (pour améliorer la qualité d'imputation en fin de chromosome, pour compléter quelques régions dépourvues de marqueurs ou pour imputer correctement les microsatellites servant au contrôle parental).
- Des marqueurs hautement prédictifs dans les modèles d'évaluation.
- Des variants fonctionnels publics issus de la littérature (dont 125 variants correspondant à des anomalies génétiques ou à des gènes d'intérêt).
- Des marqueurs privés spécifiques à chaque pays contributeur.

En ce qui nous concerne, les marqueurs privés ont été sélectionnés sur différents critères :

- Les marqueurs candidats pour des anomalies génétiques
- Les marqueurs au sein des pics GWAS
- Les marqueurs avec une annotation fonctionnelle fortement délétère.
- Les marqueurs qui pourraient être impliqués dans la régulation fonctionnelle.
- Les variants structuraux altérant la structure de certains gènes.

Le nombre de place sur la partie privative de la puce EuroG10K étant limitée, une sélection relativement sévère des variants les plus intéressants a été appliquée.

2. Inclusion des mutations causales/candidates dans les évaluations génomiques

De 2015 à 2018, j'ai coordonné avec Christèle Robert-Granié (INRA, GenPhyse) un projet du métaprogramme SelGen nommé INCOMINGS pour Inclusion des mutations causales dans les modèles d'évaluation génomique. Ce projet, impliquant trois unités du département de génétique animale de l'INRA (GenPhyse, PEGASE et GABI) s'articulait autour de 4 taches :

1. Combiner des informations génomiques obtenues sur des supports hétérogènes via des méthodes d'imputation
2. Développement d'un modèle BLUP génomique single-step
3. Développement d'un modèle d'évaluation génomique combinant gènes majeurs, mutations causales, haplotypes et part polygénique
4. Etude de la conservation des gènes majeurs entre races et de leurs effets

Mon travail a essentiellement porté sur la troisième tâche, qui, dans le cadre des bovins laitiers, a consisté à exploiter les études GWAS sur séquence dans les évaluations génomiques. Pour cela, nous avons construit des puces SNP virtuelles ou tout ou partie des SNP de la puce 50K ont été remplacés par des SNP issus des données de séquence. Les trois premières stratégies testées proposent un remplacement de tous les SNP de la puce 50K :

- PEAK : un SNP sélectionné a la plus petite p-value dans une fenêtre de ±150 SNP. Si plus de 43800 pics sont détectés (43800 correspond au nombre de SNP utiles et autosomiques sur la puce 50K), seuls les plus significatifs sont retenus. Dans cette stratégie, les SNP sélectionnés peuvent être principalement localisés sur certaines régions du génome. En conséquence, la couverture du génome, une propriété importante de la puce 50K, n'est plus respectée.
- COVER : Le génome est divisé en 43800 segments. Au sein de chaque segment, les SNP avec une MAF supérieure à 1% et la plus petite p-value sont retenus. Ce scenario permet de maintenir une bonne couverture du génome. Cependant, seul un SNP ne peut être retenu par segment et par conséquent, certaines mutations candidates importantes pourraient ne pas être retenues. Le filtre sur la MAF a été appliqué pour minimiser le risque d'exploiter des SNP qui sont faussement associés au caractère d'intérêt (en raison d'erreur d'imputation ou simplement car les SNP avec une MAF faible ont un risque accru d'être associé à un caractère à tort).
- COVER2 : Toutes les propriétés de la stratégie COVER sont maintenues mais, en plus, une priorité est donnée aux SNP qui sont localisés dans un gène. Ce scenario fait l'hypothèse que, dans une région ou plusieurs SNP significatifs sont présent, celui qui est localisé dans un gène est un meilleur candidat pour être causal.

Deux stratégies supplémentaires ont été testées mais cette fois, seule une partie des SNP de la 50K ont été remplacés :

- OPT_QTL : le génome est divisé en intervalle d'1Mb. Les SNP candidats doivent alors remplir des conditions sur la MAF (supérieure à 1, 3 ou 5%) et sur leur niveau de significativité (10^{-3}, 10^{-4} ou 10^{-5}). Sous ces conditions, le nombre de SNP retenus comme mutation candidate varie entre 169 et 747 SNP. Dans chaque région avec au moins un SNP candidat, le SNP le plus significatif remplace le SNP de la 50K avec la MAF la plus faible (qui est supposé être le moins informatif de la région).
- BOTTOM-UP : Lorsque l'on bâtit des équations de prédiction sur une population d'apprentissage, une proportion non négligeable de SNP avec un effet proche de zéro est présente quel que soit le caractère. Cette stratégie propose de remplacer ces SNP

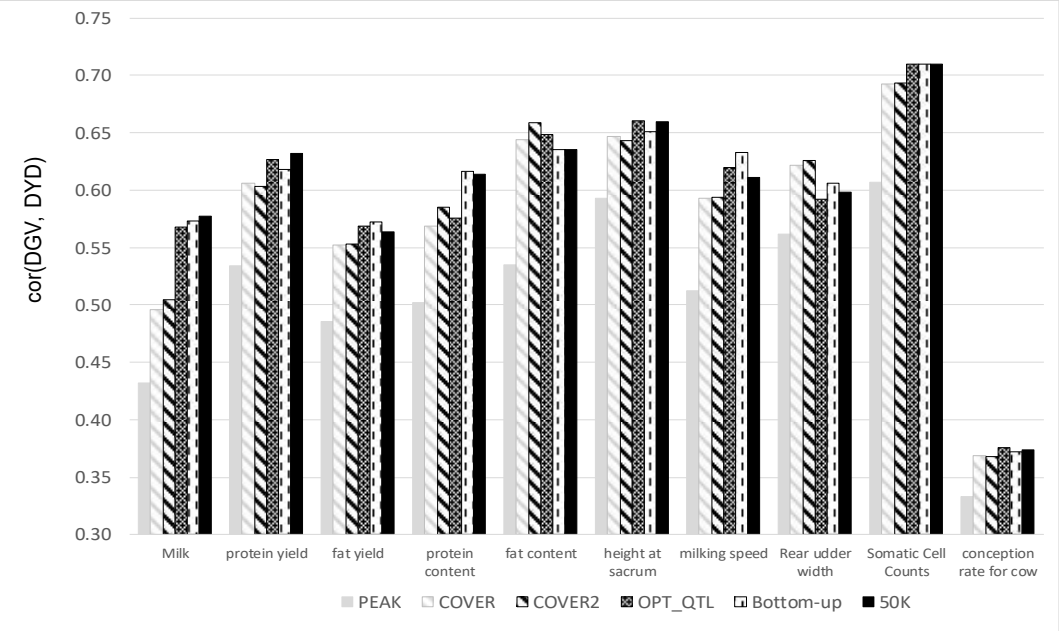
(qui n'ont pas d'intérêt pour les évaluations génomiques) par les SNP, issus des données de séquence, qui dans une région de ±0,5Mb ont la plus faible p-value. Pour chaque caractère cette stratégie a entrainé le remplacement d'environ 4000 SNP.

Ces scénarios ont été testés sur le race Montbéliarde et la race Holstein pour un panel de 10 caractères incluant les cinq caractères de production (quantité de lait, matière protéique, matière grasse, taux protéique et taux butyreux), deux caractères morphologiques (distance plancher/jarret et stature), et trois caractères fonctionnels (nombre de cellules somatiques, fertilité vache et vitesse de traite).
Les évaluations génomiques ont été réalisées en utilisant un GBLUP. La Figure 13 montre la précision de la prédiction des valeurs génétiques pour les races Montbéliarde (A) et Holstein (B) calculée à partir des corrélations entre les valeurs génomiques estimées et les performances des animaux (DYD) pour 10 caractères et pour les 5 stratégies testées ainsi que sur la puce 50K.

Pour la race Montbéliarde, la stratégie PEAK est celle qui donne les plus faibles précisions des prédictions génomiques. Les scénarios COVER et COVER2 ont permis d'améliorer l'efficacité de prédiction cependant, il n'y a que deux caractères pour lesquels l'efficacité de prédiction de ces scénarios est meilleure que celle obtenue avec la puce 50K : le taux butyreux et la stature. La stratégie OPT_QTL permet, quant à elle, d'améliorer les résultats de la 50K pour 8 des 10 caractères. Cependant, le gain de précision reste très limité avec un écart compris entre -0,5 et 2,1 points de corrélation en fonction du caractère. Enfin, la stratégie Bottom_up n'a pas permis d'améliorer la précision des prédictions puisque la précision de prédiction n'a été améliorée que pour un seul caractère (la matière grasse).
Pour la race Holstein, les résultats sont un peu plus favorables. Pour chaque caractère, au moins une puce virtuelle permet d'améliorer sensiblement les corrélations obtenues avec la puce 50K. La hiérarchie des différentes puces virtuelles est la même que pour la race Montbéliarde, ainsi, la stratégie la plus intéressante est OPT_QTL qui permet d'améliorer la précision des prédictions génomiques pour l'ensemble des caractères étudiés avec un gain de corrélation compris entre 0 et 2,4 points.

En conclusion de cette étude, l'exploitation des données de séquence à travers des puces virtuelles ne permet d'améliorer que très légèrement la précision des prédictions génomiques (Croiseau *et al.* 2017).

A.



B.

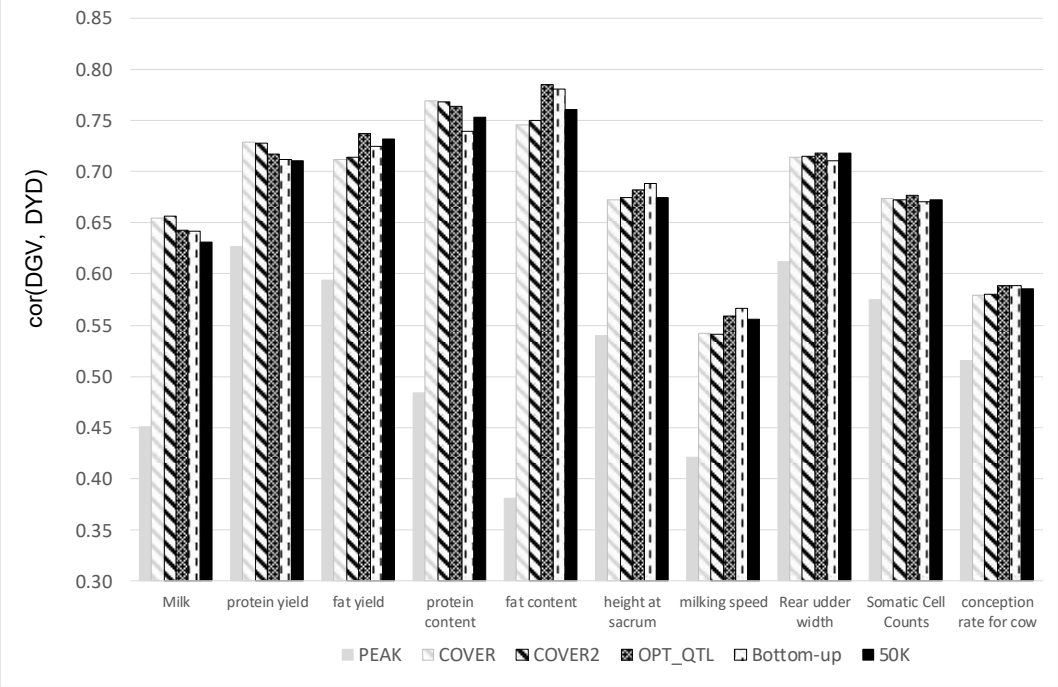


Figure 13 : Précision de la prédiction des valeurs génétiques pour les races Montbéliarde (A) et Holstein (B) calculée à partir des corrélations entre les valeurs génomiques estimées et les performances des animaux (DYD) pour 10 caractères et pour les 5 stratégies testées, en plus des résultats sur la puce 50K.

Nous avons également testé les performances du panel de SNP inclut dans la partie custom de la puce EuroG10K dans les évaluations génomiques. En moyenne, sur cette puces, 10 à 15 SNP ont été inclus par région candidate pour un total de 1759 SNP. En plus de ces marqueurs, les principales mutations causales connues et présentes sur la partie publique de la puce ont été incluses dans nos analyses. Cela représente 62 SNP issus de 36 gènes.

Nous avons testé ces marqueurs (mutations candidates de la partie custom et mutations causales de la partie publique de la puce EuroG10K) en utilisant un GBLUP en intégrant également tous les marqueurs de la puce 50K. Par ailleurs, nous avons également voulu tester

des modèles ou différentes variances génétiques additives pouvaient être attribuées en fonction des connaissances que l'on a des SNP (mutation causale, mutation candidate ou SNP de la puce 50K). Le Tableau 16 indique les différentes combinaisons de variances génétiques additives attribuées en fonction des groupes de SNP.

		genetic variance for each group of SNP		
	SNP used	50K	candidate mutations	causal mutations
same genetic variance for all SNP = GBLUP	50K	100%	-	-
	50K_all	100%		
different genetic variance according to the group of SNP = MABLUP	50K_Caus	80 -> 90%	-	10 -> 20%
	50K_Cand	80 -> 90%	10 -> 20%	-
	50K_all	60 -> 90%	0 -> 20%	0 -> 20%

Tableau 16 : Liste des SNP testés dans un GBLUP (tous les SNP ont la même variance génétique) ou un MABLUP (en fonction du statut du SNP « mutation causale, mutation candidate ou SNP de la puce 50K », différentes part de variances génétiques sont testées (par pas de 10%).

La

Figure 14 présente les précisions de prédiction génomique en race Montbéliarde (A) et Holstein (B) obtenues en utilisant tout ou partie de la puce EuroG10K en plus des SNP de la puce 50K. Dans cette étude, tous les scénarios décrits dans le Tableau 16 ont été réalisés mais seuls ceux qui maximisent la corrélation entre valeurs génétiques estimées et DYD sont présentés en

Figure 14.

A. B.

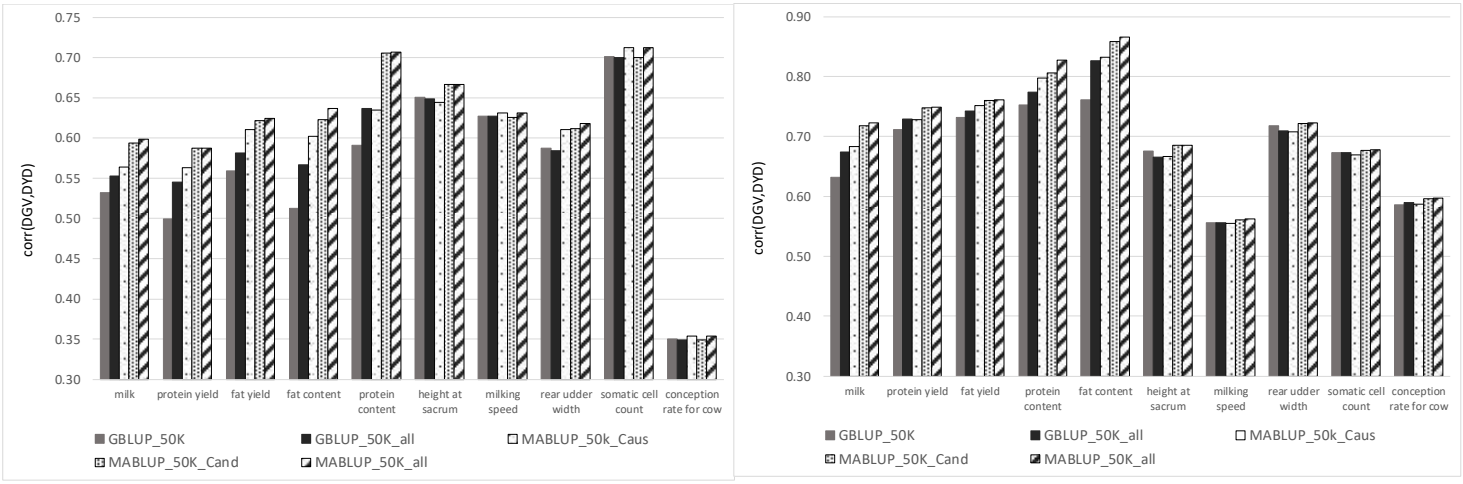


Figure 14:Corrélations entre les valeurs génomiques estimées et les performances (DYD) en race Montbéliarde (A) ou Holstein (B) pour 10 caractères et pour 4 stratégies différentes, en plus du GBLUP sur la 50K.

