Intégration de données génomiques (mutations, gènes majeurs, marqueurs SNP, haplotypes) dans les modèles d'évaluations génétiques des chèvres laitières pour améliorer l'efficacité de la sélection

Marc Teissier

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

M. MARC TEISSIER

le mardi 5 février 2019

Titre :

Intégration de données génomiques (mutations, gènes majeurs, marqueurs SNP, haplotypes) dans les modèles d'évaluations génétiques des chèvres laitières pour améliorer l'efficacité de la sélection

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GenPhySE)

Directeur(s) de Thèse :

MME CHRISTELE ROBERT-GRANIE

Rapporteurs :

Mme FLORENCE PHOCAS, INRA JOUY EN JOSAS
Mme PASCALE LE ROY, INRA RENNES

Membre(s) du jury :

Mme RACHEL RUPP, INRA TOULOUSE, Président
M. FRANCOIS GUILLAUME, EVOLUTION, Membre
M. LAURENCE MOREAU, INRA MOULON, Membre
Mme CHRISTELE ROBERT-GRANIE, INRA TOULOUSE, Membre

Remerciements

Je tiens tout d'abord à remercier la région Occitanie ainsi que le métaprogramme SELGEN (Incomings) de l'INRA qui ont financé cette thèse et sans qui, ce travail n'aurait pas été possible.

Je tiens à remercier les membres du jury de cette thèse : Pascale Le Roy, Florence Phocas qui ont accepté d'être rapporteurs et François Guillaume, Laurence Moreau, Rachel Rupp en tant qu'examinateurs.

Les membres de mon comité de thèse : Pascal Croiseau, Hélène Larroque, Andrés Legarra, Hélène Muranty et Léopoldo Sanchez ont contribué activement par leurs conseils et leurs avis bénéfiques à la réalisation de ma thèse et je les en remercie.

J'ai eu un immense plaisir à travailler dans l'unité GenPhySE, et je remercie tous les collègues pour leur accueil chaleureux.

Je remercie évidemment ma directrice de thèse Christèle Robert-Granié. Je tiens à lui faire part de toute ma reconnaissance pour m'avoir soutenu au cours de ces 3 années. Elle a su m'aider dans le développement de mon projet tout en me laissant une grande liberté. Elle a su me faire confiance pour cette thèse, qui reste une incroyable expérience dans ma vie.

Although they probably will not read this, I would like to thank all people I met at the University of Guelph. The 3 months I spent there were short, but I only have good memories in mind.

Je remercie également tous mes amis rencontrés au cours de ces 3 années : Alexandre, Claire et Claire, Émilie, Estelle, Hung, Mathieu, Morgane, Pauline, Sophie, et Tiphaine. Entre les soirées, les balades à cheval, les lasers game et les courses dans la boue, ces bons moments passés ensemble n'ont pas manqué et j'en conserverais d'excellents souvenirs.

Je passe ensuite une dédicace spéciale à ma famille. À mes parents et mon frère, sans qui je ne serais pas là devant vous aujourd'hui et qui m'ont soutenu dans mes choix et dans tous mes projets jusqu'à ce jour. J'ai une pensée pour mes grands-parents qui auraient été fiers de me voir arriver là où j'en suis aujourd'hui. Tout particulièrement pour ma grand-mère "Mamiche" qui ne m'aura pas vu arriver au bout de ces 3 années.

Résumé

Suite aux travaux de Céline Carillier (2012-2015), des évaluations ssGBLUP ont été mises en place en 2018 pour les races caprines Alpine et Saanen. L'objectif est d'améliorer les précisions des évaluations pour maximiser le progrès génétique pour les caractères d'intérêt. Pour notre première étude, nous nous sommes intéressés à l'effet de la taille de la population de référence (limitée pour ces races) sur les précisions des évaluations. L'accroissement de la population d'apprentissage ne s'est pas systématiquement accompagné d'une hausse des précisions. Le ssGBLUP présente des biais et tend à surestimer ou sous-estimer les valeurs génomiques. Des hyperparamètres ont été introduits dans la construction de la matrice génomique du ssGBLUP pour limiter ces biais. Ces hyperparamètres (α, ω et τ) peuvent améliorer les biais tout en affectant de manière limitée les précisions. Pour les races Alpine et Saanen, les biais sont proches de 1 pour un ω compris entre 0,1 et 0,3 et un τ compris entre 3 et 4. L'hyperparamètre α a peu d'effet sur les précisions et les biais, sa valeur par défaut (0,95) semble être optimale. Dans une deuxième partie, nous nous sommes intéressés à l'intégration de mutations causales ou de QTLs dans les modèles d'évaluations pour améliorer les précisions. Des mutations causales et des QTLs ont été détectés dans les races caprines. On peut citer le gène de la *caséine* α_{s1} pour le taux protéique ou *DGAT1* pour le taux butyreux. D'autres études ont identifié un QTL, localisé sur le chromosome 19, en Saanen. Il a été détecté pour les caractères : quantités de lait et de matières (grasses et protéiques), la distance plancher-jarret et pour la qualité de l'attache arrière. L'utilisation des génotypes de la *caséine* α_{s1} ou *DGAT1* dans les modèles d'évaluations (gene content) a été inefficace pour améliorer les précisions des évaluations. Le gene content est une méthode multicaractère où le « gene content » est un second caractère corrélé au caractère en sélection. Pour le taux protéique ou butyreux, les précisions avec le gene content sont entre -11 % et 0 % inférieures aux précisions du ssGBLUP. En pondérant les SNPs de manière adéquate avec un ssGBLUP (appelée Weighted ssGBLUP et notée WssGBLUP), les précisions des évaluations ont été améliorées. Cette méthode attribue des poids aux SNPs en fonction de leur association aux caractères. Ces poids sont intégrés dans la construction de la matrice de parenté génomique. Des gains jusqu'à +5 % et +14 % (Alpine et Saanen) ont été observés par rapport au ssGBLUP. Le WssGBLUP est plus adapté pour la race Saanen car des QTLs sont présents sur la majorité des caractères. Pour la race Alpine, le WssGBLUP s'est avéré intéressant pour le taux protéique. Le ssGBLUP reste la meilleure méthode lorsque le caractère a une architecture génétique polygénique. Enfin, nous nous sommes intéressés à des modèles d'évaluation génomiques haplotypiques. Les haplotypes ont été construits en regroupant plusieurs SNPs consécutifs ou en se basant sur le déséquilibre de liaison entre SNPs. Les haplotypes sont utilisés pour construire une matrice de parenté haplotypique ou convertis en pseudo-SNPs, pour construire une matrice de parenté génomique. En Alpine, les précisions du ssGBLUP haplotypiques (ou pseudo-SNPs) ont évolué entre -1 % et 19 % par rapport au ssGBLUP basé sur l'information des SNPs. En Saanen, les précisions ont évolué entre -3 % et +6 % par rapport au ssGBLUP. Nous avons appliqué le WssGBLUP avec des pseudo-SNPs. En Saanen, une amélioration des précisions jusqu'à +16 % par rapport au ssGBLUP a été observée. Les gains les plus forts (supérieurs à +10 %) sont obtenus pour les caractères avec un QTL identifié (lait, matières grasses et protéiques, taux protéique, qualité de l'attache arrière et distance entre le plancher et le jarret). En Alpine, des gains de précision entre -8 % et +5 % ont été observés par rapport au ssGBLUP selon le caractère excepté pour les matières grasses (+19 %).

Abstract

Following Céline Carillier's PhD (2012-2015), genomic evaluations based on the ssGBLUP were implemented in 2018 in the dairy goat breeds Alpine and Saanen. The objective of breeders is to improve the accuracy of genomic evaluations in order to maximize genetic gain for traits of interest. In our first study, we looked at the effect of the size of the reference population (limited for these breeds) on the accuracy of genomic evaluations. The increase of the training population was not systematically associated with an increase of genomic accuracies. The ssGBLUP has some biases and tends to overestimate or underestimate genomic value estimates. To avoid these biases, hyperparameters were introduced into the construction of the ssGBLUP genomic relationship matrix. An analysis of these hyperparameters (α, ω and τ) was carried out and we found that the choice of them improves bias while having a limited impact on genomic accuracy. For the Alpine and Saanen breeds, the biases are close to 1 for a ω between 0.1 and 0.3 and a τ between 3 and 4. The hyperparameter α has little effect on accuracy and bias and its default value (0,95) seems to be optimal. In a second part of my thesis, we focused on the integration of causal mutations or QTLs into genomic evaluation models to improve genomic accuracy. Causal mutations and QTLs were detected in the Alpine and Saanen breeds such as the α_{s1} *casein* gene for protein content or DGAT1 for fat content. Other studies have shown a QTL, located on chromosome 19, in the Saanen breed. It was detected for different traits: milk, fat and protein content, udder floor position and rear udder attachment. The use of genotypes for α_{s1} *casein* or *DGAT1* in genomic evaluation models (gene content) was inefficient in improving evaluation accuracy. The gene content is a multi-trait method where the "*gene content*" is a second trait correlated to the selected trait. Whether for protein or fat content, accuracies with gene content were between -11% and 0% lower than the ssGBLUP accuracies for the Alpine and Saanen breeds. We have shown by adequately weighting SNPs in an ssGBLUP (approach called Weighted ssGBLUP and noted WssGBLUP), the accuracy of evaluations could be improved. This method assigns weights to SNPs based on their association with traits. These weights are integrated into the construction of the genomic relationship matrix. Gains up to +5% for the Alpine breed and +14% for the Saanen breed were observed compared to the ssGBLUP. The WssGBLUP is more suitable for the Saanen breed because QTLs are present on the majority of traits. For the Alpine breed, WssGBLUP was interesting for the protein content. The ssGBLUP remained the most interesting method when the trait had a polygenic genetic architecture. Finally, in the last study, we focused on haplotype genomic evaluation models. Haplotypes were constructed either by grouping several consecutive SNPs or by using the linkage disequilibrium (LD) between SNPs. The haplotypes are then used to build a haplotypic relationship matrix or converted to pseudo-SNPs to build a genomic relationship matrix. In the Alpine breed, the accuracy of the haplotypic ssGBLUP (or pseudo-SNPs) was increased between -1% and 19% compared to an ssGBLUP based on SNP information. On the other hand, in the Saanen breed, the accuracy was increased between -3% and +6% compared to a ssGBLUP. Finally, we applied the WssGBLUP approach using pseudo-SNPs. In the Saanen breed, an improvement in accuracy up to +16% compared to a ssGBLUP was observed. The highest gains (above +10%) were obtained for traits with an identified QTL (milk, fat and protein yields, protein content, udder floor position and rear udder attachment). In the Alpine breed, accuracy gains between -8% and +5% were observed compared to ssGBLUP depending on the trait except for fat yield and fat content where the gains reach +19%.

Table des matières

Liste des abréviations

AAR: Qualité de l'attache arrière
ALOX12: Arachidonate 12-lipoxygenase
ALOX12B: Arachidonate 12-lipoxygenase, 12R type
ALOX15: Arachidonate 15-lipoxygenase
ASGR2: Asialoglycoprotein receptor 2
AVP: Forme de l'avant pis
BVSM : Modèle bayésien de sélection des variables (Bayesian Variable Selection Model)
BLUP: Meilleur prédicteur linéaire non biaisé (Best Linear Unbiased Predictor)
BSSVS: Sélection de variables de recherche stochastique bayésienne (Bayesian Stochastic Search Variable Selection)
CCS: Comptages de cellules somatiques
CGT6: Gamma-glutamyltransferase 6
CR: Call Rate
CTIG: Centre de Traitement de l'Information Génétique
DGAT1: Diacylglycerol O-Acyltransferase 1
DW: Distinct Windows (Fenêtres Distinctes)
DYD: Daughter Yield Deviation
EBV: Valeur génétique estimée (Estimated Breeding Value)
GBLUP: Meilleur prédicteur linéaire génomique non biaisé (Genomic Best Linear Unbiased Predictor)
GEBV: Valeur génomique estimée (Genomic Estimated Breeding Value)
GWAS: Etude association pangénomique (Genome Wide Association Study)
IA: Insémination animale
ICC: Index combiné caprin
IGGC: Consortium international caprin (International Goat Genome Consortium)
IMC: Index morphologique caprin
IPC: Index de production caprine
LA: Analyse de Liaison (Linkage Analyses)
LD: Déséquilibre de liaison (Linkage Disequilibrium)
LSCS: Performance à la lactation du score de cellules somatiques corrigé
MAF: Fréquence de l'allèle mineur (Minor Allele Frequency)
MG: Quantité de matière grasse
MP: Quantité de matière protéique
NGS: Nouvelles techniques de séquençage (Next Generation Sequencing)
ORT: Orientation des trayons
PLA: Distance entre le plancher de la mamelle et le jarret
PLD2: Phospholipase D2
PRM: Profil de la mamelle
QC: Contrôle Qualité (Quality Control)
QTL: Locus de caractères quantitatifs (Quantitative Trait Locus)
RARA: Retinoic Acid Receptor Alpha
RRBLUP: Ridge Regression BLUP
SCS: Score de cellules somatiques
SNP: Polymorphisme d'une seule base nucléotidique (Single Nucleotide Polymorphism)
ssGBLUP: BLUP génomique en une étape (single-step GBLUP)

STAT3: Signal transducer and activator of transcription 3
STAT5A: Signal transducer and activator of transcription 5A
STAT5B: Signal transducer and activator of transcription 5A
TABLUP: BLUP avec une matrice de parenté spécifique au caractère (Trait-specific marker-derived relationship matrix)
TB: Taux butyreux
TF: Forme des Trayons
TP: Taux protéique
WssGBLUP: ssGBLUP pondéré (Weighted single-step GBLUP)
YD: Yield Deviation

Chapitre 1 : La filière caprine à l'ère de la génomique

1. Contexte de la filière

1.1. La filière caprine dans le monde

La viande, le lait et les poils sont les principales productions caprines. En 2013, cette filière compte plus de 1 milliard de tête dans le monde (Skapetas and Vampidis, 2016) avec une augmentation régulière (+34 % entre 2000 et 2013). Cette augmentation surpasse celle des ovins (+11 %) sur la même période. Les effectifs sont inégalement répartis dans le monde (Figure 1), avec 59 % des têtes en Asie et 35 % en Afrique en 2013. L'Europe représente 1,65 % de l'effectif mondial avec 16 millions de têtes, cependant 14,20 % de la production mondiale de lait provient des pays européens. La production de viande caprine est concentrée en Asie (71 %) et en Afrique (24 %) (Skapetas and Vampidis, 2016). La production de viande en Europe représente seulement 2,22 % de la production totale (Skapetas and Vampidis, 2016). L'Afrique et l'Asie sont les principaux producteurs de poils (respectivement 77 % et 19 %). En France, la production de poils est une production de niche qui produit 30 tonnes de poils par an (Allain and Roguet, 2003).

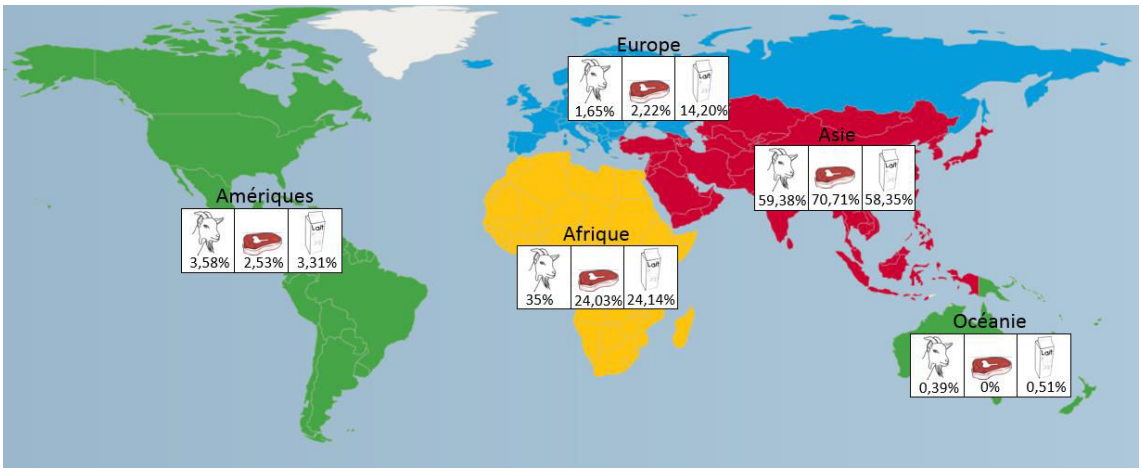


Figure 1. Répartition des effectifs et de la production de viande et de lait caprins par continent

1.2. La filière laitière caprine en France

Parmi les pays européens, la France est le premier producteur de lait avec 603 millions de litres pour l'année 2016 et est au 4ème rang européen en terme d'effectif (Institut de l'élevage, 2017a). Cette production dépasse celle de l'Espagne (503 millions de litres en 2016) ou de la Grèce (353 millions de litres en 2016) qui sont les seconds et troisièmes plus importants producteurs de lait caprin en Europe. La production de lait en France augmente (589 millions de litres de lait en 2014 contre 603 en 2016). Cette production se régionalise fortement sur le territoire français. En 2016, les plus gros bassins de production sont la région Nouvelle-Aquitaine (39 %), les Pays de la Loire (16 %), Centre Val de Loire (12 %) et Auvergne-Rhône-Alpes (11 %). La production laitière est principalement destinée à la transformation fromagère (environ 80 %). Ainsi, entre 2008 et 2013, la fabrication de fromages atteint 90 000 tonnes.

Le cheptel français diminue avec un effectif qui est passé de 1,25 million de têtes à 1,19 million de têtes entre 2014 et 2016. La filière caprine est l'une des plus petites filières agricoles puisqu'en 2016, la filière bovine comptait 19 millions de têtes et la filière ovine comptait 7,2 millions de têtes. Le nombre d'exploitations agricoles professionnelles tend à diminuer au cours du temps et ce phénomène s'observe également dans de nombreuses filières agricoles (Insee, 2013). Pour les caprins, le nombre d'exploitations de plus de 10 chèvres avec une activité laitière a chuté de 5300 en 2014 à 4900 en 2016 (Institut de l'élevage, 2017a). Parmi ces exploitations, on retrouve une grande disparité. Les exploitations agricoles spécialisées dans la

transformation du lait à la ferme (47 % des exploitations, 21 % des chèvres) possèdent des cheptels moyens de 70 chèvres tandis que le cheptel moyen atteint 237 chèvres pour les exploitations spécialisées dans la livraison de lait à des laiteries (48 % des cheptels, 72 % des chèvres).

En France, 14 races caprines laitières sont reconnues par le ministère de l'Agriculture dont 2 ont un schéma de sélection (Alpine, Saanen). Les races Alpine et Saanen sont les deux principales races françaises, elles représentent 52 % et 40 % du cheptel français et elles sont réparties sur l'ensemble du territoire. Les autres races présentes sont la chèvre des Fossés, la Lorraine, la Poitevine, la chèvre du Massif central, l'Angora, la Pyrénéenne, la chèvre provençale, la Rove et la Corse. Les races Alpine et Saanen sont originaires de Suisse et leurs effectifs en France ont augmenté après la Seconde Guerre mondiale en raison de leurs bonnes aptitudes laitières. Les résultats du contrôle laitier de 2016 montrent que la race Alpine produit 929 kg en moyenne par lactation (en 298 jours) avec un taux protéique moyen de 33,4 g/kg et un taux butyreux de 37,8 g/kg. La race Saanen produit 984 kg de lait en moyenne (en 311 jours), avec un taux protéique moyen de 32,2 g/kg et 35,9 g/kg en moyenne pour le taux butyreux.

2. La sélection génétique des races laitières Alpine et Saanen

2.1. Caractères évalués, modèles génétiques d'évaluation, organisation et efficacité de la sélection

La sélection génétique chez les caprins a débuté dans les années 1970. Le principe de cette sélection consiste à utiliser comme reproducteurs les animaux avec les valeurs génétiques (ou EBV pour Estimated Breeding Value) les plus élevées pour les caractères que l'on souhaite améliorer. Actuellement, 17 caractères (5 caractères laitiers, les comptages de cellules somatiques et 11 caractères de morphologie) sont évalués en routine, mais seuls 11 sont intégrés dans les index de synthèse. Les 5 caractères de production laitière sont les quantités de lait (LAIT : kg), les quantités de matière protéique dans le lait (MP : kg), les quantités de matière grasse dans le lait (MG : kg), le taux protéique (TP : g/kg) et le taux butyreux (TB : g/kg). Les évaluations génétiques Alpine et Saanen officielles sont réalisées conjointement pour ces 2 races (analyse multirace) pour chacun de ces 5 caractères (analyse uni-caractère). Le modèle utilisé est le suivant (Clément et al., 2015) :

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{Wp} + \boldsymbol{e} \quad \text{[Modèle 1]}$$

où $\boldsymbol{y}$ est un vecteur de phénotypes, $\boldsymbol{b}$ est un vecteur qui regroupe l'ensemble des effets d'environnement, $\boldsymbol{u}$ est un vecteur contenant les valeurs génétiques des animaux supposé distribué normalement $N(\mathbf{0}, \boldsymbol{A}\sigma_u^2)$ où $\boldsymbol{A}$ représente la matrice de parenté basée sur le pedigree, $\boldsymbol{p}$ est un vecteur des effets d'environnements permanents $N(\mathbf{0}, \boldsymbol{I}\sigma_p^2)$ et $\boldsymbol{e}$ est un vecteur des effets résiduels normalement distribué $N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$. $\boldsymbol{X}$, $\boldsymbol{Z}$ et $\boldsymbol{W}$ sont des matrices d'incidence pour les vecteurs $\boldsymbol{b}$, $\boldsymbol{u}$ et $\boldsymbol{p}$ respectivement. Les EBV des animaux sont obtenus avec la méthode Best Linear Unbiased Prediction (BLUP) par résolution des équations du modèle mixte :

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} & \boldsymbol{X'W} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \dfrac{\sigma_e^2}{\sigma_u^2}\boldsymbol{A^{-1}} & \boldsymbol{Z'W} \\ \boldsymbol{W'X} & \boldsymbol{W'Z} & \boldsymbol{Z'W} + \dfrac{\sigma_e^2}{\sigma_p^2}\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \\ \boldsymbol{Z'y} \end{bmatrix}$$

L'ensemble des caractères de production laitière sont regroupés dans un index synthétique appelé Index de production caprine (IPC) calculé comme $IPC = 0{,}12 * MG + 0{,}59 * MP + 0{,}06 * TB + 0{,}26 * TP$ (Clément et al., 2017).

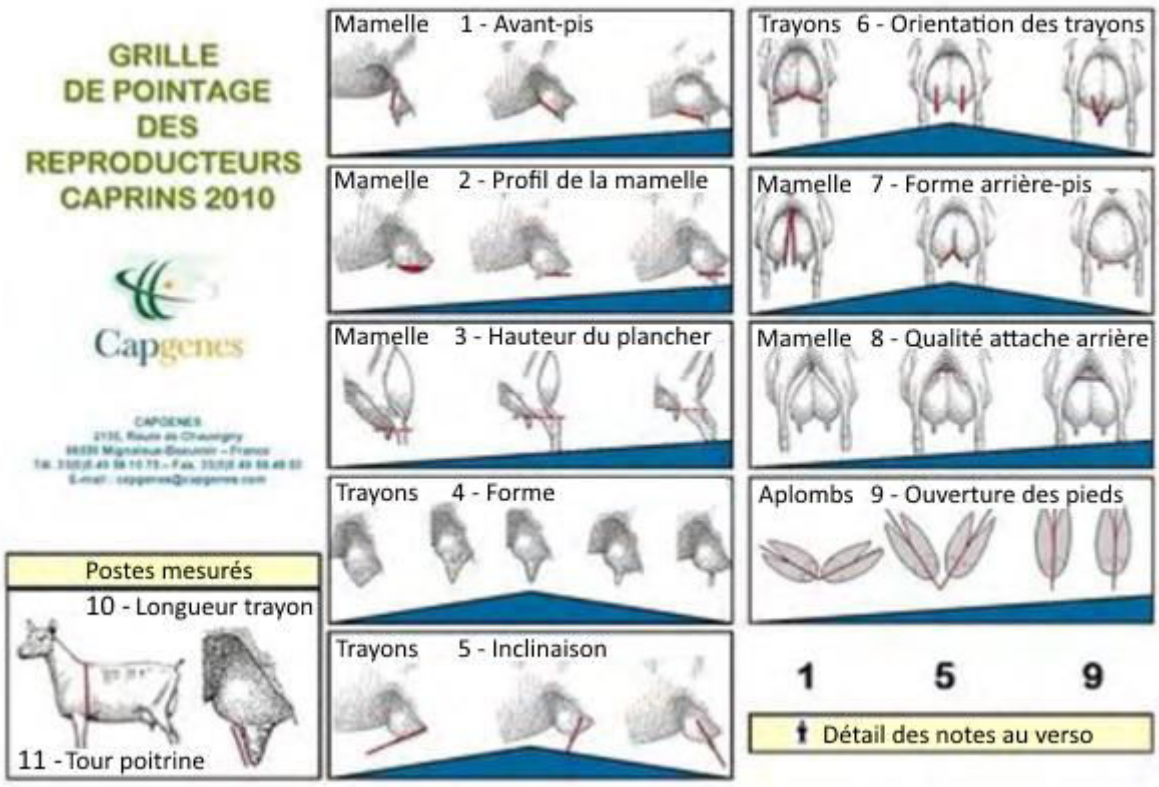


Figure 2. Grille de pointage des reproducteurs caprins (Capgènes). 1 – Forme de l'avant-pis (AVP), 2- Profil de la mamelle (PRM), 3-Hauteur du plancher (PLA), 4-Forme du trayon, 5 – Inclinaison des trayons, 6 – Orientation des trayons (ORT), 7 – Forme de l'arrière pis, 8 – Qualité de l'attache arrière (AAR), 9– Ouverture des pieds, 10 – Longueur des trayons, 11 – Tour de poitrine

Parmi les 11 caractères de la morphologie de la mamelle, pointés sur les femelles (Figure 2), 5 expliquent plus de 80 % de la variabilité totale (Clément et al., 2006) et sont utilisés dans l'indice de synthèse morphologique : la forme de l'avant-pis (AVP), le profil de la mamelle (PRM), la qualité de l'attache arrière (AAR), la distance entre le plancher de la mamelle et le jarret (PLA) et l'orientation des trayons (ORT). Ces caractères sont mesurés selon une grille de notes allant de 1 à 9. Pour les caractères AVP, PRM, PLA et AAR, des animaux avec la note maximale (9) sont recherchés tandis que la note de 5 est souhaitée pour le caractère ORT. Les caractères de morphologie de la mamelle sont évalués indépendamment dans chacune des races mais ils sont évalués simultanément (analyse multi-caractère et intra-race) (Clément et al., 2015). Le modèle d'évaluation génétique est proche du modèle 1 mais il ne comporte pas d'effet d'environnement permanent.

Comme pour les caractères de production, un index synthétique est calculé pour résumer l'information de la morphologie de la mamelle. Cet index, appelé Index morphologique caprin (IMC), est calculé comme $IMC = 0{,}2 * AVP + 0{,}2 * PRM + 0{,}2 * AAR + 0{,}2 * PLA + 0{,}2 * ORT$ (Clément et al., 2017).

Le dernier caractère considéré en sélection est le comptage de cellules somatiques/mL de lait (CCS). Il est indexé depuis 2013 afin d'intégrer la résistance aux mammites dans l'objectif de sélection (Rupp et al., 2011, 2011 ; Huau et al., 2015). Le modèle utilisé est identique au modèle 1.

L'ensemble des caractères évalués est résumé dans un index synthétique propre à chaque race, appelé Index combiné caprin (ICC), qui prend en compte les caractères de production

laitière et les caractères de morphologie de la mamelle. En Alpine, il est calculé comme $ICC_{Alpin} = IPC + 0{,}5 * IMC$ et en Saanen, comme $ICC_{Saanen} = IPC + 0{,}6 * IMC$.

Capgènes est l'organisme et l'entreprise de sélection qui gère le schéma de sélection (Figure 3). Ce dernier est identique pour les races Alpine et Saanen, mais la sélection est spécifique à chaque race. En France, la population caprine compte plus de 1 million de chèvres. Parmi ces animaux, seuls 170 000 constituent la base de sélection. Jusqu'en 2017, le schéma de sélection s'appuyait sur une évaluation génétique et était organisé comme décrit ci-après. Au sein de la base de sélection, 1000 accouplements raisonnés sont réalisés entre les meilleures femelles et les meilleurs mâles chaque année. Depuis 2006, une méthode combinée de gestion du progrès génétique et de la variabilité génétique est appliquée au moment de ces accouplements. Elle vise à minimiser l'augmentation de consanguinité pour un progrès génétique donné et fixé par Capgenes. Les mâles issus de ces accouplements sont intégrés en station de contrôle pour être évalués sur leur état sanitaire, sur leurs performances sexuelles et sur leur croissance. À la suite de cette sélection, 70 mâles environ sont testés sur descendance. Le test consiste à réaliser entre 180 et 200 inséminations par mâle. Le potentiel génétique des mâles est estimé au travers des performances des femelles issues des inséminations (70 à 80 femelles de testage par mâle). Les mâles sont triés sur leurs ICC et 40 boucs obtiendront un agrément pour que ces animaux soient utilisés comme reproducteurs.

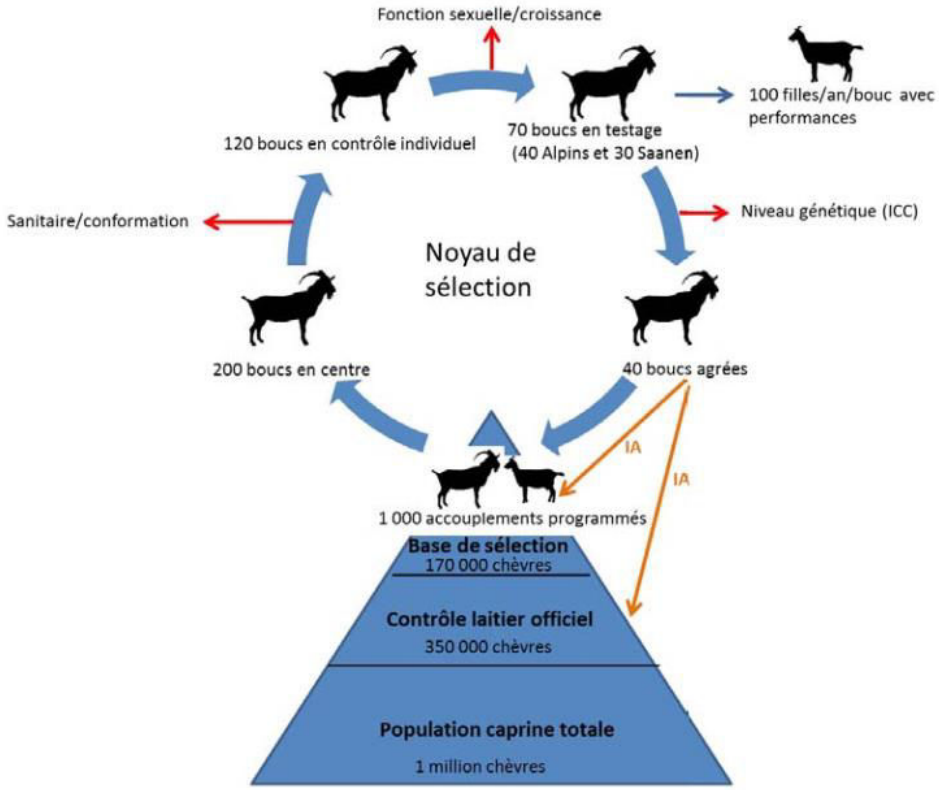


Figure 3. Organisation du schéma de sélection génétique français pour l'espèce caprine (jusqu'en 2017) (source : Capgènes)

L'efficacité du schéma de sélection peut être évaluée grâce au progrès génétique annuel (ΔG). Ce critère suit l'amélioration des valeurs génétiques des animaux au cours du temps. Il est calculé comme $\Delta G = \frac{iR\sigma_g}{\bar{T}}$. L'intensité de sélection (i) se définit comme la supériorité génotypique moyenne des animaux sélectionnés par rapport aux animaux candidats exprimés en écart-type. Elle sera plus élevée en augmentant le nombre de candidats, en diminuant le nombre d'animaux agrées ou bien en modifiant ces deux paramètres simultanément. Dans le

schéma de sélection caprin, c'est l'étape du testage sur descendance qui va limiter le nombre d'animaux candidats à la sélection (environ 70) car cette étape reste couteuse.