Pour les deux races, l'utilisation des SNP de la puce EuroG10K dans un GBLUP a permis un gain de corrélation pour la plupart des caractères (avec un maximum à +5,4 points en Montbéliarde et +6,5 points pour le taux butyreux). Lorsque la part de la variance génétique attribuée au

SNP dépend de la catégorie du SNP (mutation candidate, causale ou SNP de la puce 50K), le gain de corrélation augmente drastiquement. Lorsque seules les mutations candidates sont incluses dans le modèle, le gain de corrélation atteint 11,5 points pour le taux protéique en race Montbéliarde et 9,7 points pour le taux butyreux en race Holstein. Lorsque seules les mutations causales sont incluses dans le modèle, le gain de corrélation atteint 9 points de corrélation en race Montbéliarde et 7,1 points de corrélation en race Holstein pour le taux butyreux. Enfin, lorsque les deux catégories de SNP sont exploitées, la précision de la prédiction génomique est améliorée pour tous les caractères avec un maximum de +12,5 points en race Montbéliarde et 10,5 points en race Holstein pour le taux butyreux.

Le choix des SNP, la densification des régions candidates et l'utilisation de variance génétique adaptée permettent donc une amélioration sensible de la qualité de prédiction des valeurs génomiques (Croiseau *et al.* 2017).

3. Exploitation des données de séquence dans un contexte multiracial

Dans le chapitre précédent, nous avons abordé l'apport des données de séquence sur la recherche de mutations candidates ou causales et leur inclusion dans les modèles d'évaluation génomique. Ces études ont été conduites en intra-race or la densité de marqueurs que nous apporte les données de séquence nous permet de travailler avec un déséquilibre de liaison populationnel et non race spécifique, ce qui est une première car, même avec la puce HD, ce n'est pas le cas.
Nous pouvons donc maintenant envisager des études d'association pangénomiques multiraciales ou des méta-analyses d'études d'association intra-race pour gagner en puissance de détection des mutations candidates. C'est que nous avons fait dans le cadre du projet INCOMINGS à travers un stage de M2, réalisé par Marc Teissier en 2015 portant sur « les études d'association au niveau de la séquence : Comparaison des méthodes de méta-analyses et analyses multi-race chez les bovins laitiers » (Teissier *et al.* 2018).

Pour cette étude, nous disposions d'un panel de 6262 taureaux Holstein, 2434 taureaux Montbéliards et 2175 taureaux Normands avec données de séquences après imputation et phénotypés pour les cinq caractères laitiers (quantité de lait, matière grasse, matière protéique, taux protéique et taux butyreux). Les analyses d'associations intra-race ont été conduites en utilisant un modèle GRAMMAR.

En ce qui concerne les analyses multi-races, différentes approches ont été testées :

- L'analyse jointe : il s'agit également d'un modèle GRAMMAR mais avec un effet race :

$$\boldsymbol{y} = \boldsymbol{Wa} + \boldsymbol{Zu} + \boldsymbol{w}s + \boldsymbol{e}$$

où **a** est le vecteur d'effets de la race et **W** est la matrice d'incidence correspondante, **u** est le vecteur d'effets polygéniques, **w** est un vecteur de covariables codées (0,1,2) pour chaque génotype au SNP étudié et s est l'effet de substitution au marqueur, enfin, **e** est le vecteur de résiduelles.

- La méta-analyse des études d'association intra-races en utilisant l'approche FIXED : Dans cette approche, l'effet des SNP est une moyenne pondérée des effets intra-races :

$$\boldsymbol{\beta} = \frac{\sum_{i=1}^{k} w_i * \beta_i}{\sum_{i=1}^{k} w_i}$$

où l'effet β_i de la race *i* et w_i correspond au poids qui s'exprime de la manière suivante :

$$w_i = \frac{1}{SE_i^2}$$

où SE_i est l'écart type d'erreur de β_i :

$$\boldsymbol{SE} = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}$$

- La méta-analyse des études d'association intra-races en utilisant l'approche RANDOM : Cette approche autorise aux SNP d'avoir un effet différent en fonction de la race. Ainsi, le poids est calculé en prenant en compte deux sources de variation possible : une issue de l'erreur résiduelle intra-race (celle qui est prise en compte dans le modèle FIXED), et une issue de l'erreur résiduelle entre populations. Cette hétérogénéité de variance est définie de la manière suivante :

$$\alpha_i = \frac{1}{SE_i^2 + \tau^2}$$

avec $\tau^2 = \frac{(Q-(k-1))}{(\sum_{i=1}^{k} w_i - \left(\frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}\right))}$

et $Q = \sum_{i=1}^{k} w_i * (\beta_i - \beta)$

où w_i correspond au poids défini dans le modèle FIXED pour la race *i*, k correspond au nombre de races présentes et β correspond à l'estimation de l'effet du SNP dans l'approche FIXED. L'effet pondéré et son écart type s'expriment alors de la manière suivante :

$$\boldsymbol{\beta} = \frac{\sum_{i=1}^{k} \alpha_i * \beta_i}{\sum_{i=1}^{k} \alpha_i}$$

$$\boldsymbol{SE} = \sqrt{\frac{1}{\sum_{i=1}^{k} \alpha_{i_i}}}$$

- La méta-analyse des études d'association intra-races en utilisant l'approche Z-SCORE : Cette approche n'exploite pas les estimations des effets de SNP mais les p-value nominales (p_i) pour la race *i* et la direction de l'effet β (Δ_i) pour calculer une p-value multiraciale suivant la statistique suivante :

$$\boldsymbol{Z} = \frac{\sum_{i=1}^{k} Z_i * \delta_i}{\sqrt{\sum_{i=1}^{k} \delta_i^2}}$$

Avec $Z_{i=}\, \phi^{-1}\left(\frac{p_i}{2}\right) * \Delta_i$; $\delta_{i=}\sqrt{n_i}$ et n_i le nombre d'animaux dans la race *i*.
Cette approche permet de combiner des analyses réalisées dans des unités différentes.

La Figure 15 nous montre un Manhattan plot pour la production de lait en utilisant soit des analyses intra-races (A) ou des analyses multiraciales (B) directes (analyse jointe) ou indirectes (méta-analyses).

Pour ce caractère, un seul QTL est identifié chez les trois races (chromosome 6) et un seul autre QTL est partagé en race Holstein et Normande (chromosome 5). Les autres QTL étant uniquement détectés en race Holstein. En analyse multi-races, tous les QTL détectés en intra-race (exceptés ceux présents sur le chromosome 15 et 29 en race Holstein) sont identifiés dans au moins une des approches. Dans cette étude, la population Holstein a une taille bien plus importante que les deux autres races, ce qui lui procure une puissance de détection des QTL plus forte. Le fait de travailler à une échelle multiraciale permet d'accroitre la puissance de détection nominale des trois races, c'est la raison pour laquelle une grande partie des QTL identifiés en intra-race, chez la Holstein, sont également détectés dans les analyses multiraciales. Par ailleurs, en ce qui concerne les méta-analyses, qui sont une alternative aux analyses jointes, les différents modèles testés ne fournissent pas les mêmes résultats. Les modèles FIXED et ZSCORE détectent le même nombre de QTL que l'analyse jointe par contre, le modèle RANDOM détecte nettement moins de QTL car l'inclusion de l'hétérogénéité entraine une perte de puissance.

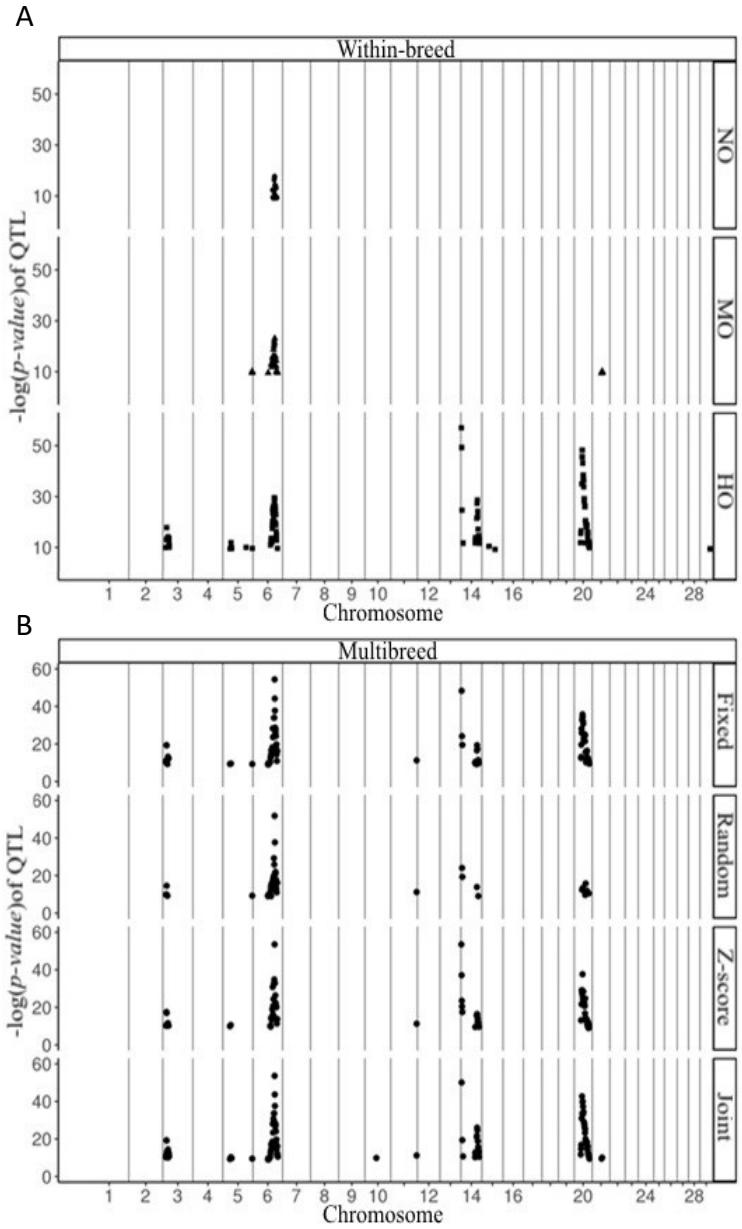


Figure 15: localisation des QTL associés à la production de lait le long du génome définie à partir d'études d'association intra-races pour la Normande (NO), la Montbéliarde (MO) et la Holstein (HO) (A) ou multi-races à travers des méta-analyses ou une analyse jointe (B).

Pour évaluer la capacité des différentes approches multiraciales à localiser les QTL, des listes de QTL détectées, soit par les études d'association intra-races, ou bien par les approches multiraciales, ont été testées à travers un MABLUP en validation croisée. Chaque MABLUP a été réalisé indépendamment dans chaque race et avec les différentes listes de QTL obtenues. Les résultats sont présentés sur la Figure 16.

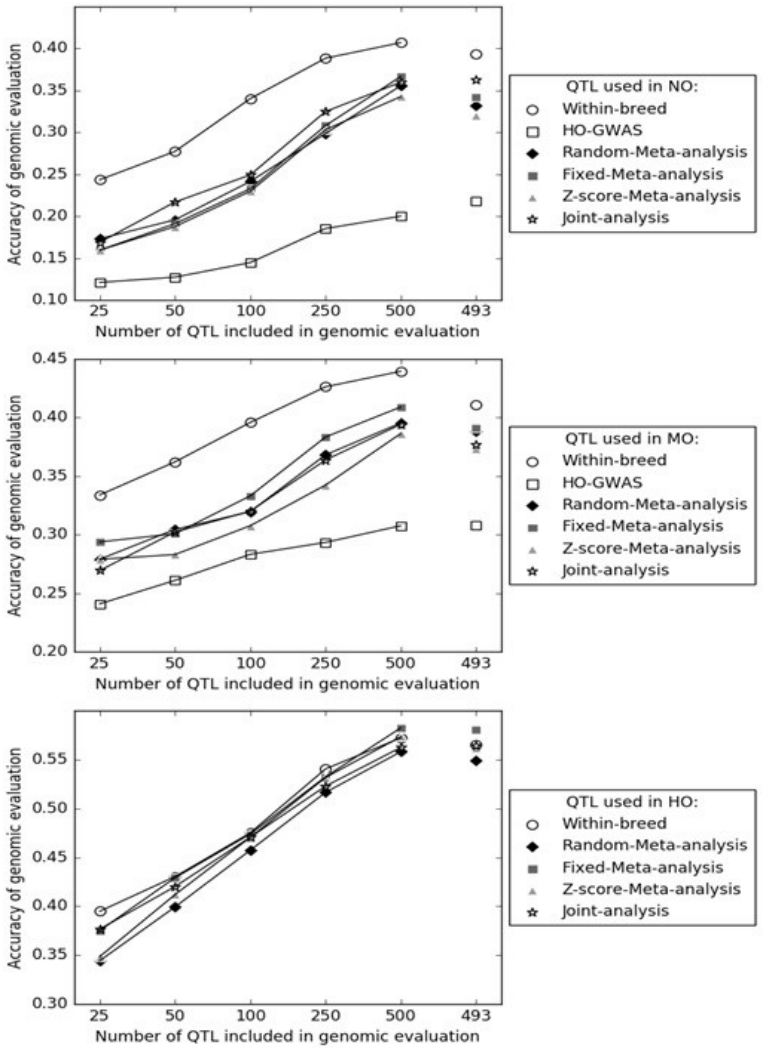


Figure 16: Précision de l'évaluation génomique en rance Normande (NO), Montbéliarde (MO) ou Holstein (HO) en fonction de la liste de SNP (constituée d'un nombre de SNP compris entre 25 et 500) utilisée pour estimer les valeurs génomiques de jeunes animaux. Ces SNP ont été sélectionnés après des études d'association intra-races, des analyses jointes ou après des méta-analyses sous différents modèles (FIXED, RANDOM ou Z-score).

Dans cette étude, seuls les QTL ont été inclus dans le modèle, nous ne cherchons donc pas à maximiser la corrélation entre les valeurs génomiques prédites et les DYD. L'objectif est seulement de mesurer la capacité des listes de QTL à prédire les performances.