La précision de l'évaluation génétique (R) correspond à la corrélation entre les valeurs génétiques « vraies » et les valeurs génétiques prédites des animaux. Plus les précisions seront élevées et plus le progrès génétique sera important, on recherche donc des évaluations génétiques les plus précises possibles. Les méthodes utilisées pour les évaluations permettent, entre autres, d'améliorer les précisions des évaluations.

L'écart-type génétique (σ_g) indique la variabilité génétique d'un caractère. Un caractère avec une grande variabilité aura un progrès génétique annuel plus important qu'un caractère peu variable. Si un caractère est très variable, il devient possible de sélectionner des animaux avec des valeurs génétiques extrêmes. Un critère complémentaire de l'écart-type génétique est l'héritabilité (h^2). L'héritabilité se définit comme la part de la variabilité phénotypique (σ_P^2) expliquée par la part de la variabilité génétique (σ_G^2) :

$$h^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

La sélection pour les caractères avec une forte héritabilité (supérieur à 0,5) est très efficace. Au contraire, plus l'héritabilité diminue, plus l'amélioration des caractères est compliquée. Les caractères sélectionnés en caprin ont des héritabilités assez variables (Tableau 1). Les héritabilités les plus élevées sont observées pour les taux (TB et TP), moyennes pour les matières (MG et MP) et pour les caractères de morphologie de la mamelle (AVP, PRM, AAR, PLA et ORT). Les héritabilités les plus faibles concernent le comptage des cellules somatiques (CCS). Les estimations des héritabilités sont proches entre la race Alpine et la race Saanen. Les écarts les plus importants s'observent pour les caractères PRM (0,36 en Alpine et 0,27 en Saanen), AAR (0,26 en Alpine et 0,33 en Saanen), et ORT (0,33 en Alpine et 0,25 en Saanen).

Tableau 1. Héritabilité des caractères en sélection pour les races caprines Alpine et Saanen

	Alpine	**Saanen**
LAIT	0,3	0,3
MG	0,3	0,3
MP	0,3	0,3
TB	0,5	0,5
TP	0,5	0,5
AVP	0,34	0,33
PRM	0,36	0,27
AAR	0,26	0,33
PLA	0,26	0,28
ORT	0,33	0,25
LSCS	0,19	0,21

L'amélioration génétique de l'ensemble de ces caractères est complexe, car certains caractères sont corrélés entre eux (Tableau 2 et Tableau 3). La sélection sur les quantités de lait (LAIT) aura un effet positif sur les quantités de matières (corrélation allant de 0,76 à 0,92), mais un effet négatif sur les taux (corrélation allant de -0,10 à -0,29) (Manfredi and Ådnøy, 2012). De fortes corrélations sont également observées pour les caractères de morphologie de la mamelle (Tableau 3). Les plus fortes corrélations sont entre PLA et AAR (0,77 en Alpine et 0,72 en Saanen) ou entre AVP et AAR (0,62 en Alpine et 0,65 en Saanen). A contrario, certains

caractères sont peu corrélés : PRM et AVP, PRM et AAR, PRM et PLA avec des corrélations génétiques comprises entre -0,09 et 0,08.

Tableau 2. Corrélations génétiques entre les caractères de production laitière pour les races Alpine (triangulaire supérieur) et Saanen (triangulaire inférieur) (Manfredi and Ådnøy, 2012)

	LAIT	**MG**	**MP**	**TB**	**TP**
LAIT	1	0,77	0,89	-0,18	-0,28
MG	0,76	1	0,86	0,49	N.A.
MP	0,92	0,83	1	N.A.	0,19
TB	-0,1	0,49	N.A.	1	0,61
TP	-0,29	N.A.	0,1	0,51	1

Tableau 3. Corrélations génétiques entre les caractères de morphologie de la mamelle pour les races Alpine (triangulaire supérieur) et Saanen (triangulaire inférieur) (Manfredi et al., 2001)

	AVP	**PRM**	**AAR**	**PLA**	**ORT**
AVP	1	0,03	0,62	0,51	0,11
PRM	-0,09	1	0,08	0,02	0,63
AAR	0,65	0,16	1	0,77	0,37
PLA	0,58	0,06	0,72	1	0,17
ORT	0,17	0,61	0,25	0,12	1

Enfin, l'intervalle entre générations ($\bar{\bar{T}}$) correspond à l'âge moyen des parents au moment de la naissance de leurs descendants. La biologie de l'espèce ainsi que l'organisation de la sélection génétique chez les caprins vont impacter ce progrès génétique. Le testage sur descendance est, par exemple, une étape chronophage pour l'espèce caprine (en moyenne 3 ans (Lamaix, 2004)), car il faut attendre les performances des filles de testage. Il contribue à l'allongement de l'intervalle entre générations ($\bar{\bar{T}}$), qui est en moyenne de 4,5 ans dans le schéma classique caprin.

Pour les races Alpine et Saanen, le progrès génétique a été évalué entre la période 2003 et 2012 pour l'IPC et l'IMC (Palhière et al., 2014). Il est de 0,15 écart-type génétique par an pour la race Alpine et de 0,20 écart-type génétique par an pour la race Saanen sur l'IPC. Le progrès génétique sur l'IMC est quant à lui proche de 0 car l'IMC a un poids modéré dans les objectifs de sélection. Le progrès génétique a également été analysé pour les caractères de production laitière entre 1990 et 2010 (Danchin-Burge et al., 2012). Celui-ci est de 13 kg/an pour le LAIT ou de 0,40 kg (0,20 écart-type génétique) pour la MP pour les races Alpine et Saanen.

2.2. Données nécessaires et disponibles pour les évaluations génétiques

Les évaluations génétiques nécessitent des informations de diverses origines : phénotypes, pedigrees, effets de milieux ou d'environnement. L'ensemble des informations sont centralisées au Centre de Traitement de l'Information Génétique (CTIG, INRA, Jouy-en-Josas) dans la base de données nationale caprine. Sont présentés ci-dessous un état des données phénotypiques et de pedigrees disponibles en janvier 2016, au démarrage de ma thèse.

2.2.1. Phénotypes

Les caprins sont évalués en routine sur 17 caractères (production laitière, morphologie de la mamelle et comptage de cellules somatiques). Les caractères de production laitière et le comptage des cellules somatiques sont enregistrés dans le cadre du Contrôle Laitier Officiel. Le suivi des performances est réalisé grâce à 4 protocoles (A, AT, AZ et CZ). Pour les 4

protocoles, les enregistrements des performances se font toutes les 4 à 5 semaines entre le début de la lactation et le tarissement :

- Le protocole A (28,3 % des contrôles en 2016 (Institut de l'élevage, 2017b)) mesure toutes les traites comprises dans un délai de 24 heures (1, 2 ou 3 traites). Ces contrôles concernent la quantité de lait, les matières grasses et protéiques ainsi que la numération des cellules somatiques du lait.
- Le protocole AT (45,5 % des contrôles en 2016 (Institut de l'élevage, 2017b)) enregistre une seule des deux traites quotidiennes, alternativement celle du matin et celle du soir. Ce protocole est réalisable uniquement pour des troupeaux où les chèvres sont traites 2 fois par 24 heures.
- Le protocole AZ (19,7 % des contrôles en 2016 (Institut de l'élevage, 2017b)) mesure les quantités de lait pour les 2 traites quotidiennes. Les mesures pour les quantités de matières (grasses et protéiques) ainsi que la numération des cellules somatiques du lait sont réalisées en alternance sur la traite du matin et la traite du soir. Comme pour le protocole AT, ce protocole n'est réalisable que pour les troupeaux dont les chèvres sont traites 2 fois par 24 heures.
- Le protocole CZ (6,5 % des contrôles en 2016 (Institut de l'élevage, 2017b)) est identique au protocole AT, avec les mêmes enregistrements. En revanche, l'éleveur peut enregistrer seul la traite où les matières ne sont pas évaluées, sans l'aide d'un contrôleur laitier.

Les performances mensuelles enregistrées par le contrôle laitier sont ensuite transformées en variables à la lactation, pour être utilisées dans les évaluations génétiques. Pour les caractères de production laitière, la méthode Fleischmann (méthode par interpolation est utilisée afin d'obtenir une quantité de lait, de matière grasse et protéique par animal pour une durée de lactation de référence de 250 jours. Les phénotypes utilisés pour les évaluations génétiques pour les CCS sont le score de comptage somatique corrigé (LSCS). Pour cela, un score de comptage somatique non corrige (SCS) est calculé à partir des CCS comme : $SCS = Log_2\left(\frac{CCS}{100000}\right) + 3$ pour chaque contrôle. Cette transformation est réalisée afin d'obtenir une performance normalement distribuée. Les SCS sont ensuite corrigés pour les effets du rang de lactation ainsi que le stade de lactation. Afin d'obtenir une performance à la lactation par animal (LSCS), les SCS corrigés d'un animal sont pondérés et moyennés (Clément et al., 2008).

En 2016, 1 134 troupeaux pour la race Alpine et 836 troupeaux pour la race Saanen étaient enregistrés au contrôle laitier. Au total, 151 566 lactations pour l'Alpine et 104 409 lactations ont été qualifiées (Institut de l'élevage, 2017b), c'est-à-dire respectant le protocole d'enregistrement des performances et pouvant être utilisées pour les évaluations génétiques officielles (B.O.agri, 2014).

Les caractères de morphologie de la mamelle sont enregistrés par l'entreprise Capgènes. Des techniciens réalisent des visites d'exploitations pour pointer les animaux reproducteurs. Le pointage consiste à évaluer individuellement des femelles en première lactation issue d'insémination animale (IA) (Martin, 2016). Les femelles sont évaluées sur un ensemble de postes. Ces postes sont des mesures directes (tour de poitrine, longueur des trayons,…) ou bien des notes attribuées par les techniciens (qualité de l'attache arrière, orientation des trayons,…) (Figure 2). Au cours de ma thèse, 5 caractères de morphologie de la mamelle, pointés par les techniciens ont été analysés (Tableau 4). Les femelles sont pointées une seule fois au cours de leur vie, ce qui représente environ 25 000 chèvres pointées chaque année.

Le Tableau 4 présente le nombre de phénotypes, le nombre de femelles ainsi que la moyenne et l'écart-type de chaque caractère en sélection pour les races Alpine et Saanen. Le nombre de

phénotypes est plus important pour les caractères de production laitière et pour les LSCS car un animal peut avoir plusieurs lactations au cours de sa carrière. On dispose de plus de phénotypes pour les caractères de production laitière que pour les LSCS car l'enregistrement de ces phénotypes est plus ancien.

Tableau 4. Synthèse du nombre de phénotypes, nombre de femelles, moyenne et écart-type pour les caractères en sélection pour les races Alpine et Saanen (source : fichier des performances, évaluation génétique janvier 2016)

	Alpine				Saanen			
	Nb phénotypes	**Nb femelles**	**Moyenne**	**Ecart-type**	**Nb phénotypes**	**Nb femelles**	**Moyenne**	**Ecart-type**
LAIT	4 655 038	1 679 740	818,73	253,83	3 519 470	1 291 684	836,01	265,68
MG			28,60	10,80			27,85	10,21
MP			25,21	8,22			24,99	8,07
TB			34,86	7,38			33,36	6,10
TP			30,83	3,36			30,01	2,93
AVP	243 401	243 401	3,14	1,00	160 141	160 141	3,29	1,16
PRM			5,82	1,41			6,28	1,34
AAR			4,47	1,49			4,83	1,67
PLA			6,43	1,03			6,23	1,14
ORT			3,60	0,92			4,03	0,86
LSCS	1 824 623	869 658	8,59	1,42	1 450 121	705 965	8,82	1,34

2.2.2. Effets de milieux ou d'environnements

En complément des mesures de production laitière, du pointage ou du contrôle du nombre de cellules somatiques, des paramètres sur l'animal et son milieu de vie sont enregistrés. Ces éléments sont pris en compte dans les modèles d'évaluation génétique pour corriger les performances des animaux des effets de milieux et mieux prédire la valeur génétique de tous les animaux. Les effets d'environnements enregistrés sont : le numéro d'élevage (autant de niveaux que d'élevage), l'année du contrôle, la parité de l'animal (1, 2 ou ≥ 3), l'âge à la mise bas, le mois de mise bas (12 niveaux), la région du contrôle (4 régions en France) et la durée du tarissement.

Ces effets sont combinés entre eux et inclus dans les modèles d'évaluation génétique officielle. Pour les caractères laitiers par exemple, ces effets sont :

1) *Effet troupeau :* il combine le numéro d'élevage avec l'année du contrôle et la parité de l'animal
2) *Âge à la mise bas :* il combine l'âge à la mise bas avec l'année du contrôle et la région du contrôle
3) *Mois de la mise bas :* il combine le mois de la mise bas avec l'année du contrôle et la région du contrôle
4) *Période du tarissement :* il combine la durée du tarissement avec l'année du contrôle et la région du contrôle

2.2.3. Pedigree

L'enregistrement des pedigrees conditionne en partie la réussite d'un schéma de sélection. En effet, les estimations des valeurs génétiques des animaux se basent sur le principe que deux individus apparentés partagent une partie de leur génome. L'absence de cette information pour un individu empêche de prédire correctement sa valeur génétique. Le pedigree contient 4 974 865 animaux (2 874 136 Alpine et 2 100 729 Saanen) nés entre 1914 et 2016 (Figure 4).

Avant les années 1970 (non montrés sur la Figure 4), seules 2 197 Alpine et 478 Saanen sont présents dans le pedigree. Après les années 1970, de plus en plus d'animaux sont enregistrés dans le pedigree chaque année pour atteindre aujourd'hui entre 150 000 et 200 000 enregistrements par an.

La proportion de parents inconnus totale (père et mère) s'élève à 29 % pour la race Alpine. Lorsque l'on décompose par parent, elle est de 40 % pour les pères et de 18 % pour les mères. Pour la race Saanen, la proportion de parents inconnus totale est légèrement supérieure (33 %). On obtient au total, 46 % de pères inconnus et 20 % de mères inconnues. La proportion de femelles avec généalogie (dont le père est connu) était de 52 % en 2005 (Carillier-Jacquin, 2015), a chuté jusqu'en 2013 (42 %) (Carillier-Jacquin, 2015 ; Palhière, 2015) pour ensuite rester stable (42% dans les fichiers de pedigree de 2016).

La baisse observée de la connaissance des généalogies des femelles a deux origines (Isabelle Palhière, communication personnelle). Le premier facteur est la taille des troupeaux qui ne cessent d'augmenter. Le deuxième facteur est l'utilisation plus importante de grands lots avec plusieurs boucs plutôt que d'utiliser des lots de lutte (1 seul bouc) qui est une pratique chronophage. En 2016, 69 497 inséminations animales (IA) ont été réalisées pour les races Alpine (42 992 IA) et Saanen (26 505 IA). Le taux de mise bas suite à l'IA est de 83,6 % pour la race Alpine et de 80 % pour la race Saanen. (Capgènes - Institut de l'élevage, 2017).

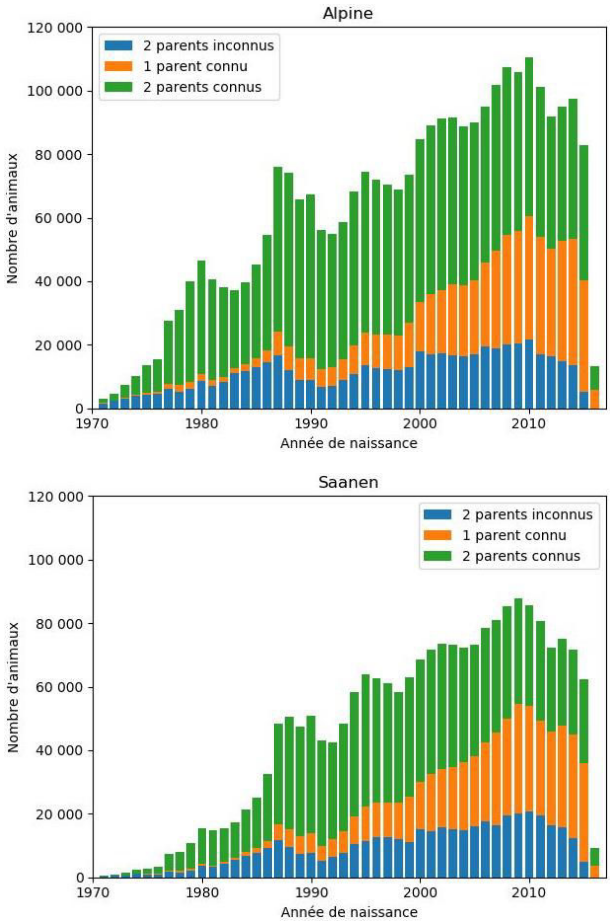


Figure 4. Enregistrement des pedigrees en fonction de l'année de naissance des animaux et de la connaissance des parents

3. Arrivée de la sélection génomique chez les caprins

Les avancées des outils moléculaires et des technologies nouvelles générations de ces dernières années permettent d'analyser plus finement le génome. Un consortium international

caprin (IGGC, www.goatgenome.org) s'est créé en 2010, dont l'objectif était de mutualiser les efforts de recherche pour établir une séquence de référence caprine. Cette séquence a été mise à disposition de la communauté scientifique en 2012 (Dong et al., 2013). Pour de nombreuses espèces d'élevage, la disponibilité d'une séquence de référence a permis le développement de puce à ADN comme en bovins (Fan et al., 2010), ovins (Fan et al., 2010 ; Auvray et al., 2011), porcins (Ramos et al., 2009 ; Fan et al., 2010), ou volailles (Fan et al., 2010 ; Groenen et al., 2011). Ces puces à ADN permettent de déterminer le génotype des animaux à certains points précis du génome. Aujourd'hui, ces puces utilisent des SNPs comme marqueurs et donnent une vue d'ensemble du génome. En 2011, une puce caprine 50K (Illumina GoatSNP50 BeadChip), contenant 53347 SNPs, est devenue disponible pour la communauté scientifique (Tosser-Klopp et al., 2014) permettant d'ouvrir de nouvelles perspectives dans le domaine de la sélection animale. L'information moléculaire dense de ces marqueurs SNP permet une estimation plus précise de la valeur génétique d'un individu dès la naissance. Les programmes d'évaluation génétique des reproducteurs doivent donc être revisités pour prendre en compte ces nouvelles informations.

La sélection génomique consiste dans un premier temps à génotyper et phénotyper un grand nombre d'individus (en général, et pour des raisons de coût, des mâles testés sur descendance) et à établir une relation statistique (décrite par une équation de prédiction) entre les génotypes aux marqueurs et les phénotypes. Les animaux génotypés et phénotypés sur lesquels se base l'équation de prédiction forment la «population de référence». Une fois la relation entre les génotypes aux marqueurs et les phénotypes établie, il est alors possible de prédire une valeur génomique pour les jeunes animaux (sans phénotypes), uniquement sur la base de leur génotype aux marqueurs SNP. Ces nouvelles avancées permettent ainsi de prédire la valeur génétique d'un animal dès sa naissance, avant même de connaître ses performances ou celles de ses descendants avec des niveaux de précision relativement élevés (mais généralement plus faibles qu'un testage sur descendance) et en tout cas plus grande que celle permise par l'information sur ascendance. Ceci peut entraîner potentiellement un raccourcissement de l'intervalle de génération et un gain de précision des valeurs génétiques estimées pour les femelles. La validité de cette prédiction repose toutefois sur la fiabilité de la relation entre performances et génotypes aux marqueurs SNP (Robert-Granie et al., 2011).

Les bovins laitiers ont été les premiers animaux d'élevage à bénéficier d'évaluations génomiques. C'est en effet pour ces races que les premières puces à ADN ont fait leur apparition (en 2007 avec la création d'une puce Illumina BovineSNP50 BeadChip, Matukumalli et al., 2009). L'arrivée de la sélection génomique a ainsi bouleversé l'organisation des schémas de sélection classiques: fin du testage sur descendance en 2010, utilisation de mâles reproducteurs à un âge plus précoce, diminution de l'intervalle de génération, augmentation du progrès génétique, … .

La réussite de la sélection génomique chez les bovins laitiers a questionné les autres filières et sa mise en œuvre a été étudiée dès 2012 chez les caprins laitiers.

3.1. Les grandes familles de méthodes des évaluations génomiques

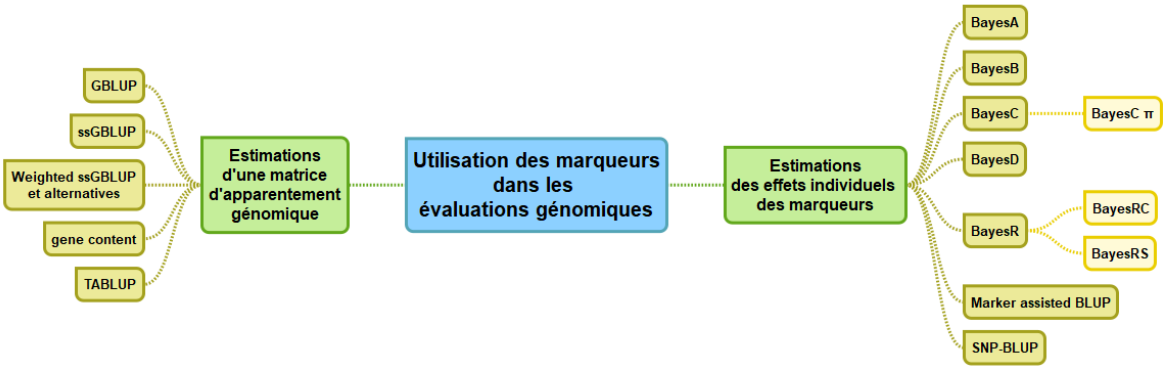


Figure 5. Mindmap des modèles d'évaluations génomiques selon l'utilisation des marqueurs

De nombreux modèles d'évaluation génomique ont été proposés dans la littérature. Actuellement, deux grandes familles de modèles existent (Figure 5). La première famille consiste à estimer les effets individuels des marqueurs SNP. Le modèle classiquement utilisé est :

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Xg} + \boldsymbol{e} \quad \text{[Modèle 2]}$$

où **y** est un vecteur de phénotypes corrigés pour les effets de milieux, **μ** est la moyenne des phénotypes, **X** est une matrice d'incidence des effets des SNPs (**g**) et **e** est le vecteur des effets résiduels. Une fois les effets des SNPs estimés, on peut prédire la valeur génétique d'un animal « i » uniquement à partir de son génotype :

$$GEBV_i = \sum_{j=1}^{m} x_{ij} g_j$$

où m correspond aux nombres de SNPs, x_{ij} est le génotype au marqueur j pour l'individu i et $g_j$est l'effet estimé du SNP j. Plusieurs modèles et méthodes ont été développés dans la littérature pour estimer les effets de SNP. Le modèle appelé SNP-BLUP suppose que les effets de marqueurs ($\boldsymbol{g}$) sont identiquement et indépendamment distribués de moyenne 0 et de variance σ_g^2 (Meuwissen et al., 2001 ; Koivula et al., 2012 ; Shen et al., 2013). Plusieurs variantes à ce modèle existent et sont basées sur des approches bayésiennes (BayesA, BayesB, BayesC, BayesR, ,…). Elles se distinguent par les hypothèses faites sur la distribution des effets de SNP et reposent sur des modèles hiérarchiques : on décrit par exemple la forme générale de la distribution d'un effet qui dépend d'un paramètre, par exemple une variance, qui provient elle-même d'une distribution générale des variances d'effets des SNPs, etc. Nous ne décrirons pas en détail chacun de ces modèles puisqu'ils n'ont pas été abordés en détail dans la thèse. Cependant, la performance de ces modèles en sélection génomique a été soulignée dans de nombreux articles lorsque le caractère étudié présente une architecture génétique non polygénique (Meuwissen et al., 2001; Gianola, 2013; Hayashi and Iwata, 2013).

L'utilisation de tels modèles suppose le génotypage et phénotypage de nombreux individus (1) pour avoir une estimation précise des effets des marqueurs et (2) seuls les individus ayant simultanément un phénotype et génotype contribuent à ces prédictions. Cela peut poser des problèmes pour les espèces pour lesquelles le caractère étudié ne peut être mesuré que sur une

seule partie des animaux (caractères laitiers, caractères de composition de la carcasse par exemple).

La deuxième famille de modèles d'évaluations génomiques (Figure 5) utilise les génotypes des animaux afin d'estimer l'apparentement entre pairs d'individus. Le BLUP génomique (GBLUP), développé par VanRaden, (2008), fait partie de cette famille et le modèle sous-jacent est identique au Modèle 1 présenté dans le paragraphe 2.1 (Chapitre 1). La différence entre le BLUP et le GBLUP est l'utilisation de la matrice de parenté génomique (**G**) en remplacement de la matrice de parenté basée sur le pedigree (**A**). Cette matrice **G** se calcule comme :

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{2 * \sum_{i=1}^{m}(p_i * (1 - p_i))}$$

$$\text{Avec } \boldsymbol{Z} = \boldsymbol{M} - \boldsymbol{P}$$

où m est le nombre total de SNP, p_i est la fréquence de l'allèle i, $\boldsymbol{M}$ est la matrice centrée des génotypes, **P** est une matrice où chaque colonne contient : $2 * p_i(1 - p_i)$. La matrice **G** mesure ainsi la ressemblance entre les individus d'un point de vue du génotype puisqu'elle estime la proportion moyenne d'allèles partagés par deux individus pondérée par leurs fréquences alléliques. L'apparentement mesuré par **G** est ainsi plus précis que celui estimé à partir du pedigree puisque l'information génomique permet de capter l'aléa de méiose.

L'avantage de la méthode GBLUP est que son application est simple et demande peu de modifications par rapport à un BLUP classique, mais nécessite beaucoup d'individus génotypés pour avoir une bonne estimation des apparentements entre individus. Pour les espèces laitières, des pseudo-performances étaient construites pour les mâles puisqu'en général ils étaient les seuls à être génotypés. Enfin, l'utilisation de tels modèles engendre souvent une grande perte d'information et induit des biais (Patry and Ducrocq, 2011). Un moyen de contourner ces problèmes serait d'utiliser des évaluations génomiques utilisant toutes les informations simultanément (phénotypes, pedigrees et génotypes). Cette méthode, appelée single-step GBLUP (ssGBLUP), a été développé par Legarra et al., (2009). Elle est basée sur la construction d'une matrice de parenté ($\boldsymbol{H}$) combinant les informations des génotypes et de pedigree pour l'ensemble des animaux, génotypés ou non. La matrice $\boldsymbol{H}$ est définie sous la forme suivante :

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{A_{11} + A_{12} + A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21}} & \boldsymbol{A_{12}A_{22}^{-1}G} \\ \boldsymbol{GA_{22}^{-1}A_{21}} & \boldsymbol{G} \end{bmatrix}$$

où les indices 1 et 2 représentent les animaux non génotypés et génotypés respectivement. La matrice **A** correspond à la matrice de parenté basée sur le pedigree et la matrice **G** est la matrice de parenté génomique. L'inverse de la matrice **H**, utilisée dans les équations du modèle mixte a une expression relativement simple :

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

où $\boldsymbol{A^{-1}}$ est l'inverse de la matrice de parenté, $\boldsymbol{A_{22}^{-1}}$ est l'inverse de la matrice de parenté pour le bloc des animaux génotypés et $\boldsymbol{G^{-1}}$ est l'inverse de la matrice de parenté génomique légèrement différente de celle présentée par VanRaden, (2008). En effet pour le ssGBLUP, la matrice **G** est calculée comme :

$$\boldsymbol{G} = 0{,}95 * \frac{\boldsymbol{ZZ'}}{\sum_{i=1}^{m} p_i * (1 - p_i)} + 0{,}05 * \boldsymbol{A_{22}}$$

car les matrices **G** et **A** doivent rester compatibles entre elles (Legarra et al., 2014b). Cette modification permet également de rendre la matrice **H** inversible. L'avantage du ssGBLUP est qu'il est possible d'obtenir des GEBV pour tous les animaux (qu'ils soient ou non génotypés, et/ou qu'ils aient ou non des phénotypes).

3.2. Critères d'évaluation de la qualité des méthodes d'évaluation génomique

3.2.1. La précision des évaluations génomiques

Le critère le plus couramment utilisé dans la littérature pour comparer les différentes méthodes d'évaluations génomiques est la précision des évaluations génomiques. Il est calculé comme le coefficient de corrélation de Pearson entre les valeurs génétiques « vraies » et les valeurs génétiques prédites (EBV ou GEBV). La valeur génétique « vraie » d'un animal est en général inconnue, on utilise alors une approximation de cette valeur à l'aide du calcul des Daughter Yield Deviation (DYD) (VanRaden and Wiggans, 1991).

Les DYD sont calculés à partir des Yield Deviation (YD) des femelles qui sont des performances brutes corrigées des effets fixes, des effets aléatoires autres que génétiques. Les YD s'expriment comme (Szyda et al., 2008) :

$$\boldsymbol{YD} = \boldsymbol{y} - (\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{p}})$$

où **y** est un vecteur de phénotypes, $\widehat{\boldsymbol{\beta}}$ un vecteur représentant l'estimation des effets fixes, et $\widehat{\boldsymbol{p}}$ l'estimation des effets d'environnement permanent. Les DYD sont alors une moyenne pondérée des YD de toutes les filles. Pour un mâle, son DYD est calculé comme :

$$DYD = \frac{\sum_{i=1}^{n} \left(\widehat{YD}_i - 0{,}5 * g_{mère}\right) * w_i}{\sum_{i=1}^{n} w_i}$$

où YD_i est le YD de la femelle i, $g_{mère}$ est l'effet polygénique de la mère de la femelle i et w_i un poids associé au nombre de lactations de la femelle. Les DYD sont des estimations précises des valeurs génétiques « vraies » des animaux car ils moyennent un nombre important de performances (les YD).

La méthode d'évaluation génomique la plus intéressante sera celle pour laquelle la corrélation entre les valeurs génétiques « vraies » et les valeurs génétiques prédites sera la plus proche de 1. Dans notre étude, les valeurs génétiques « vraies » ont été des DYD estimés à partir des évaluations génétiques officielles de janvier 2016. La précision des évaluations génomiques est calculée à l'aide d'une méthode de validation croisée appelée « forward validation » (Figure 6). Elle consiste à diviser la population de référence en 2 sous-populations en fonction de l'année de naissance des animaux. La première sous-population, appelée population d'apprentissage, est constituée des animaux les plus anciens. Pour lesquels phénotypes (performances des filles), pedigrees et génotypes sont connus et utilisés. La seconde sous population, appelée population de validation, contient les animaux les plus jeunes pour lesquels seuls génotypes et ascendance (pedigree) sont connus. La population de validation simule la population actuelle, nous excluons donc toutes les données qui concernent la descendance de ces animaux.