Sachant que la race Holstein, de par son effectif, disposait d'une puissance de détection des QTL plus importante que les autres races, nous nous demandions si les QTL détectés par les analyses multiraces n'était pas, la plupart du temps, un copié/collé de ce qui avait été trouvé en race Holstein. Pour répondre à cette question, la liste de QTL Holstein a été testée en races Normandes et Montbéliarde et il apparait clairement que cette liste dégrade drastiquement les performances de la liste intra-race correspondant à chaque race, mais elle se situe également en dessous des performances obtenues par les listes multiraciales.
Nous avons donc pu montrer qu'un certain nombre de QTL étaient bien partagés entre races et que cela couvre une bonne partie de la variabilité génétique puisqu'en race Normande et Montbéliarde la perte de corrélation est d'environ 5 points lorsque des listes de 500 QTL sont utilisés. En race Holstein, le bilan est encore plus positif puisque cet écart n'est que d'environ 2 points de corrélation.
En ce qui concerne les méta-analyses, la comparaison des différents modèles est compliquée car le comportement diffère en fonction du nombre de QTL inclus dans la liste et parfois de la

race étudiée. Toutefois, le modèle FIXED semble être le plus efficace, ce qui diffère des conclusions d'une étude menée par (Van den Berg *et al.* 2016) en analysant des résultats de GWAS uniquement.

Enfin, nous avons réalisé des comparaisons par paires de listes de QTL afin de vérifier les proportions de QTL communs ou proches entre les différentes races. Ces résultats sont présentés sur la Figure 17.

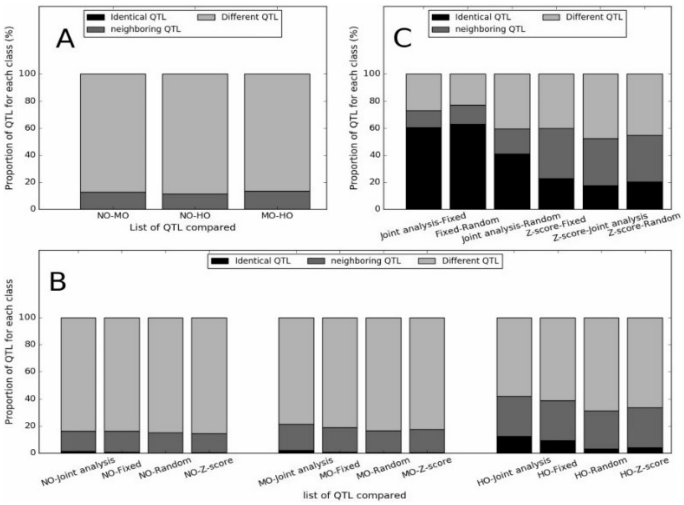


Figure 17: Identification des QTL communs, voisins et différents en comparant les listes de QTL deux à deux. Les comparaisons ont été effectuées entre les listes intra-races (A), entre les listes intra-races et multi-races (B) et entre les listes multi-races (C).

La proportion de QTL communs et voisins est clairement plus élevée lorsque l'on compare les listes intra-races et multi-races (B) qu'entre les listes intra-races seules (A). Ce résultat était attendu car les approches multi-races, à la différence des analyses intra-races, sont censées résumer les informations provenant des trois populations. Par contre, le nombre de QTL communs reste faible.

Cette étude ouvre des portes pour une extension de ces travaux à des races régionales qui ne disposent que de très peu d'animaux séquencés et qui pourraient profiter des mutations candidates et causales identifiées chez les grandes races laitières. Ce sujet sera abordé dans les chapitres suivants.

IV. Quelles évolutions pour les évaluations génomiques de demain ?

Dans ce chapitre, nous aborderons les défis méthodologiques à initier pour améliorer les performances et la durabilité des évaluations génomiques. Le passage au single-step, l'inclusion des mutations causales, la mise à disposition d'une évaluation génomique pour les races régionales, le développement d'une évaluation génomique en croisement ou encore l'exploitation de données épigénétiques sont des défis importants à mettre en place dans les prochaines années. Par ailleurs, l'agroécologie se place aujourd'hui au cœur des préoccupations de toute la filière agricole (de l'éleveur au consommateur). A notre niveau, cela se traduit par le maintien d'une diversité génétique dans nos populations d'élevage afin d'avoir des animaux capables de s'adapter aux objectifs de sélection du futur. Ces objectifs de sélection seront certainement orientés vers des nouveaux caractères en lien avec le réchauffement climatique tels que la production de méthane ou la thermotolérance.

A. Prise en compte du biais de présélection dans les évaluations génomiques

Les évaluations génomiques mises en place aussi bien en France qu'à l'international reposent toutes sur une approche multi-étapes. La première étape est un modèle polygénique incluant tous les individus avec phénotypes et tous leurs ancêtres connus, qu'ils soient génotypés ou non. La seconde étape n'analyse que les individus génotypés. Enfin, une troisième étape peut éventuellement combiner les résultats des deux premières étapes (« blending ») lorsque la quantité d'information perdue entre les deux premières étapes est importante.

La première étape fournit des évaluations polygéniques, des effets des autres facteurs du modèle, et permet de construire des phénotypes concentrant le maximum d'information sur les animaux génotypés. Pour une vache avec performance, on construit une performance corrigée (ou YD) qui est la moyenne de ses performances corrigées pour tous les effets non génétiques du modèle. Pour un taureau, la performance corrigée, ou DYD, est la moyenne des performances des filles, corrigées pour tous les effets non génétiques ainsi que pour la valeur génétique de leurs mères, c'est-à-dire des conjointes du taureau. Ces DYD et YD, accompagnés de leurs poids mesurant la quantité d'information associée, sont les ingrédients d'entrée des évaluations génomiques de l'étape 2.

Cependant, les jeunes animaux génotypés et sélectionnés sur la base de leur évaluation génomique qui rentrent dans les évaluations polygéniques ne sont plus représentatifs de la population générale. Plus précisément, le modèle polygénique suppose que l'espérance de valeur génétique d'un individu est égale à la moitié de celles de ses parents et que l'espérance de l'aléa de méiose est nulle. La sélection génomique rend cette condition fausse (on a conservé les meilleurs sur les critères sélectionnés et leur aléa de méiose est en moyenne positif) et cette présélection n'est pas prise en compte dans les équations du modèle mixte. En quelques générations, cela induit ce qu'on a appelé un « biais de présélection » qui rend difficile la comparaison des animaux entre eux et conduit à une sous-estimation du niveau génétique des animaux sélectionnés par la génomique (Patry and Ducrocq 2009).

Pour résoudre ce problème, une méthode d'évaluation génomique nommée « single-step » a été proposée. Cette méthode fait aujourd'hui l'unanimité et sera prochainement mise en production dans la plupart des pays disposant d'une évaluation génomique de routine.

1. L'approche single-step

L'approche single-step consiste à réaliser une évaluation génomique intégrant à la fois les animaux non génotypés et les animaux génotypés. Le single step considère une matrice d'apparentement entre individus estimée à partir des marqueurs pour les animaux génotypés, à partir du pedigree pour les autres. La prise en compte explicite de l'apparentement génomique entre parents et produits typés permet de prédire l'aléa de méiose reçu par le produit. La sélection basée sur cet aléa de méiose est donc bien pris en compte et le modèle est donc supposé non biaisé par la sélection génomique. Par ailleurs, le single step rend possible l'utilisation des informations issues de l'ensemble des animaux avec performances, indépendamment de leur apparentement aux animaux génotypés, et sans nécessité de construire des phénotypes intermédiaires (DYD ou YD) sous-optimaux. Il fournit des index uniques, sans distinction entre génotypique et polygénique. Enfin, il ouvre la possibilité d'une sélection génomique pour les races qui n'en disposent pas encore malgré plusieurs centaines d'individus génotypés (races Simmental, Jersiaise, Aubrac, Salers). Bien sûr, dans ces races, le gain apporté par la génomique dépendra de la taille de la population de référence, mais d'un point de vue mécanique, le passage du modèle polygénique au modèle single step peut s'opérer de façon continue à partir du premier animal génotypé.
Dans cette approche single step, la matrice de parenté est scindée en deux groupes, celui des animaux non génotypés et celui des animaux génotypés afin de diffuser l'information génomique aux apparentés non génotypés. Cette matrice de parenté (**H**) doit être inversée mais il a été montré que $\mathbf{H}^{-1}$ a une forme assez simple où n'interviennent que les inverses de la matrice de parenté attendue entre animaux génotypés ($\mathbf{A_{22}}$), la matrice de parenté génomique entre animaux génotypés (**G**) et la matrice de parenté complète (**A**) :

$$\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G^{-1}} - \boldsymbol{A_{22}^{-1}} \end{bmatrix}$$

Cette matrice $\mathbf{H^{-1}}$ est ensuite introduite dans le modèle mixte. La partie la plus difficile est le calcul de $\mathbf{G}^{-1}$ car **G** est une matrice dense, difficile à inverser si le nombre d'animaux génotypés dépasse quelques dizaines ou centaines de milliers, ce qui est le cas en bovins. Cependant, les développements algorithmiques et méthodologiques ont été nombreux sur le sujet ces dernières années et les temps de calcul sont aujourd'hui compatibles avec des évaluations génomiques de routine. Parmi les algorithmes proposés, notons :

- L'algorithme APY (Misztal *et al.* 2014) qui distribue les animaux typés en deux catégories, un « cœur » et les autres. Une bonne approximation de $\mathbf{G}^{-1}$ est obtenue à partir de l'inverse de la matrice $\mathbf{G_c}$ des animaux du cœur.
- L'utilisation d'un modèle équivalent considérant non pas la valeur génétique des individus mais les effets des SNP. Dans ce cas, le nombre de SNP est fixé, même quand le nombre d'animaux typés devient très grand, de sorte que la taille du système d'équations est maitrisée. Toutefois, cet algorithme, avec diverses variantes, impose d'imputer le génotype (à la volée) des animaux non typés.
- Un modèle intermédiaire où le cœur est remplacé par les principaux vecteurs propres de la matrice d'apparentement.

2. Utilisation du single-step pour les évaluations génomiques françaises

Un logiciel d'évaluation Single-Step est actuellement en cours de développement par Thierry Tribout dans le cadre du projet INCOMINGS. D'autres programmes existent à l'international, mais l'intérêt de développer et de maîtriser un outil interne est de pouvoir répondre aux futures demandes du terrain, réaliser les évolutions nécessaires sans contraintes tout en conservant les spécificités des modèles français, quels que soient la filière et le caractère. Un développement interne est également une garantie de maintien des compétences dans le domaine de l'évaluation génétique et des collaborations historiques avec les filières. Cet outil développé dans l'équipe est un single step hybride, estimant les effets SNP et les effets génétiques des animaux non typés, selon l'algorithme proposé par Fernando *et al.* (2016). Comme cet outil remplacera tous les programmes d'évaluation précédents, il devra inclure toutes les options nécessaires pour traiter des modèles variés (multicaractères, effets directs et maternels, variances hétérogènes, groupes de parents inconnus, etc), ce qui en complique le développement.

Pour mettre en production ce nouvel outil, nous avons mis en place deux projets : un projet d'évaluation du single-step à travers le logiciel développé en interne (projet ASAP financé par APISGENE et accepté en février 2019) et un projet de déploiement et de mise en production de l'outil (projet UNIGENO financé par CASDAR et accepté en aout 2019).

L'évaluation de l'approche single-step implémentée en France est prévue à différents niveaux. Tout d'abord, il s'agit de tester le logiciel pour s'assurer qu'il donne des résultats exacts. Compte tenu de la complexité de cette phase de validation, nous avons choisi de comparer les résultats avec ceux obtenus avec d'autres approches single-step (et notamment le logiciel BLUPF90, (Aguilar *et al.* 2014). Compte tenu de la dimension importante des données chez les bovins et en particulier les bovins laitiers, une attention particulière sera apportée aux temps de calcul et aux besoins en mémoire.
Une comparaison entre les performances du MABLUP (méthode utilisée à ce jour pour les évaluations génomiques officielles) et celles du single-step devra être réalisée. L'objectif est de vérifier que progrès génétique réalisé avec le single-step est au moins aussi bon qu'avec le MABLUP. Cette comparaison pose le problème du juge de paix. En effet, avec une approche en 2 étapes, ce sont les évaluations polygéniques qui fournissent les DYD qui servent de juge de paix dans les études de validation des évaluations génomiques. Avec l'arrivée du single-step, seules les performances propres seront utilisées et il n'y aura plus production de DYD. Or, utiliser les DYD des évaluations polygéniques n'est pas envisageable pour mesurer les performances du single-step. Il faudra donc faire produire au single step des DYD.

L'objectif est également de mesurer l'ampleur des variations d'index observées par rapport au MABLUP ainsi que la stabilité des index obtenus entre deux traitements. Il est tout à fait possible qu'un logiciel assure des corrélations entre performances observées et estimées comparable pour un caractère donné mais que la hiérarchie des taureaux ne soit pas conservée. Cela a déjà été constaté dans le cadre du single-step. Beaucoup d'algorithmes ont été développés pour limiter les temps de calcul et cela passe parfois par des astuces algorithmiques et méthodologiques qui nécessitent quelques approximations. C'est le cas de l'algorithme APY qui propose une approximation pour inverser la matrice **G** et qui permet d'utiliser cette approche pour des grandes populations génotypées (Fernando *et al.* 2014; Liu *et al.* 2014; Misztal *et al.* 2014; Taskinen *et al.* 2017). Cette approche a permis de réduire considérablement les temps de calcul mais au détriment de la stabilité des index dans le temps

(Misztal *et al.* 2019). Il nous parait donc important de contrôler qu'un tel scenario n'arrive pas avec notre logiciel.
Enfin, ce projet vise à contrôler que l'intérêt du passage d'une approche en 2 étapes à une approche en 1 étape est bien mesurable. Il convient donc de vérifier que l'approche single-step permet bien de maitriser le biais de présélection. Pour cela, une étude sur jeux de données simulées sera menée.

B. Vers une meilleure prise en compte des mutations candidates et causales dans les modèles d'évaluation génomique

L'amélioration des prédictions génomiques des bovins laitiers et allaitants restera une thématique forte de mes activités de recherche dans les années qui viennent. Beaucoup reste à faire pour maintenir une évaluation génomique compatible avec une application en routine tout en incorporant dans ces évaluations un panel d'informations biologiques toujours plus important.

1. Mieux pondérer les variances génétiques des mutations candidates et causales

Nous avons pu montrer dans le projet INCOMINGS qu'en s'éloignant des hypothèses du GBLUP (où la même part de variance génétique est attribuée à tous les marqueurs) et en donnant une variance génétique plus importante aux mutations candidates et causales, il était possible d'améliorer les prédictions génomiques. Ces travaux ont été confirmés dans notre équipe à travers une étude sur la fromageabilité du lait (Sanchez *et al.* 2018). Cependant, les variances génétiques additives attribuées aux mutations candidates et causales restent très imprécises. Cela consiste pour le moment à définir une part de la variance dédiée aux mutations et de partager équitablement cette variance entre les marqueurs. Avec cette stratégie, un gène majeur comme DGAT1, bien connu des espèces bovines et caprines laitières impliqué dans la production laitière et composition en matière grasse et protéique du lait, se verrait attribuer la même part de variance génétique qu'un petit QTL d'effet plus modeste.
Pour améliorer la précision des prédictions génomiques, utiliser une variance génétique spécifique à chaque mutation candidate et causale est une voie intéressante. Intégrer des variances connues est simple en routine. Comme il est impossible en pratique d'estimer ces variances en routine, il convient de les estimer en amont, ce qui représente un travail considérable compte tenu du nombre de caractères et de populations.
Pour cela, plusieurs options sont envisageables :

- On peut se baser sur les effets estimés. Cette méthode est connue pour être biaisée, soit dans le sens dans la sous-estimation si l'estimation est par GBLUP, soit dans le sens d'une surestimation si l'estimation est par GWAS. Mais c'est une méthode qui donne des premières valeurs.
- On peut regrouper les variants par catégories et estimer une variance expliquée par chacune des catégories. On en déduit ensuite la variance d'un SNP en supposant la même part pour chaque SNP intra catégorie. En pratique, chaque catégorie définit une matrice d'apparentement génomique que l'on importe dans un logiciel REML.
- L'estimation des variances génétiques des mutations candidates et causales se fait au sein d'un modèle Bayésien. Cette stratégie a déjà été mise en place à travers le Bayes R proposé par (Erbe *et al.* 2012). Par rapport à la stratégie précédente dans laquelle les catégories étaient constituées a priori, dans cette approche, l'allocation des SNP

dans les groupes est un résultat de l'analyse, ainsi que la variance de chaque groupe. On peut ainsi faire la distinction entre des petits, des moyens et de gros QTL, en plus du groupe de SNP avec un effet nul. Chaque groupe est défini par sa part de variance génétique qu'il partage entre ses SNP. Toutefois, cette approche est trop lourde pour pouvoir être utilisée en routine. Son utilisation serait dans un cadre expérimental puis, une fois les SNP identifiés et leur classe déterminée, on exploite cette information comme dans les options précédentes.

On pourrait aller un peu plus loin en faisant estimer la part de variance génétique de toutes les mutations candidates et causales par le modèle puis, partager le reste de la variance génétique entre tous les SNP de la puce.

2. Prise en compte d'information d'annotation fonctionnelle pour améliorer les prédictions génomiques

Cette thématique est au cœur de deux projets européens qui démarrent fin 2019 et dans lesquels je vais être impliqué à différents niveaux :

- Le projet BovReg (Identification de régions génomiques fonctionnellement actives et pertinentes pour la diversité phénotypique et la plasticité des bovins) est piloté par Christa Kuehn (Allemagne) et Dominique Rocha (INRA, équipe G2B). Dans ce projet, une tâche est dédiée à la mise en place d'une évaluation génomique guidée par la biologie. L'objectif sera de développer des méthodes capables d'exploiter les informations issues de données d'annotation fonctionnelle au sein d'une évaluation génomique et de vérifier que ces approches peuvent améliorer la prédiction des phénotypes et de la valeur génétique des animaux.

  En effet, les sources d'information biologique disponibles aujourd'hui sont croissantes grâce aux données RNAseq qui nous permettent d'identifier des régulateurs d'expression (eQTL), métabolique (mQTL) ou encore d'épissage (sQTL)… Ces données ne sont disponibles que sur un faible panel d'individus et l'objectif de cette tâche est d'inférer ces informations pour des animaux phénotypés et séquencés dépourvus de données d'annotation fonctionnelle.

- le projet GeneSwitch (Le génome régulationnel du porc et du poulet : annotation fonctionnelle pendant le développement) est piloté par Elisabetta Giuffra (INRA, GABI). L'équipe G2B ne participe pas à ce projet européen. Toutefois, je vais m'y investir à travers le co-encadrement d'une thèse avec Andréa Rau (INRA, GABI). Cette thèse, portée par l'équipe Gibbs de GABI, sera réalisée par Fanny Mollandin, a débuté en octobre 2019 et porte sur « l'incorporation de connaissances d'annotation fonctionnelle dans des modèles de prédiction génomique Bayésiens ». L'objectif de cette thèse est de développer et de valider des modèles de prédiction génomique pondérant les SNP dans les modèles d'évaluation en fonction des informations d'annotation fonctionnelle. En d'autres mots, l'apport d'un SNP pour un phénotype donné se décomposerait en une somme de composantes issues d'annotation fonctionnelle. Pour estimer les effets de ces composantes, un modèle Bayésien sera utilisé.

  Si une première étude consistera à valider notre modèle sur un jeu de données simulées, les tests sur données réelles seront planifiés dans un second temps sur des données porcines et de poulet. Par ailleurs, les liens étant relativement étroits avec les délivrables du projet BovReg, des tests en bovins laitiers seront également réalisés.

3. Exploiter les informations d'annotation fonctionnelle des grandes races pour l'amélioration des prédictions génomiques de races régionales

En comparaison avec les grandes races bovines laitières, l'accumulation des génotypes se fait de manière très différente chez les races bovines régionales car les volumes de génotypage en races régionales ne sont en rien comparables à ceux des grandes races. Si nous avons pu montrer dans les chapitres précédents que l'accès aux évaluations génomiques ne leur était pas pour autant fermé, il est clair que toutes les études qui ont aujourd'hui été menées sur les grandes (GWAS à l'échelle de la séquence, données d'annotation fonctionnelle) ne leur sont pas accessibles.
Il sera intéressant de tester l'intérêt d'exploiter les informations obtenues grâce aux grandes races laitières (mutations causales et candidates, données RNAseq) sur ces races régionales.

C. Evaluations génomiques en croisement

En France, le croisement se développe (15% des inséminations environ) mais il est majoritairement terminal, les animaux nés n'étant pas mis à la reproduction. Le croisement continu, c'est-à-dire pour procréer des animaux futurs reproducteurs, représente un peu moins de 2% des inséminations artificielles (IA) dans les races laitières mais connait une progression rapide. L'intérêt des éleveurs pour les animaux croisés est lié à leur robustesse, leur rusticité, leurs capacités d'adaptation supérieures (à des changements de système notamment) à celles des animaux de race pure. Le croisement permet de compenser les défauts d'une race (faible fertilité, problèmes de santé, format) ou de combiner les avantages de plusieurs races (haute production et taux intéressants, par exemple). Ceci en conservant de bonnes performances de production. Des travaux de simulation et des études technico-économiques montrent que la robustesse et la longévité des vaches Holstein croisées est meilleure que celle des vaches de race pure. Le lancement d'une démarche de croisement permet l'obtention d'un avantage dès la première génération, grâce à la complémentarité entre les races impliquées dans un croisement et à l'effet d'hétérosis (supériorité des animaux croisés par rapport à la moyenne des deux races concernées), qui est d'autant plus important que les caractères sont peu héritables. Enfin, le croisement permet une réduction drastique puis un contrôle efficace des taux de consanguinité et l'augmentation de la diversité génétique au sein d'un troupeau, ce qui en renforce la résilience. Fournir aux éleveurs engagés dans une démarche en croisement les moyens d'optimiser le renouvellement de leur troupeau et d'améliorer les progrès en robustesse et durabilité permis par l'utilisation d'un schéma de sélection exploitant le croisement apparait donc important pour la filière bovine. Or aujourd'hui, aucun outil technique n'est disponible pour piloter la génétique des croisés. Les croisés ne sont pas évalués, les seuls index disponibles sont en race pure et non comparables entre races, et aucune recommandation d'accouplement n'est disponible. A fortiori, aucun index génomique n'est actuellement prédit pour les bovins croisés.

Le problème peut se décomposer en plusieurs parties :

1) Les effets des marqueurs peuvent être différents d'une race à l'autre. Ceci s'explique par exemple par des associations marqueur – qtl différentes entre races. Même si le marqueur est la mutation causale, ses effets peuvent être différents entre races, du fait de différences de fréquences alléliques ou d'épistasie ou d'interaction avec le background génétique. Il conviendrait donc d'estimer des effets alléliques différents

selon la race d'origine de l'allèle. Cela rend l'évaluation bien plus complexe que l'évaluation intra race qui ne compte qu'un seul effet additif par marqueur.

2) Les croisés bénéficient davantage d'effets non additifs, au travers de l'hétérosis. Ces effets non additifs ne se transmettent pas aussi simplement que les effets additifs. Les omettre du modèle peut limiter la précision du modèle et induire des biais dans l'estimation des effets additifs. Par ailleurs, c'est un enjeu important que de prédire l'hétérosis, en particulier pour l'orientation des accouplements.

De manière générale, la prise en compte d'effets non-additifs peut permettre de contribuer à l'amélioration des prédictions génomique (Toro and Varona 2010; Aliloo *et al.* 2016; Duenk *et al.* 2017) mais, dans un contexte de croisement, elle peut permettre de mieux gérer les plans d'accouplement entre candidats à la sélection (Toro and Varona 2010; Mäki-Tanila and Hill 2014; Aliloo *et al.* 2017).

Les effets non additifs sont de plusieurs natures :

- La dominance : L'effet d'un marqueur ne dépend pas du nombre de copies des allèles portés mais simplement de la présence ou non de l'allèle dominant dans le génotype.
- La dépression de consanguinité (hétérosis) : Il s'agit d'une dominance dirigée ou il y a une proportion plus importante d'effets de dominance positifs que négatifs (Falconer and Mackay 1996).
- L'empreinte parentale : ce phénomène implique une inactivation partielle ou totale des allèles paternels et maternels (Reik *et al.* 2001).
- L'épistasie : Il s'agit de la forme la plus complexe de variation génétique non additive qui correspond à l'interaction entre plusieurs gènes.

La modélisation de ces effets non additifs dans les modèles d'évaluation génomique sont bien documentés (Varona *et al.* 2018) cependant, dans un contexte de croisement, la composante raciale ajoute une complexité qu'il n'est pas facile de prendre en compte. Il existe deux écoles en ce qui concerne la méthodologie d'évaluation en croisement, l'une est basée sur un modèle animal tandis que l'autre est basée sur un modèle SNP.
Dans le modèle animal, proposé par (Wei and van der Werf 1994; Christensen *et al.* 2015), la valeur génétique des animaux croisés correspond à la somme des valeurs génétiques partielles race-spécifiques plus une valeur génétique de ségrégation. Il est donc nécessaire de construire une matrice de parenté pour chaque race plus une matrice de parenté de ségrégation. Pour relier les races entre elles, une solution est de recourir à des métafondateurs, c'est-à-dire des pseudo-individus qui représentent les populations de base intra-race. Les métafondateurs étant apparentés entre eux, il n'y a donc au final qu'une matrice de parenté unique à manipuler.
Toutefois, cette approche, qui semble pertinente dans le cas de croisement terminaux, parait peu intéressante dans un schéma rotatif au vu du nombre de paramètres à estimer.
L'alternative à cette approche est le modèle SNP proposé par (Zeng *et al.* 2013). Dans ce modèle, chaque SNP possède un effet estimé différent pour chacune des races étudiées. Ce modèle nécessite donc la connaissance de l'origine raciale des allèles mais en terme d'estimation, la complexité est bien plus faible que pour le modèle animal. Mis à part l'intérêt purement algorithmique, cette approche offre également la possibilité de travailler sous un modèle dominant ce qui peut se révéler important dans un modèle en croisement.

Dans le cadre du projet européen GenTORE, nous proposons de développer un logiciel d'évaluation génomique en croisement basé sur un modèle SNP. Par ailleurs, un projet

complémentaire financé par Apis-Gene nommé Evagenoc nous apportera le génotypage pour plus de 8500 animaux croisés issus de 4 races, ce qui nous permettra, à terme, la mise en place et le test d'une évaluation génomique en croisement en situation réelle.

D. Exploitation des marques épigénétiques dans les évaluations génomiques

Alors que la génétique correspond à l'étude des gènes, l'épigénétique s'intéresse à une "couche" d'informations complémentaires qui définit comment ces gènes vont être utilisés par une cellule ou ne pas l'être. En d'autres termes, l'épigénétique correspond à l'étude des changements dans l'activité des gènes, n'impliquant pas de modification de la séquence d'ADN et pouvant être transmis lors des divisions cellulaires. Contrairement aux mutations qui affectent la séquence d'ADN, les modifications épigénétiques sont réversibles. Les modifications épigénétiques sont d'une part soumises à un programme génétique (qui pilote la différenciation des cellules) et d'autre part induites par l'environnement au sens large. La cellule reçoit en permanence toutes sortes de signaux l'informant sur son environnement, de manière à ce qu'elle se spécialise au cours du développement, ou ajuste son activité à la situation. Ces signaux, y compris ceux liés par exemple au mode d'alimentation ou à un stress (climatique, restriction alimentaire, …), peuvent conduire à des modifications dans l'expression des gènes. Le phénomène peut être transitoire, mais il existe des modifications épigénétiques pérennes, qui persistent lorsque le signal qui les a induites disparaît. Concrètement, ces modifications sont matérialisées par des marques biochimiques, les histones, qui sont des protéines qui s'associe à l'ADN pour le compacter et former la chromatine. Pour qu'un gène conduise à la synthèse d'une protéine, il doit être lisible, c'est-à-dire accessible à différents complexes protéiques qui interviennent dans ce processus. Les marques de méthylation localisées sur l'ADN vont obstruer les aires d'arrivée de ces histones, conduisant ainsi à l'inactivation des gènes concernés.
L'utilisation de marques épigénétiques en sélection nécessite que celles-ci soient transmises à la descendance. La transmission intergénérationnelle de marques épigénétiques est très documentée chez les plantes (Sudan *et al.* 2018). Chez les mammifères, l'étude du phénomène est beaucoup plus complexe et fait encore l'objet de controverses. La formation des gamètes implique un effacement des marques épigénétiques nécessaires pour la pluripotence des cellules. Toutefois, certains gènes semblent y échapper (Heard and Martienssen 2014).

Aujourd'hui, le séquençage devenant plus accessible, il peut être utilisé différemment, pour étudier l'épigénétique. En traitant l'ADN au bisulfite, il est possible d'étudier de façon systématique et à grande échelle la méthylation des cytosines de l'ADN. Cela permettra d'analyser le déterminisme de la transmission de la méthylation entre générations, et par conséquent son impact sur la transmission des caractères. On pourra également mesurer le pouvoir prédictif des méthylations sur les performances et cela pourrait donc aussi impacter la précision des prédictions génomiques.
Pour analyser le déterminisme de la transmission de la méthylation, le taux de méthylation sera utilisé en tant que phénotype dans une étude d'association avec les marqueurs du génome. Les résultats de cette étude nous apporteront la connaissance des régions du génome pour lesquelles les variations de marques épigénétiques sont sous contrôle génétique.
Par ailleurs, dans une population de taille suffisante où chaque individu est génotypé, épigénotypé et phénotypé, on pourra rechercher la contribution de la méthylation sur la similarité entre individus.

Un projet européen nommé RUMIGEN est en cours de montage sur cette thématique avec comme délivrable la production d'une puce d'épigénotypage de ces marques épigénétiques trans-générationnelle.

L'exploitation de cette puce d'épigénotypage nous ouvre la possibilité d'exploiter cette source d'information dans les modèles d'évaluation génomique. Pour cela, nous devrons disposer d'une population de référence solide.