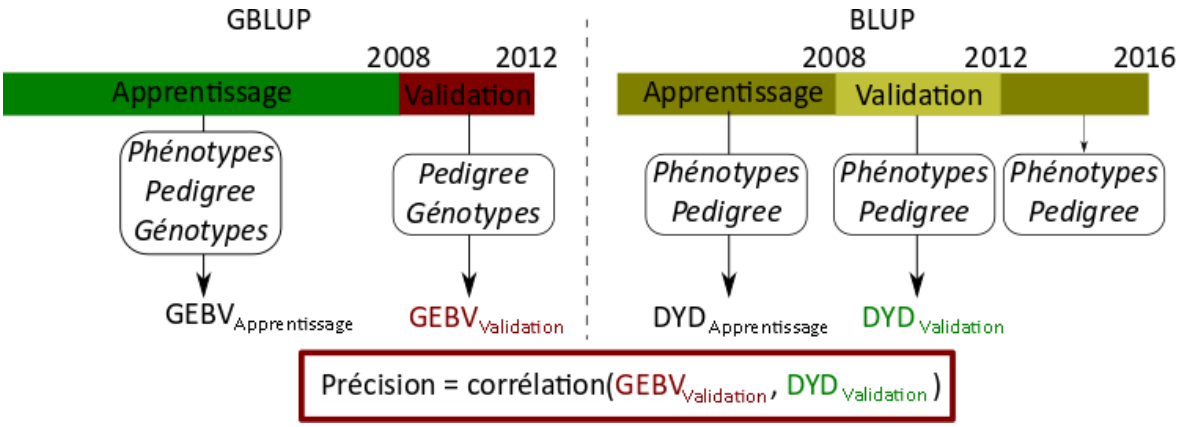


Figure 6. Principe de la validation croisée (forward validation) utilisée pour les évaluations génomiques caprines mises en œuvre dans la thèse

Une fois les évaluations génomiques réalisées, on peut calculer une précision génomique pour chacune des 2 sous-populations (apprentissage et validation). Pour l'ensemble des travaux de ma thèse, les précisions des évaluations sur la population de validation ont été comparées.

3.2.2. Les biais des évaluations génomiques

La sélection des animaux mis à la reproduction induit des biais dans les évaluations génomiques (Vitezica et al., 2011 ; Gowane et al., 2018). En effet, les modèles d'évaluations génomiques font l'hypothèse que les animaux génotypés ne sont pas sélectionnés (Hayes et al., 2009 c), ce qui est loin d'être le cas en pratique. De plus, le ssGBLUP repose sur le postulat que la matrice de parenté $\boldsymbol{A}$ et la matrice de parenté génomique $\boldsymbol{G}$ sont compatibles. Cette compatibilité se traduit par une égalité des valeurs génétiques moyennes et de la variance génétique de la population de base pour les matrices $\boldsymbol{A}$ et $\boldsymbol{G}$ (Garcia-Baccino et al., 2017)**.** Dans le cas inverse, ce sont des sources potentielles de biais pour les évaluations génomiques (Christensen, 2012). La connaissance des généalogies, le nombre de générations connues pour la matrice $\boldsymbol{A}$ (Misztal, 2017), le nombre de SNPs, la qualité des SNPs et la fréquence des SNPs pour la matrice $\boldsymbol{G}$ (Misztal, 2017) sont autant de paramètres qui pourront réduire ou augmenter la compatibilité des matrices $\boldsymbol{A}$ et $\boldsymbol{G}$. L'analyse des biais des évaluations génomiques est donc importante car une estimation des valeurs génétiques non biaisée est nécessaire pour suivre le progrès génétique et pour comparer les valeurs génétiques des animaux entre générations (Henderson et al., 1959).

En complément des précisions des évaluations, nous avons analysé les biais des évaluations génomiques. Le biais est calculé par une régression des valeurs génétiques vraies (DYD) en fonction des GEBV à l'aide de l'équation suivante :

$$DYD = a * GEBV + b$$

On recherche des valeurs génétiques prédites et des vraies valeurs génétiques identiques. Dans ce cas, les coefficients de la régression entre les DYD et les GEBV sont de 1 pour a et 0 pour b. Lorsque la pente est supérieure à 1, les GEBV sont sous-estimées par rapport au DYD. À l'inverse, une pente inférieure à 1 signifie que les GEBV sont surestimées par rapport au DYD.

3.3. Les génotypes disponibles en race caprine Alpine et Saanen

3.3.1. Les génotypes de la puce Illumina GoatSNP50 BeadChip

Le projet européen 3SR (http://www.3srbreeding.eu/, 2010-2012) et le projet national PhénoFinlait (http://idele.fr/linstitut-de-lelevage/sites-partenaires/phenofinlait.html, 2008-2013) ont financé les premiers génotypages en race Alpine et Saanen. Le choix des animaux à

génotyper a été réalisé au regard de deux objectifs : une étude de primo détection de QTL et une étude sur l'intérêt d'une sélection génomique en caprin. Un dispositif QTL de 20 familles de pères (11 Alpins et 9 Saanen) comprenant au minimum 100 filles par père nées entre 2008 et 2009 a été génotypé, ainsi que les mâles testés sur descendance nés à partir de 1998 jusqu'en 2011. De 2012 à aujourd'hui, tous les mâles testés sur descendance sont systématiquement génotypés (soit 30 à 40 mâles dans chaque race par an).

En 2012, la population de référence comprenait au total 2 955 animaux (905 mâles et 2050 femelles (Figure 7 et Figure 8) : 1 749 animaux de race Alpine (512 mâles et 1 237 femelles) et 1206 animaux de race Saanen (393 mâles et 813 femelles).

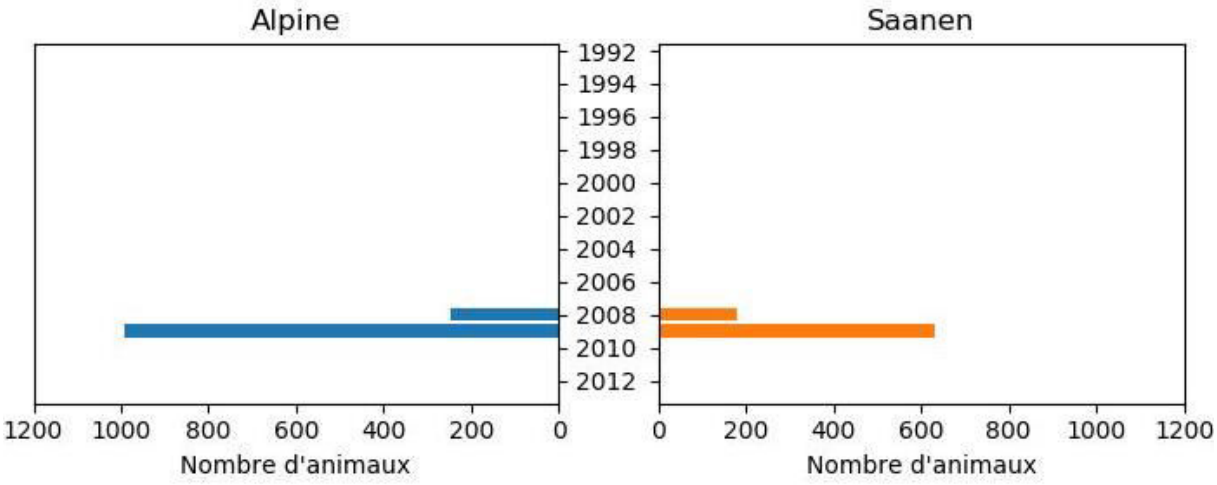


Figure 7. Nombre de femelles génotypées avec la puce 50K par année de naissance pour les races Alpine et Saanen

Les sources d'erreurs de génotypages restent nombreuses (ADN de mauvaise qualité,…) (Forneris et al., 2015), c'est pourquoi un contrôle qualité des données (QC) est réalisé afin d'éliminer les animaux ou les SNPs de faible qualité. Pour un SNP bi-allélique (A et B), on comptabilise 3 génotypes différents : AA, AB et BB. S'il y a une ambiguïté sur la détection, le génotype est déclaré comme manquant (Huentelman et al., 2005). Un premier filtre appelé Minor Allele Frequency (MAF) consiste à supprimer les SNPs dont la fréquence de l'allèle le plus rare est inférieure à un seuil (fréquence inférieure à 1 %). Une fréquence trop faible peut conduire à des estimations peu précises des effets des SNPs (Wray et al., 2013). Le deuxième filtre est le SNP Call Rate (CR) et consiste à calculer le ratio entre le nombre d'animaux avec un génotype connu (AA, AB ou BB) sur le nombre total d'animaux pour chaque SNP. Un SNP CR trop faible indique un problème de génotypage pour ce SNP en particulier. Dans nos analyses, nous avons éliminé les SNPs avec un SNP CR inférieur à 95 %. Un test d'équilibre de Hardy-Weinberg est réalisé sur chacun des SNPs. Ce test compare la fréquence des génotypes observés avec la fréquence des génotypes prédits selon la loi d'Hardy-Weinberg. Un écart de fréquences peut avoir plusieurs causes : erreurs de génotypages, une stratification de la population. Les SNPs qui ne respectent pas l'équilibre de Hardy-Weinberg sont éliminés des analyses. Dans nos analyses, tous les SNPs avec une statistique de test pour le test de Hardy-Weinberg supérieur à 24 ont été supprimés. Le dernier filtre appliqué est l'animal CR et consiste à calculer la proportion de SNPs dont le génotype est connu (AA, AB ou BB) sur le nombre total de SNP pour chaque animal. Si l'animal CR est trop faible, cela suggère que. le génotypage de cet animal devient incertain. Nous avons éliminé les animaux avec un CR inférieur à 90 %.

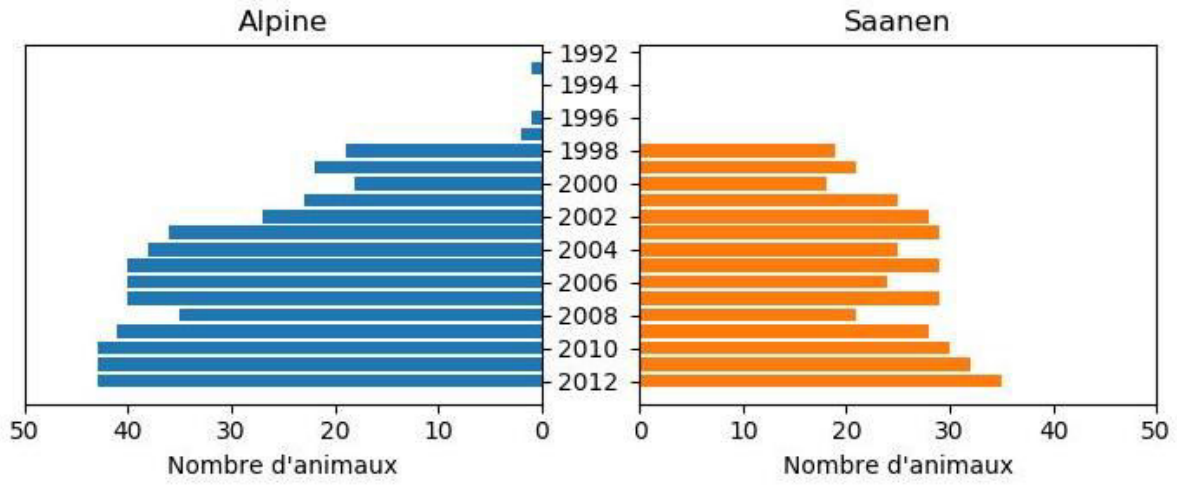


Figure 8. Nombre de mâles génotypés avec la puce 50K par année de naissance pour les races Alpine et Saanen

3.3.2. Autres génotypes disponibles en race caprine Alpine et Saanen

À partir des années 80, les futures mères à boucs ainsi que les mâles candidats à la sélection ont été génotypés pour le gène majeur de la *caséine* α_{s1} (Grosclaude et al., 1987). Ce gène affecte particulièrement le taux protéique du lait pour les races caprines (Grosclaude et al., 1987). Ce gène est polymorphe puisque 6 allèles différents sont observés dans les populations caprines françaises. Les allèles A, B et C sont associés à une forte synthèse de *caséine* α_{s1}, les allèles E et F sont associés à une synthèse intermédiaire. On peut observer une absence de synthèse de *caséine* α_{s1} avec l'allèle O (Grosclaude et al., 1987 ; Carillier-Jacquin et al., 2016). L'ensemble des combinaisons donne 21 génotypes possibles, mais ils ne sont pas tous observés dans les populations Alpine et Saanen. Les génotypes CO et OO sont absents des 2 races étudiées. Pour la sélection, les accouplements sont raisonnés en fonction du génotypage pour augmenter la fréquence des allèles forts au sein de la population. La fréquence des allèles faibles était élevée principalement dans la race Saanen (Piacère et al., 1997).

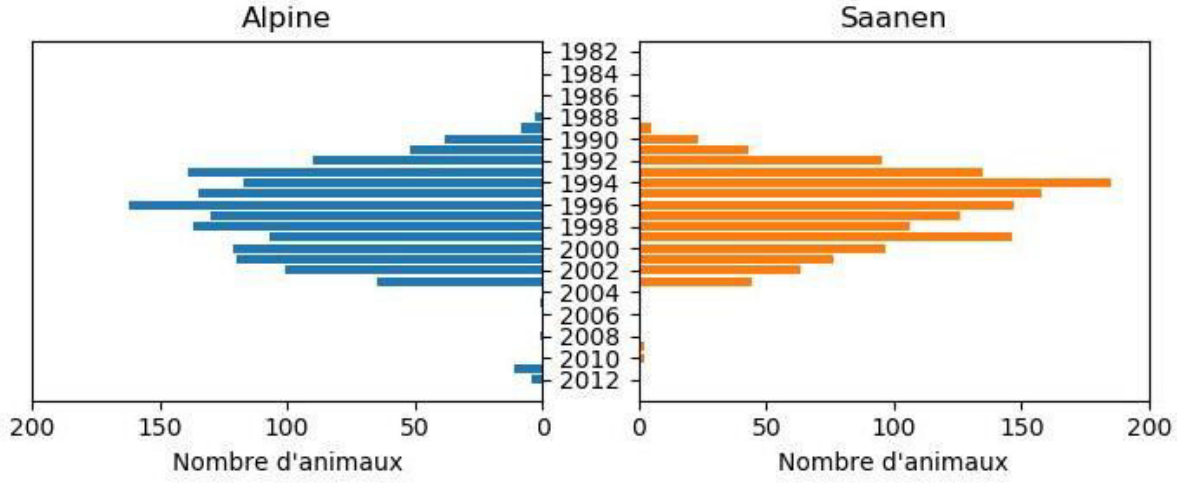


Figure 9. Nombre de génotypages caséine α_{s1} selon la race et l'année de naissance des animaux pour les femelles

Depuis les années 1990, en moyenne 100 femelles sont génotypées chaque année jusqu'en 2003 (Figure 9). Ces femelles font partie du noyau de sélection et sont destinées à devenir des mères à boucs. Depuis les années 2003, le génotypage des femelles s'est quasiment stoppé, seules quelques femelles sont génotypées (16 en Alpine et 6 en Saanen depuis 2008). Le même

effort de génotypages a été fait pour les mâles (Figure 10). Quelques animaux ont été génotypés avant les années 1985 (18 Alpine et 3 Saanen). Ensuite, tous les animaux entrés en station ont été génotypés (entre 100 et 200 génotypages par an). À partir de 2007, seuls les animaux candidats à la sélection sont génotypés en race Alpine (environ 40 génotypages par an). Les génotypages en race Saanen continuent, car, les fréquences des allèles avec une faible synthèse de la *caséine* α_{s1} sont plus importantes qu'en race Alpine. Au total, 7 202 génotypages *caséine* α_{s1} sont disponibles en 2012 pour les deux races.

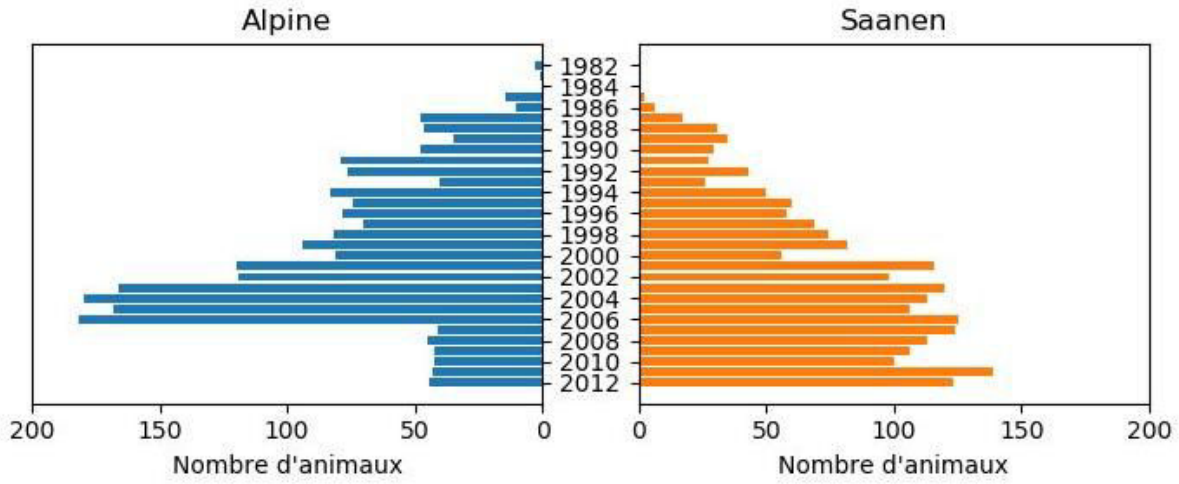


Figure 10. Nombre de génotypages caséine α_{s1} selon la race et l'année de naissance des animaux pour les mâles

Les effets des génotypes *caséine* α_{s1} sur le TP ont pu être estimés au cours de l'étude de (Carillier-Jacquin et al., 2016). Ils sont présentés dans le Tableau 5 pour la race Alpine et pour la race Saanen. Les génotypes AA, AB, AC, BB et BC sont les génotypes avec les effets les plus importants sur le TP (entre 1,6 et 3,7 g/kg). Ces génotypes sont principalement recherchés dans la population caprine. Avec un effet entre 0.5 et 1 g/kg, les allèles AE, AF, BE, BF, CE et CF ont un effet intermédiaire sur le TP. Enfin, les allèles EE, EF ont un effet faible voir négatif sur le TP (entre -0,9 et 0,2 g/kg).

Tableau 5. Effet des génotypes caséine α_{s1} sur le TP pour les populations Alpine et Saanen

Effet sur le TP	*Génotype caséine α_{s1}*	Saanen (g/kg)	Alpine (g/kg)
Positif fort	*AA*	2.2	2.5
	AB	2.5	1.7
	AC	1.6	*
	BB	2.4	*
	BC	3.7	*
	CC	*	*
Intermédiaire	*AE*	1.0	1.0
	AF	0.5	0.7
	AO	*	*
	BE	0.5	1.1
	BF	0.6	0.9
	BO	*	*
	CE	1.0	*
	CF	0.6	*
	CO	*	*
Faible	*EE*	−0.7	0.2
	EF	−0.9	−0.4
	EO	*	*
	FF	*	*

***** Les effets n'ont pu être estimés en raison d'un trop faible ou d'une absence d'animaux ayant ces génotypes

Les fréquences des génotypes *caséine α_{s1}* sont différentes entre les deux races (Figure 11). On observe que les génotypes les plus fréquents pour la race Alpine sont les génotypes AA (39 %), AE (24 %), AB (11 %), AF (10 %) et AC (6 %). Ces 5 génotypes représentent 90 % des génotypes observés et sont associés à des effets positifs forts sur le TP (AA, AB et AC) ou à des effets positifs intermédiaires (AE et AF). Les autres génotypes sont minoritaires, avec des fréquences inférieures à 2 %. Pour la race Saanen, les génotypes les plus importants sont AE (42 %), EE (24 %), EF (11 %), AA (6 %) et AF (5 %). La fréquence cumulée de ces 5 génotypes atteint 88 %. Alors que le génotype AA est associé à un fort effet positif sur le TP, les génotypes AE, EE et AF sont associés à un effet positif intermédiaire et le génotype EE est associé à un effet négatif sur le TP.

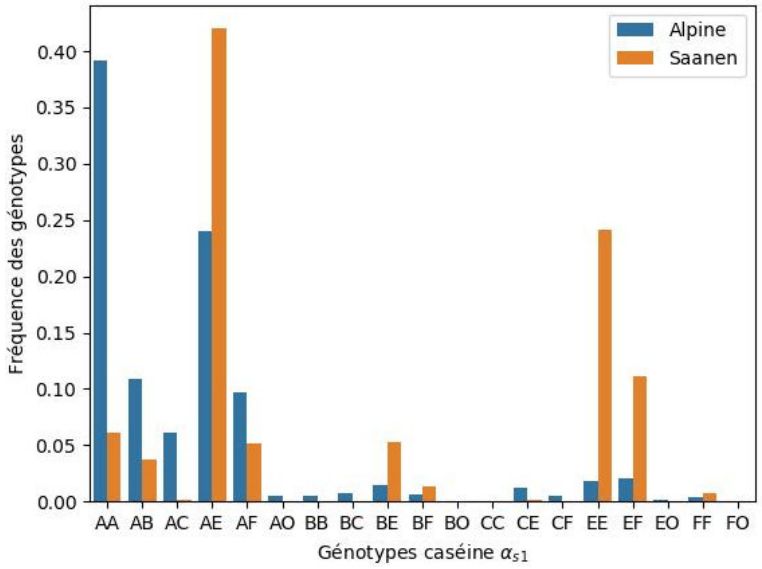


Figure 11. Fréquence des génotypes caséine α_{s1} selon la race

Les animaux génotypés pour le gène de la *caséine* α_{s1} ne sont pas nécessairement génotypés avec la puce 50K. La Figure 12 montre le nombre d'animaux pour lesquels on dispose à la fois de l'information sur la puce 50K et sur le génotype de la *caséine* α_{s1}. Les animaux qui sont génotypés avec la 50K et pour la *caséine* α_{s1} (510 Alpine et 393 Saanen) sont des mâles. Seuls 2 mâles Alpins génotypés sur la puce 50K ne sont pas génotypés pour le gène de la caséine α_{s1}.

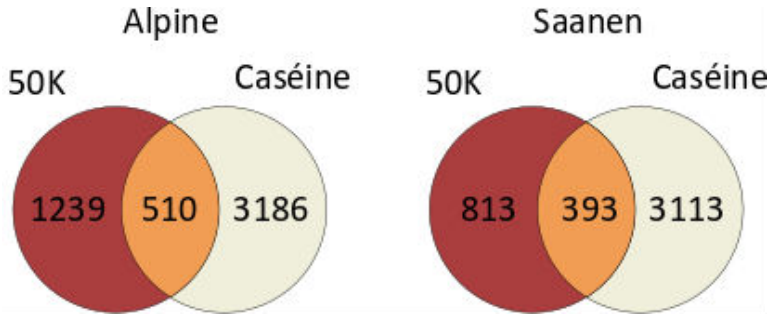


Figure 12. Répartition des génotypes 50K et caséine α_{s1} pour les analyses multirace (Alpin + Saanen), Alpine et Saanen

Le gène Diacylglycerol O-Acyltransferase 1 (*DGAT1*) est connu pour affecter la MG et le TB du lait (Martin et al., 2017). Des génotypes pour cette mutation sont également disponibles pour les races Alpine et Saanen. Deux mutations existent dans ce gène majeur et provoquent des modifications de la protéine : la mutation R251L et la mutation R396W. La mutation R251L est caractérisée par une substitution d'une arginine (R) en leucine (L) pour le 251^{e} acide aminé. Cette substitution est provoquée par la présence d'un SNP dans la séquence nucléotidique avec la substitution d'un nucléotide G par T. Pour la mutation R396W, on observe une substitution de l'arginine (R) en tryptophane (W) sur le 396^{e} acide aminé de la protéine. Cette fois-ci, un allèle C est substitué par l'allèle T. L'essentiel des génotypes femelles disponibles provient d'animaux nés entre 2008 et 2009 (Figure 13), soit au total 1315 Alpine et 937 Saanen, correspondant essentiellement aux femelles impliquées dans le dispositif QTL (PhénoFinLait).

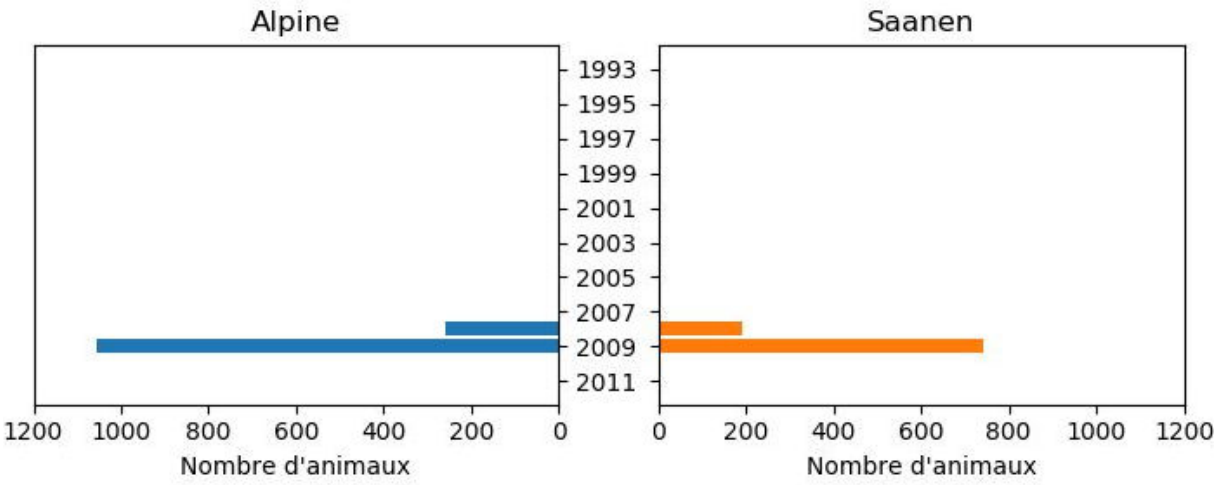


Figure 13. Nombre de génotypages DGAT1 des femelles selon la race et l'année de naissance

Pour les mâles, les génotypes ont été réalisés depuis plus longtemps (Figure 14). Les premiers animaux génotypés pour le gène *DGAT1* sont nés en 1993 pour la race Alpine et 1998 pour la race Saanen. En moyenne, 30 génotypages sont réalisés dans la race Alpine et 20 pour la race Saanen par année de naissance. Au total, 442 et 333 génotypages sont disponibles pour la race Alpine et Saanen respectivement.

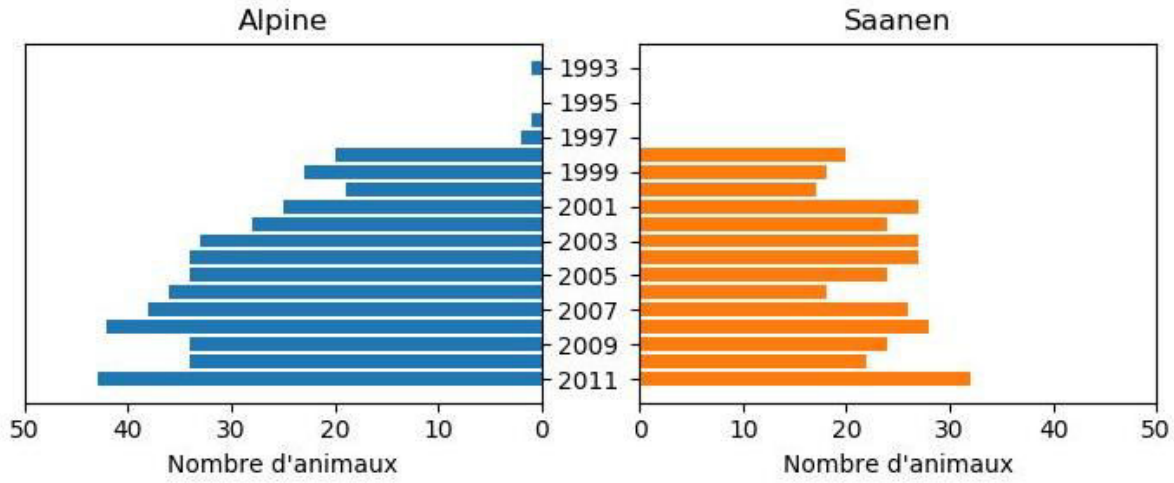


Figure 14. Nombre de génotypages DGAT1 des mâles selon la race et l'année de naissance

Les 2 mutations pour *DGAT1* (R251L et R396W) étant des SNPs, 3 génotypes sont possibles pour chacune des mutations (Figure 15). Pour la mutation R251L, 99 % et 92,4 % des génotypes sont G/G pour la race Alpine et la race Saanen respectivement. Les autres génotypes ont une fréquence très faible pour la race Alpine (0,9 % pour G/T et 0,1 % pour T/T) et Saanen (7,3 % pour G/T et 0,3 % pour T/T). Pour la mutation R396W, le génotype C/C prédomine dans la population (85,7 % pour la race Alpine et 76,0 % pour la race Saanen). Comme pour la mutation R251L, les autres génotypes (C/T et T/T) sont minoritaires, 13,7 % et 0,6 % pour la race Alpine et 22,8 % et 1,2 % pour la race Saanen respectivement.

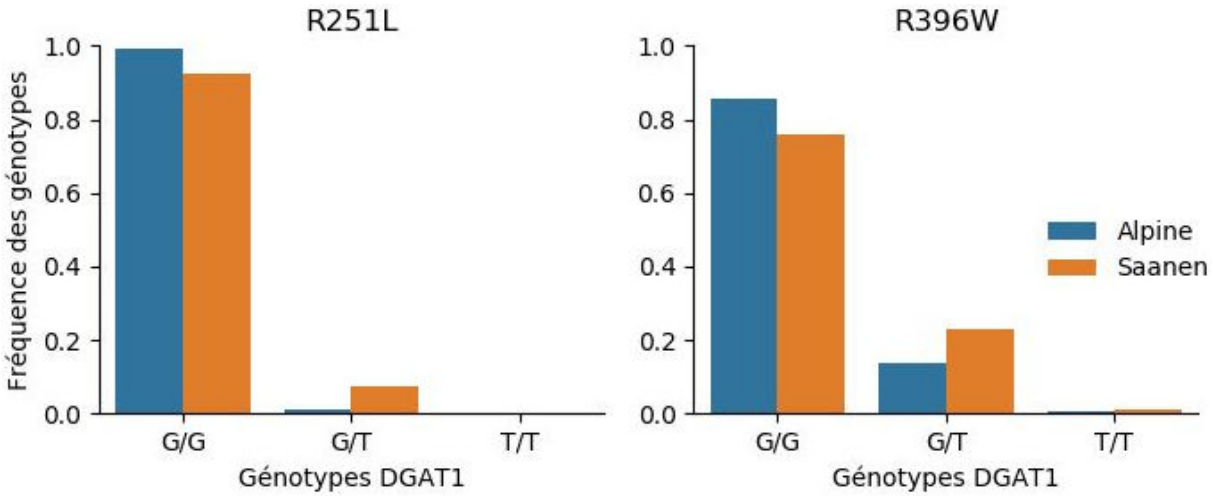


Figure 15. Fréquence des génotypes DGAT1 selon la mutation (R251L et R396W) et selon la race

Tous les animaux génotypés pour les mutations *DGAT1* ne sont pas génotypés avec la puce 50K (Figure 16). On observe que la proportion d'animaux génotypés pour la puce 50K et pour les mutations *DGAT1* est bien supérieure au cas de la *caséine* α_{s1}. Pour *DGAT1*, 91 % et 94 % des animaux génotypés pour la puce 50K sont également génotypés pour les mutations *DGAT1* en race Alpine et Saanen respectivement.

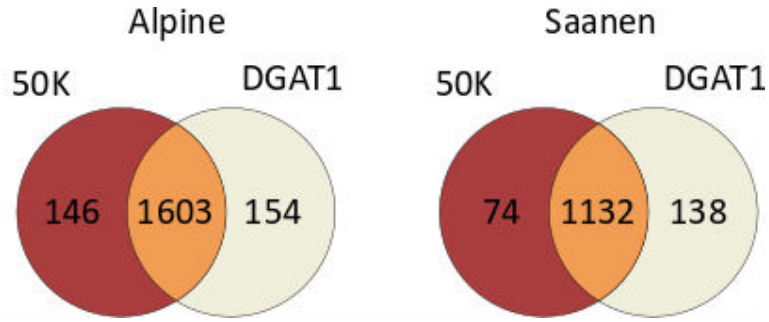


Figure 16. . Répartition des génotypes 50K et DGAT1 pour les analyses multirace, Alpine et Saanen

3.4. Facteurs influençant la précision des évaluations génomiques

Les précisions des évaluations génomiques sont influencées par différents facteurs intrinsèquement liés à la population étudiée (Meuwissen et al., 2001 ; Goddard and Hayes, 2007 ; Daetwyler et al., 2008 ; Hayes et al., 2009 b). Ces facteurs peuvent être 1) le déséquilibre de liaison (LD) entre les marqueurs, 2) le nombre d'animaux de la population d'apprentissage, 3) l'apparentement entre population de référence et population des candidats ou encore 4) l'architecture génétique du caractère (Goddard and Hayes, 2007; Erbe et al., 2012).

Certains de ces facteurs ont été étudiés en détail pour les races Alpine et Saanen par Carillier-Jacquin, (2015). Leurs analyses ont porté sur une population de 2 810 animaux (1 645 Alpine et de 1 165 Saanen nés entre 1993 et 2011) (Tableau 6), scindée en 3 populations :

- La population d'apprentissage contient 430 animaux mâles (236 Alpine et 194 Saanen nés entre 1993 et 2005). Ces mâles étaient génotypés et avaient des filles phénotypées.
- La population de validation contient 247 animaux mâles (148 Alpine et 99 Saanen) nés entre 2006 et 2009, et 420 femelles (239 Alpine et 629 Saanen), tous génotypés. Cette population a été utilisée pour calculer la précision des évaluations génomiques. Ces animaux avaient des filles avec phénotypes, mais ils ont été utilisés uniquement pour l'estimation des DYD des mâles.

- La population des candidats contient 148 animaux (86 Alpine et 62 Saanen nés entre 2010-2011). Au moment de l'étude de Carillier-Jacquin, (2015), ces animaux n'avaient aucune fille. Cette population a été utilisée pour calculer la précision des évaluations génomiques théorique. Ces précisions ont été comparées à la précision sur ascendance pour évaluer l'intérêt d'une sélection génomique en caprins.

L'ensemble des animaux de la population d'apprentissage et de validation constituait la population de référence.

Tableau 6. Composition de la population de référence caprine des analyses de Carillier-Jacquin, (2015)

	Mâles			Femelles		
	Total	**Alpine**	**Saanen**	**Total**	**Alpine**	**Saanen**
Apprentissage	430	236	194	0	0	0
Validation	247	148	99	420	239	181
Candidats	148	86	62	1565	936	629
Total	825	470	355	1985	1175	810

3.4.1. La taille de la population de référence

La taille de la population de référence peut influencer la précision des évaluations génomiques. À partir de données simulées, Meuwissen, (2009) a montré que la précision des évaluations augmentait avec une taille de population d'apprentissage plus importante, quelle que soit la densité en SNPs (Tableau 7). La densité en marqueurs est calculée en fonction de la taille efficace de la population (Ne) et de la taille du génome (en Morgan M). En augmentant la taille de la population de référence de 500 à 2000 animaux, la précision des évaluations génomique avec un GBLUP est passée de 0,73 à 0,88 pour une densité en SNP de 20 Ne/M. Des tendances similaires étaient observées avec d'autres méthodes, en particulier l'approche BayesB.

Tableau 7. Précisions du GBLUP et du BayesB en fonction de la densité en SNPs et de la taille de la population d'apprentissage (Meuwissen, 2009)

Densité en SNPs (Ne/M)	**Taille population d'apprentissage**	**GBLUP**	**BayesB**
1	500	0,66	0,70
1	1000	0,72	0,73
1	2000	0,76	0,77
20	500	0,73	0,83
20	1000	0,82	0,88
20	2000	0,88	0,93

Sur des données réelles, les mêmes conclusions ont été constatées. Par exemple, dans le cadre du consortium bovin EuroGenomics, qui a permis de constituer une population de référence Holstein européenne (Lund et al., 2011) de grande taille, les précisions des évaluations génomiques sont 9 % plus précises en utilisant une population de référence européenne contenant 12 078 animaux, qu'en utilisant une population de référence nationale française avec 3 071 animaux.

Bien qu'à ce jour la population de référence caprine française soit la plus grande à l'échelle mondiale (Carillier-Jacquin, 2015), elle reste de taille limitée. Les premières études de

faisabilité de mise en œuvre d'une sélection génomique en caprins ont été réalisées sur une population de référence constituée de 825 mâles et de 1 985 femelles génotypées pour un peu plus de 50 000 SNPs. La précision moyenne pour les caractères de quantités laitières (LAIT, MG et MP) était de 0,40 en moyenne. Les précisions génomiques sont moindres que celles obtenues chez les bovins laitiers (de 0,50 à 0,57 pour les caractères de quantités laitières pour (Hayes et al., 2009a ; Erbe et al., 2012)). Ces précisions plus faibles s'expliquent par la taille limitée de la population de référence caprine, mais également en raison d'autres critères qui sont détaillés dans la suite de ce document. Afin d'améliorer les précisions des évaluations génomiques, la filière caprine poursuit ses efforts en génotypant 30 à 40 mâles chaque année dans chacune des races pour accroître la taille de la population de référence.

3.4.2. Le déséquilibre de liaison (LD)

Le LD est défini comme une association préférentielle entre 2 allèles à des positions différentes (loci) sur le génome. En équilibre de liaison, les fréquences alléliques dans la population pour les différents haplotypes (groupe d'allèles à différents loci) seront les mêmes (25 % pour chaque haplotype) (Tableau 8). Tout écart de cet équilibre de liaison est le signe d'un LD (Tableau 8). Dans le cas où un LD existe entre un marqueur (SNP) et un QTL, on s'attend à ce que le SNP capte une partie de l'information du QTL, les évaluations génomiques seront alors plus précises si le LD entre SNPs et QTL est fort (Calus et al., 2008; Solberg et al., 2008; Hayes and Goddard, 2010).

Tableau 8. Exemple de fréquence d'haplotype pour deux loci en équilibre de liaison ou en déséquilibre de liaison

		Equilibre de liaison Allèle au locus 2		**Déséquilibre de liaison Allèle au locus 2**	
		B	**b**	**B**	**b**
Allèle au locus 1	**A**	0,25	0,25	0,10	0,45
	a	0,25	0,25	0,40	0,05

Il existe plusieurs approches pour estimer le LD entre marqueurs. Une de ces mesures, appelée r^2 (Rogers and Huff, 2009), est bien adaptée au cas des populations animales puisqu'elle ne nécessite pas de connaitre les phases alléliques. Cette mesure correspond à un calcul de corrélation entre génotypes à 2 loci donnés pour tous les individus, elle est calculée comme :

$$r^2 = \frac{\left(cov(g_i, g_j)\right)^2}{var(g_i) * var(g_j)}$$

où g_i et g_j sont les génotypes (codé 0,1 ou 2) pour le SNP i et j. En pratique, sont ensuite construits des intervalles de 20 kb pour estimer un r^2 moyen (Figure 17). Deux loci en équilibre de liaison auront un r^2 égal à 0 tandis que deux loci en LD total auront un r^2 proche de 1.

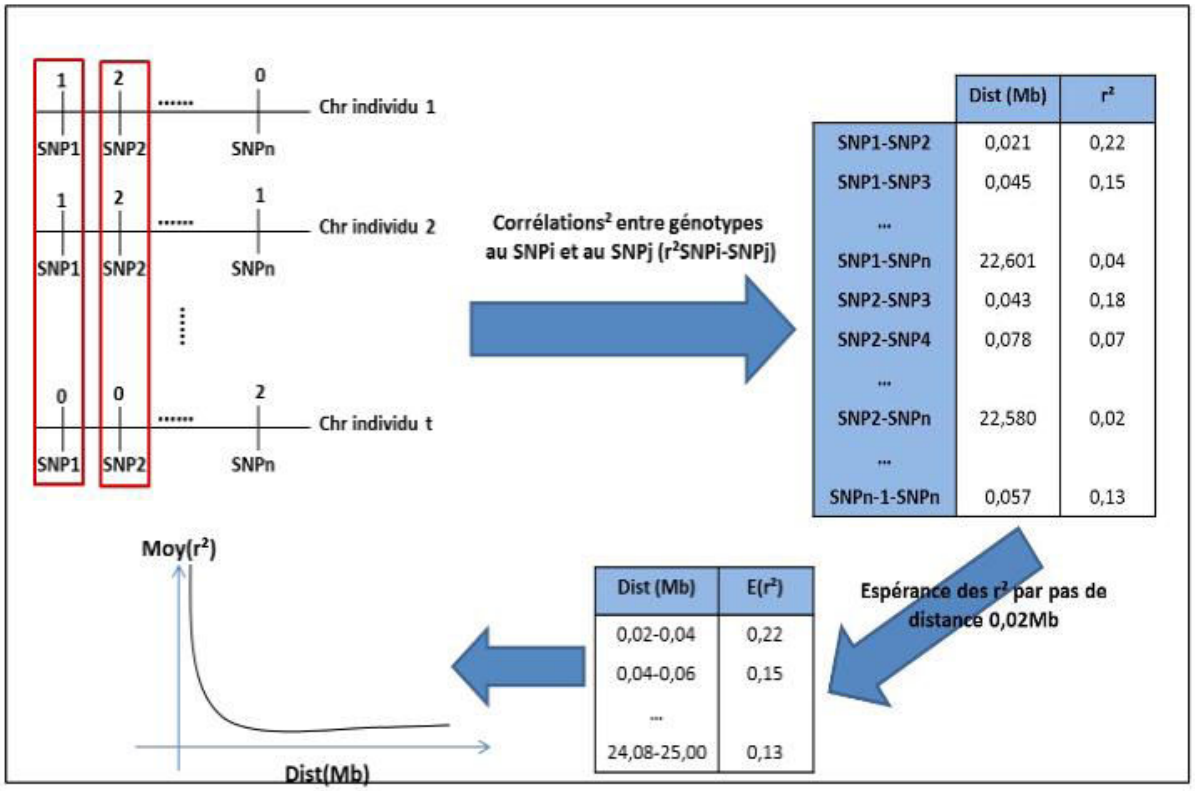


Figure 17. Calcul du déséquilibre de liaison avec la méthode de Roger et Huff (2009) (source : Carillier-Jacquin, 2015)

En races Alpine et Saanen, le niveau moyen de LD obtenu pour des marqueurs distants de 50 kb est de 0,17 sur l'ensemble de la population de référence (Carillier et al., 2013). La distance de 50 kb correspond à la distance moyenne entre 2 SNPs sur la puce 50K. En comparaison, cette valeur est proche de celles estimées en races caprines canadiennes (0,14 en Alpine et 0,15 en Saanen (Brito et al., 2015)), en bovins allaitants (entre 0,13 pour la race Brahman et 0,23 pour la race Hereford (Porto-Neto et al., 2014)) et en ovins laitiers (entre 0,13 et 0,14 (Baloche et al., 2014)). Elle est, par contre, inférieure à celle observée en races bovines laitières (entre 0,18 et 0,23 (de Roos et al., 2008)) ou en porcin (LD supérieur à 0,50 (Wang et al., 2013 ; Grossi et al., 2017).

3.4.3. L'apparentement entre la population d'apprentissage et la population de validation

Les évaluations génétiques reposent sur l'apparentement entre animaux afin de prédire les (G)EBV des animaux. Pour estimer l'apparentement, on utilise le coefficient de parenté entre 2 individus (i et j) calculé à partir du pedigree. Le coefficient de parenté (F_{ij}), au sens de Malécot, (1948), est défini comme la probabilité qu'un allèle pris au hasard chez l'individu i pour un locus donné soit identique à un allèle pris au hasard chez l'individu j pour le même locus. Il est calculé comme :

$$F_{ij} = \sum_{c=1}^{chaînes} \left(\frac{1}{2}\right)^{n} * (1 + F_A)$$

où $chaînes$ correspond aux nombres de chaînes de parenté qui existent entre les individus i et j, n est le nombre d'animaux dans la chaîne de parenté, et F_A est le coefficient de consanguinité de l'ancêtre commun.

On s'attend à ce que les précisions des évaluations génomiques soient plus élevées si les populations d'apprentissage et de validation sont apparentées (Habier et al., 2010 ; Moser et al., 2010). Habier et al., (2010) ont analysé la précision des évaluations génomiques en contrôlant l'apparentement entre la population d'apprentissage et la population de validation. L'apparentement entre 2 individus a été défini comme 2 fois le coefficient de parenté F_{ij}. Les animaux de la population d'apprentissage avec un apparentement supérieur à un seuil (nommé a_{max}) avec les animaux de la validation ont été éliminés. Les seuils de 0,6 ; 0,49 ; 0,249 ; 0,1249 pour a_{max} ont été testés. Avec un seuil de 0,6 la population d'apprentissage contenait les pères, les frères, les sœurs, les demi-frères et les demi-sœurs des animaux de la population de validation. Avec un seuil de 0,49, seuls les demi-frères et demi-sœurs étaient présents dans la population d'apprentissage. Avec un seuil de 0,249 et 0,1249, les animaux de la population de validation n'étaient pas apparentés avec les animaux de la population d'apprentissage. En analysant 4 caractères (LAIT, MG, MP et CCS) avec un GBLUP et un BayesB, ils observent que les précisions diminuent lorsque le seuil a_{max} diminue quel que soit le caractère.

Les coefficients de parenté moyens entre population de référence et population des candidats ont été analysés par Carillier-Jacquin, (2015) pour les races Alpine et Saanen. Ils ont comparé aussi le coefficient de parenté estimé à partir des génotypes ($F_{ij(géno)}$) (Manichaikul et al., 2010):

$$F_{ij(géno)} = \frac{N_{i\ et\ j\ hétérozygotes} - 2N_{i\ et\ j\ homozygotes}}{N_{i\ hétérozygote} + N_{j\ hétérozygote}}$$

où $N_{i\ et\ j\ hétérozygotes}$ est le nombre de SNP pour lesquels i et j sont hétérozygotes, $N_{i\ et\ j\ homozygotes}$ est le nombre de SNP pour lesquels i et j sont homozygotes, $N_{i\ hétérozygote}$ est le nombre de SNP où l'individu i est hétérozygote (respectivement pour l'individu j avec $N_{j\ hétérozygote}$. Ils ont observé que la parenté entre individus est en moyenne inférieure à 3 % (Tableau 9). La parenté est plus élevée pour la race Saanen que pour la race Alpine (2 % pour la race Saanen contre 1,7 % pour la race Alpine pour la population d'apprentissage). L'apparentement moyen pour les races Alpine et Saanen est inférieur à ce qui est observé en bovins laitiers : entre 3 % et 10 % (Pszczola et al., 2012 ; Wientjes et al., 2012)

Tableau 9. Coefficient de parenté moyen (en %) estimé avec des données de pedigree ou génomiques pour la population de référence des races Alpine et Saanen

	Alpine			**Saanen**		
	Référence	**Candidats**	**Entre référence et candidats**	**Référence**	**Candidats**	**Entre référence et candidats**
$\boldsymbol{F_{ij(géno)}}$	1,7	0,7	1,1	2,0	2,3	2,4
$\boldsymbol{F_{ij(pedigree)}}$	1,9	1,1	1,2	2,1	1,5	1,5

3.5. Mise en place de la sélection génomique dans les races Alpine et Saanen françaises

La population caprine (Alpine et Saanen) possède des caractéristiques moins favorables à la mise en place d'évaluations génomiques que les populations de grande taille en bovins laitiers (LD plus faible, petite population de référence, apparentement entre population de référence et candidats plus faibles,…). Les premières études en sélection génomique pour les deux principales races caprines françaises ont été réalisées dans le cadre de la thèse de C. Carillier (2012-2015). La population d'étude est celle présentée au paragraphe 3.4 (Chapitre 1). Plusieurs modèles et méthodes génomiques ont été mis en œuvre et testés, soit sur une

population multirace (en poolant les données Alpine et Saanen dans une même population), soit sur chacune des populations. Les animaux de la population d'apprentissage ont été utilisés afin de prédire les GEBV de 252 mâles de la population de validation nés entre 2006 et 2009 (151 Alpine et 100 Saanen). Ces évaluations ont porté sur les 11 caractères présentés au paragraphe 2.1 (Chapitre 1). La précision des évaluations a été calculée comme la corrélation de Pearson entre les valeurs génétiques « vraies » (DYD calculé en 2013) et les (G)EBV. Les méthodes génomiques (Bayésiennes), basées sur des performances pré-corrigées (DYD ou EBV derégressés) n'ont pas permis une amélioration significative des précisions des évaluations génomiques par rapport aux évaluations BLUP. Le choix des phénotypes analysés (DYD des animaux génotypés, DYD de l'ensemble des animaux génotypés ou non, EBV derégressés pondérés ou non) n'a pas eu d'impact sur les précisions obtenues. En revanche, des gains de précisions plus importants ont été observés (entre +3,40 % et +21,30 %) sur la population multirace entre un GBLUP et un BLUP (Tableau 10). Les résultats obtenus avec les évaluations intra-races étaient différents de ceux obtenus avec l'évaluation multirace. Dans le cas des analyses intra-races, la taille de la population de référence était divisée de moitié, malgré tout, les corrélations de validation n'étaient pas si mauvaises et pouvaient même être supérieures à celles obtenues avec les évaluations multiraciales (Carillier-Jacquin, 2015). Les corrélations de validation obtenues en caprins restent toutefois inférieures à celles obtenues dans la littérature en ovins ou bovins laitiers, ce qui peut être expliqué en partie par les petites tailles des populations de référence en caprins.

Tableau 10. Précisions des évaluations BLUP et GBLUP multirace sur la population de validation (2006-2009) caprines (Carillier et al., 2013)

	BLUP	**GBLUP**	**Gain (%)**
LAIT	0,37	0,39	5,10
MG	0,35	0,37	6,20
MP	0,35	0,36	4,90
TB	0,50	0,53	7,70
TP	0,50	0,52	3,40
PLA	0,30	0,37	20,70
PRM	0,28	0,34	21,10
AAR	0,40	0,43	7,30
AVP	0,27	0,33	21,30
TA	0,32	0,35	8,60
CCS	0,31	0,32	5,20

À la suite de ces premiers résultats et à l'implémentation en 2009 de la méthode single-step GBLUP (ssGBLUP, Legarra et al., (2009)) dans le logiciel blup90iod (Misztal et al, 2002), la méthode ssGBLUP a été appliquée aux données caprines afin d'améliorer les précisions des évaluations (Carillier et al., 2014). Les évaluations ont été réalisées en multirace et intra-race (Tableau 11). L'avantage du ssGBLUP est, comme pour le BLUP, de pouvoir considérer dans le modèle les performances de tous les animaux qu'ils soient génotypés ou non, et de combiner les informations moléculaires disponibles (génotypages issus de la puce 50K) et les pedigrees. Les précisions génomiques obtenues dans les analyses intra-race sont identiques à celles de l'évaluation multirace pour le LAIT, la MP, PLA et les CCS. Les analyses intra-race sont en moyenne légèrement plus précises pour les caractères MG, MP, TB, PRM, AAR et AVP (entre +0,01 et +0,03). Les précisions sont légèrement inférieures pour les analyses intra-races pour les caractères TP et TA. Les résultats montrent que les précisions des analyses multirace avec le ssGBLUP sont meilleures que celles obtenues avec les analyses GBLUP (Carillier et al.,

2013, 2014). Les améliorations observées vont de +10 % pour le LAIT à +74 % pour TA. Seule une baisse de 8 % a été observée pour la MP. Suite aux travaux de thèse de C. Carillier, les évaluations génomiques sont appliquées en routine pour les races Alpine et Saanen avec la méthode ssGBLUP depuis janvier 2018.

Tableau 11. Précisions des évaluations génomiques ssGBLUP sur la population de validation (2006-2009) pour les analyses intra-races et multiraces (Carillier et al., 2014)

	Multirace	**Intra-race (moyenne)**
LAIT	0,43	0,43
MG	0,44	0,46
MP	0,33	0,36
TB	0,61	0,63
TP	0,70	0,69
PLA	0,59	0,59
PRM	0,55	0,56
AAR	0,64	0,66
AVP	0,50	0,51
TA	0,61	0,59
CCS	0,47	0,47

3.6. Résultats complémentaires obtenus dans le cadre d'analyse de détection de QTL chez les caprins

L'arrivée des outils moléculaires (puces SNP 50K) a permis également d'effectuer ou de revisiter des analyses sur l'architecture génétique des caractères à partir des données disponibles issues de dispositifs QTL. Les travaux de thèse de Maroteau (2014) et Martin (2016) ont mis en évidence des régions d'intérêt sur le génome caprin sur de nombreux caractères. En particulier, deux régions ont été confirmées sur les chromosomes 6 et 14, elles agissent sur la composition du lait :

- Sur le chromosome 6 autour de 82 Mb pour le TP. Cette région contient le cluster des gènes déjà connus de la caséine (α_{s1}, β et κ (Martin et al., 2002)). Le gène de la *caséine* α_{s1} est également présent chez d'autres ruminants laitiers : en bovins (Grosclaude, 1988) ou en ovins (Barillet et al., 2005). Chez les caprins, le gène α_{s1} explique 38,2 % de la variance génétique en race Alpine et 24,4 % de la variance génétique en race Saanen (Carillier-Jacquin et al., 2016).
- Sur le chromosome 14 autour de 11 Mb pour le TB. Cette région contient le gène *DGAT1*, qui se retrouve également dans les races bovines laitières (Grisart, 2002). Chez les caprins, deux mutations causales sont identifiées : la mutation R251L et la mutation R396W. La mutation R251L explique 6 % de la variance génétique alors que la mutation R396W explique 46 % de la variance génétique pour des analyses multirace (Martin et al., 2017).

D'autres régions contenant des QTLs avec des effets plus faibles ont été mises en évidence par Maroteau, (2014) mais aucun gène candidat n'a pu être identifié à ce jour. En race Alpine, des QTLs sur les chromosomes 1 et 7 ont été détectés pour la MP ; sur le chromosome 2 pour le TB et le TP et sur les chromosomes 1 et 8 pour AAR. En race Saanen, des QTLs ont été identifiés sur le chromosome 22 pour la MP et sur le chromosome 21 pour le TB et le TP.

Enfin, Maroteau, (2014) et Martin, (2016) ont mis en évidence un important QTL sur le chromosome 19 qui affecte plusieurs caractères : LAIT, MG, MP, AAR, PLA, la forme des trayons (TF) et CCS uniquement en Saanen. La Figure 18 présente les résultats des analyses d'association de Martin et al., (2018) pour le LAIT, MP, MG, PLA, AAR et TF sur le chromosome 19 en Saanen. On observe un pic significatif pour ces caractères autour de 24,5 Mb et 27 Mb avec des rapports de vraisemblance atteignant 80 pour PLA, 63 pour AAR, 57 pour la MP, 46 pour la MG et 45 pour le LAIT. Cette région a été identifiée comme ayant un effet pléiotropique, elle améliore les caractères de production laitière (LAIT, MG et MP) alors qu'elle détériore les caractères de morphologie de la mamelle (AAR et PLA) (Martin et al., 2018). Cette région est très riche en gènes annotés (192 gènes entre 22^{e} Mb et le 27,6e MB) (Martin et al., 2018). Elle contient en moyenne 34 gènes/Mb alors que le génome complet a une moyenne de 9,6 gènes/Mb (Martin et al., 2018). Cette situation rend difficile la détection de gènes candidats. Pour les caractères de production laitière, plusieurs gènes candidats, liés au métabolisme des lipides et des acides gras, ont été proposés. C'est le cas des gènes phospholipase D2 (PLD2), gamma-glutamyltransferase 6 (CGT6) et l'arachidonate lipoxygenase (ALOX12, ALOX12B, ALOX15). Mais il est peu probable que les fonctions de ces gènes aient des effets sur les caractères de morphologie de la mamelle (Martin et al., 2018). Pour les CCS, une autre région entre 33 et 42 Mb a été identifiée (Martin et al., 2018), région pour laquelle des gènes candidats (*RARA*, *STAT3*, *STAT5A*, and *STAT5B*) liés à la réponse aux infections intra-mammaires sont présents. Mucha et al., (2018) ont réalisé des analyses GWAS similaires sur des animaux croisés (Alpine, Saanen et Toggenbourg) au Royaume-Uni. Leurs travaux mettent en évidence la présence du QTL sur le chromosome 19 pour le LAIT et pour les caractères PLA et AAR. Ils suggèrent que le gène asialoglycoprotein receptor 2 (ASGR2) pourrait être impliqué dans l'expression du caractère AAR. Ce gène avait déjà été identifié par des GWAS en bovins laitiers pour le caractère AAR (Schrooten et al., 2000), il est impliqué dans l'homéostasie des lipides et la stabilité des protéines.

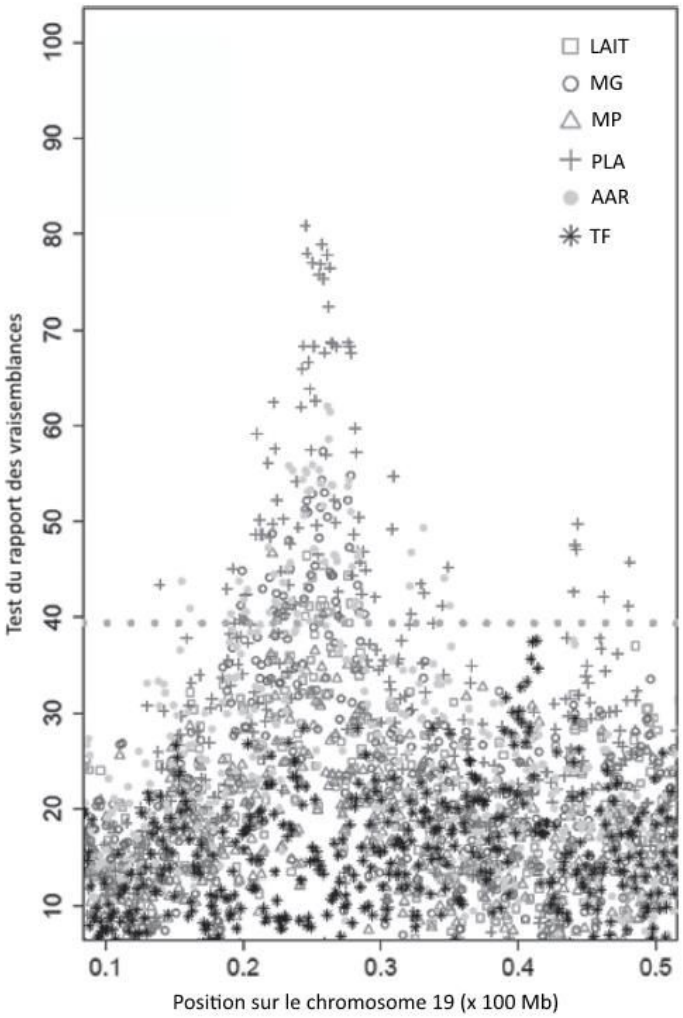


Figure 18. Rapport de vraisemblance globale des analyses d'associations pour la race Saanen sur le chromosome 19 pour les caractères : quantité de lait (LAIT), matières grasses (MG), matières protéiques (MP), distance plancher-jarret (PLA) et qualité de l'attache arrière (AAR) (Martin et., 2018)

4. Objectifs de la thèse

Les travaux de Carillier-Jacquin (2015) ont montré la supériorité du ssGBLUP sur le BLUP pour prédire les valeurs génétiques des animaux. Cette méthode suppose que le déterminisme génétique du caractère est polygénique, ce qui se traduit par une variance génétique expliquée pour chaque SNP identique. Des études d'association (GWAS) réalisées dans les races caprines indiquent que certains caractères s'éloignent de cette hypothèse. C'est le cas du taux protéique avec le gène de la *caséine* α_{s1} ou du taux butyreux avec le gène *DGAT1*. D'autres QTLs avec des effets importants ont été détectés pour les quantités de matières du lait (LAIT, MG et MP) en Saanen ainsi que pour certains caractères de morphologie de la mamelle (PLA et AAR). Le ssGBLUP n'intègre pas *a priori* la connaissance de ces gènes majeurs/QTLs. Il serait possible d'améliorer les précisions des évaluations génomiques en utilisant des modèles et des méthodes adaptés, pouvant intégrer une information sur l'architecture génétique des caractères.

Les objectifs de ma thèse ont été d'étudier l'évolution des précisions génomiques avec le ssGBLUP en fonction de l'accroissement des données disponibles et d'améliorer les précisions des évaluations en testant de nouveaux modèles et méthodes qui prennent en compte l'architecture génétique des caractères en sélection chez les caprins.

Le chapitre 2 est consacré à la description des modèles et des méthodes d'évaluations génomiques que j'ai pu mettre en place au cours de ma thèse. Ces méthodes permettent notamment de prendre en compte la présence de QTLs ou de gènes majeurs pour les caractères sélectionnés en routine. Des méthodes utilisant l'information individuelle des SNPs ou basées sur la construction d'haplotypes ont été testées.

Le chapitre 3 s'intéresse à l'application du ssGBLUP pour les races caprines en intégrant les nouvelles données de génotypages 50K disponibles. Une première partie étudie l'effet de la taille de la population d'apprentissage et de validation sur les précisions des évaluations génomiques. Dans une deuxième partie, nous nous sommes intéressés aux hyperparamètres du ssGBLUP et de leurs effets sur les précisions et sur les biais des évaluations génomiques caprines.

Le chapitre 4 s'intéresse à l'utilisation de différents modèles et méthodes d'évaluations génomiques pour intégrer la connaissance de gènes majeurs, notamment le gène de la *caséine* α_{s1} pour le taux protéique et le gène *DGAT1* pour le taux butyreux. Sur l'ensemble des autres caractères, des approches plus globales telles que le ssGBLUP pondéré (utilisant uniquement les génotypes de la puce 50K mais en pondérant les effets des SNPs) ont été testées pour tenir compte de l'existence possible de QTLs et voir l'effet de ces méthodes sur les précisions génomiques en fonction du déterminisme génétique du caractère (polygénique ou non).

Le chapitre 5 présente l'implémentation de modèles d'évaluation génomique basés sur la construction d'haplotypes. Ces derniers sont utilisés, soit sous forme d'haplotypes ou de pseudo-SNP, pour construire la matrice de parenté génomique.

Enfin, des perspectives sont proposées et discutées dans le chapitre 6.