E. L'agroécologie au cœur de la sélection

La sélection génomique a permis une accélération considérable du progrès génétique, en particulier pour des caractères difficilement mesurables. Cependant, l'utilisation des évaluations génomiques à travers les programmes de sélection peut être un atout (si on sélectionne sur l'aléa de méiose, on exploite moins l'information familiale qu'un BLUP sur pedigree et la sélection génomique permet de sélectionner sur plus de caractères ce qui favorise des aptitudes différentes) mais, si elle n'est pas maitrisée, elle peut conduire à une perte de diversité génétique tout aussi rapide (la réduction de l'intervalle de génération et l'augmentation possible de l'intensité de sélection conduisent à une augmentation de la consanguinité annuelle). Cette perte de diversité, caractérisée par une augmentation de la consanguinité, s'accompagne d'une baisse générale de la fitness (Leroy 2014), que l'on attribue principalement à l'augmentation du « fardeau génétique », c'est-à-dire la hausse de fréquence d'apparition de tares génétiques récessives (Hedrick and Garcia-Dorado 2016). Lutter contre la perte de diversité génétique et plus particulièrement contre le fardeau génétique contribue à la pérennité de l'élevage en maintenant une sélection efficace tout en évitant ses contreparties défavorables.
C'est ce que nous planifions de faire au sein du projet GdivSelGen (metaprogramme SelGen) qui a démarré fin 2017 et que je porte avec Hélène Muranty dont l'acronyme signifie : Gestion de la diversité en Sélection Génomique. Ce projet embrasse un large éventail des problématiques concernant la diversité dans le contexte de la sélection génomique tel que son estimation au moyen de l'information génomique, sa prise en compte dans les évaluations génomiques, sa valorisation par la quantification du fardeau génétique, l'utilisation des ressources génétiques conservées dans les centres de ressources biologiques ou encore sa gestion dans les schémas de sélection. Ce projet rassemble un grand nombre d'équipes et associe généticiens animaux et végétaux, méthodologistes et gestionnaires de populations.
Notre contribution à ce projet se fait majoritairement au travers d'une thèse CIFRE réalisée par Anna-Charlotte Doublet que je co-encadre avec Denis Laloë sur la prise en compte de la diversité génétique en sélection génomique. Les trois parties de cette thèse, démarrée fin 2017 sont détaillées ci-dessous.

1. Prise en compte de la diversité génétique dans les schémas de sélection

a) Evolution de la diversité génétique depuis la mise en place de la sélection génomique

La mise en place de la sélection génomique en France s'est effectuée de manière très rapide, et différente suivant les races considérées. Or, la principale étude sur la gestion de la diversité (Colleau *et al.* 2015), n'a traité que le cas d'une grande race, la Montbéliarde, avec le schéma UMOTEST alors que les recommandations peuvent varier en fonction de la taille du noyau de sélection, de l'organisation de la sélection ou des contraintes économiques propre

à chaque race ou entreprise. La première partie de la thèse consiste donc à mener une enquête au sein de plusieurs entreprises partenaires afin de recenser, pour différentes races, les pratiques et les schémas réellement mis en place et de comprendre les contraintes inhérentes à chacune d'elles, et donc les leviers d'action possibles. Ce sera, à notre connaissance, le premier bilan des pratiques de la sélection génomique au sein des populations françaises 8 ans après sa mise en place. Il servira à alimenter le modèle de simulation mis en place dans la deuxième partie de la thèse. De plus, cela pourra servir de base de données pour nos partenaires.
Dans cette étude, les trois grandes races laitières ont été analysées à travers l'ensemble des taureaux mis en marché et génotypés en France. Cette population a l'avantage de représenter les taureaux réellement utilisés par les éleveurs.
Au cours de cette étude, nous nous sommes intéressés à la consanguinité mesurée à partir du pedigree ou à partir de l'information moléculaire à travers l'identification de Run of Homozygosity (ROH) (McQuillan *et al.* 2008). Les ROH correspondent à de longs segments ininterrompus homozygotes qui permettent l'identification de segments IBD. La longueur des ROH devient alors un marqueur de datation de la consanguinité. Les longs segments représentent une consanguinité récente tandis que les courts segments représentent une consanguinité plus ancienne. Ainsi, l'évolution des ROH nous fournit une vision de l'évolution de la population dans le temps. Toutefois, une identification empirique des ROH dépend de différents paramètres : taille minimale, nombre de SNP minimal, densité minimale, distance maximale entre deux SNP homozygotes, nombre maximal de génotypes manquants et nombre d'hétérozygotes permis maximal. D'autres modèles, basés sur des chaines de markov cachées permettent une modélisation probabiliste complète du processus IBD le long des chromosomes mais ces méthodes sont difficiles à exploiter sur des grandes populations (Druet and Gautier 2017).

Ainsi, en plus de la consanguinité sur pedigree (F_{ped}) (Figure 18) et sur ROH (F_{ROH}, défini comme la part du génome marqué couvert par des ROH) (Figure 19), l'évolution de la longueur des ROH (Figure 20) et du progrès génétique (Figure 21) ont été évalués. Nous disposons de ces mesures sur un intervalle de temps de 10 ans allant de 2005 à 2015. Dans ces trois populations, la sélection génomique a été mise en place en 2010. Une période de transition a donc été définie entre 2010 et 2012 pour ne pas fausser les résultats. Les résultats en races Normande et Montbéliarde étant relativement proches, seuls les résultats en race Montbéliarde seront présentés (Doublet *et al.* 2019).
Pour pouvoir interpréter ces figures, nous nous sommes intéressés à la comparaison des pentes de régression avant et après le démarrage de la sélection génomique en utilisant la formule suivante :

$$Changement\ relatif = \frac{a - b}{|b|}$$

où *a* et *b* correspondent aux pentes de régression respectivement avant et après sélection génomique. Les changements relatifs et leur niveau de significativité sont présentés sur le Tableau 17.

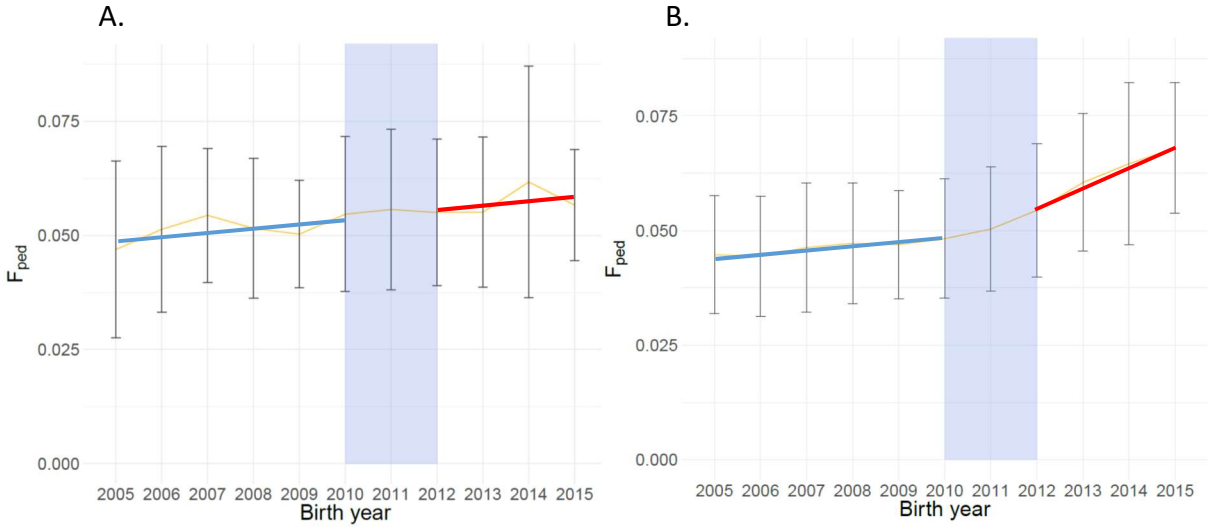


Figure 18: Evolution de la consanguinité mesurée à partir du pedigree entre 2010 et 2015 en race Montbéliarde (A) et Holstein (B). Les droites bleues et rouges correspondent aux pentes de régression entre 2005 et 2010 (évaluations génétiques) et entre 2012 et 2015 (évaluations génomiques) respectivement.

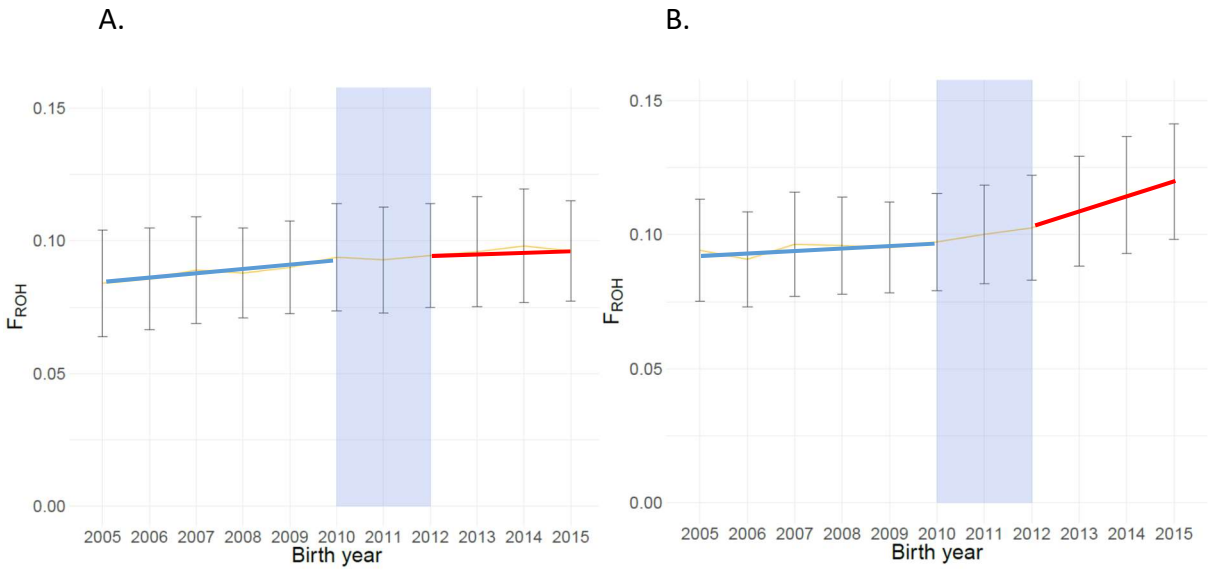


Figure 19: Evolution de la consanguinité mesurée à partir des ROH entre 2010 et 2015 en race Montbéliarde (A) et Holstein (B). Les droites bleues et rouges correspondent aux pentes de régression entre 2005 et 2010 (évaluations génétiques) et entre 2012 et 2015 (évaluations génomiques) respectivement.

Tout d'abord, nous avons pu constater que quelle que soit la race, la consanguinité sur pedigree (Figure 18) comme la consanguinité sur ROH (Figure 19) a augmenté entre 2005 et 2015. Toutefois, nous avons pu montrer que la sélection génomique conduit à un accroissement de la perte de diversité en race Holstein (changement relatif de 4,75 et 5,65 pour la consanguinité sur pedigree et sur ROH respectivement) alors qu'elle est stable pour les deux autres races.

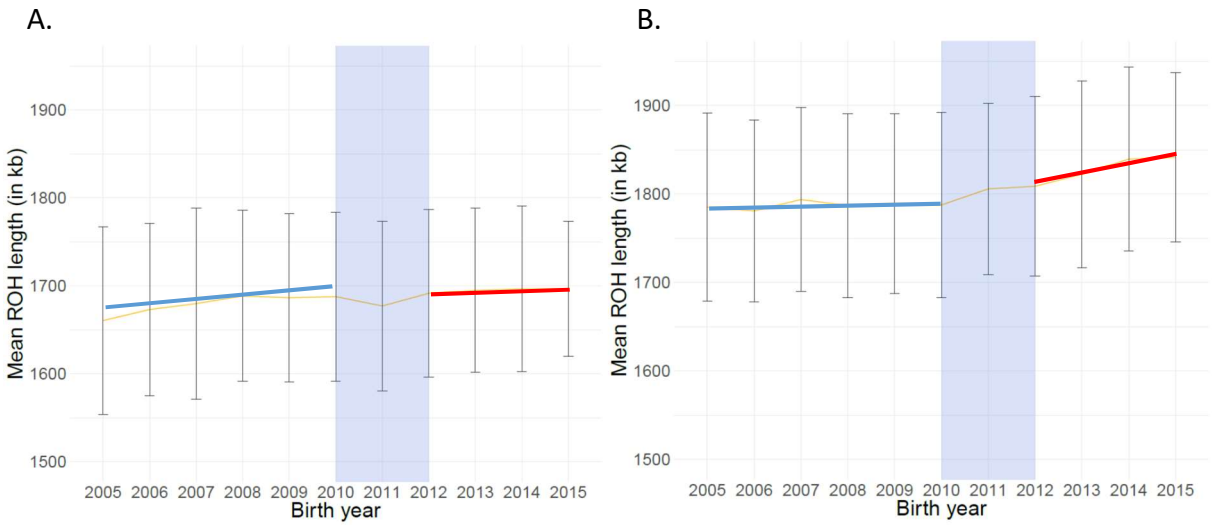


Figure 20: Evolution de la longueur des ROH entre 2010 et 2015 en race Montbéliarde (A) et Holstein (B). Les droites bleues et rouges correspondent aux pentes de régression entre 2005 et 2010 (évaluations génétiques) et entre 2012 et 2015 (évaluations génomiques) respectivement.

Concernant la longueur des ROH (Figure 20), celle-ci a augmenté significativement en race Holstein depuis l'arrivée de la sélection génomique (changement relatif de 10,65). Ce n'est pas le cas pour les deux autres races pour lesquelles les changements relatifs sont non significatifs.

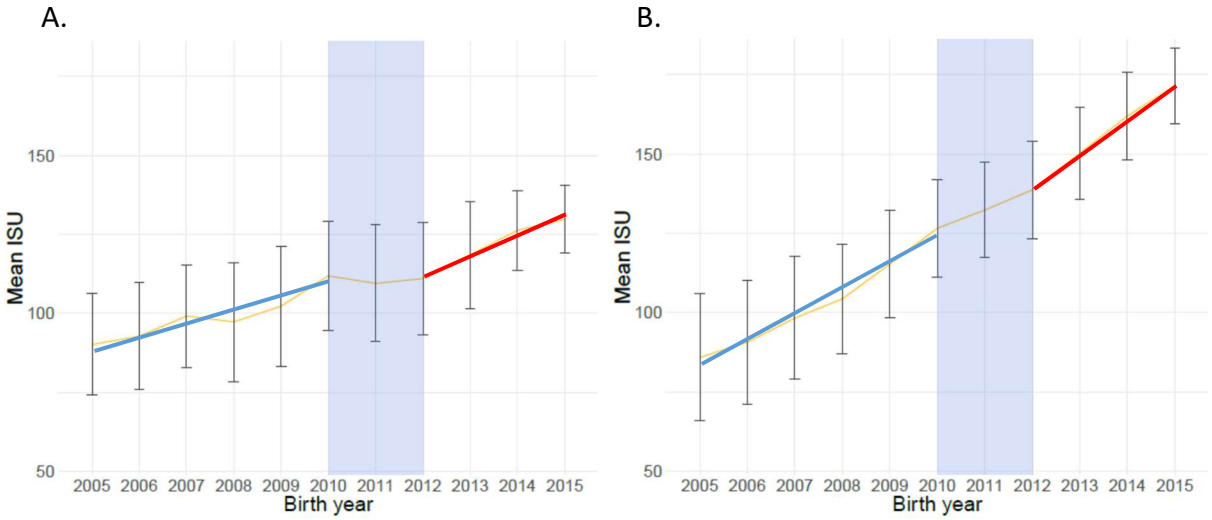


Figure 21: Evolution du progrès génétique mesuré à partir d'un index synthétique (ISU) entre 2010 et 2015 en race Montbéliarde (A) et Holstein (B). Les droites bleues et rouges correspondent aux pentes de régression entre 2005 et 2010 (évaluations génétiques) et entre 2012 et 2015 (évaluations génomiques) respectivement.

En ce qui concerne le progrès génétique (Figure 21), l'index synthétique ISU a eu une croissance nettement plus importante depuis l'arrivée de la sélection génomique que sur la

période de sélection génétique et ce pour les trois races étudiées (changement relatif compris entre 0,38 pour la Holstein et 0,71 pour la Normande).

Breed	Genetic gain TMI	Genetic diversity F_{ped}	F_{ROH}	ROH length
Montbéliarde	**0.69 ****	0.38 ns	-0.49 ns	-0.66 ns
Normande	**0.71 ***	2.86 ns	-0.12 ns	-18.33 ns
Holstein	**0.38 ****	**4.75 ****	**5.65 ****	**10.65 ****

*Tableau 17: Changement relatif des pentes de régression pour le gain génétique, la diversité génétique mesurée sur pedigree ou à travers les ROH ainsi que sur la taille des ROH pour les races Montbéliarde, Normande et Holstein. Le niveau de significativité des changements relatifs est fourni (** : p-value <0.001 ; * : p-value < 0.05 et ns : p-value >0.05)*

A travers cette étude, nous avons pu montrer que la sélection génomique a bien porté ses fruits sur le progrès génétique et cela se confirme dans la plupart des pays ayant mis en place une évaluation génomique officielle (García-Ruiz *et al.* 2016; Howard *et al.* 2017). Cela s'explique principalement par une diminution drastique de l'intervalle de génération (proche de 50%). Cependant, ce progrès génétique accru peut s'accompagner d'une perte de diversité génétique si on n'y prend pas garde. Aussi, les races Montbéliardes et Normande, qui ont intégré dans leur schéma de sélection des outils pour limiter la contribution d'un taureau obtiennent des résultats sur la diversité génétique plus favorables que la race Holstein placée dans un contexte de concurrence internationale exacerbé.

b) Prise en compte de la diversité génétique dans les schémas de sélection

En considérant les spécificités de chaque race, le but de cette deuxième partie de thèse est de développer un logiciel d'aide à la décision permettant d'évaluer les schémas de sélection. L'évaluation se fera selon des critères de progrès et de diversité génétiques en intégrant une contrainte économique. La diversité sera mesurée non seulement via l'augmentation annuelle de consanguinité mais aussi par des mesures moléculaires plus contemporaines à l'échelle du génome (ex : hétérozygotie et sa distribution le long du génome, apparentement entre individus). Les contraintes imposées pourront être le coût ou la durée de mise en station des animaux à chaque étape de la sélection génomique, ces critères ayant un impact direct sur l'intensité de sélection, le renouvellement des taureaux et leur diffusion. Ainsi le choix du meilleur scénario pourra se faire à budget imposé et répondra aux attentes de chaque organisme de sélection et de chaque race.
Le logiciel que nous utiliserons se fondera sur des simulations de génotypes qui évolueront en réponse aux schémas proposés et sous contrainte économique (coût, nombre maximum de taureaux…). Pour une première étude, nous nous intéresserons à l'impact du transfert embryonnaire sur le progrès et la diversité génétique. Les techniques d'ovulation multiple couplées à du transfert embryonnaire sont de plus en plus utilisées chez les bovins laitiers. Grâce à cette technologie, il est possible, en moyenne, de faire produire à une vache 60 embryons viables ce qui permet la naissance de 15 fils pendant sa carrière là ou, en moyenne, elle en aurait difficilement produit plus de 2 dans un schéma classique. Avec l'insémination artificielle, toute la génétique était orientée par la voie mâle, le transfert embryonnaire utilisé sur des femelles génotypées ouvre la possibilité d'exploiter des femelles d'élite et ainsi

orienter la génétique par la voie femelle. En France, une grande partie des entreprises de sélection pratiquent le transfert embryonnaire (Evolution, UMOTEST, MIDATEST, …) cependant, à l'instar de l'insémination artificielle, le transfert embryonnaire a un coût en terme de diversité génétique qu'il convient de mesurer et de prendre en compte dans les schémas de sélection même si le facteur limitant reste le nombre de mâles mis en marché.

c) Mise en place d'un modèle d'évaluation génomique considérant le fardeau génétique

La dépression de consanguinité génère des baisses de performance (fertilité notamment, (Leroy 2014; Dezetter *et al.* 2015)) et donc des coûts importants pour la filière. Actuellement la gestion passe par l'évitement des accouplements entre apparentés et l'augmentation du nombre de taureaux afin de limiter l'augmentation inévitable de la consanguinité. Nous proposons ici de ne pas seulement considérer la consanguinité moyenne mais aussi le fardeau de mutation. En effet, l'augmentation de la consanguinité n'est délétère qu'en présence de celui-ci, ainsi son estimation et sa gestion sont cruciales. Cela permet d'accepter des niveaux de consanguinités moyens plus élevés tout en limitant la dépression de consanguinité. Nous proposons donc d'estimer l'intensité du fardeau génétique le long du génome et de l'inclure celui-ci dans l'évaluation des individus en pondérant l'effet des marqueurs impliqués dans ce fardeau lors de l'évaluation génomique des candidats à la sélection.
Ce travail se fondera à la fois sur des données réelles et des simulations afin de tester la capacité de détection et d'évaluation du fardeau génétique le long du génome. Nous nous appuierons sur des méthodes existantes (Pryce *et al.* 2014) mais aussi sur des travaux qui ont eu lieu dans notre unité lors du stage de M2 d'Anna-Charlotte Doublet.
Dans un deuxième temps, ces modèles seront programmés et intégrés dans les routines d'évaluations génomiques existantes (estimation conjointe en GBLUP, estimation indépendante et utilisation du BayesCpi…) Nous pourrons alors étudier le comportement des index ainsi calculés (stabilité des classements et des performances estimées). Finalement, ces nouvelles routines seront directement incluses dans le logiciel d'aide à la décision afin de déterminer les schémas de sélection les plus aptes à maintenir un progrès et une diversité génétique importante.

d) Evolution temporelle de la distribution de ROH le long du génome

Dans le cadre de la thèse d'Anna-Charlotte Doublet, nous avons pu étudier, à travers l'exploitation des ROH, l'évolution de la diversité génétique au cours du temps. Cependant, la localisation des ROH nous donne accès aux régions du génome sujets à une perte de diversité. L'exploitation de ces informations et l'évolution dans le temps de ces régions peuvent nous apprendre beaucoup sur l'évolution de la structure du génome. Je me suis intéressé à ce sujet à travers l'encadrement de Katy Paul dans le cadre d'un stage de M2.
Dans un premier temps, on s'est intéressé à l'analyse du partage des régions ROH entre individus en essayant de comprendre le lien entre la taille des ROH et leur partage entre individus afin de comprendre à quelle vitesse la consanguinité s'accumule dans le génome. Nous nous sommes également demandé si l'évolution de la taille et des fréquences des ROH était homogène le long du génome ou non.
Pour comprendre si certaines régions du génome sont plus susceptibles de présenter des ROH, nous avons besoin d'une mesure à l'échelle de la population. Or, les ROH sont définis par individu, ainsi, deux individus peuvent avoir une région du génome couverte par un ROH mais

ce ROH peut avoir des positions de début et de fin différente selon l'individu. Pour exploiter une mesure à l'échelle de la population, nous avons calculé les fréquences de présence de ROH sur l'ensemble des individus d'une population pour chaque position du génome (PPROHsnp) (Figure 22).

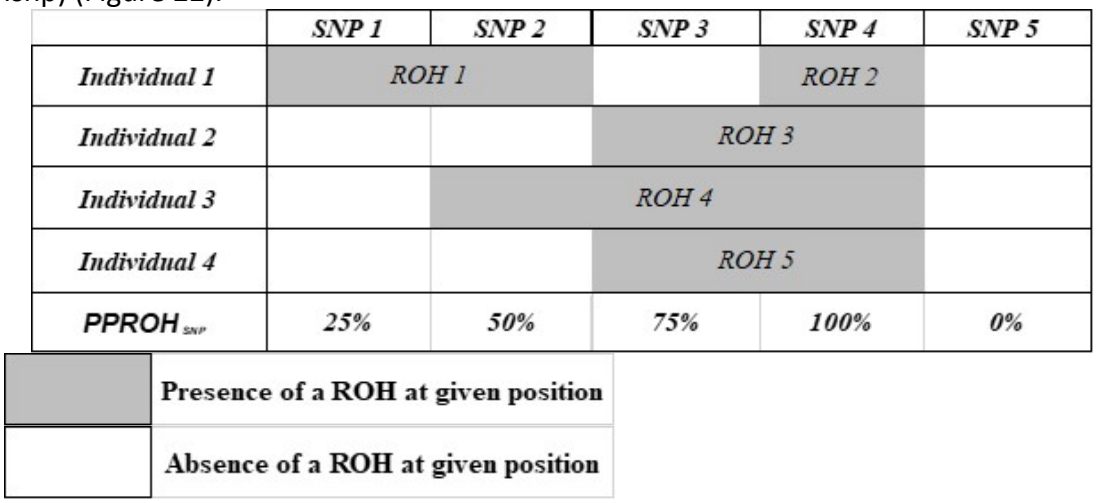


Figure 22: représentation schématique de la proportion de la population partageant un ROH à une position donnée (SNP) sur le génome.

Puis, pour pouvoir exploiter les résultats, nous avons catégorisé les valeurs de PPROH tel que présenté sur le

Tableau 18. L'évolution de la proportion de SNP au sein de chaque catégorie est présentée sur la Figure 23.

Catégories	*PPROH (%)*
1	] 0 to 5
2	5 to 10
3	10 to 15
4	15 to 20
5	20 to 30
6	≥ 30

Tableau 18 : Liste des catégories de PPROH.

Pour toutes les races, la proportion de SNP au sein des catégories 1 et 2 tend à diminuer tandis que celle des catégories 4, 5 et 6 tend à augmenter au cours du temps. En ce qui concerne la catégorie 3, pour la race Montbéliarde, la proportion tend à diminuer alors qu'elle reste stable pour la race Normande et qu'elle tend à augmenter pour la Holstein.
L'évolution des catégories de PPROH dans le temps est à mettre en relation avec l'accroissement de la part du génome avec un statut ROH. Un enrichissement de certaines régions du génome sous sélection en ROH long ou une augmentation du partage de petits ROH (plus anciens) entre individus pourraient expliquer le renforcement des catégories 4, 5 et 6 au détriment des catégories 1 et 2.

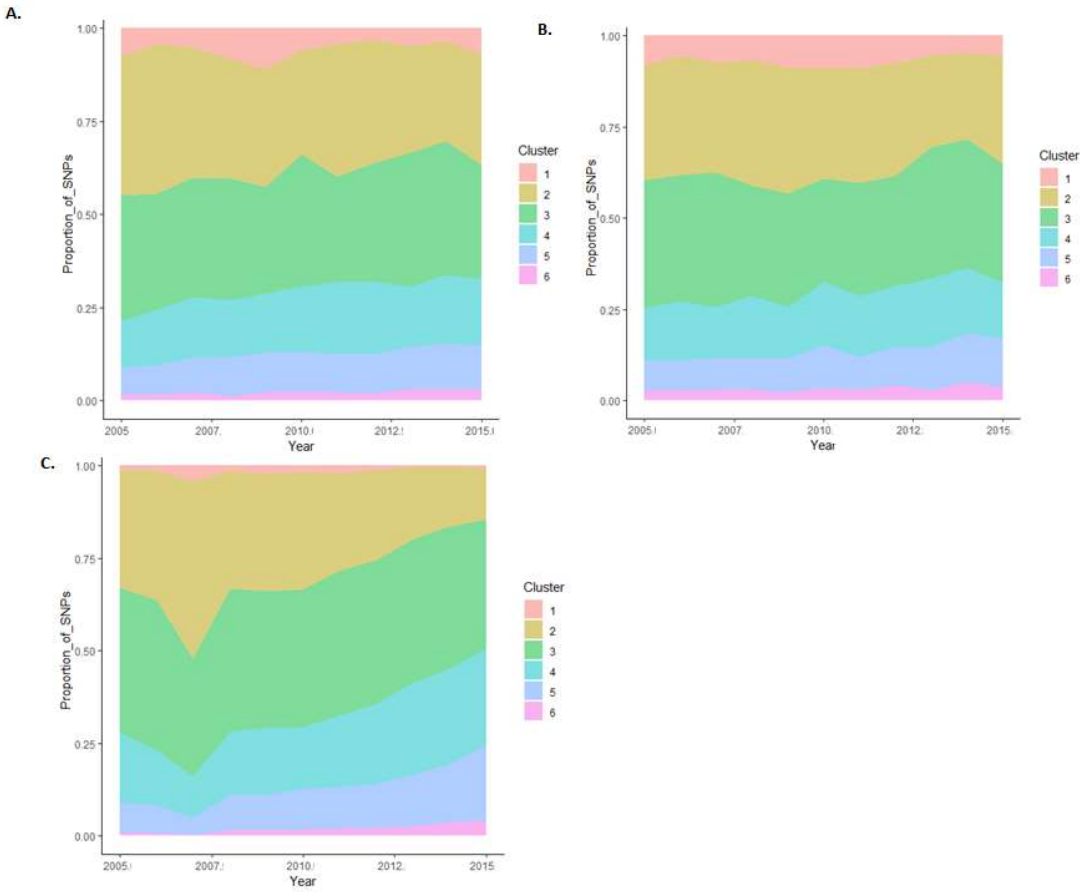


Figure 23: Proportion de SNP dans chaque catégorie de PPROH entre 2005 et 2015 pour les races Montbéliarde (A), Normande (B) et Holstein (C).

A travers la Figure 24, nous avons étudié l'impact de la position des SNP sur la relation entre le PPROH du SNP et la taille du ROH dans lequel ce SNP est situé.