Chapitre 2 : Modèles et méthodes d'évaluations génomiques permettant de prendre en compte des mutations, gènes majeurs ou QTLs

1. Méthodes d'évaluation génomique basées sur l'utilisation des marqueurs SNP

1.1. Le Weighted ssGBLUP (WssGBLUP)

Le WssGBLUP (Wang et al., 2012) est une approche itérative basée sur la méthode ssGBLUP. Comme pour le ssGBLUP, le WssGBLUP prédit les GEBV des animaux en utilisant l'apparentement entre individus estimé à partir des informations du pedigree et des génotypes. Cependant, il utilise une matrice de parenté génomique pondérée (***G****) contrairement à la matrice ***G*** utilisée avec le ssGBLUP. Des poids sont attribués aux SNPs en fonction de leur association au caractère, ils seront pris en compte pour l'estimation de l'apparentement entre individus. Plus un SNP aura un effet important sur le caractère et plus son poids sera élevé. La matrice génomique ***G**** est construite comme suit :

$$\boldsymbol{G}^{*} = 0{,}95\frac{\boldsymbol{ZDZ'}}{\sum_{i=1}^{m} p_i * (1-p_i)} + 0{,}05 * \boldsymbol{A_{22}}$$

où $\boldsymbol{Z}$ est la matrice des génotypes corrigés pour la fréquence des SNPs (paragraphe 3.1, Chapitre 1), $\boldsymbol{D}$ est une matrice diagonale ($m * m$) contenant les poids de SNPs, m est le nombre total de SNPs, p_i est la fréquence du SNP i et $\boldsymbol{A_{22}}$ représente la matrice de parenté basée sur le pedigree. Le WssGBLUP nécessite l'estimation des effets de chacun des SNPs ($\hat{\boldsymbol{a}}$) ainsi que leurs poids ($\hat{\boldsymbol{d}}$). Cette étape est réalisée après l'obtention des GEBV des animaux avec un ssGBLUP classique. Une fois les poids estimés, la matrice ***G**** peut être calculée. L'algorithme itératif mis en œuvre pour le WssGBLUP est le suivant et est illustré sur la Figure 19 :

$$\lambda = \frac{1}{\sum_{i=1}^{m} p_i * (1-p_i)}$$

1) $it = 1\ (numéro\ de\ l'itération\ courante),\ phase\ d'initialisation$:
   $\boldsymbol{D_{(1)}} = \boldsymbol{I}, \quad \boldsymbol{G^{*}_{(1)}} = 0{,}95 * \lambda \boldsymbol{Z D_{(1)} Z'} + 0{,}05 * \boldsymbol{A_{22}}$
2) Lancer ssGBLUP avec $\boldsymbol{G^{*}_{(1)}}$ pour obtenir $\hat{u}_{\boldsymbol{g(1)}}$
   Où $\hat{u}_{\boldsymbol{g(1)}}$ est le vecteur des valeurs génétiques pour l'itération 1
3) $it = 2$
4) Estimer les effets des SNPs $\hat{a}_{\boldsymbol{(it)}} = \lambda \boldsymbol{D_{(it-1)} Z' G^{*-1}_{(it-1)}} \hat{u}_{\boldsymbol{g\ (it-1)}}$
5) Estimer les poids des SNPs $d_i{}^{*} = \hat{a}_i^2 * 2p_i(1-p_i)$ pour tous les SNPs « i » (les éléments $d_i{}^{*}$ sont ensuite stockés dans la diagonale de la matrice $D^{*}_{(it)}$)
6) Normaliser les poids des SNPs $D_{(it)} = \frac{tr(D_{(1)})}{tr(D^{*}_{(it)})} * D^{*}_{(it)}$
7) Construire $\boldsymbol{G^{*}_{(it)}} = 0{,}95 * \lambda \boldsymbol{Z D_{(it)} Z'} + 0{,}05 * \boldsymbol{A_{22}}$
8) Lancer WssGBLUP avec $\boldsymbol{G^{*}_{(it)}}$ pour obtenir $\hat{u}_{\boldsymbol{g(it)}}$
9) $it = it + 1$
10) Fin, ou retourner à l'étape 4

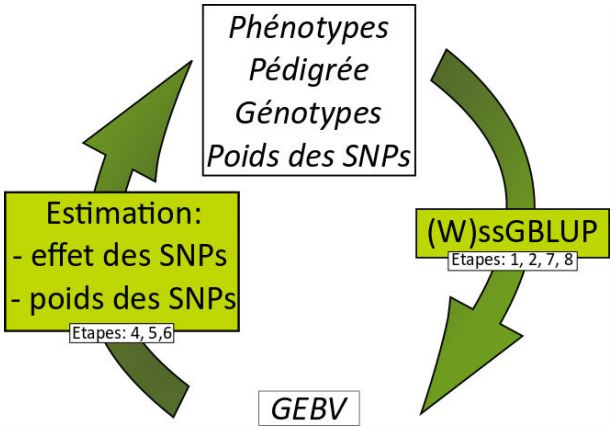


Figure 19. Principe général du Weighted single-step GBLUP

Le WssGBLUP autorise que chaque SNP explique une part de variance génétique différente. Cette méthode permet d'intégrer la présence de QTL ou de gènes majeurs dans les évaluations génomiques si les SNPs sont en LD avec les QTLs ou les mutations causales (Hayes and Goddard, 2010). Plusieurs études dans la littérature ont montré la supériorité du WssGBLUP par rapport au ssGBLUP pour des caractères régis par un ou plusieurs QTLs. Les travaux de Wang et al., (2012), sur données simulées, montrent que dès la deuxième itération du WssGBLUP, les précisions des évaluations sont supérieures à celles du ssGBLUP. Cependant, les précisions chutent au-delà de 2 itérations. Des améliorations de précisions ont été observées chez la truite arc-en-ciel pour un caractère de survie à une bactérie (Vallejo et al., 2017), les précisions des évaluations génomiques sont passées de 0,63 en ssGBLUP à 0,67 avec un WssGBLUP. Des QTLs ont pu être identifiés pour ce caractère sur les chromosomes 8 et 11 (Palti et al., 2015 ; Vallejo et al., 2017). En bovins laitiers, des précisions supérieures entre +3 et +5 points ont été observées avec le WssGBLUP par rapport au ssGBLUP pour le TB et le TP (Lourenco et al., 2014). Ces caractères sont connus pour présenter un gène majeur : le gène de la *caséine* α_{s1} pour le TP sur le chromosome 6 (Khatkar et al., 2004) et le gène *DGAT1* pour le TB sur le chromosome 14 (Grisart, 2002).

Au cours des itérations, le défaut majeur de l'approche WssGBLUP est que les poids des SNPs les plus faibles ont tendance à tendre vers 0 et les poids des SNPs les plus forts ont tendance à exploser (Wang et al., 2012). Pour limiter ce phénomène, des alternatives à l'approche WssGBLUP ont été proposées. Zhang et al. (2016) propose, à partir de la première itération d'un WssGBLUP, d'attribuer le même poids à plusieurs SNPs consécutifs (Figure 20). Pour cela, on considère *n* SNPs adjacents et on leur attribue tous le même poids. 3 approches ont ainsi été définies :

– $\text{WssGBLUP}_{\text{Max}}$: le poids le plus important observé sur l'intervalle est attribué aux n SNPs adjacents

– $\text{WssGBLUP}_{\text{Sum}}$: la somme des poids des n SNPs adjacents est attribuée aux n SNPs adjacents

– $\text{WssGBLUP}_{\text{Mean}}$: la moyenne des poids des n SNPs adjacents est attribuée aux n SNPs adjacents

Les méthodes $\text{WssGBLUP}_{\text{Max}}$, $\text{WssGBLUP}_{\text{Sum}}$ et $\text{WssGBLUP}_{\text{Mean}}$ suivent le même algorithme que le WssGBLUP. Une seule étape est ajoutée entre les étapes 5 et 6 du WssGBLUP pour calculer les nouveaux poids et construire la matrice $\mathbf{G}^*$.

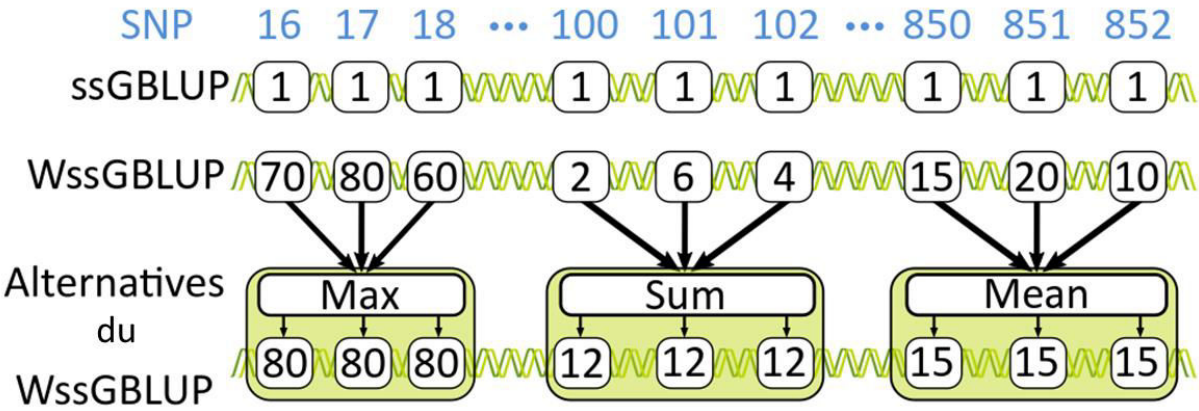


Figure 20. Pondération des SNPs selon les méthodes ssGBLUP, WssGBLUP et les alternatives du WssGBLUP (Max, Sum et Mean)

Zhang et al., (2016) indiquent que ces approches permettent de mieux capter le signal d'une région du génome. Cependant, il est nécessaire de calculer la taille optimale de la fenêtre. L'étude de Zhang et al., (2016) explique que les meilleures précisions sont obtenues pour des fenêtres de 20 SNPs consécutifs. Ce paramètre dépend bien sûr de la structure de la population elle-même (LD, consanguinité,…) (Su et al., 2014) et du caractère étudié. Zhang et al., (2016) ont simulé des caractères avec la présence de 5, 100 et 500 QTLs répartis sur 20 chromosomes (contenant au total 45 000 SNPs). Ils observent de légers gains en utilisant les approches $WssGBLUP_{Max,}$ $WssGBLUP_{Sum}$. Ces gains sont de +1 à + 2 points par rapport au WssGBLUP utilisant des poids différents pour chaque SNP.

1.2. Le gene content

Le gene content est une méthode permettant d'intégrer l'information d'un gène majeur et pour lequel des génotypages de ce gène existent et sont disponibles (Gengler et al., 2007; Legarra and Vitezica, 2015). Le gene content, est défini comme le nombre de copies de chaque allèle pour un génotype donné (Figure 21).

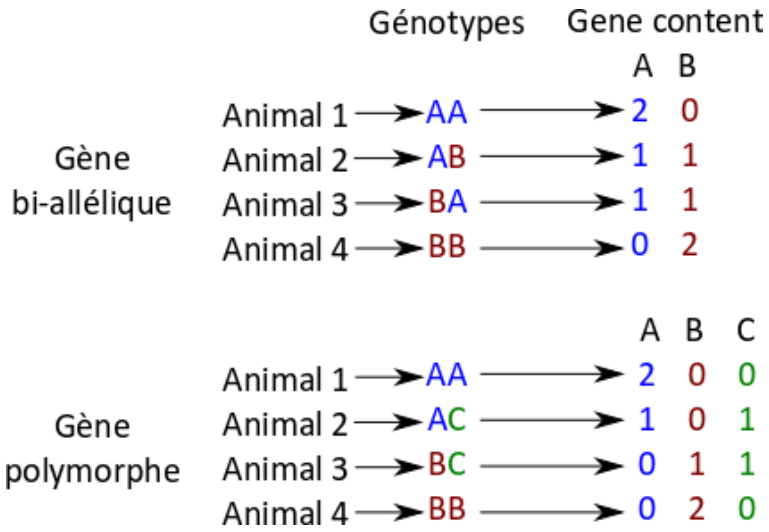


Figure 21. Exemple de l'approche gene content pour un marqueur bi-allélique (type SNP) et un marqueur polymorphe

L'approche gene content est mise en œuvre à l'aide d'un modèle multicaractère dans lequel on évalue la valeur génétique des animaux pour le caractère étudié et la valeur génétique des animaux pour le « *gene content* » de chaque allèle. Le modèle multicaractère pour un gène polymorphe, contenant 3 allèles, s'écrit comme suit :

$$\begin{cases} \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{+e} \\ \boldsymbol{y_A} = \mu_{\boldsymbol{A}} + \boldsymbol{Z}_A\boldsymbol{u}_A + \boldsymbol{e}_A \\ \boldsymbol{y_B} = \mu_{\boldsymbol{B}} + \boldsymbol{Z}_B\boldsymbol{u}_B + \boldsymbol{e}_B \\ \boldsymbol{y_C} = \mu_{\boldsymbol{C}} + \boldsymbol{Z}_C\boldsymbol{u}_C + \boldsymbol{e}_C \end{cases} \text{ [Modèle 3]}$$

où $\boldsymbol{y}$ est le vecteur de performances du caractère étudié, $\boldsymbol{\beta}$ est le vecteur d'effets fixes, $\mathbf{u}$ et $\mathbf{e}$ sont les vecteurs aléatoires pour les effets génétiques additifs et la résiduelle respectivement. $\mathbf{X}$ et $\mathbf{Z}$ sont des matrices d'incidence pour les vecteurs $\boldsymbol{\beta}$ et $\mathbf{u}$. Les vecteurs $\mathbf{y_A}$ à $\mathbf{y_C}$ contiennent les « *gene content* » pour les allèles A, B et C du gène d'intérêt (codés sous la forme 0, 1 ou 2). $\boldsymbol{\mu_A}$ à $\boldsymbol{\mu_C}$ sont les effets moyens de chaque allèle. $\mathbf{Z_A}$ à $\mathbf{Z_C}$ sont des matrices d'incidences pour les effets génétiques aléatoires ($\mathbf{u_A}$ à $\mathbf{u_O}$) du « *gene content* » pour chaque allèle.

La matrice de variances/covariances du modèle multicaractère est définie de la manière suivante (Legarra and Vitezica, 2015; Carillier-Jacquin et al., 2016) :

$$cov\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{u_A} \\ \boldsymbol{u_B} \\ \boldsymbol{u_C} \end{pmatrix} = \begin{pmatrix} \boldsymbol{H\sigma^2_u} & \boldsymbol{H\sigma_{u,u_A}} & \boldsymbol{H\sigma_{u,u_B}} & \boldsymbol{H\sigma_{u,u_C}} \\ \boldsymbol{H\sigma_{u,u_A}} & \boldsymbol{H\sigma^2_{u_A}} & \boldsymbol{H\sigma_{u_A,u_B}} & \boldsymbol{H\sigma_{u_A,u_C}} \\ \boldsymbol{H\sigma_{u,u_B}} & \boldsymbol{H\sigma_{u_A,u_B}} & \boldsymbol{H\sigma^2_{u_B}} & \boldsymbol{H\sigma_{u_B,u_C}} \\ \boldsymbol{H\sigma_{u,u_C}} & \boldsymbol{H\sigma_{u_A,u_C}} & \boldsymbol{H\sigma_{u_B,u_C}} & \boldsymbol{H\sigma^2_{u_C}} \end{pmatrix}$$

La variance génétique ($\boldsymbol{u}$) est calculée comme :

$$Var(\boldsymbol{u}) = \boldsymbol{H\sigma^2_u} = \boldsymbol{H}\left[\sigma^2_e + 2 * \sum_i p_i q_j \alpha_i^2 - 2 * \sum_i \sum_{j \neq i} p_i q_j \alpha_i \alpha_j\right]$$

Pour i et j $\in \{A, B, C\}$ dans notre exemple, où p_A, p_B, p_C correspondent aux fréquences des allèles A, B et C respectivement et $q_i = 1 - p_i$, les valeurs α_A, α_B, α_C sont les effets des allèles A, B et C respectivement, enfin σ^2_e est la variance résiduelle. Les variances pour chaque allèle i, i $\in \{A, B, C\}$, sont calculées comme :

$$var(\boldsymbol{u_i}) = \boldsymbol{H\sigma^2_{u_i}} = p_i q_i$$

Les covariances entre les valeurs génétiques ($\mathbf{u}$) et les valeurs génétiques du « *gene content* » ($\mathbf{u}_A$, $\mathbf{u}_B$, $\mathbf{u}_C$) sont calculées comme :

$$cov(\boldsymbol{u}, \boldsymbol{u_i}) = \boldsymbol{H\sigma_{u,u_i}} = 2\boldsymbol{H}p_i q_j \alpha_i - 2\boldsymbol{H}\sum_i \sum_{j \neq i} p_i q_j \alpha_i \alpha_j$$

Enfin, les covariances entre allèles, pour i et j $\in \{A, B, C\}$, sont calculées comme :

$$cov\left(\boldsymbol{u_i}, \boldsymbol{u_j}\right) = \boldsymbol{H\sigma_{u_i,u_j}} = -p_i p_j$$

L'avantage de ce modèle est de pouvoir gérer les données manquantes, c'est-à-dire que tous les animaux génotypés sur la puce 50K ne sont pas nécessairement génotypés pour le gène majeur et inversement (Legarra and Vitezica, 2015). La méthode peut facilement s'étendre si l'on souhaite intégrer plusieurs gènes majeurs dans les évaluations. En revanche, la taille du système à résoudre (nombre de caractères à évaluer simultanément) augmente rapidement. Dans la littérature, on trouve très peu d'applications de cette approche dans le cadre des évaluations génomiques. Dans les races françaises Alpine et Saanen, Carillier-Jacquin et al., (2016) a réalisé une première étude en analysant le TP et en exploitant les données de génotypages disponibles sur le gène de la *caséine* α_{s1}. Les analyses ont porté sur les populations présentées dans le paragraphe 3.4 (Chapitre 1). Les résultats du gene content ont été comparés avec une évaluation

ssGBLUP. Les précisions des évaluations étaient supérieures avec le gene content que ce soit pour les analyses multirace, Alpine ou Saanen (Tableau 12). Les plus fortes améliorations de précisions sont observées pour la race Saanen avec +13 points pour les évaluations avec le gene content par rapport au ssGBLUP. Pour la race Alpine, le gain est de +5 points pour le gene content et de +3 points pour les analyses multirace.

Tableau 12. Précisions des évaluations génomiques caprines pour le TP pour des évaluations ssGBLUP et gene content (Carillier-Jacquin et al., 2016)

	ssGBLUP	Gene content
Multirace	0,72	0,75
Alpine	0,63	0,68
Saanen	0,75	0,86

1.3. Le Trait-specific marker-derived relationship matrix (TABLUP)

Le TABLUP décrit par Zhang et al., (2011), consiste à réaliser des évaluations génomiques ssGBLUP en utilisant une matrice de parenté génomique restreinte. Les SNPs pris en compte pour construire la matrice génomique G sont présélectionnés, seuls les SNPs associés au caractère y sont inclus. Zhang et al., (2011) ont testé le TABLUP sur des données simulées contenant 3 226 animaux et un génome de 5 chromosomes (10 031 SNPs au total). Ils ont sélectionné les SNPs en fonction de leurs effets estimés au préalable par une approche BayesB ou un RRBLUP. Ils ont conservé les 100, 200, 500, 1 000, 2 000 et 5 000 SNPs avec les effets les plus forts pour construire la matrice génomique ***G***. Ils n'ont observé aucune différence entre les précisions génomiques obtenues avec un BayesB (0,676) en considérant l'ensemble des 10 031 SNPs et un TABLUP ne considérant qu'un sous-ensemble présélectionnés de SNPs à l'aide d'un BayesB (0,672-0,678). En revanche, les précisions étaient légèrement améliorées en utilisant un TABLUP, dont les SNPs étaient présélectionnés à partir d'un RRBLUP (Tableau 13). La précision maximale est atteinte en utilisant uniquement 200 ou 500 SNPs (0,647) contre 0,608 pour les analyses RRBLUP avec l'ensemble des SNPs.

Tableau 13. Précisions des évaluations génomiques estimées avec un RRBLUP et un TABLUP (pré-sélection de SNPs avec un RRBLUP) dans les analyses de Zhang et al., 2011

	Nb SNP	Précision
RRBLUP	10031	0,608
	10031	0,626
	5000	0,632
	2000	0,640
TABLUP	1000	0,643
	500	0,647
	200	0,647
	100	0,626

Les SNPs utilisés dans la méthode TABLUP peuvent être sélectionnés selon différentes approches : RRBLUP, BayesB, GWAS, WssGBLUP, ….

1.4. Le BayesR

Le BayesR fait partie des méthodes qui estiment directement les effets des SNPs (Figure 5), ils sont ensuite sommés pour prédire les GEBV. Le BayesR fait appel à des méthodes de Gibbs sampling pour résoudre les systèmes d'équations. Il consiste à repartir les SNPs en 4 classes

distinctes, de sorte à autoriser que l'ordre de grandeur potentiel des effets des marqueurs soit variable (Erbe et al., 2012). L'effet d'un SNP dans une classe donnée pourra être échantillonné à partir des paramètres de cette classe. Le modèle sous-jacent est le suivant :

$$\boldsymbol{y} = \mu + \boldsymbol{Zu} + \boldsymbol{Xg} + \boldsymbol{e} \text{ [Modèle 4]}$$

où y est le vecteur des phénotypes, μ est l'effet moyen, $\boldsymbol{u}$ est le vecteur des effets génétiques aléatoires supposé normalement distribué $N(\mathbf{0}, \boldsymbol{A}\sigma_u^2)$ avec **A** la matrice de parenté (construite à partir du pedigree) et σ_u^2 la variance génétique pour l'effet polygénique, **g** est le vecteur des effets des SNPs. Chaque SNP i est distribué selon une loi normale $g_i \sim N(\mathbf{0}, \sigma_k^2)$ k représentant une des 4 classes et **e** est le vecteur des résidus distribué selon une loi normale $N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$. Les matrices **Z** et **X** sont des matrices d'incidence pour les effets **u** et **g** respectivement.

Chaque SNP appartient à une des 4 classes, chacune ayant des variances différentes :

$$classe\ 1: \sigma_1^2 = 0$$

$$classe2: \sigma_2^2 = 0{,}001\sigma_g^2$$

$$classe3: \sigma_3^2 = 0{,}005\sigma_g^2$$

$$classe\ 4: \sigma_4^2 = 0{,}01\sigma_g^2$$

où σ_g^2 représente la variance génétique totale.

Chaque SNP est attribué à une des 4 classes puis son effet sera échantillonné selon la distribution de cette classe. Initialement, la distribution des proportions des SNPs (**pr**) pour chaque classe k ($\boldsymbol{pr}_k$) est tirée selon une loi de Dirichlet **pr** ~*Dir*(**α**) avec $\boldsymbol{\alpha} = [1, 1, 1, 1]$. Ce choix d'*a priori* présente l'avantage d'être peu informatif (Erbe et al., 2012), les SNPs seront uniformément répartis dans les 4 classes. Une distribution a posteriori des proportions des SNPs est calculée comme **pr** ~*Dir* (**α** + **β**), où **β** est un vecteur contenant le nombre de SNPs dans chaque classe estimé à partir des données. Pour obtenir le vecteur **β**, plusieurs étapes sont nécessaires :

(1) calculer les vraisemblances qu'un SNP i appartienne à chacune des classes k, pour k ∈ {1, 2, 3, 4}. Le calcul de la vraisemblance ($LogL(i,k)$) dépend des données (phénotypes, pedigree et génotypes) mais également de la distribution *a priori* de la proportion de SNP (**pr**) dans chaque classe :

$$LogL(i,k) = -0{,}5\ log|\boldsymbol{V}| - \frac{0{,}5\left(\boldsymbol{y}^{*\prime}\boldsymbol{y}^{*} - \boldsymbol{y}^{*\prime}\boldsymbol{Z}^{*}\widehat{\boldsymbol{u}}\right)}{\sigma_e^2 + \log(\boldsymbol{pr}_k)}$$

où **y*** est le vecteur des phénotypes corrigés pour tous les autres marqueurs, la moyenne ainsi que les effets polygéniques ; **Z*** est un vecteur colonne qui contient tous les génotypes pour le SNP i, **V** est la matrice de variance-covariance d'un modèle réduit qui inclut uniquement l'effet du SNP i et une résiduelle. $\boldsymbol{pr}_k$est la proportion de SNP de la classe k échantillonné à partir d'une distribution de Dirichlet.

(2) Une fois la vraisemblance calculée pour les k classes, la probabilité que le SNP i appartienne à la classe k est déterminée comme :

$$P(classe(SNP_i) = k) = \frac{1}{\sum_{l=1}^{4} exp[L(i,l) - L(i,k)]}$$

(3) La classe du SNP est alors choisie en tirant une variable aléatoire dans une loi uniforme. En connaissant les valeurs du vecteur **pr**, la classe du SNP est choisie (Figure 22). Son effet peut être ensuite estimé à l'aide de la distribution de l'effet du SNP associé à sa classe.

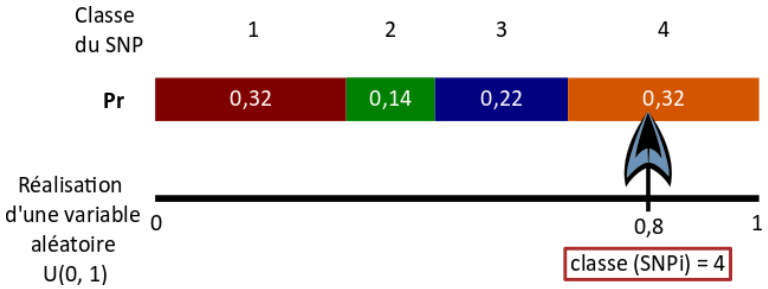


*Figure 22. Sélection d'une classe pour le SNP i en fonction du vecteur **pr** avec le BayesR*

Erbe et al., (2012) ont appliqué le BayesR pour prédire des caractères de production laitière (LAIT, MG et MP) à l'aide d'une puce 50K en race Holstein et Jersiaise. En Holstein, les précisions étaient supérieures de +4, +8 et +1 points pour le LAIT, la MG et la MP respectivement en faveur du BayesR par rapport au GBLUP (Tableau 14). Ces gains étaient de +18, +3, +6 pour le LAIT, la MG et la MP en race Jersiaise.

Tableau 14. Précisions des évaluations génomiques GBLUP et BayesR pour des populations Holstein et Jersey (Erbe et al., 2012)

	Méthode	LAIT	MG	MP	Moyenne
Holstein	GBLUP	0,58	0,58	0,56	0,57
	BayesR	0,62	0,66	0,57	0,62
Jersey	GBLUP	0,33	0,46	0,40	0,40
	BayesR	0,51	0,49	0,46	0,49

Kemper et al., (2015) ont obtenu des résultats similaires. Ils ont amélioré les précisions des évaluations génomiques pour des caractères de production laitière (LAIT, MG, MP, TB et TP) de +3 points en moyenne pour les Holstein et de +6 points pour les Jersiaises (Tableau 15).

Tableau 15. Précisions des évaluations génomiques GBLUP et BayesR pour des populations Holstein et Jersey (Kemper et al., 2015)

	Méthode	LAIT	MG	MP	TB	TP	Moyenne
Holstein	*GBLUP*	0,60	0,58	0,59	0,71	0,83	0,66
	BayesR	0,63	0,62	0,58	0,81	0,83	0,69
Jersey	*GBLUP*	0,56	0,62	0,67	0,63	0,75	0,65
	BayesR	0,56	0,70	0,72	0,77	0,79	0,71

2. Méthodes d'évaluation génomique haplotypique

Les méthodes d'évaluations génomiques utilisant l'information des données moléculaires (puce SNPs 50K) présentent quelques limites. La première est que les SNPs présents sur la puce sont sélectionnés pour avoir des MAF modérées à élevées. Cette sélection favorise des mutations « anciennes » qui ont eu le temps de se propager dans la population (Meuwissen et al., 2014a ; Zahra, 2018). L'estimation de l'apparentement entre individus se base sur des évènements de mutations lointains et ignore des changements dus à une sélection récente (Zahra, 2018). De plus, les SNPs sont des marqueurs peu polymorphes (bi-allélique). Leur utilisation dans des méthodes d'évaluations génomiques n'est pas optimale pour capter le LD entre marqueurs et QTL. Pour surpasser ces contraintes, l'utilisation d'haplotypes peut

améliorer les précisions des évaluations génomiques (Calus et al., 2008 ; Cuyabano et al., 2014 ; Ferdosi et al., 2016 ; Jónás et al., 2016 ; Zahra, 2018).

Un haplotype est défini comme un groupe de SNPs proches, situés sur le même chromosome et souvent transmis ensemble. L'avantage des haplotypes est qu'ils identifient mieux des évènements liés à une sélection récente (Meuwissen et al., 2014b ; a ; Zahra, 2018). Ils présentent aussi l'avantage d'être plus polymorphes que des SNPs.

Il existe différentes méthodes de construction d'haplotypes : regroupement des SNPs selon leur position, selon leur LD, en prenant en compte les MAF des SNPs,… (Hayes et al., 2007; Villumsen and Janss, 2009; Cuyabano et al., 2014; Luan et al., 2014; Ferdosi et al., 2016). Nous nous focaliserons ici sur deux méthodes de construction d'haplotypes : l'approche appelée Distinct Windows (DW) et celle utilisant le LD (LD). Nous présenterons ensuite les modèles et méthodes d'évaluation génomique haplotypiques mises en œuvre au cours de ma thèse.

2.1. Construction des haplotypes

La construction d'haplotypes nécessite que les génotypes soient phasés, cela signifie que l'on connait quel SNP a été transmis par la mère et par le père. Plusieurs outils/algorithmes sont disponibles dans la littérature, notamment les logiciels Fimpute (Sargolzaei et al., 2014) et Beagle (Browning et al., 2018). Ces logiciels utilisent à la fois des informations familiales (si le pedigree est disponible) et des informations populationnelles pour déterminer le plus précisément possible les phases de chaque SNP. Une fois les génotypes phasés, il devient possible de construire des haplotypes.

La méthode DW, proposée par Hickey et al., (2013), découpe chaque chromosome en un nombre donné de segments, chaque segment comportant un nombre donné et identique de SNPs consécutifs. La Figure 23 illustre l'approche DW à partir d'un chromosome comportant 5 SNPs et la construction d'haplotype comportant 2 SNPs (cet exemple est issu de l'article de Ferdosi et al., (2016)). En fin de chromosome, le nombre de SNPs restants peut être inférieur à la taille de l'haplotype. Dans ce cas, on constitue un haplotype en utilisant des SNPs déjà présents dans l'haplotype précédent (Figure 23).

Pour déterminer la taille optimale des haplotypes, une approche consiste à lancer des évaluations avec des haplotypes de différentes tailles. La taille optimale des haplotypes sera définie comme celle pour laquelle la précision des évaluations sera maximale, cette approche est utilisée par Ferdosi et al., (2016) et Zahra, (2018).

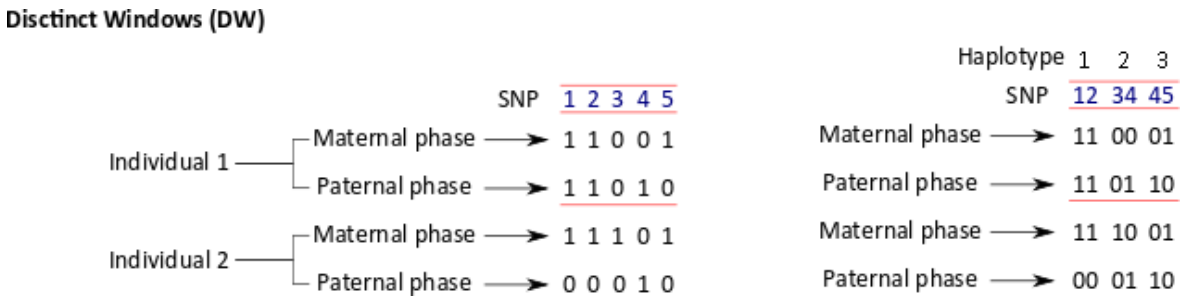


Figure 23. Construction des haplotypes à partir de la méthode Dinstinct Windows (DW) (Ferdosi et al., 2016)

L'approche des haplotypes construits sur la base du LD consiste à calculer le LD entre toutes les paires de SNPs (Figure 24). On forme un haplotype uniquement si tous les SNPs entre eux ont un LD supérieur à un seuil donné (les SNPs dont le LD est supérieur au seuil sont

représentés en rouge sur la Figure 24). Les SNPs regroupés forment ce qu'on appelle un haploblock. Dans notre exemple (Figure 24), on peut noter que le LD entre les SNPs 2 et 4 est supérieur au seuil de LD, mais le LD entre les SNPs 1 et 4 d'une part et entre les SNPs 3 et 4 d'autre part sont inférieurs au seuil. Dans ce cas, le premier haploblock construit est constitué uniquement des SNPs 1, 2 et 3. Enfin, le LD entre les SNPs 4 et 5 est inférieur au seuil, ces deux SNPs ne sont pas regroupés dans un haploblock. Chacun des SNPs 4 et 5 forme à eux seuls des haplotypes. Avec cette approche, la taille des haplotypes ainsi construits n'est pas constante le long du chromosome.

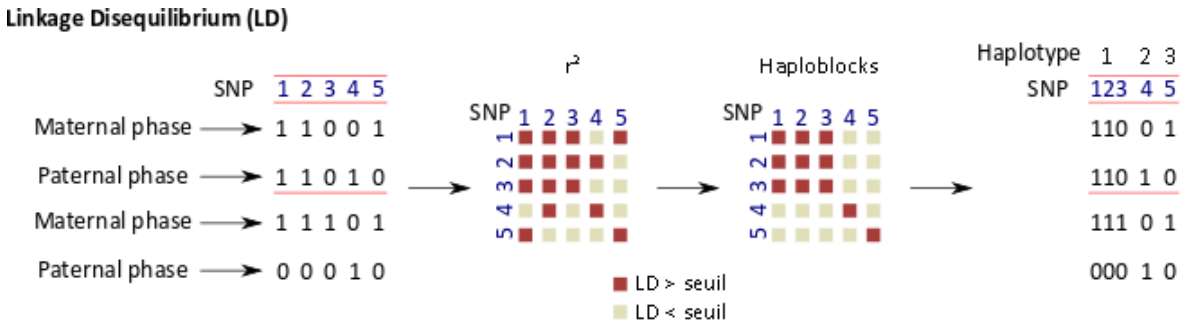


Figure 24. Construction des haplotypes sur la base du LD (Cuyabano et al., 2014)

2.2. Utilisation d'une matrice de parenté haplotypique dans les évaluations ssGBLUP

Les évaluations génomiques haplotypiques selon la méthode de Ferdosi utilisent le modèle d'évaluation suivant :

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{Wp} + \boldsymbol{e} \quad \text{[Modèle 5]}$$

où $\boldsymbol{y}$ est un vecteur de phénotypes, $\boldsymbol{b}$ est un vecteur qui regroupe l'ensemble des effets d'environnement, $\boldsymbol{u}$ est un vecteur contenant les valeurs génétiques des animaux supposé distribué normalement $N(\mathbf{0}, \boldsymbol{H_{hap}}\sigma_u^2)$ où $\boldsymbol{H_{hap}}$ est la matrice de parenté qui combine la matrice de parenté **A** et une matrice de parenté haplotypique $\mathbf{G_{hap}}$, $\boldsymbol{p}$ est un vecteur des effets d'environnements permanents $N(\mathbf{0}, \boldsymbol{I}\sigma_p^2)$ et $\boldsymbol{e}$ est un vecteur des effets résiduels normalement distribué $N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$. $\boldsymbol{X}$, $\boldsymbol{Z}$ et $\boldsymbol{W}$ sont des matrices d'incidence pour les vecteurs $\boldsymbol{b}$, $\boldsymbol{u}$ et $\boldsymbol{p}$ respectivement.

La construction de la matrice $\mathbf{G_{hap}}$ passe par un comptage des haplotypes identiques entre deux phases (Figure 25). Ces comptages sont stockés dans la matrice $\boldsymbol{\Gamma}$, contenant le nombre d'haplotypes sur la diagonale et le nombre d'haplotypes identiques entre deux phases hors diagonale. La matrice $\boldsymbol{\Gamma}$ peut être construite à partir des haplotypes construits avec les méthodes DW et LD. La matrice de parenté génomique ($\boldsymbol{G_{hap}}$) est alors définie comme suit :

$$\boldsymbol{G_{hap}} = \frac{\boldsymbol{K\Gamma K'}}{2}$$

où $\boldsymbol{K} = \boldsymbol{I} \otimes [11]$ avec I une matrice identité de dimension 2n * 2n où n est le nombre d'animaux génotypés. Cette construction revient à découper la matrice $\boldsymbol{\Gamma}$ en bloc, permettant ainsi de distinguer l'appartenance des phases à un animal (bloc rouge et bleu dans la Figure 25). Puis, tous les éléments de chaque bloc sont additionnés et divisés par 2.

La méthode de Ferdosi et al, (2016) est relativement simple à mettre en œuvre. En effet, la matrice $\mathbf{G}_{hap}$ peut être intégrée facilement dans la matrice **H** d'un ssGBLUP. La principale

contrainte réside dans la création des haplotypes (phasage + construction des haplotypes) qui nécessite des temps de calcul relativement longs.

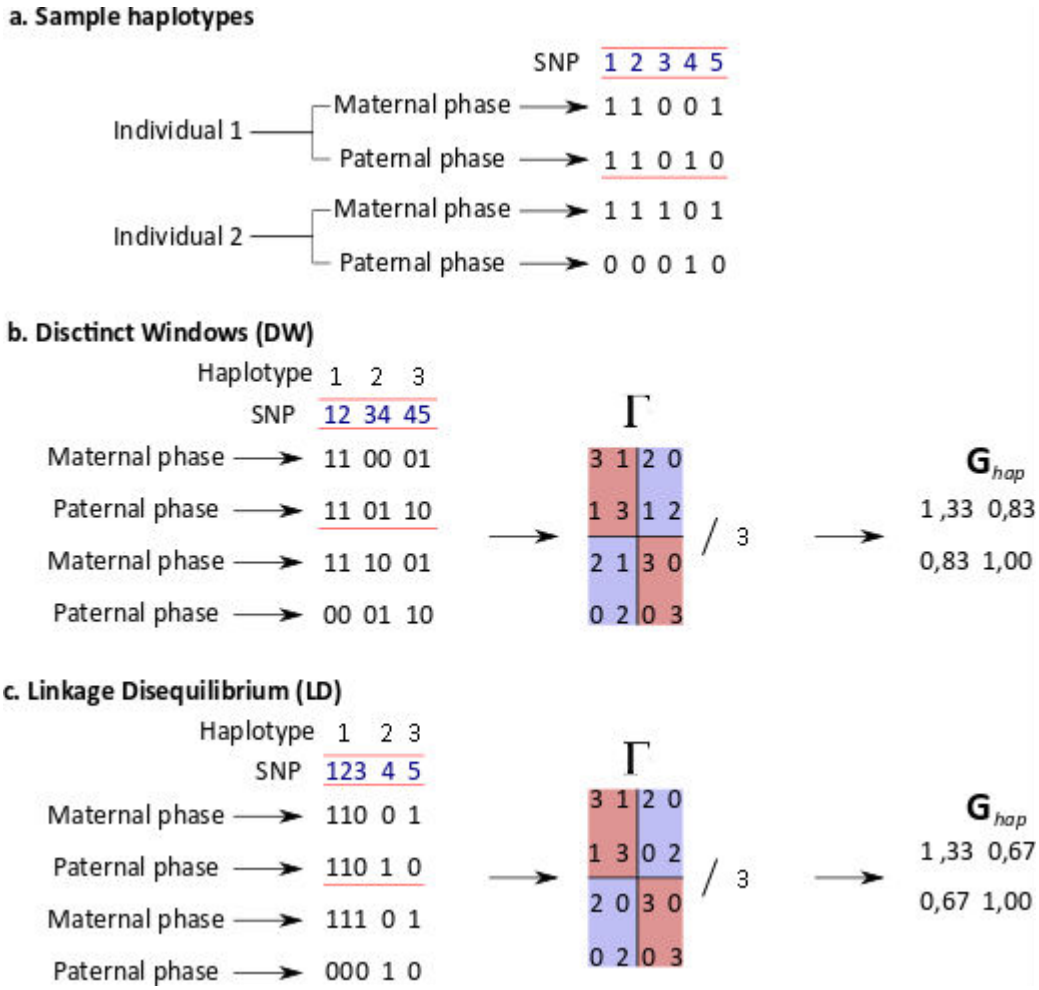


Figure 25. Construction de la matrice de parenté génomique haplotypique ($\boldsymbol{G_{hap}}$) selon la méthode DW (b) ou LD (c) (Ferdosi et al., 2016)

Ferdosi et al. (2016) ont analysé 3 caractères (la circonférence du scrotum, l'âge à la puberté et le poids du premier corps jaune) chez des populations Brahman à l'aide de l'approche DW. Ils disposaient d'environ 1 000 animaux génotypés pour 50 000 SNPs. Ils ont testé des haplotypes de 2, 5, 10, 20, 40, 80, 100, 120, 180, 200 et 240 SNPs. Ils ont observé une amélioration des précisions des évaluations génomiques avec l'utilisation de la matrice $\boldsymbol{G_{hap}}$ comparée à un GBLUP (Tableau 16). Les précisions sont supérieures de +5 points pour la circonférence du scrotum, +2 points pour l'âge à la puberté et de +2 points pour le poids du premier corps jaune. Les meilleures précisions ont été obtenues pour des fenêtres relativement petites, contenant 16 SNPs, 7 SNPs et 11 SNPs pour la circonférence du scrotum, l'âge à la puberté et le poids du premier corps jaune respectivement.

Tableau 16. Précisions des évaluations génomiques GBLUP et des évaluations génomiques haplotypiques (DW) pour 3 caractères de reproduction chez les bovins Brahman (Ferdosi et al., 2016)

	GBLUP	DW	Meilleure fenêtre
Circonférence du scrotum	0,38	0,43	16
Age à la puberté	0,31	0,33	7
Poids du premier corps jaune	0,40	0,42	11

2.3. Utilisation de pseudo-SNP dans les évaluations ssGBLUP et WssGBLUP

Zahra (2018) propose d'intégrer les haplotypes dans le modèle d'évaluation génomique sous la forme de pseudo-SNPs. Chaque haplotype (ou pseudo-SNP) est codé de la même manière que les SNPs, c'est-à-dire codé sous forme de 0,1 ou 2. Les méthodes telles que le ssGBLUP et le WssGBLUP peuvent alors être implémentées facilement. La Figure 26 illustre la conversion des haplotypes sous forme de pseudo-SNP, qu'ils soient construits à l'aide de l'approche DW ou celle basée sur le LD. Chaque allèle de chaque haplotype est transformé en pseudo-SNP. Il suffit ensuite de compter le nombre d'allèles dont l'animal est porteur pour chaque allèle (valeur qui peut être 0,1 ou 2).

Par exemple avec l'approche DW (Figure 26), l'haplotype 1 a deux allèles différents (11 et 00). Ces deux allèles sont transformés en pseudo-SNP : l'animal 1 aura un génotype de 2 pour l'allèle 11 et 0 pour l'allèle 00. Si plus de 2 allèles sont présents (cas de l'haplotype 2 dans l'approche DW ou de l'haplotype 1 dans l'approche LD, Figure 26), l'opération est identique.

Une fois les pseudo-SNPs construits, la matrice de parenté génomique (VanRaden, 2008) est définie comme:

$$\boldsymbol{G_{pseudo-SNP}} = 0{,}95 * \frac{\boldsymbol{Z_{pseudo-SNP} Z'_{pseudo-SNP}}}{\sum_{i=1}^{m} p_i * (1 - p_i)} + 0{,}05 * \boldsymbol{A_{22}}$$

où $\boldsymbol{Z_{\text{pseudo-SNP}}}$ sera la matrice des pseudo-SNPs. Par extension, on peut également utiliser ces haplotypes dans la méthode WssGBLUP avec comme matrice de parenté génomique pondérée ($\boldsymbol{G^*_{pseudo-SNP}}$) :

$$\boldsymbol{G^*_{pseudo-SNP}} = 0{,}95 * \frac{\boldsymbol{Z_{pseudo-SNP} D Z'_{pseudo-SNP}}}{\sum_{i=1}^{m} p_i * (1 - p_i)} + 0{,}05 * \boldsymbol{A_{22}}$$

Les étapes qui consistent à estimer les effets des haplotypes et les poids associés sont identiques à la méthode WssGBLUP présentée au paragraphe 1.1 (Chapitre 2).

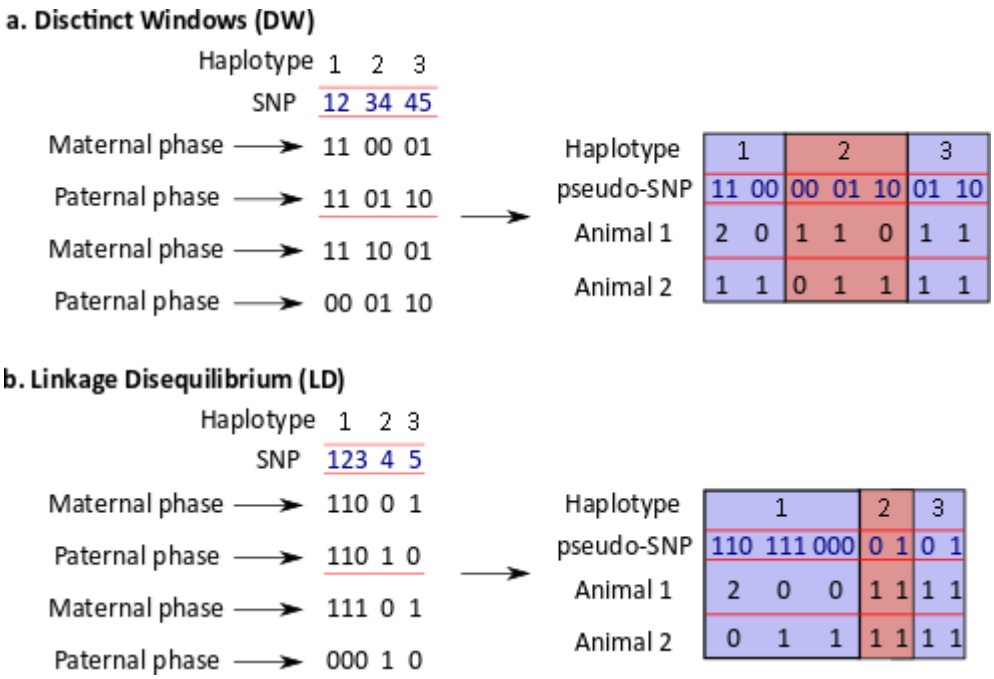


Figure 26. Conversion des haplotypes en pseudo-SNPs selon les méthodes DW (a) et LD (b)

Zahra (2018) a analysé un ensemble de 57 caractères avec des héritabilités variables, allant de 0,003 à 0,53 pour la race Holstein (Figure 27). Les haplotypes ont été construits avec la méthode DW en considérant des longueurs d'haplotypes de 5, 10, 15 et 20 SNPs consécutifs. Ils ont comparé les précisions avec un GBLUP pseudo-SNPs, dont les haplotypes ont des longueurs de 5, 10, 15 et 20 SNPs (G_{hap5}, G_{hap10}, G_{hap15}, G_{hap20}) et un GBLUP utilisant des SNPs (G_{SNP}). Ils ont également comparé les précisions des 57 caractères en les classant selon leurs héritabilités. La classe H1 contenait tous les caractères avec une héritabilité comprise entre 0 et 0,15, la classe H2 contenait tous les caractères avec une héritabilité comprise entre 0,15 et 0,30. La classe H3 était réservée aux caractères avec une héritabilité supérieure à 0,30. Les précisions des évaluations avec G_{hap5} étaient supérieures aux précisions avec G_{hap10}, supérieures aux précisions avec G_{hap15} elles-mêmes supérieures à G_{hap20}. Ces différences sont faibles pour les caractères de la classe H1 et H2, en revanche elles sont plus importantes pour les caractères de la classe H3. L'utilisation de la matrice G_{hap5} améliore légèrement les précisions des évaluations par rapport aux évaluations G_{SNP} pour des caractères appartenant à la classe H3.

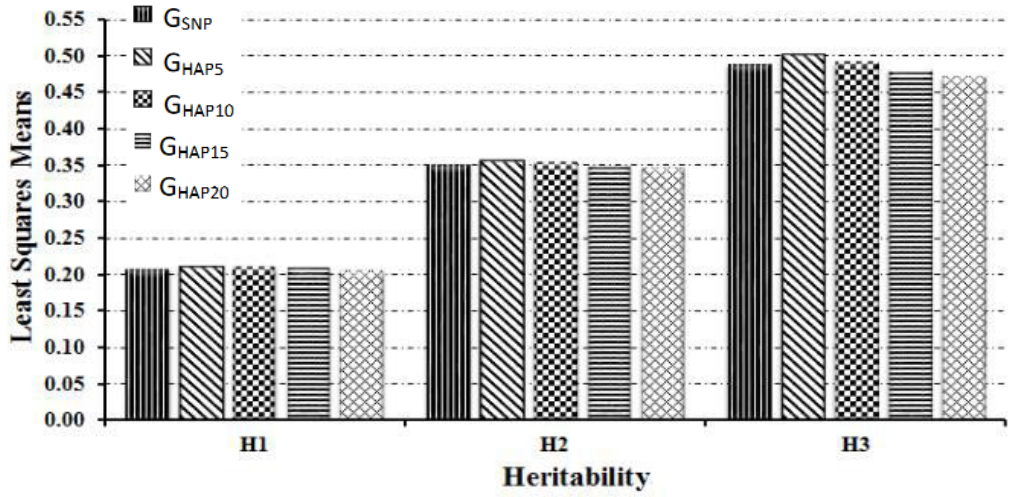


Figure 27. Précisions des évaluations génomiques GBLUP et des évaluations génomiques haplotypique (DW) avec l'utilisation de pseudo-SNPs chez la race Holstein (Zahra, 2018). Le groupe H1 comprend des caractères avec $h^2 \in [0, 0,15]$, $h^2 \in]0,15$ 0,30] pour le groupe H2 et $h^2 > 0,30$ pour le groupe H3. Les méthodes utilisées sont un GBLUP avec des SNP (G_{SNP}), un GBLUP avec des pseudo-SNP (G_{hap5}, G_{hap10} G_{hap15} G_{hap20}) où les pseudo-SNP sont construit avec des fenêtres de 5, 10, 15 et 20 SNPs respectivement.

Chapitre 3 : Précisions des évaluations génomiques avec le ssGBLUP

Depuis les travaux de thèse de Carillier-Jacquin, (2015), de nouveaux génotypages sont disponibles et ont enrichi la population de référence. Des évaluations génétiques BLUP et ssGBLUP avec des populations d'apprentissage et de validation de tailles différentes ont ainsi pu être mises en œuvre. L'objectif était d'étudier l'évolution des précisions génétiques et génomiques en fonction de la taille de ces populations. Nous nous sommes également intéressés aux hyperparamètres du ssGBLUP et leurs effets sur les précisions et les biais des évaluations génomiques en caprins. Les données utilisées sont celles présentées dans le chapitre 1, c'est-à-dire les données des évaluations génétiques officielles de janvier 2016.

1. Effet de la taille de la population de référence sur les précisions des évaluations

1.1. Construction des populations d'apprentissage et de validation

Les populations d'apprentissage et de validation ont été construites en fonction de l'année de naissance des animaux. Pour définir ces populations, nous nous sommes basés sur les animaux génotypés. Nous avons testé 5 situations différentes (A, B, C, D et E) (Figure 28). Les populations A sont équivalentes aux populations utilisées par Carillier et al. (2013), avec une population d'apprentissage constituée d'animaux nés entre 1993 et 2005, et une population de validation comprenant les animaux nés entre 2006 et 2009. Nous avons augmenté les populations d'apprentissage et de validation d'une année jusqu'au scénario où la population d'apprentissage comprenait tous les animaux nés entre 1993 et 2008 et tous les animaux nés entre 2009 et 2012 pour la population de validation (scénario B, C et D). Un dernier scénario (E) a été testé avec une taille de la population de validation plus conséquente : la population d'apprentissage comprenait les animaux nés entre 1993 et 2007 tandis que la population de validation comprenait les animaux nés entre 2008 et 2012. Nous avons conservé les phénotypes des femelles jusqu'en 2008 pour le scénario A, 2009 pour le scénario B, 2010 pour le scénario C, 2011 pour le scénario D et 2010 pour le scénario E. Les performances des filles des mâles de validation ont été exclues. Les généalogies ont été entièrement conservées pour les animaux de la population d'apprentissage. La descendance des animaux de la population de validation a été exclue du pedigree.

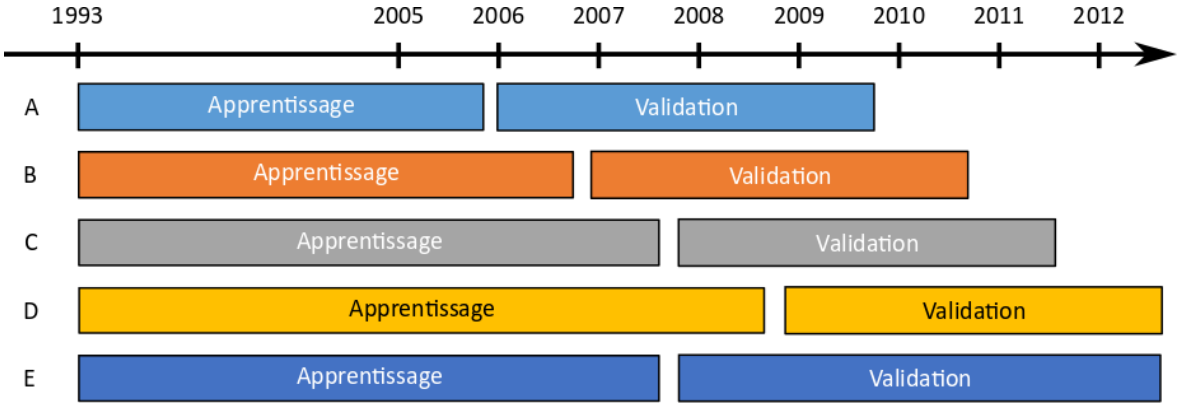


Figure 28. Composition des populations d'apprentissages et de validations pour les évaluations génomiques caprines en fonction de l'année de naissance

Les évaluations BLUP et ssGBLUP ont été réalisées avec une population multirace (données Alpine et Saanen analysées conjointement) et des populations intra-race (Alpine et Saanen analysées séparément). Les nombres de phénotypes, d'animaux dans le pedigree et d'animaux dans les populations d'apprentissage et de validation pour chaque scénario sont présentés dans le Tableau 17 pour les analyses multirace, le Tableau 18 pour les analyses en race Alpine et le Tableau 19 pour les analyses en race Saanen.

Tableau 17. Nombre de phénotypes, pedigree et génotypes pour chaque scénario considéré pour les analyses multirace

				Génotypes			
				Apprentissage		Validation	
Populations	**Caractères**	**Phénotypes**	**Pedigree**	**Males**	**Femelles**	**Males**	**Femelles**
A	Production laitière	6 596 647	2 505 111	421	0	258	2050
	Morphologie mamelle	212 166	412 535				
	Cellules somatiques	1 847 085	1 242 371				
B	Production laitière	6 923 445	2 628 544	485	0	267	2050
	Morphologie mamelle	235 093	446 581				
	Cellules somatiques	2 067 867	1 351 698				
C	Production laitière	7 254 649	2 755 783	554	0	273	2050
	Morphologie mamelle	257 067	478 469				
	Cellules somatiques	2 293 636	1 463 478				
D	Production laitière	7 576 556	2 875 984	610	427	295	1623
	Morphologie mamelle	280 985	512 529				
	Cellules somatiques	2 514 737	1 570 251				
E	Production laitière	7 254 649	2 755 861	554	0	351	2050
	Morphologie mamelle	257 067	478 547				
	Cellules somatiques	2 293 636	1 463 556				

Entre le scénario A et D, les nombre de phénotypes et d'animaux dans le pedigree augmentent avec l'arrivée de nouveaux contrôles chaque année que ce soit pour les analyses multirace, Alpine ou Saanen. Pour le scénario E, la population d'apprentissage est la même que pour le scénario C, le nombre de phénotypes est identique. Le pedigree diffère légèrement entre les scénarios C et E car les populations de validation sont différentes.

Tableau 18. Nombre de phénotypes, pedigree et génotypes pour chaque scénario considéré pour les analyses en race Alpine

				Génotypes			
				Apprentissage		Validation	
Populations	**Caractères**	**Phénotypes**	**Pedigree**	**Males**	**Femelles**	**Males**	**Femelles**
A	Production laitière	3 487 189	1 313 706	227	0	156	1237
	Morphologie mamelle	123 833	235 431				
	Cellules somatiques	1 013 581	668 310				
B	Production laitière	3 663 575	1 379 382	267	0	159	1237
	Morphologie mamelle	137 613	255 494				
	Cellules somatiques	1 135 855	727 551				
C	Production laitière	3 844 314	1 446 253	307	0	162	1237
	Morphologie mamelle	150 676	290 613				
	Cellules somatiques	1 262 187	788 533				
D	Production laitière	4 021 755	1 512 074	342	246	170	991
	Morphologie mamelle	166 088	296 103				
	Cellules somatiques	1 386 923	848 866				
E	Production laitière	3 844 314	1 446 296	307	0	205	1237
	Morphologie mamelle	150 676	290 656				
	Cellules somatiques	1 262 187	788 576				

Les effectifs des mâles augmentent dans les populations d'apprentissage et de validation entre le scénario A et D (pour les analyses multirace, Alpine et Saanen), dû à l'augmentation des animaux candidats à la sélection génotypés chaque année. On note dans le scénario D que les femelles du dispositif QTL nées en 2009 se retrouvent dans la population d'apprentissage (246 Alpine et 181 Saanen). La population d'apprentissage est la même pour les scénarios C et E, et la taille de la population de validation est plus conséquente pour le scénario E (43 Alpine et 35 Saanen en plus).

Tableau 19. Nombre de phénotypes, pedigree et génotypes pour chaque scénario considéré pour les analyses en race Saanen

				Génotypes			
				Apprentissage		Validation	
Populations	**Caractères**	**Phénotypes**	**Pedigree**	**Males**	**Femelles**	**Males**	**Femelles**
A	Production laitière	2 643 672	987 539	194	0	102	813
	Morphologie mamelle	85 201	174 934				
	Cellules somatiques	833 504	548 883				
B	Production laitière	2 783 459	1 041 561	218	0	108	813
	Morphologie mamelle	94 189	188 180				
	Cellules somatiques	932 012	598 133				
C	Production laitière	2 923 419	1 097 349	247	0	111	813
	Morphologie mamelle	102 967	206 119				
	Cellules somatiques	1 031 450	648 426				
D	Production laitière	3 057 810	1 148 286	268	181	125	632
	Morphologie mamelle	111 334	212 504				
	Cellules somatiques	1 127 817	694 457				
E	Production laitière	2 923 419	1 097 384	247	0	146	813
	Morphologie mamelle	102 967	206 154				
	Cellules somatiques	1 031 450	648 461				

1.2. Précisions des évaluations avec différentes populations de référence

1.2.1. Comparaison des précisions des évaluations BLUP et ssGBLUP

1.2.1.1. Précisions des évaluations en population multirace

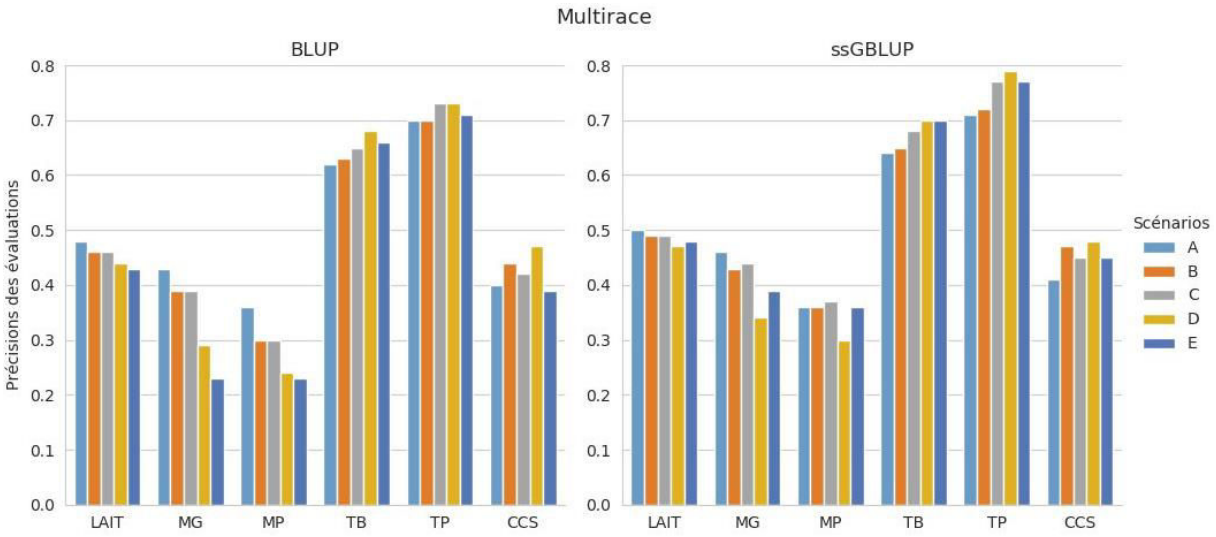


Figure 29. Précisions des évaluations génomiques multirace BLUP et ssGBLUP des caractères de productions laitières et le comptage des cellules somatiques pour différentes populations d'apprentissages et de validations

La Figure 29 présente les précisions des évaluations en population multirace BLUP et ssGBLUP pour les caractères de productions laitières et CCS, selon les différents scénarios (A, B, C, D et E). En moyenne, le ssGBLUP est plus précis que le BLUP de +2 points pour le scénario A, de +3 points pour le scénario B et de +4 points pour les scénarios C et D. Les évolutions des précisions entre les scénarios A, B, C et D permettent d'identifier trois profils de caractères : (i) les précisions augmentent entre les scénarios A et D, (ii) les précisions des évaluations restent relativement constantes entre les scénarios A et D et (iii) les précisions diminuent entre les scénarios A et D. Les résultats du ssGBLUP montrent que le TB et le TP appartiennent au profil (i). Les précisions des évaluations pour le TB sont passées de 0,64 à 0,70 entre les scénarios A et D, et de 0,71 à 0,79 pour le TP. Le LAIT et CCS appartiennent au

profil (ii). Pour ces caractères, la précision moyenne entre les scénarios A et D est de 0,48 pour le LAIT et de 0,45 pour CCS. On retrouve les caractères MG et MP dans le profil (iii). Les précisions chutent entre les populations A et D de 0,46 à 0,34 pour la MG et de 0,36 à 0,30 pour la MP. Les résultats des évaluations BLUP montrent les mêmes tendances que les résultats du ssGBLUP.