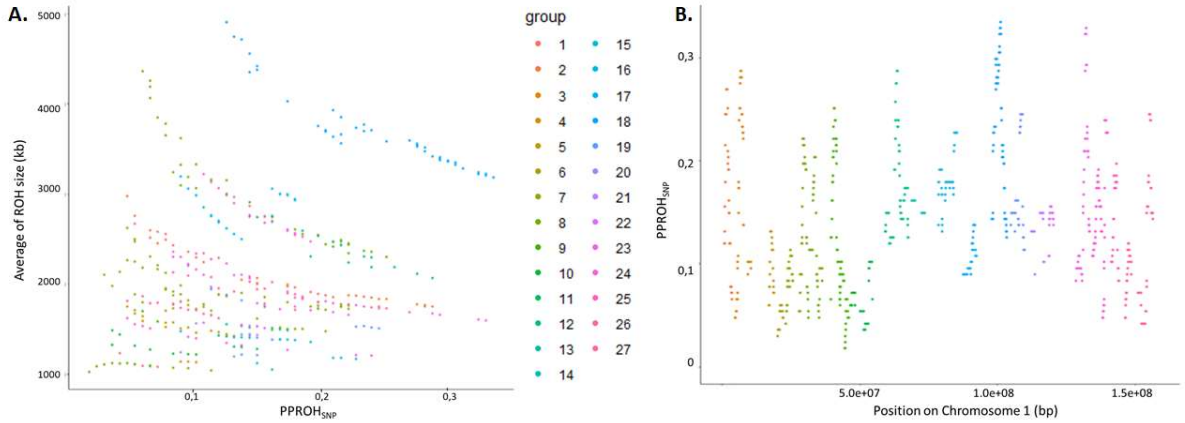


*Figure 24: Distribution des SNP en fonction de la taille moyenne des ROH dans lequel ils sont présents et de leur PPROH, (**A**) et distribution des SNP en fonction de leur position sur le chromosome et de leur PPROH (**B**). Les figures traitent du chromosome 1 de la race Montbéliarde en 2005. Chaque point représente un SNP et chaque ROH est représenté par des couleurs différentes.*

Nous avons pu montrer que plus la taille du ROH dans lequel est situé un SNP est grande, plus le PPROH de ce SNP est faible (Figure 24 A). Nous avons également étudié l'impact de la position du SNP sur la relation entre la fréquence et la longueur des ROH (Figure 24 B) et nous

avons observé une interaction significative entre les positions sur le génome et les fréquences des ROH (PPROH) sur la longueur moyenne des ROH. Cette mesure s'apparente au taux de recombinaison le long du génome.

A partir de ces résultats préliminaires, nous pensons qu'il est possible de modéliser la vitesse de décroissance des ROH dans le temps en fonction de la position des ROH sur le génome. Sachant que la diminution de la taille des ROH est principalement due aux recombinaisons, cette modélisation pourrait nous donner accès à une nouvelle métrique pour mesurer les profils de taux de recombinaison le long du génome. La relation entre la taille des ROH et leur proportion dans la population pourrait permettre de détecter et de dater la consanguinité et donc l'histoire de la population (ce qui est particulièrement intéressant pour les populations n'ayant pas de généalogie connue). L'étude des ROH conservés peut également nous renseigner sur l'existence de régions génomiques soumises à sélection de longue date. Des travaux complémentaires portant sur l'évolution de la consanguinité à proximité de ces régions sélectionnées et sur la possibilité de détecter facilement les empreintes de sélection à l'aide de ROH sont à prévoir.

e) Prise en compte de la diversité génétique dans les schémas de sélection des races régionales

Pour les races régionales, la sélection génomique a offert la possibilité de réduire l'intervalle de génération tout en augmentant le nombre de taureaux diffusés. Les races régionales, de par leur petite taille de population, sont plus susceptibles de subir la dérive génétique et une augmentation de la consanguinité.
Dans le cadre d'un projet européen nommé RUMIGEN, en cours de montage, nous proposons d'étudier l'impact de 5 ans de sélection génomique en race Abondance, Tarentaise et Vosgienne sur le progrès et la diversité génétique.
Pour ces races, la population de référence est encore en cours de construction et, en fonction des résultats de ces analyses, nous serons force de proposition pour que cette population de référence assure une qualité de prédiction génomique intéressante tout en participant au maintien de la diversité génétique (Eynard *et al.* 2018). En particulier, nous nous intéresserons aux animaux à génotyper à chaque génération (en incluant des critères de diversité génétique et de connexion en plus de leur potentiel génétique).
Par ailleurs, la dépression de consanguinité a un coût élevé en termes de performance et de fitness. Cette dépression de consanguinité repose à la fois sur le taux de consanguinité et sur le fardeau génétique. Les petites races sont fortement exposées à une augmentation de la consanguinité et il est donc utile d'identifier et de localiser les régions génomiques ou s'exercent ce fardeau génétique. Deux approches différentes sont possibles à cette fin : soit relier le niveau de consanguinité aux performances, soit utiliser l'exploration de données génomiques pour détecter des régions supposées à effet délétère (annotation de variants, déficit en homozygote, estimation des effets de dominance, …). Une fois ces régions identifiées, nous pourrons les comparer à celles détectées dans les grandes races laitières en terme de localisation, de nombre, de répartition ou d'effet.

2. Prise en compte du réchauffement climatique en sélection : réduction des émissions de méthane et adaptation à la chaleur

a) Réduction des émissions de méthane

En France, l'activité agricole représente près de 20% des émissions de gaz à effet de serre (GES) nationales, dont 10% liés directement aux exploitations bovines. De même, à l'échelle européenne, une étude conduite par le Joint Research Center évalue la contribution du secteur de l'élevage aux émissions de GES à 9,1% (Dollé *et al.* 2011). En 2007, les émissions totales annuelles de CH4 entérique par les animaux d'élevage en France s'élevaient à 1,4 million de tonnes, avec une contribution des bovins de 90%, dont 58% liés aux vaches (laitières et allaitantes) et mâles reproducteurs (Popova *et al.* 2011). S'il est indéniable que la fermentation entérique des bovins contribue significativement à l'ensemble des GES européens, son impact doit être relativisé, puisqu'elle ne représente que 5 % des émissions de gaz à effet de serre nationaux.

Toutefois, vus l'impact écologique des GES, le contexte réglementaire (paquet Climat Energie visant une réduction de 30% des GES d'ici 2030, COP21…) et le contexte de mise en accusation médiatique de l'élevage, il est essentiel pour la filière d'élaborer des solutions à moyen et long termes pour proposer une alternative à la seule réduction des effectifs de ruminants élevés sur le territoire européen. Réduire la contribution de l'élevage aux émissions de GES est un enjeu d'acceptabilité sociétale et de durabilité de l'élevage.

Outre l'impact sur le réchauffement climatique, la production de CH4 représente une perte énergétique de l'ordre de 2 à 12% de l'énergie ingérée par l'animal. Or il n'existe pas actuellement en Europe de stratégie efficace à long terme visant à atténuer les émissions de CH4 entérique des vaches laitières. Le projet Methabreed propose de développer des outils de phénotypage, de sélection génomique et d'aide à la décision pour permettre aux producteurs de lait de maitriser les émissions de méthane entérique dans leur troupeau.

Le projet ambitionne de mobiliser et valider les connaissances et mettre en place les outils permettant de réduire les émissions de CH4 entérique des bovins laitiers, sans nuire à la durabilité et à l'efficience économique des élevages.

Pour cela, il s'agira de relever 5 défis principaux :

1. Développer et déployer une méthode de phénotypage des émissions de CH4 à grande échelle à l'aide de prédiction à partir de spectres infra-rouge du lait
2. Mettre en place une évaluation génomique de la production de CH4 en races Holstein, Montbéliarde et Normande
3. Inclure la production de méthane dans des objectifs de sélection durables et équilibrés et mettre en place des outils d'aide à la décision faciles à utiliser
4. Quantifier par simulations le niveau de réduction de CH4 atteignable, le gain économique escompté au niveau des élevages à travers une étude de réponse à la sélection sur tous les caractères.
5. Fournir des indicateurs d'aide au pilotage des élevages

Ma contribution à ce projet, qui a démarré en 2019, sera principalement sur la mise en place d'une évaluation génomique. Les données de prédiction du méthane devraient permettre de

constituer rapidement une population de référence de grande ampleur. Ces données de phénotypage, couplées au génotypage réalisé en routine dans les élevages permettront la mise en place d'une évaluation génomique sur ce caractère.

b) Analyse génomique de la tolérance à la chaleur

Les grands ruminants sont sensibles à la chaleur, et ce d'autant plus que l'humidité de l'air est élevée. Ceci s'explique par leur format important et donc leur rapport surface / volume faible peu favorable à la dissipation de chaleur, d'une part, et par la production de chaleur importante dans leur rumen, d'autre part. Dans les conditions actuelles, les bovins sont déjà au-dessus de leur température de confort (13°) environ la moitié de l'année. L'évolution du climat attendue au cours des prochaines décennies dans les zones tempérées va induire un stress de chaleur dans la plupart des conditions d'élevage françaises. Ce stress peut être aigu lors de pics de chaleur et de canicule, ou chronique lorsque les animaux sont exposés à une température au-delà de leur zone de confort sur de longues durées. On peut anticiper que ce stress aura plusieurs conséquences.

Tout d'abord, un impact défavorable, aigu ou chronique selon les conditions, sur les quantités d'aliments ingérées et sur les performances de production, de fertilité et de santé. On peut imaginer des effets immédiats mais aussi sans doute des effets rémanents plusieurs mois après une vague de chaleur. Au-delà des effets moyens du stress sur ces caractères, la littérature rapporte une variabilité de réponse entre individus qui se traduit par un reclassement des individus selon le milieu, autrement dit par des interactions génotype x milieu. Si les conditions de milieu varient de façon continue, comme c'est le cas pour les données climatiques, le déterminisme d'un même caractère peut varier également de façon continue le long d'une trajectoire qui combine le potentiel intrinsèque de l'individu et sa capacité d'adaptation au milieu. Par ailleurs, ce stress induira des situations de trade-off entre caractères, complexes à gérer pour l'animal. Il est connu que les caractères de production et de fitness sont généralement en opposition plus ou moins marquée. Mais il est vraisemblable que les corrélations génétiques entre ces caractères deviennent de plus en plus défavorables en situation de stress croissant, en particulier de chaleur. Enfin, Il est également soupçonné qu'un stress appliqué à certains stades critiques de gestation peut induire des effets sur l'ensemble de la carrière du veau à naitre. Ces effets intergénérationnels pourraient avoir une base épigénétique. Mieux connaitre ces effets permettrait de mieux les maitriser en élevage.

Les évolutions du climat vers une augmentation de la température étant inéluctables, il est important d'étudier leurs effets pour les anticiper. D'une part, on pourra adapter certaines conditions d'élevage pour limiter les effets. D'autre part, on peut envisager de sélectionner pour une meilleure tolérance à la chaleur. Une telle sélection prendra du temps et il est important d'anticiper au maximum les efforts à réaliser. Toutefois, depuis 10 ans, la sélection bovine est devenue génomique et elle a gagné beaucoup en efficacité et en délais de réalisation. Elle peut être un outil de choix pour accélérer la transition vers des animaux plus tolérants.

Dans le cadre du projet européen RUMIGEN en cours de conception, l'équipe se positionne pour travailler sur ce caractère, par contre, à titre personnel, je ne serai pas impliqué dans cette étude. Pour rendre ce projet possible, une convention a été signée avec Météo France. Cela donne accès à la base qui contient les données climatiques élaborées. Les élevages sont géolocalisés, ce qui permet de les caractériser au niveau climatique. Dix variables

météorologiques sont disponibles sur une base journalière et sur une profondeur de plusieurs dizaines d'années, avec une maille géographique de 8 x 8 km. Les amplitudes de température et d'humidité sont fortes, tant entre régions qu'entre saisons, ce qui donne un sens à cette étude. En terme de génotypage, nous disposons aujourd'hui de plus de 400 000 vaches génotypées et phénotypées réparties sur le territoire national.

Ainsi, après avoir défini plus finement les caractères à analyser, une détection des QTL impliqués dans la tolérance à la chaleur pourra être réalisée. A terme, cela offrira la possibilité de placer ces QTL importants sur la partie privative d'une puce SNP. En parallèle, une évaluation génomique de ces caractères pourra être entrepris. Sachant qu'il n'existe pas une condition climatique mais probablement des profils très variés, les approches utilisées reposeront sur des modèles de normes de réaction, des modèles extrêmement flexibles qui permettent de définir autant de caractères que de type de condition climatique. L'utilisation de ces modèles en évaluation génomique est un challenge algorithmique.

V. Discussion et conclusion

Au cours de ces dix dernières années, les technologies de génotypage et de séquençage ont connu un essor considérable. Chez les bovins et plus précisément chez les bovins laitiers dans un premier temps, cet essor technologique, qui a commencé avec la mise à disposition d'une puce 50K, a été accompagné d'une mobilisation massive des entreprises de sélection pour l'exploitation de ces technologies. Cela s'est traduit par une force de génotypage considérable pour les trois grandes races de bovins laitiers (Holstein, Normande et Montbéliarde) qui ont été accessibles à l'INRA à la fois pour remplir ses missions de réalisation des évaluations génétiques mais également à des fins de recherche. Ces données, couplées à un dispositif de contrôle de performances efficace, nous ont permis de lancer des projets de grande ampleur pour déployer une évaluation génomique efficace chez ces grandes races dès 2010. Ensuite, la puce HD nous a été proposée et, d'un point de vue recherche, nous y avons vu l'opportunité de réaliser des études multi-races et ainsi, de faire profiter aux plus petites races de la population de référence des grandes races. Malheureusement, cela n'a pas pu être exploité car la densité de marqueurs de la puce HD, que nous pensions suffisante pour exploiter un déséquilibre de liaison entre marqueurs populationnel ne l'était finalement pas. Toutefois, le génotypage HD d'une population conséquente, qui a été rendu possible grâce au projet ANR GEMBAL, nous a offert la possibilité de mieux imputer l'intégralité des animaux génotypés sur la puce 50K jusqu'à la séquence complète avec une bonne précision d'imputation. En effet, ces génotypages HD nous ont permis de réaliser cette imputation en deux étapes : d'abord de la 50K vers la HD puis de la HD vers la séquence, procédure qui maximise la qualité de l'imputation. Enfin, ces dernières années nous avons contribué à la naissance du consortium 1000 génomes bovins et pour nous, cela signifie que l'intégralité des animaux génotypés en France et disponible à la recherche peut maintenant être imputée au niveau de la séquence.

Pour faire le bilan de cette période, je dirais que pendant ces dix dernières années, nous avons vécu au rythme de ces évolutions technologiques. Cela a eu des cotés très positifs : la satisfaction de travailler avec les outils les plus modernes, nous avons eu à notre disposition une quantité de données très importante et, si nous couplons cette quantité de données à la structuration de nos populations bovines, nous disposons d'un dispositif très puissant pour l'analyse du génome sous toutes ses formes. Nous avons également vécu et participé à la mise en place de modèles d'évaluation génomique et cela s'est accompagné d'un lien très fort avec la profession pour l'accompagner vers cette transition génomique. Si cette évolution a d'abord été réservée aux plus grandes races, nous avons souhaité la généraliser au plus grand nombre de populations, ce qui évite d'accroitre le différentiel de compétitivité et assure aux races régionales une meilleure garantie de pérennité. Cette période a également été très faste en terme de projets de grande ampleur avec des collaborations, y compris au-delà du monde strictement bovin. La contrepartie de cette dynamique marquée par ces évolutions rapides est que nous sommes passés d'une technologie à l'autre trop rapidement pour nous permettre d'exploiter en profondeur les données à notre disposition.