Pour le scénario E, le ssGBLUP est plus précis que le BLUP en moyenne de +6 points pour les caractères de productions laitières et CCS. Les précisions des évaluations obtenues avec les scénarios C et E sont comparées car les populations d'apprentissage sont les mêmes. Les précisions sont plus faibles pour le scénario E pour les caractères MG (-5 points) et MP (-1 points) ; sont identiques pour le LAIT (0,48), le TP (0,77) et CCS (0,45). Pour le TB, les précisions sont supérieures pour le scénario E (0,70 contre 0,68). Les résultats sont différents pour le BLUP : on observe une baisse des précisions avec le scénario E comparé au scénario C pour tous les caractères (-3 points pour le LAIT, -16 points pour la MG, -7 points pour la MP, -2 points pour le TP et -2 points pour CCS) sauf pour le TB (+1 point). Nous avons aussi comparé les scénarios D et E car le scénario D a la population d'apprentissage la plus importante. Les précisions des évaluations ssGBLUP sont plus élevées pour le scénario E pour le LAIT (+2 points), la MG (+5 points) et la MP (+6 points) ; plus faibles avec le scénario E pour le TP (-1 point) et CCS (-3 points). Pour les évaluations BLUP, tous les caractères ont des précisions plus faibles avec le scénario E.

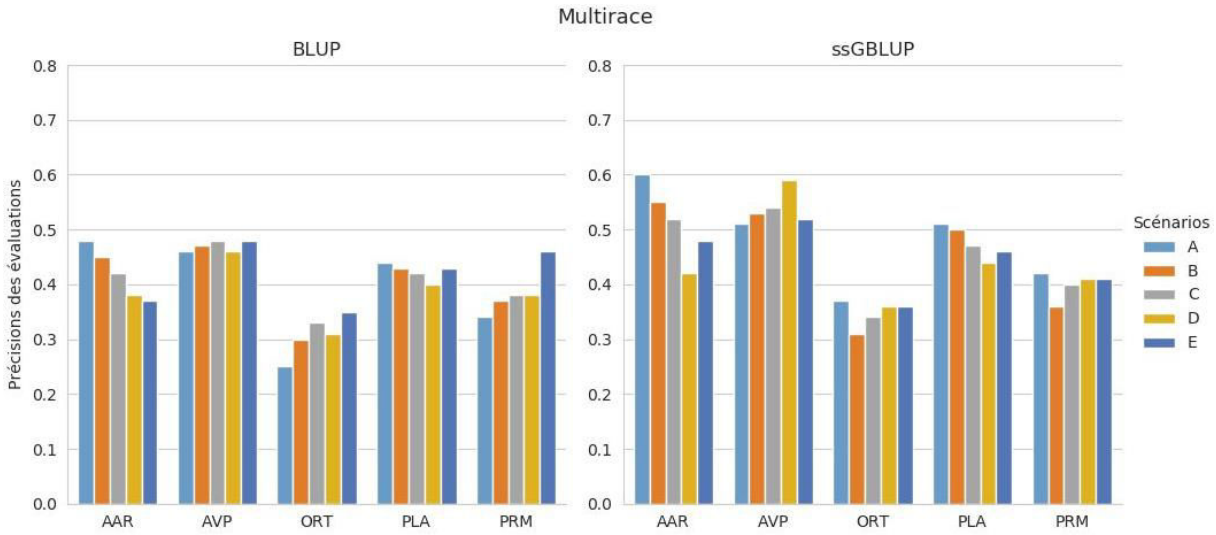


Figure 30. Précisions des évaluations génomiques multirace BLUP et ssGBLUP des caractères de morphologie de la mamelle pour différentes populations d'apprentissage et de validation

La Figure 30 présente les précisions des évaluations multirace BLUP et ssGBLUP pour les caractères de morphologie de la mamelle pour les différents scénarios (A, B, C, D et E). En moyenne, le ssGBLUP est plus précis que le BLUP de +9 points pour le scénario A, de +5 points pour les scénarios B et C et de +6 points pour le scénario D. Pour le ssGBLUP, l'AVP est le seul caractère qui appartient au profil (i) avec des précisions allant de 0,51 à 0,59 entre les scénarios A et D. Les caractères ORT et PRM appartiennent au profil (ii) avec des précisions moyennes de 0,35 et de 0,40 respectivement entre les scénarios A et D. AAR et PLA appartiennent au profil (iii) avec des chutes de précisions de 0,60 à 0,42 pour AAR et de 0,51 à 0,44 pour PLA entre les scénarios A et D.

Pour le scénario E, le ssGBLUP est plus précis que le BLUP en moyenne de +3 points. Les écarts les plus importants entre les scénarios C et E sont observés pour AAR (0,52 contre 0,48 respectivement). Pour le caractère AVP, les précisions sont plus élevées pour le scénario C

(0,54 contre 0,52). On observe des précisions identiques pour PLA (0,46) et plus fortes pour le scénario E pour PRM (+2 points) et ORT (+2 points). Des tendances similaires sont observées pour les évaluations BLUP entre les scénarios C et E. Entre les scénarios D et E, les précisions ssGBLUP sont similaires pour PRM (0,41). On observe une chute importante de la précision pour AVP (0,59 pour le scénario D contre 0,52 pour le scénario E). Les précisions sont améliorées avec le scénario E pour AAR (0,42 contre 0,48), PLA (0,42 contre 0,44) et ORT (0,35 contre 0,36). Pour le BLUP, de meilleures précisions sont observées pour le scénario D pour tous les caractères sauf pour AAR. Les gains sont compris entre +3 points et +8 points.

1.2.1.2. Précisions des évaluations en race Alpine

Les profils identifiés dans le paragraphe 1.2.1.1 (Chapitre 3) seront réutilisés pour les analyses en race Alpine.

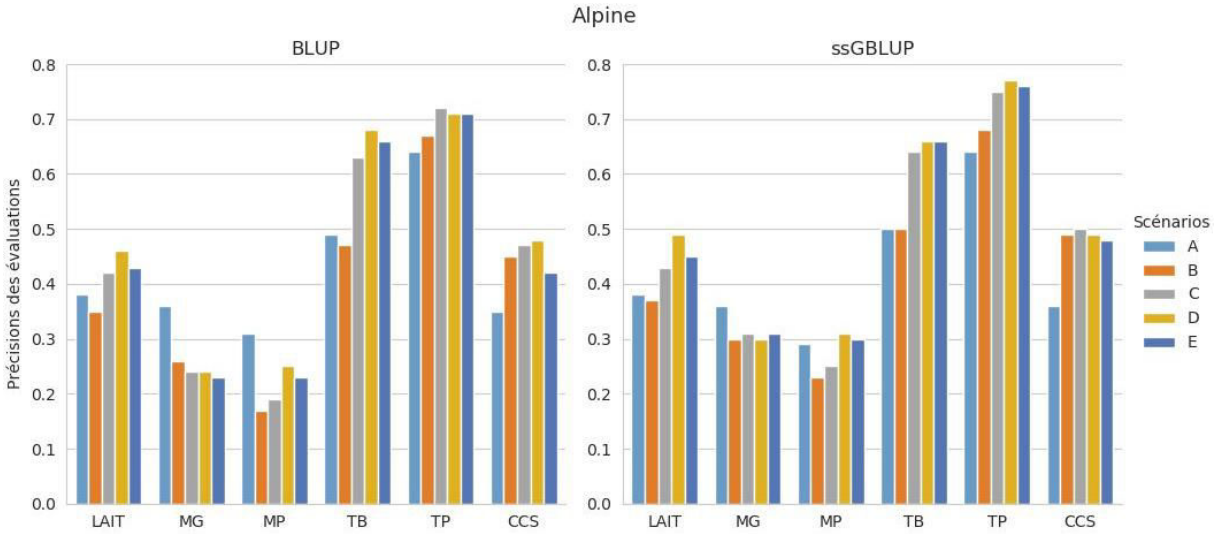


Figure 31. Précisions des évaluations génomiques Alpine BLUP et ssGBLUP des caractères de productions laitières et le comptage des cellules somatiques pour différentes populations d'apprentissages et de validations

Les évaluations ssGBLUP Alpine pour les caractères de production laitière et CCS sont en moyenne plus précises que les évaluations BLUP de +3 points pour les scénarios B, C et D (Figure 31). Les évaluations ssGBLUP avec le scénario A sont aussi précises que les évaluations BLUP. Pour le ssGBLUP, les caractères LAIT, TB, et CCS appartiennent au profil (i). Les précisions augmentent de 0,38 à 0,49 pour le LAIT, de 0,50 à 0,66 pour le TB et de 0,36 à 0,48 pour CCS entre les scénarios A et D. Pour la MP, les précisions s'améliorent légèrement entre les scénarios A et D (0,29 et 0,31 respectivement) et sont plus faibles pour les populations B et C (0,23 et 0,25 respectivement). La MG se classe dans le profil (iii) avec une précision passant de 0,36 à 0,31. Toutefois, les précisions pour les scénarios B, C et D sont proches (0,31; 0,30 et 0,31). Les mêmes tendances sont retrouvées pour les analyses BLUP.

Pour le scénario E, les évaluations ssGBLUP sont en moyenne plus précises de +3 points par rapport aux évaluations BLUP. On observe de meilleures précisions pour le scénario E comparé au scénario C pour le LAIT (+2 points), la MP (+5 points), le TB (+2 points), le TP (+1 point) sauf pour CCS (-2 points) et les précisions sont identiques pour la MG (0,31). Les précisions sont inférieures pour le scénario E comparé au scénario D : -4 points pour le LAIT et -1 point pour la MP, le TB, le TP et CCS. On observe les mêmes tendances pour les évaluations BLUP entre les scénarios C, D et E que les évaluations ssGBLUP.

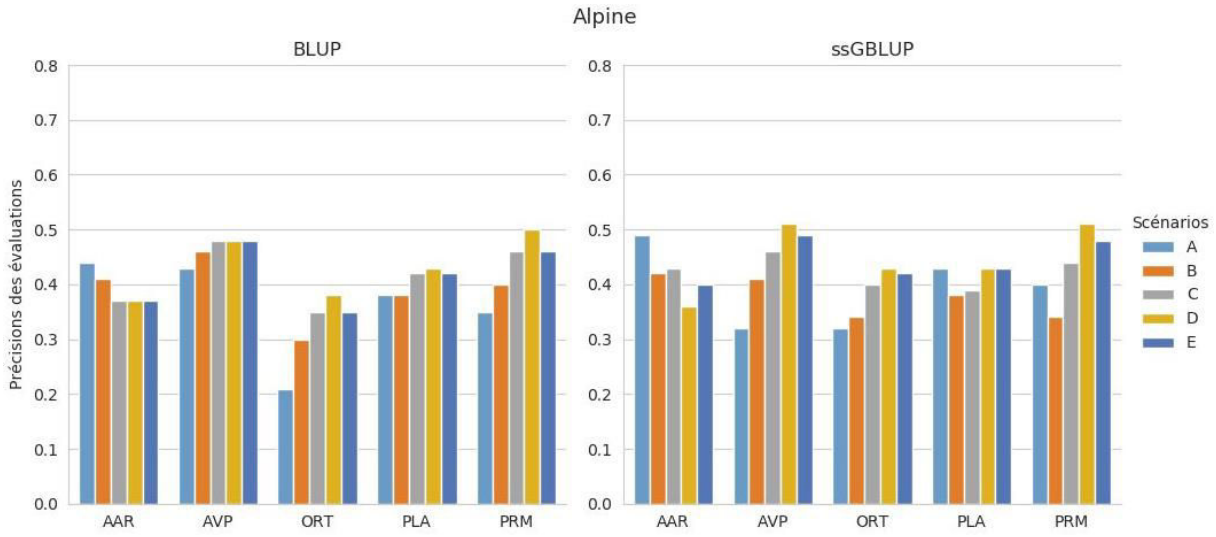


Figure 32. Précisions des évaluations génomiques Alpine BLUP et ssGBLUP des caractères de morphologie de la mamelle pour différentes populations d'apprentissages et de validations

Pour les caractères de morphologie de la mamelle, le ssGBLUP améliore en moyenne les précisions des évaluations comparées au BLUP (Figure 32) pour les scénarios A (+1 point), C (+1 point) et D (+2 points) mais est moins précis pour le scénario B (-1 point). Pour le ssGBLUP, nous retrouvons les caractères AVP, ORT et PRM dans le profil (i) avec des précisions allant de 0,32 à 0,51, de 0, 32 à 0,46 et de 0,40 à 0,51 respectivement. Le caractère PLA appartient au profil (ii) avec une précision moyenne de 0,41. Enfin, le caractère AAR correspond au profil (iii) avec une précision chutant de 0,49 à 0,36. On retrouve des profils similaires pour les évaluations BLUP.

Avec le scénario E, les évaluations ssGBLUP sont en moyenne plus précises de +3 points que les évaluations BLUP. Avec le ssGBLUP, on observe de meilleures précisions pour le scénario E comparé au scénario C pour AVP (+3 points), ORT (+2 points), PLA (+4 points) et PRM (+4 points). Les précisions des évaluations BLUP entre les scénarios C et E sont identiques pour tous les caractères. Avec le ssGBLUP, on observe des précisions plus faibles pour le scénario E comparé au scénario D pour tous les caractères de morphologie de la mamelle (entre –1 points et –3 points) sauf pour AAR (+4 points). Dans le cas des évaluations BLUP, les précisions entre le scénario D et E sont identiques pour AAR (0,37) et AVP (0,48). Pour les caractères ORT, PLA et PRM, les précisions sont plus faibles avec le scénario E que le scénario D (entre -1 et -4 points).

1.2.1.3. Précisions des évaluations en race Saanen

Pour les évaluations en race Saanen, nous utiliserons les profils identifiés dans le paragraphe 1.2.1.1 (Chapitre 3).

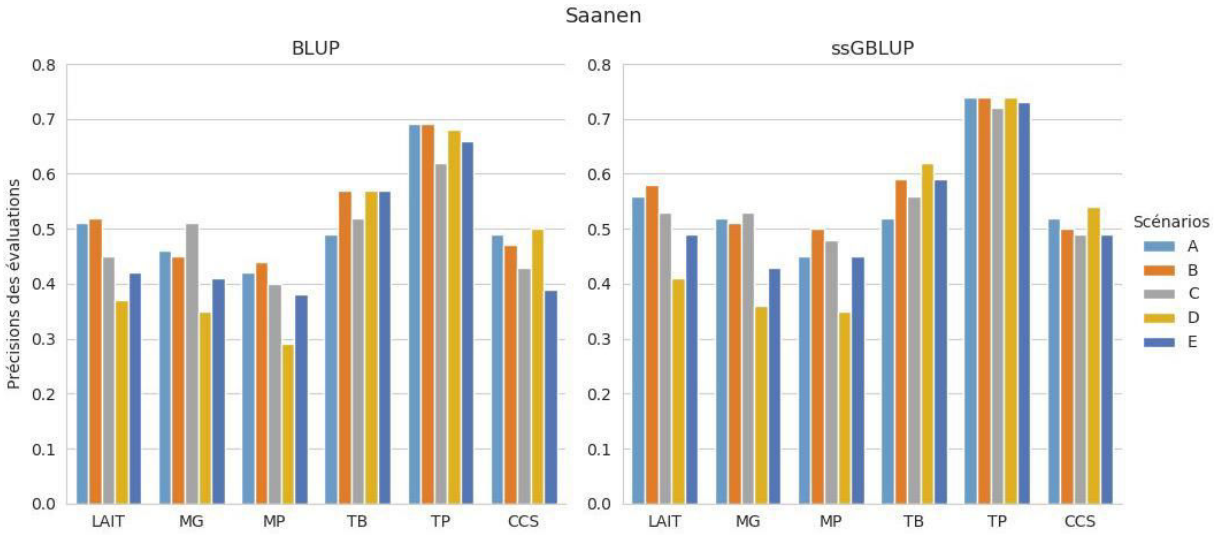


Figure 33. Précisions des évaluations génétiques Saanen (BLUP et ssGBLUP) des caractères de productions laitières et le comptage des cellules somatiques avec différentes populations d'apprentissages et de validations

La Figure 33 présente les précisions des évaluations Saanen BLUP et ssGBLUP pour les caractères de productions laitières et les CCS. Les tendances observées dans les analyses multirace et Alpine se retrouvent pour la race Saanen. Les évaluations ssGBLUP sont en moyenne plus précises que les évaluations BLUP pour les scénarios A, B, C et D (+4 points, +5 points, +6 points et +4 points respectivement). Avec le ssGBLUP, le TB se classe dans le profil (i) avec des précisions passant de 0,52 à 0,62 entre les populations A et D. Le TP et les CCS se classent dans le profil (ii) avec des précisions en moyenne de 0,73 et 0,51 respectivement. Le LAIT, la MG et la MP appartiennent au profil (iii), cependant, les précisions restent proches pour les scénarios A, B et C (en moyenne de 0,53; 0,51 et 0,48 respectivement) et chutent pour le scénario D (0,41; 0,36; 0,35 respectivement). Les observations pour les évaluations BLUP sont similaires aux évaluations ssGBLUP.

Les précisions des évaluations ssGBLUP pour le scénario E sont en moyenne supérieures de +6 points par rapport aux évaluations BLUP. Pour les évaluations ssGBLUP, les précisions avec le scénario E pour le LAIT, la MG et la MP sont inférieures aux précisions avec le scénario C (-4 points, -10 points et -3 points respectivement). Pour le TB et le TP, les précisions sont plus élevées pour le scénario E (+3 et +1 points respectivement). Les précisions sont plus élevées pour le scénario E comparé au scénario D pour le LAIT (+8 points), la MG (+7 points), la MP (+10 points). On observe une tendance inverse entre le scénario D et E pour le TB (-3 points), le TP (-1 points) et les CCS (-5 points). Les résultats des évaluations BLUP suivent les mêmes tendances que les évaluations ssGBLUP.

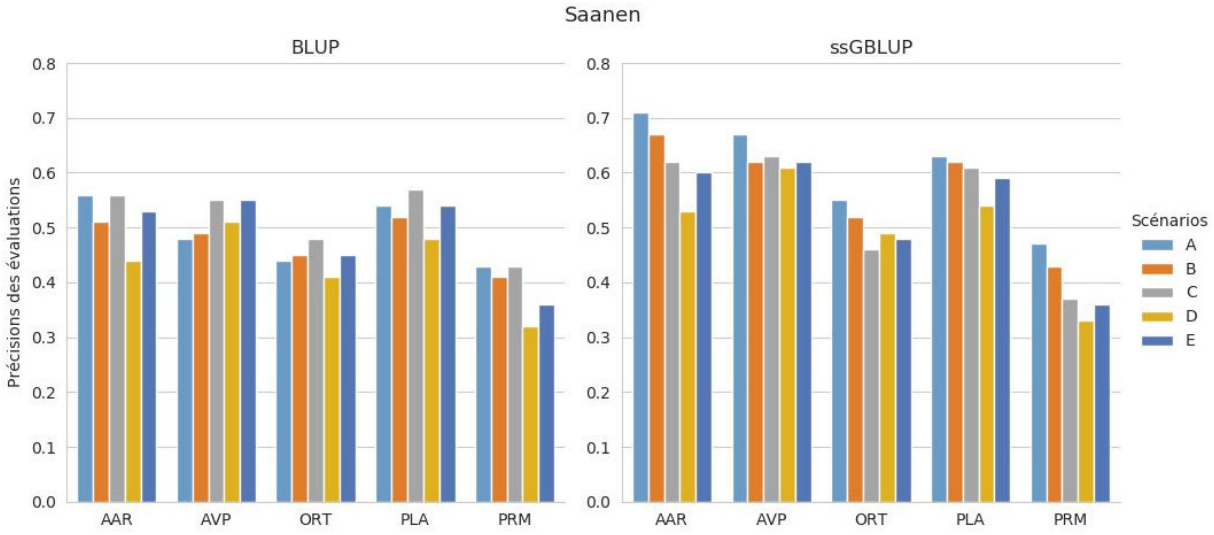


Figure 34. Précisions des évaluations génétiques Saanen (BLUP et ssGBLUP) des caractères de morphologie de la mamelle avec différentes populations d'apprentissages et de validations

Les évaluations ssGBLUP sont plus précises que les évaluations BLUP pour les caractères de morphologie de la mamelle pour la race Saanen (Figure 34). Cette supériorité est de +12 points pour le scénario A, de +9 points pour le scénario B, de +2 points pour le scénario C et de +7 points pour le scénario D. Tous les caractères de morphologie de la mamelle avec le ssGBLUP appartiennent au profil (iii), les précisions des évaluations chutent progressivement entre les populations A et D. Les précisions chutent de 18 points pour AAR, de 6 points pour AVP, de 6 points pour ORT, de 9 points pour PLA et de 14 points pour PRM.

Les évaluations ssGBLUP pour le scénario E sont plus précises que les évaluations BLUP en moyenne de +5 points. Pour le ssGBLUP, la comparaison entre les scénarios C et E montrent que seules les précisions pour ORT sont plus élevées pour le scénario E (+3 points) ; pour les autres caractères les précisions sont identiques ou inférieures (entre 0 et -2 points). Pour les évaluations BLUP, les tendances observées sont les mêmes que pour les évaluations ssGBLUP. Les précisions avec le ssGBLUP sont plus élevées pour le scénario E comparé au scénario D pour AAR (+7 points), l'AVP (+1 point), le PLA (+5 points) et PRM (+3 points). Peu de différence existe entre le scénario D et E pour ORT (0,49 et 0,48 respectivement). Pour les évaluations BLUP, les précisions sont supérieures pour tous les caractères avec le scénario E comparé au scénario D (entre +3 points et +8 points).

1.2.2. Évolution de la précision des évaluations ssGBLUP en fonction de l'année de naissance des animaux pour les analyses multirace

La baisse des précisions pour certains caractères n'était pas attendue, nous avons alors analysé les précisions en fonction de l'année de naissance des animaux pour voir s'il n'y avait pas une année en particulier pour laquelle les précisions étaient faibles. Pour les évaluations ssGBLUP multirace, Alpine et Saanen, nous avons étudié les corrélations entre GEBV et DYD par année de naissance des mâles de la population de référence pour chacun des scénarios A, B, C et D pour l'ensemble des 11 caractères. Les analyses multirace, Alpine et Saanen présentant des résultats similaires, seuls les résultats en multirace sont illustrés. De même, seuls sont présentés les résultats obtenus sur les caractères Lait et MP, les autres caractères présentant des tendances similaires à ces 2 caractères.

Les corrélations entre GEBV et DYD en analyse multirace pour le Lait sont représentées pour les scénarios A, B, C et D sur la Figure 35. Pour les animaux de la population d'apprentissage, comme attendu, les corrélations sont très proches de 1 (jusqu'en 2005, 2006,

2007, 2008 pour les scénarios A, B, C et D respectivement). Pour les animaux de la population de validation, les corrélations chutent et d'autant plus que l'année considérée est éloignée de la population de validation. Ce phénomène est bien illustré avec le scénario D (0,52 pour 2009, 0,50 pour 2010, 0,42 pour 2011 et 0,45 pour 2012). On peut toutefois observer des variations d'une année à l'autre pour les autres scénarios. Pour le scénario A, on observe des corrélations égales à 0,41 en 2006, 0,56 en 2007, 0,42 en 2008 et 0,51 en 2009. On remarque également que les corrélations pour une même année s'améliorent lorsque la population d'apprentissage devient plus grande (passage du scénario A au scénario B, etc.). Des résultats similaires ont été obtenus sur la MG, le TB, le TP, AVP, PRM, PLA, AAR et CCS.

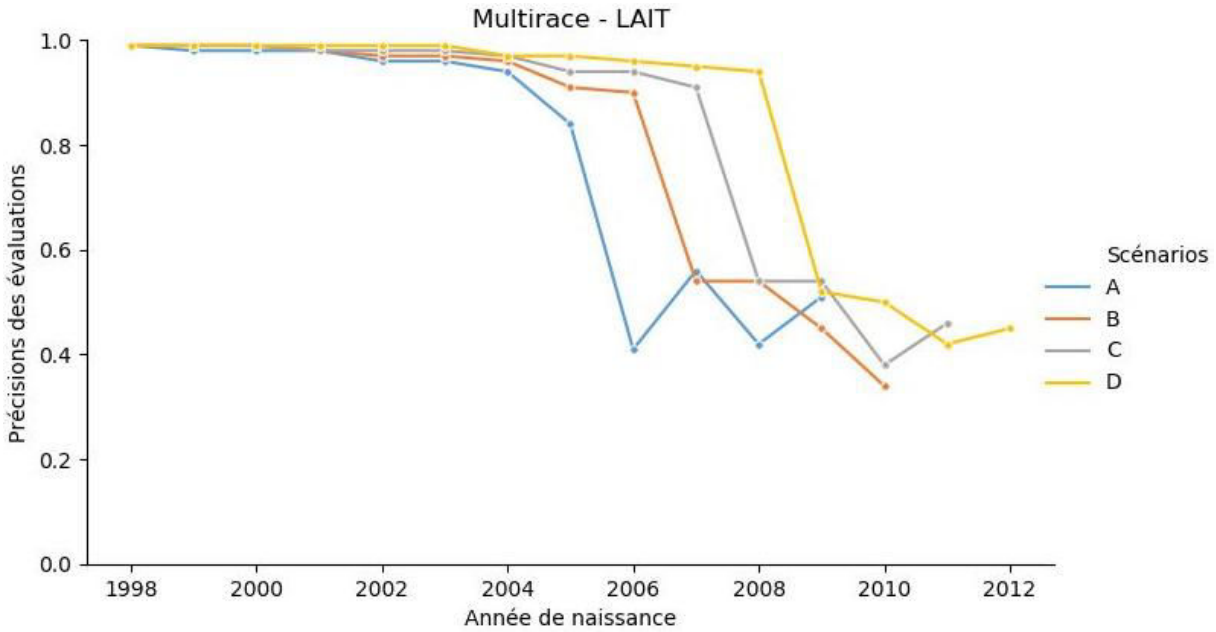


Figure 35. Corrélations entre GEBV et DYD avec le ssGBLUP multirace selon l'année de naissance des animaux pour le LAIT pour les scénarios A, B, C et D

On observe des variations de corrélations entre années beaucoup plus importantes sur la MP et l'AAR (Figure 36). Les corrélations restent proches de 1 pour les animaux de la population d'apprentissage. En revanche, les corrélations présentent d'importantes variations pour les animaux de la population de validation, notamment pour les années 2006 et 2009. Pour le scénario A, les corrélations chutent à 0,17 et 0,13 pour 2006 et 2009 alors que les corrélations sont de 0,48 et 0,43 pour les années 2007 et 2008 respectivement. On observe également une forte baisse des corrélations pour l'année 2009 pour les populations B (0,01), C (0,15) et D (0,16).

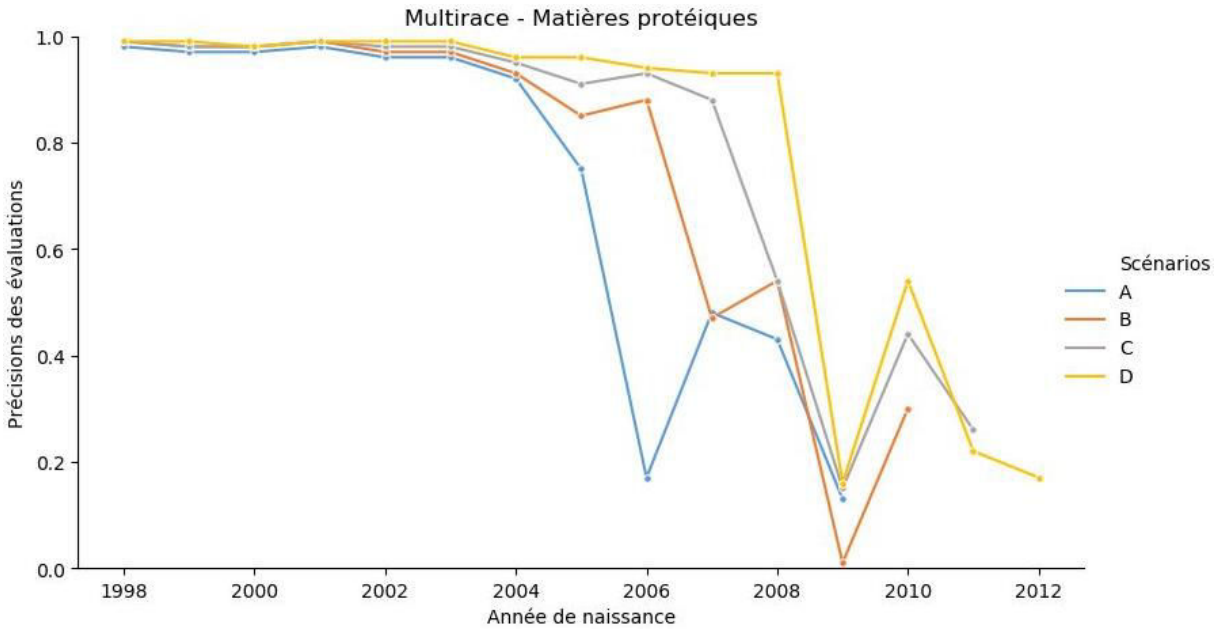


Figure 36. Corrélations entre GEBV et DYD avec le ssGBLUP multirace selon l'année de naissance des animaux pour la MP pour les scénarios A, B, C et D

1.3. Discussion

Quel que soit le scénario (A, B, C, D et E), les précisions obtenues avec le ssGBLUP sont supérieures à celles du BLUP. Ces résultats sont en accord avec les conclusions de Carillier-Jacquin, (2015). Pour les analyses multirace, ils ont observé des précisions supérieures de +16 % à +91 % avec un ssGBLUP par rapport à un BLUP excepté pour la MP (-6 %), les caractères de morphologie de la mamelle ayant les plus fortes augmentations de précisions. Les gains entre BLUP et ssGBLUP sont plus faibles dans notre étude. Carillier et al., (2013) ont réalisé des évaluations BLUP à partir de pseudo-phénotypes pour les mâles (des DYD), alors que nos évaluations BLUP ont été réalisées avec l'ensemble des phénotypes des femelles. Il en résulte que les précisions obtenues pour les évaluations BLUP ne sont pas comparables dans les 2 études.

L'objectif de cette première étude était de comparer les précisions des évaluations ssGBLUP obtenues avec le scénario A (populations comparables à celles étudiées dans Carillier-Jacquin, (2015)) et avec une population de référence plus importante (scénario D comprenant une population d'apprentissage augmentée de 189 mâles et de 427 femelles dans le cas d'analyse multirace). Il était attendu qu'entre les scénarios A et D, les précisions restent stables ou s'améliorent avec une population de référence plus importante. Les résultats montrent que pour certains caractères, des baisses importantes des précisions sont observées pour tous les caractères en Saanen sauf pour les taux (TB et TP) et les CCS ; en MG, MP et AAR en Alpine et multirace. Cela pose question, en particulier pour certains caractères d'intérêt important tel que la MP pour la filière caprine (Hélène Larroque, communication personnelle). L'origine de cette baisse a donc été recherchée par une analyse détaillée des fichiers de performances, reproducteurs et de pedigree. Cependant, nous n'avons pas pu mettre en évidence des différences de traitements, de sélection particulière, de nombre de descendants par mâle, de niveau moyen d'apparentement entre les millésimes. Les résultats et tendances observés avec le ssGBLUP sont également observés dans les analyses BLUP. La méthode d'évaluation n'est donc pas à l'origine de ces baisses de précisions.

Une analyse sur l'apparentement entre la population d'apprentissage et la population de validation a été réalisée pour chacun des scénarios. Comme mentionné au paragraphe 3.4.3

(Chapitre 1), plus l'apparentement est fort entre les populations d'apprentissage et de validation, plus les précisions sont élevées. Nous avons calculé la moyenne des coefficients hors-diagonaux de la matrice de parenté génomique entre les animaux de la population d'apprentissage, entre les animaux de l'apprentissage et de validation et entre les animaux de la population de validation (résultats non montrés). Les apparentements moyens sont similaires pour l'ensemble des scénarios (A, B, C, D ou E). De plus, aucune modification n'a été apportée au schéma de sélection caprine ces dernières années pouvant justifier une baisse de l'apparentement entre populations d'apprentissage et de validation. Ce critère ne semble pas être à l'origine des variations de précisions observées.