Après les innovations de rupture du génotypage, de la sélection génomique et du séquençage, la période à venir connaitra sans doute, au moins pour notre équipe, une évolution plus incrémentale et une moindre pression technologique, même si de nouvelles sources d'informations sont à prévoir (nouvel assemblage du génome bovin et apport du séquençage « long reads », épigénétique et utilisation de puces d'épigénotypage, phénomique, phénotypage à haut débit, données d'annotation du génome, variant structuraux, édition du

génome, …). Aussi, si l'amélioration des méthodes de prédiction restera d'actualité avec la volonté d'enrichir nos modèles de données biologiques de plus en plus précise, si l'afflux de données de génotypage à plus ou moins grande densité persistera, ce qui nous contraindra à savoir gérer des volumes de données de plus en plus grands, nous constatons que l'émergence de nouveaux phénotypes s'intensifie. Cela a déjà été le cas ces dernières années avec des projets forts sur la composition fine du lait, la fromageabilité du lait, mais aussi avec des données de santé (données de parage, acétonémie, sensibilité à la paratuberculose, …) et cela va s'intensifier avec des caractères à visée agro-écologique tel que la production de méthane, la tolérance à la chaleur, l'efficience alimentaire, …

L'ensemble de ces travaux nous permettra d'améliorer nos connaissances du génome bovin et par conséquent d'envisager des modèles de prédiction de plus en plus précis, répondant à la fois aux besoins de la sélection pour des animaux de plus en plus adaptés à leur condition d'élevage mais répondant également aux demandes sociétales que ce soit d'un point de vue environnemental que d'un point de vue « bien-être » animal.

VI. Bibliographie

Aguilar, I., I. Misztal, S. Tsuruta, A. Legarra, and H. Wang, 2014 PREGSF90–POSTGSF90: computational tools for the implementation of single-step genomic selection and genome-wide association with ungenotyped individuals in BLUPF90 programs, in *Proceedings of the 10th world congress of genetics applied to livestock production*,.

Aliloo, H., J. E. Pryce, O. González-Recio, B. G. Cocks, M. E. Goddard *et al.*, 2017 Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. Journal of dairy science 100: 1203–1222.

Aliloo, H., J. E. Pryce, O. González-Recio, B. G. Cocks, and B. J. Hayes, 2016 Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genetics Selection Evolution 48: 8.

Aulchenko, Y. S., D.-J. De Koning, and C. Haley, 2007 Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177: 577–585.

Beissinger, T. M., G. J. Rosa, S. M. Kaeppler, D. Gianola, and N. De Leon, 2015 Defining window-boundaries for genomic analyses using smoothing spline techniques. Genetics Selection Evolution 47: 30.

Blott, S., J.-J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet *et al.*, 2003 Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163: 253–266.

Boichard, D., M. Boussaha, A. Capitan, D. Rocha, C. Hozé *et al.* Experience from large scale use of the Eutogenomics custom SNP chip in cattle. Proceedings of the World Congress on Genetics Applied to Livestock Production 675.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol *et al.*, 2012 Genomic selection in French dairy cattle. Animal Production Science 52: 115.

Boichard, D., L. Maignel, and E. Verrier, 1996 Analyse généalogique des races bovines laitières françaises. Productions Animales 5 (9), 323-335.(1996).

Boichard, D., L. Maignel, and E. Verrier, 1997 The value of using probabilities of gene origin to measure genetic variability in a population. Genetics Selection Evolution 29: 5.

Bouquet, A., G. Renand, and F. Phocas, 2009 Evolution de la diversité génétique des populations françaises de bovins allaitants spécialisés de 1979 à 2008. Productions Animales 22: 317.

Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei *et al.*, 2018 Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nature genetics 50: 362.

Brøndum, R. F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen *et al.*, 2011 Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. Journal of Dairy Science 94: 4700–4707.

Browning, B. L., and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics 84: 210–223.

Calus, M. P. L., and R. F. Veerkamp, 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. J. Anim. Breed. Genet. 124: 362–368.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4: 7.

Chen, L., C. Li, S. Miller, and F. Schenkel, 2014 Multi-population genomic prediction using a multi-task Bayesian learning model. BMC genetics 15: 53.

Christensen, O. F., A. Legarra, M. S. Lund, and G. Su, 2015 Genetic evaluation for three-way crossbreeding. Genetics Selection Evolution 47: 98.

Colleau, J. J., S. Fritz, F. Guillaume, A. Baur, D. Dupassieux *et al.*, 2015 Simulation des potentialités de la sélection génomique chez les bovins laitiers. INRA Prod. Anim 28: 251–258.

Coop, G., X. Wen, C. Ober, J. K. Pritchard, and M. Przeworski, 2008 High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. science 319: 1395–1398.

Croiseau, P., M.-N. Fouilloux, D. Jonas, S. Fritz, A. Baur *et al.*, 2014 Extension to Haplotypes of Genomic Evaluation Algorithms. 10th World Congress of Genetics Applied to Livestock Production.

Croiseau, P., A. Legarra, F. Guillaume, S. Fritz, A. Baur *et al.*, 2011 Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. Genetics research 93: 409–417.

Croiseau, P., T. Tribout, D. Boichard, M. P. Sanchez, and S. Fritz, 2017 Use of whole sequence GWAS to improve genomic evaluation in dairy cattle.

Dezetter, C., H. Leclerc, S. Mattalia, A. Barbat, D. Boichard *et al.*, 2015 Inbreeding and crossbreeding parameters for production and fertility traits in Holstein, Montbéliarde, and Normande cows. Journal of Dairy Science 98: 4904–4913.

Dollé, J. B., J. Agabriel, J.-L. Peyraud, P. Faverdin, V. Manneville *et al.*, 2011 Les gaz à effet de serre en élevage bovin: évaluation et leviers d'action. Productions Animales 24: 415.

Doublet, A.-C., P. Croiseau, S. Fritz, A. Michenet, C. Hozé *et al.*, 2019 The impact of genomic selection on genetic diversity and genetic gain in three French dairy cattle breeds. Genetics Selection Evolution 51: 52.

Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume *et al.*, 2008 Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178: 2227–2235.

Druet, T., and M. Gautier, 2017 A model-based approach to characterize individual inbreeding at both global and local genomic scales. Molecular ecology 26: 5820–5841.

Duenk, P., M. P. Calus, Y. C. Wientjes, and P. Bijma, 2017 Benefits of dominance over additive models for the estimation of average effects in the presence of dominance. G3: Genes, Genomes, Genetics 7: 3405–3414.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of dairy science 95: 4114–4129.

Eynard, S. E., P. Croiseau, D. Laloë, S. Fritz, M. P. Calus *et al.*, 2018 Which individuals to choose to update the reference population? Minimizing the loss of genetic diversity in animal genomic selection programs. G3: Genes, Genomes, Genetics 8: 113–121.

Falconer, D. S., and T. F. C. Mackay, 1996 Introduction to quantitative genetics. 1996. Harlow, Essex, UK: Longmans Green 3:.

Fernando, R. L., H. Cheng, B. L. Golden, and D. J. Garrick, 2016 Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. Genetics Selection Evolution 48: 96.

Fernando, R. L., J. C. Dekkers, and D. J. Garrick, 2014 A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genetics Selection Evolution 46: 50.

Fikse, W. F., and G. Banos, 2001 Weighting factors of sire daughter information in international genetic evaluations. Journal of dairy science 84: 1759–1767.

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds, 2015 minimac2: faster genotype imputation. Bioinformatics 31: 782–784.

García-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-López *et al.*, 2016 Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proceedings of the National Academy of Sciences 113: E3995–E4004.

Gianola, D., R. L. Fernando, and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761–1776.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics 12: 186.

Hayes, B. J., A. J. Chamberlain, and M. E. Goddard, 2006 Use of markers in linkage disequilibrium with QTL in breeding programs., pp. 30–06 in *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil, 13-18 August, 2006*, Instituto Prociência.

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman *et al.*, 2007 Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genet. Res. 89: 215–220.

Heard, E., and R. A. Martienssen, 2014 Transgenerational epigenetic inheritance: myths and mechanisms. Cell 157: 95–109.

Hedrick, P. W., and A. Garcia-Dorado, 2016 Understanding inbreeding depression, purging, and genetic rescue. Trends in Ecology & Evolution 31: 940–952.

Hoggart, C. J., J. C. Whittaker, M. De Iorio, and D. J. Balding, 2008 Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS genetics 4: e1000130.

Howard, J. T., J. E. Pryce, C. Baes, and C. Maltecca, 2017 Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. Journal of dairy science 100: 6009–6024.

Hozé, C., M.-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville *et al.*, 2013 High-density marker imputation accuracy in sixteen French cattle breeds. Genetics Selection Evolution 45: 33.

Hoze, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq *et al.*, 2014a Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. Journal of dairy science 97: 3918–3929.

Hoze, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq *et al.*, 2014b Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds, pp. 17–22 in *Proceedings of the 10th world congress of genetics applied to livestock production*,.

Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers, and P. Donnelly, 2005 Human recombination hot spots hidden in regions of strong marker association. Nature genetics 37: 601.

Jonas, D., V. Ducrocq, and P. Croiseau, 2015 Haplotype construction methods to enhance genomic evaluation.

Jónás, D., V. Ducrocq, M.-N. Fouilloux, and P. Croiseau, 2016 Alternative haplotype construction methods for genomic evaluation. Journal of dairy science 99: 4537–4546.

Jónás, D., C. Hozé, D. Boichard, and P. Croiseau, 2014 Application of a three-haplotype LDLA model to the French Holstein population, in *Proceedings, 10th World Congress of*

Genetics Applied to Livestock Production. 2014; 10. World Congress of Genetics Applied to Livestock Production, Vancouver, CAN, 2014-08-17-2014-08-22, 1-3,.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nature genetics 42: 348.

Larmer, S. G., M. Sargolzaei, and F. S. Schenkel, 2014 Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. Journal of dairy science 97: 3128–3141.

Legarra, A., P. Croiseau, M. P. Sanchez, S. Teyssèdre, G. Sallé *et al.*, 2015 A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. Genetics Selection Evolution 47: 6.

Legarra, A., A. Ricard, and O. Filangi, 2013 GS3—Genomic selection, Gibbs Sampling, Gauss Seidel and Bayes Cπ.

Leroy, G., 2014 Inbreeding depression in livestock species: review and meta-analysis. Animal genetics 45: 618–628.

Liu, Z., M. E. Goddard, F. Reinhardt, and R. Reents, 2014 A single-step genomic model with direct estimation of marker effects. Journal of dairy science 97: 5833–5850.

Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendaño, 2007 Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet. 124: 377–389.

Ma, L., J. R. O'Connell, P. M. VanRaden, B. Shen, A. Padhi *et al.*, 2015 Cattle sex-specific recombination and genetic control from a large pedigree analysis. PLoS genetics 11: e1005387.

Mäki-Tanila, A., and W. G. Hill, 2014 Influence of gene interaction on complex trait variation with multilocus models. Genetics 198: 355–367.

McQuillan, R., A.-L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic *et al.*, 2008 Runs of homozygosity in European populations. The American Journal of Human Genetics 83: 359–372.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Meuwissen, T. H., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161: 373–379.

Misztal, I., A. Legarra, and I. Aguilar, 2014 Using recursion to compute the inverse of the genomic relationship matrix. Journal of dairy science 97: 3943–3952.

Misztal, I., S. Tsuruta, Y. Masuda, I. Pocrnic, A. Legarra *et al.*, 2019 Fluctuations in genomic predictions with APY inversion. Interbull Meeting.

Mrode, R. A., and G. J. T. Swanson, 2004 Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livestock Production Science 86: 253–260.

Muir, W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124: 342–355.

Patry, C., and V. Ducrocq, 2009 Bias due to genomic selection. Interbull Bulletin 77–77.

Popova, M., D. P. Morgavi, M. Doreau, and C. Martin, 2011 Production de méthane et interactions microbiennes dans le rumen. Productions Animales 24: 447.

Pryce, J. E., M. Haile-Mariam, M. E. Goddard, and B. J. Hayes, 2014 Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genetics Selection Evolution 46: 71.

Reik, W., W. Dean, and J. Walter, 2001 Epigenetic reprogramming in mammalian development. Science 293: 1089–1093.

Sanchez, M. P., M. El Jabri, S. Minéry, V. Wolf, E. Beuvier *et al.*, 2018 Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectra in a large data set of Montbéliarde cows. Journal of dairy science 101: 10048–10061.

Sanchez, M. P., D. Jonas, A. Baur, V. Ducrocq, C. Hoze *et al.*, 2016 Mise en place d'une évaluation génomique en races Abondance, Tarentaise et Vosgienne, pp. np in *23. Rencontres autour des Recherches sur les Ruminants*,.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel, 2014 A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15: 478.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447–2454.

Sorensen, D., and D. Gianola, 2007 *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer Science & Business Media.

Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand *et al.*, 2012 Comparison of genomic predictions using medium-density (∼ 54,000) and high-density (∼ 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science 95: 4657–4665.

Taskinen, M., E. A. Mäntysaari, and I. Strandén, 2017 Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. Genetics Selection Evolution 49: 36.

Teissier, M., M. P. Sanchez, M. Boussaha, A. Barbat, C. Hoze *et al.*, 2018 Use of meta-analyses and joint analyses to select variants in whole genome sequences for genomic evaluation: An application in milk production of French dairy cattle breeds. Journal of dairy science 101: 3126–3139.

Teyssèdre, S., J.-M. Elsen, and A. Ricard, 2012 Statistical distributions of test statistics used for quantitative trait association mapping in structured populations. Genetics Selection Evolution 44: 32.

Toro, M. A., and L. Varona, 2010 A note on mate allocation for dominance handling in genomic selection. Genetics Selection Evolution 42: 33.

Van Binsbergen, R., M. C. Bink, M. P. Calus, F. A. Van Eeuwijk, B. J. Hayes *et al.*, 2014 Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 46: 41.

Van den Berg, I., D. Boichard, and M. S. Lund, 2016 Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. Journal of dairy science 99: 8932–8945.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker *et al.*, 2013 Genomic imputation and evaluation using high-density Holstein genotypes. Journal of dairy science 96: 668–678.

VanRaden, P. M., and G. R. Wiggans, 1991 Derivation, Calculation, and Use of National Animal Model Information. Journal of Dairy Science 74: 2737–2746.

Varona, L., A. Legarra, M. A. Toro, and Z. G. Vitezica, 2018 Non-additive effects in genomic selection. Frontiers in genetics 9: 78.

Wakefield, J., 2009 Bayes factors for genome-wide association studies: comparison with P-values. Genetic epidemiology 33: 79–86.

Wei, M., and J. H. van der Werf, 1994 Maximizing genetic response in crossbreds using both purebred and crossbred information. Animal Science 59: 401–413.

Weng, Z.-Q., M. Saatchi, R. D. Schnabel, J. F. Taylor, and D. J. Garrick, 2014 Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genetics Selection Evolution 46: 34.

Zeng, J., A. Toosi, R. L. Fernando, J. C. Dekkers, and D. J. Garrick, 2013 Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genetics Selection Evolution 45: 11.

Zou, H., and T. Hastie, 2003 Regression shrinkage and selection via the elastic net, with applications to microarrays. JR Stat Soc Ser B 67: 301–20.

Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67: 301–320.

VII. Table des illustrations

A. Liste des figures

B. Liste des tableaux