En comparant les scénarios A, B, C et D, les précisions des évaluations évoluent (soit à la baisse, soit à la hausse). Ces variations ne sont pas spécifiques au scénario E, mais semblent être une tendance. D'autre part, les corrélations entre GEBV et DYD des évaluations multirace ssGBLUP en fonction de l'année de naissance des animaux montrent pour la majorité des caractères que plus on s'éloigne de la population d'apprentissage et plus les précisions diminuent. Ces résultats sont conformes à ceux de la littérature. Solberg et al., (2009) ont simulé une population sur 1 000 générations (reproduction aléatoire). À la génération 1 000, la taille efficace de la population était de 100 avec un génome de 10 chromosomes (chacun de longueur 100 cM). Plusieurs densités de marqueurs ont été testées (1, 2, 4 et 8 Ne/M). À la génération 1 001, des évaluations génomiques avec un BayesB ont permis de prédire les GEBV des animaux de la génération 1002 à 1006. Les précisions ont chuté de 0,801 à 0,717 (situation similaire à une population bovine) lorsqu'ils ont prédit les GEBV des animaux les plus éloignés de la génération 1 001. Enfin, plus la densité du génome est faible et plus la chute de précision est importante. Sur des données réelles, la baisse des précisions sur plusieurs générations a été observée par Wolc et al., (2011) chez des poules pondeuses pour 16 caractères économiquement importants. Ces animaux ont été génotypés avec une puce Illumina personnalisée contenant 23 356 SNPs. Pour calculer la persistance des précisions, ils ont accumulé les données de la population d'apprentissage jusqu'à une génération donnée, la génération 1 étant la première génération sans phénotype enregistré. Ils ont ensuite prédit les GEBV des animaux de la génération 1 à 5 en utilisant un BLUP, un GBLUP, un BayesA et un BayesCπ. Entre les générations 1 à 5, les précisions chutent de 0,33 à 0,08 pour le BLUP, de 0,41 à 0,35 pour le GBLUP, de 0,43 à 0,37 pour le BayesA et de 0,43 à 0,37 pour le BayesCπ. Les mêmes évaluations ont été réalisées en accumulant les données de la population d'apprentissage (prédiction de la génération 2 avec des phénotypes enregistrés jusqu'à la génération 1,…). Les précisions pour la génération 1 sont en moyenne de 0,34 pour le BLUP, 0,42 pour le GBLUP, 0,43 pour le BayesA et de 0,43 pour le BayesCπ. Les précisions restent stables pour le BLUP et passe à 0,37 pour la génération 5. Pour les autres méthodes les précisions s'améliorent pour atteindre 0,49 pour le GBLUP, 0,53 pour le BayesA et 0,52 pour le BayesCπ. Ces résultats sont en accord globalement avec ceux que nous avons obtenus, excepté pour les caractères MP et AAR, pour lesquels des précisions proches de 0 ont été observés pour les années 2006 et 2009. Il faut toutefois rester prudent sur les résultats obtenus par millésime, car ils sont calculés avec un nombre limité d'animaux (environ 70 animaux par an pour les 2 races confondues). Suite à cette étude, nous avons choisi de conserver le scénario E comme population de référence pour la suite de nos travaux (avec une taille de population de validation plus importante, et des précisions ssGBLUP globalement plus élevées que celles observées avec le scénario D).

2. Effets des hyperparamètres de la matrice **H** sur les précisions du ssGBLUP

2.1. Présentation des différents hyperparamètres : α, ω et τ

Plusieurs auteurs ont proposé d'introduire des hyperparamètres (α, ω et τ) dans la construction de la matrice de parenté génomique **H**. Ils permettent de réduire certains biais

constatés dans les évaluations génomiques et d'améliorer la convergence des algorithmes (Martini et al., 2018). Ils sont intégrés dans la matrice $\boldsymbol{H^{-1}}$ de la façon suivante:

$$\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha \boldsymbol{G} + (1-\alpha)\boldsymbol{A_{22}})^{-1} - \omega \boldsymbol{A_{22}^{-1}} \end{bmatrix}$$

Les valeurs par défaut fixées dans le logiciel blupf90 (Misztal et al., 2002) sont : $\alpha = 0{,}95$, $\tau = 1$ et $\omega = 1$. Un $\tau > 1$ donne plus de poids à la matrice **G** tandis qu'un $\omega < 1$ donne plus de poids au pedigree (Lourenco et al., 2014). L'introduction de l'hyperparamètre α évite des problèmes de singularité de la matrice **G,** il peut également améliorer les précisions (Lourenco et al., 2014). Un $\alpha > 0.95$ signifie qu'une part plus importante de la variance polygénique n'est pas expliqué par les SNPs (Croué et Ducrocq, 2017).

Nous avons étudié l'influence des valeurs attribuées aux hyperparamètres α, ω et τ sur les précisions des évaluations génomiques ssGBLUP. Ces évaluations ont été réalisées en populations multirace, Alpine et Saanen avec le scénario E (décrit dans le paragraphe 1.1 (Chapitre 3)) et sur les caractères LAIT, TP, AAR, ORT et CCS, caractères avec des héritabilités fortes à moyennes (0,20 à 0,50) et présentant des déterminismes génétiques différents. Le LAIT, AAR et CCS sont des caractères pour lesquels un QTL sur le chromosome 19 a été détecté en Saanen mais pas en Alpine. Le TP est lui, un caractère pour lequel le gène de la *caséine* α_{s1} sur le chromosome 6 a un effet fort pour les deux races. Le caractère ORT est quant à lui polygénique pour les deux races.

Pour chaque scénario étudié, nous avons fixé 2 des hyperparamètres à leurs valeurs par défaut ($\alpha = 0{,}95$, $\tau = 1$ et $\omega = 1$), et avons fait varier la valeur du troisième hyperparamètre. Pour les hyperparamètres α et ω, nous avons testé les valeurs 0 ; 0,1 ; 0,2 ; 0,3 ; 0,4 ; 0,5 ; 0,6 ; 0,7 ; 0,8 ; 0,9 et 1. Pour l'hyperparamètre τ, nous avons testé les valeurs 0 ; 0,3 ; 0,6 ; 1 ; 1,5 ; 2 ; 2,5 ; 3 ; 4 ; 5 et 6. 4 critères ont été étudiés : les précisions des évaluations génomiques, les biais des évaluations (ou pentes), les écarts-types des GEBV dans la population de validation et le nombre d'itérations pour atteindre la convergence.

2.2. Etude des hyperparamètres α, ω et τ

2.2.1. Effets des hyperparamètres α, ω et τ sur les évaluations génomiques multirace

La Figure 37 présente les précisions des évaluations et les biais des évaluations ssGBLUP pour les analyses multirace en fonction des hyperparamètres α, ω et τ. Les hyperparamètres α et ω ont un effet limité sur les précisions des évaluations génomiques. Cette précision chute de 4 points en moyenne pour un α variant de 0,95 à 0 et de 2 points en moyenne pour un ω variant de 1 à 0. La baisse des précisions est observable pour tous les caractères. Pour l'hyperparamètre τ, les précisions des évaluations évoluent en trois étapes : elles s'améliorent, atteignent un maximum et diminuent. Avec un τ de 0, les précisions convergent vers 0 et augmentent rapidement au-delà de cette valeur. Les meilleures précisions sont atteintes avec un τ de 1 (la valeur par défaut) pour le TP (0,77), AAR (0,48) et CCS (0,45). Le maximum est atteint pour un τ de 2,5 pour le LAIT (0,50) et un τ de 0,30 pour ORT (0,38). Au-delà de ce maximum, les précisions baissent pour atteindre 0,49 pour le LAIT, 0,72 pour le TP, 0,29 pour l'ORT, 0,43 pour AAR et 0,41 pour CCS avec un τ de 6.

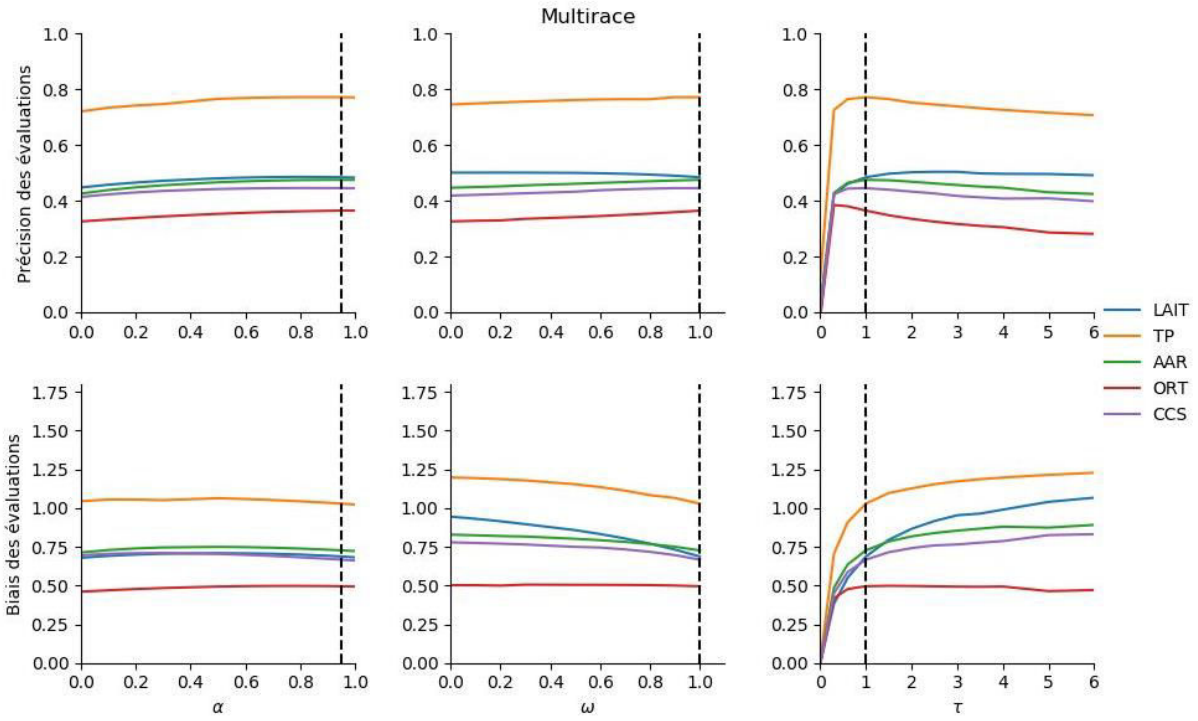


Figure 37. Évolution des précisions génomiques et des biais en fonction des valeurs des hyperparamètres α, ω et τ dans les évaluations génomiques ssGBLUP multirace pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS)

L'hyperparamètre α a un effet très limité sur les biais (pente). Ils sont en moyenne de 0,70 pour le LAIT, de 1,05 pour le TP, 0,74 pour AAR, 0,48 pour ORT et 0,69 pour CCS. Les pentes sont plus fortement impactées avec l'hyperparamètre ω. Le passage d'un ω de 1 à 0 améliore les biais pour le LAIT (pente de 0,69 à 0,95), AAR (pente de 0,73 à 0,83) et CCS (pente de 0,67 à 0,78). Les pentes sont égales à 0,50 pour toutes les valeurs de ω pour ORT. Diminuer la valeur de ω augmente les biais pour le TP (pente de 1,03 pour un ω de 1 et 1,20 pour un ω de 0). Pour l'hyperparamètre τ, les pentes avoisinent les 0 pour un τ de 0. Les pentes s'améliorent quand τ augmente. Pour le caractère ORT, un maximum est atteint pour un τ de 1,5 puis les pentes restent autour de 0,50. Pour les autres caractères, plus la valeur de τ est grande, plus les pentes sont élevées. Des pentes de 1,06 pour le LAIT, 0,89 pour l'AAR et 0,83 pour les CCS sont obtenues avec un τ de 6. Pour le TP, une pente de 1,02 est obtenue avec un τ de 1 puis ces pentes atteignent 1,23 pour un τ de 6.

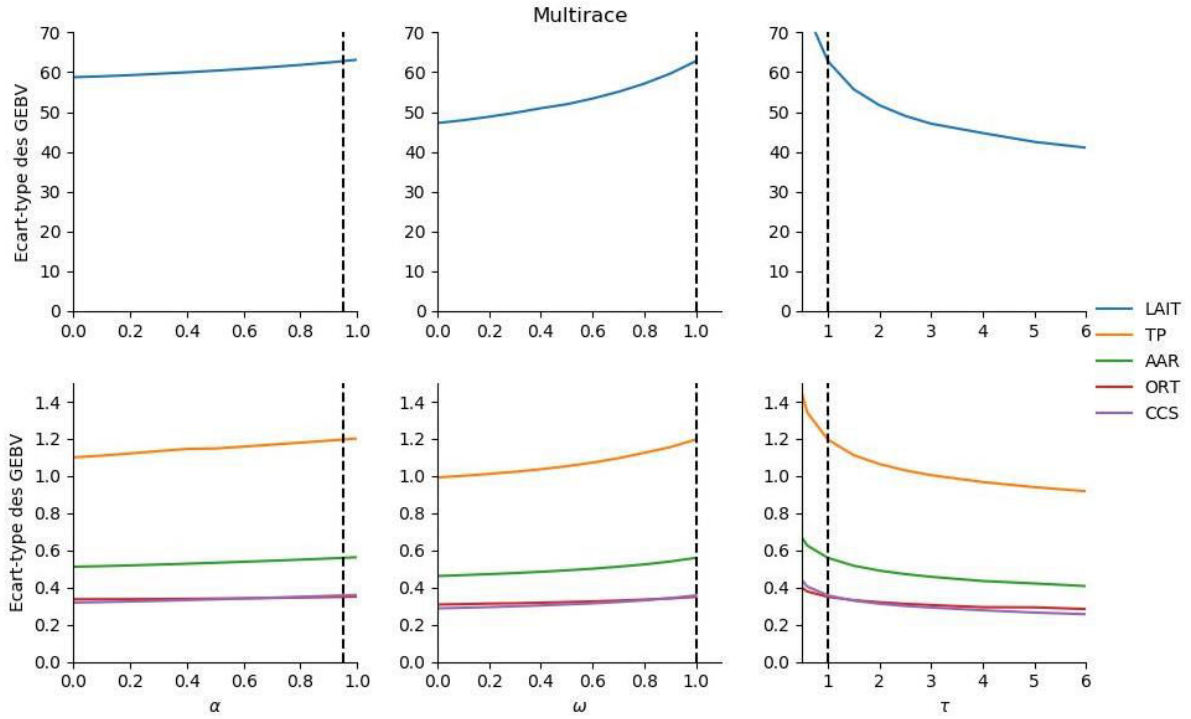


Figure 38. Ecart-type des GEBV de la population de validation en fonction des valeurs des hyperparamètres α, ω et τ pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS) en analyse multirace

La Figure 38 représente l'écart-type des GEBV des animaux de validation pour les 5 caractères (LAIT, TP, AAR, ORT et CCS) en fonction des valeurs des hyperparamètres α, ω et τ. Les hyperparamètres ω et τ ont les effets les plus importants sur l'écart-type des GEBV. Lorsque α tend vers 0, en les écarts-types des GEBV diminuent légèrement (diminution de 4,4 kg pour le LAIT, 0,10 g/kg pour le TP, 0,05 point pour AAR, 0,02 point pour ORT et 0,05 point pour CCS). La diminution de l'écart-type des GEBV est plus importante avec l'hyperparamètre ω (variant de 1 à 0) : perte de 15 kg pour le LAIT, 0,21 g/kg pour le TP, 0,09 point pour AAR, 0,05 point pour ORT et 0,07 point pour les CCS. Pour l'hyperparamètre τ, plus sa valeur est élevée, plus les écarts-types des GEBV sont faibles. Pour un τ de 0, les écarts-types des GEBV sont très élevés (non représentés sur la Figure 38) : 8 626 kg pour le LAIT, 12 g/kg pour le TP, 5,60 points pour l'AAR, 4,4 points pour l'ORT et 10 points pour les CCS. Avec un τ de 6, les écarts-types sont de 41 kg pour le LAIT, 0,92 g/kg pour le TP, 0,40 point pour AAR, 0,28 point pour ORT, 0,26 point pour les CCS.

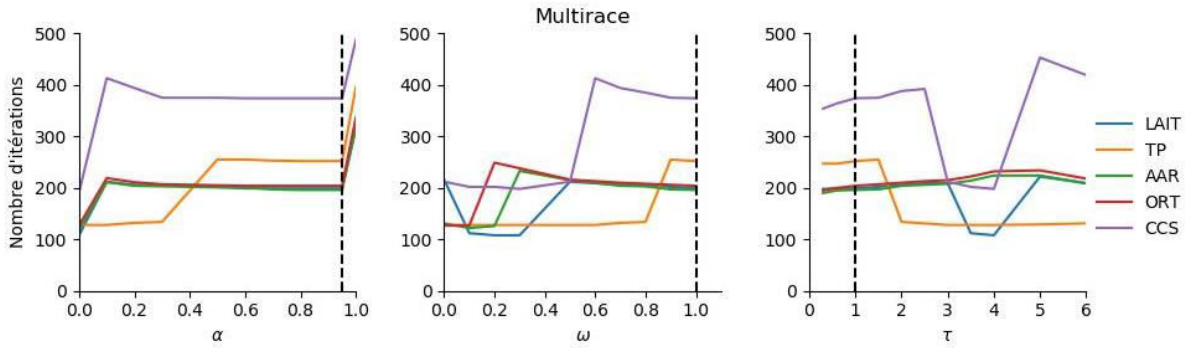


Figure 39. Nombre d'itérations pour atteindre la convergence du ssGBLUP en fonction des valeurs des hyperparamètres α, ω et pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS) en analyse multirace

La Figure 39 présente le nombre d'itérations pour atteindre la convergence des évaluations ssGBLUP multirace en fonction des valeurs des hyperparamètres α, ω et τ. Pour le paramètre α, le nombre d'itérations reste quasiment constant pour un α entre 0,1 et 0,95 sauf pour le TP. Pour le LAIT, l'AAR, l'ORT et le CCS, le nombre d'itérations moyen est de 250. Le nombre d'itération atteint 365 en moyenne pour un α de 1 et chute à 135 en moyenne pour un α de 0. Pour le TP, le nombre d'itération forme deux plateaux. Le premier avec environ 250 itérations en moyenne pour un α entre 0,5 et 0,95 et un deuxième plateau avec 130 itérations en moyenne pour un α entre 0 et 0,4. Les hyperparamètres ω et τ provoquent des modifications plus grandes sur le nombre d'itérations. Pour l'hyperparamètre ω, la convergence est atteinte plus rapidement lorsque ω est faible (ω < à 0,5) où le nombre d'itération est en moyenne inférieur à 200. Au-dessus de 0,5, le nombre d'itération augmente et dépasse 230 itérations en moyenne. Cependant, les caractères AAR et ORT semblent être moins sensibles à cet hyperparamètre. Pour l'hyperparamètre τ, les caractères AAR et ORT semblent peu sensibles puisque le nombre d'itération moyen est de 205 pour AAR et de 210 pour ORT. Les résultats pour un τ de 0 ne sont pas montrés, car ils dépassaient les 1000 itérations. Pour les autres caractères, un minimum est obtenu pour un τ entre 3,5 et 4 où l'on observe 150 itérations en moyenne pour obtenir la convergence contre 250 itérations en moyenne pour un τ en dehors de cet intervalle.

2.2.2. Effets des hyperparamètres α, ω et τ sur les évaluations génomiques Alpine

La Figure 40 présente les précisions des évaluations génomiques et les biais des évaluations ssGBLUP Alpine pour chaque valeur de α, ω et τ testée. Pour les précisions, nous observons les mêmes tendances que les analyses multirace. Les précisions sont modérément impactées par les hyperparamètres α et ω. Les précisions des évaluations chutent en moyenne de 4 points pour ces 5 caractères (0,54 à 0,50) avec un α entre 0,95 et 0. La diminution de l'hyperparamètre ω (de 1 à 0) entraine une baisse des précisions de 1 point en moyenne. Tous les caractères ne se comportent pas de la même manière puisque les précisions s'améliorent pour le LAIT (0,45 à 0,50) et très légèrement pour l'AAR (0,40 à 0,41). Pour l'hyperparamètre τ, on retrouve le profil en 3 étapes observé pour les évaluations multirace. Les précisions maximums sont atteintes pour un τ de 1 pour le TP (0,76), pour un τ de 4 pour l'AAR (0,43), pour un τ entre 1 et 2 pour ORT (0,42) et pour un τ entre 1 et 4 pour le CCS (0,48). Avant ce seuil, les précisions sont plus faibles et atteignent 0 pour un τ égal à 0. Au-dessus de ce seuil, les précisions baissent légèrement pour atteindre 0,71 pour le TP, 0,41 pour l'AAR, 0,38 pour l'ORT et 0,47 pour le CCS avec un τ de 6. Pour le LAIT, les précisions fluctuent plus fortement. Elles s'améliorent avec un τ entre 0 et 3,5 (de 0 à 0,50), avant de diminuer pour un τ de 4 et 5 (0,49), enfin un maximum est atteint pour un τ de 6 (0,51).

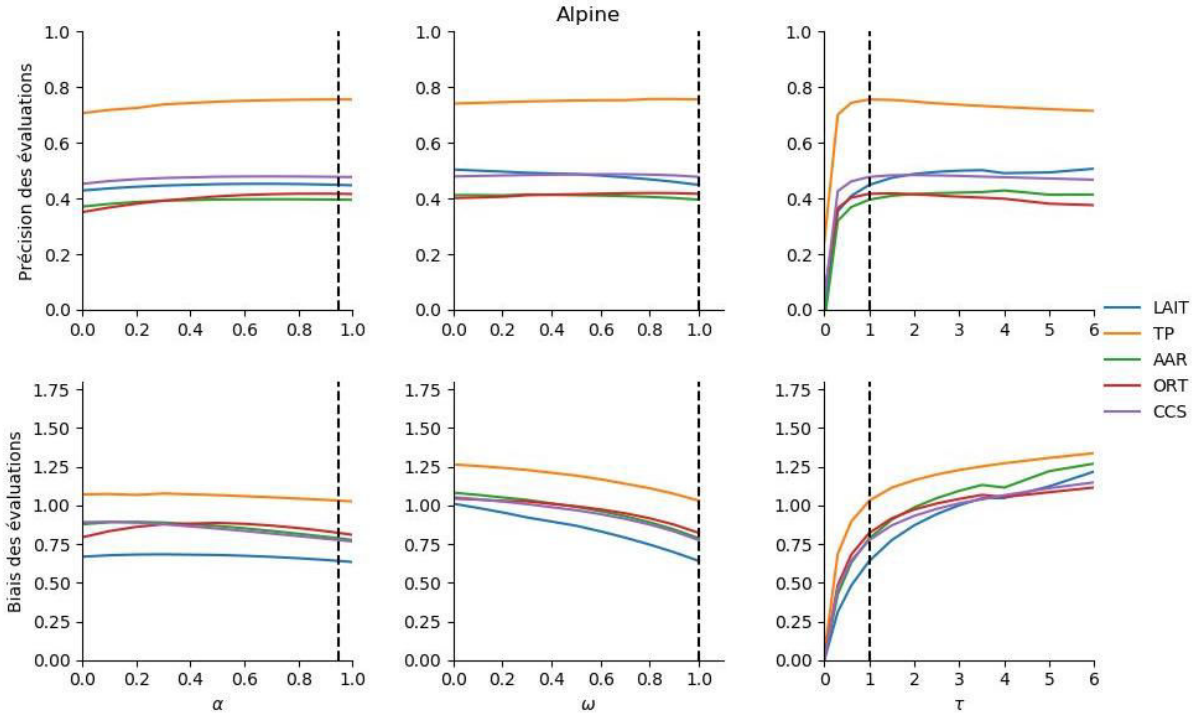


Figure 40. Évolution des précisions génomiques et des biais en fonction des valeurs des hyperparamètres α, ω et τ dans les évaluations génomiques ssGBLUP Alpine pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS)

L'hyperparamètre α a un impact plus important sur les biais pour les évaluations Alpine que les évaluations multirace. On observe 2 profils parmi les 5 caractères. Pour le TP et les CCS, plus la valeur de α est faible, plus les pentes sont grandes. Les pentes passent de 1,02 à 1,07 pour le TP et de 0,77 à 0,89 pour les CCS avec des valeurs de α allant de 1 à 0. Pour le LAIT, l'AAR et l'ORT, la diminution de la valeur α entraîne une augmentation des pentes jusqu'à un maximum avant de rediminuer. Un maximum est atteint avec un α de 0,3 pour le LAIT (pente de 0,68), un α de 0,5 pour l'ORT (pente de 0,89), et un α de 0,2 pour AAR (pente de 0,89). Pour l'hyperparamètre ω, tous les caractères suivent la même tendance : la diminution de ω augmente les pentes. Des pentes de 1 (à 0,03 point près) sont obtenues pour un ω de 0 pour le LAIT, un ω de 1 pour le TP, un ω de 0,5 pour AAR, un ω de 0,5 pour ORT et un ω de 0,3 pour CCS. Pour tous les caractères, plus τ est élevé et plus les pentes sont élevées. On observe des pentes de 1 avec un τ de 3 pour le LAIT, un τ de 1 pour le TP, un τ de 2 pour AAR, un τ de 2,5 pour ORT et un τ de 3 pour CCS. Les pentes augmentent pour atteindre au maximum pour un τ égal à 6 : 1,21 pour le LAIT, 1,34 pour le TP, 1,27 pour AAR, 1,15 pour ORT et 1,15 pour CCS avec un τ de 6.

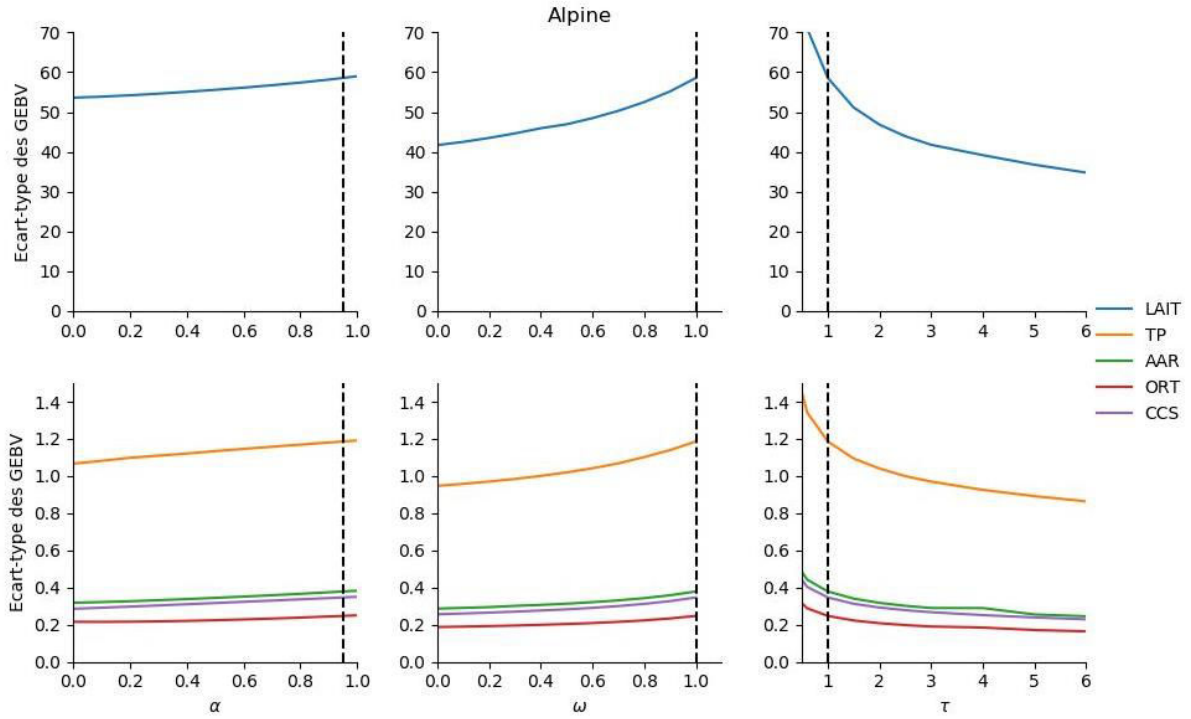


Figure 41. Ecart-type des GEBV de la population de validation en fonction des valeurs des hyperparamètres α, ω et τ pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS) en analyse Alpine

La Figure 41 présente les écarts-types des GEBV pour les évaluations en race Alpine en fonction des valeurs des paramètres α, ω et τ. On obtient des résultats comparables à ceux observés dans les analyses multirace. La modification des hyperparamètres ω et τ entraîne une plus grande variation des écarts-types des GEBV. Les écarts-types diminuent légèrement lorsque α passe de 1 à 0 : de 5 kg pour le LAIT, de 0,13 g/kg pour le TP, de 0,07 point pour AAR, de 0,04 point pour ORT et de 0,07 point pour CCS. Lorsque ω passe de 1 à 0, on observe une baisse des écarts-types de 17 kg pour le LAIT, de 0,24 g/kg pour le TP, de 0,09 point pour AAR, de 0,06 point pour ORT et de 0,09 point pour CCS. Pour l'hyperparamètre τ, plus τ augmente, plus les écarts-types diminuent passant de 5 432 kg à 35 kg pour le LAIT, de 8,54 g/kg à 0,86 g/kg pour le TP, de 8,34 points à 0,24 point pour AAR, de 2,89 points à 0,16 point pour ORT et de 5,79 points à 0,23 point pour les CCS.

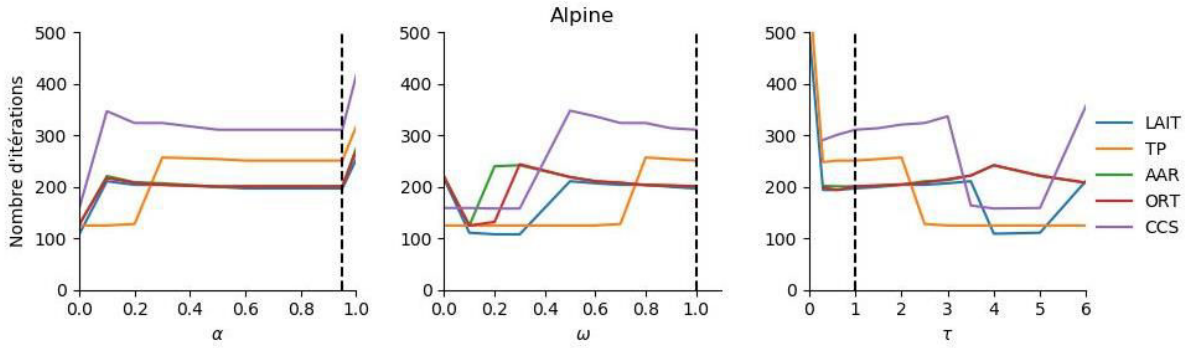


Figure 42. Nombre d'itérations pour atteindre la convergence du ssGBLUP en fonction des valeurs des hyperparamètres α, ω et pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS) en analyse Alpine

Le nombre d'itération pour atteindre la convergence en fonction des valeurs des hyperparamètres α, ω et τ dans les évaluations Alpine est présenté dans la Figure 42. Comme pour les évaluations multirace, l'hyperparamètre α affecte peu le nombre d'itération sauf pour le TP. Le nombre moyen d'itération se situe autour de 230 itérations sauf pour un α de 0 où il atteint 130 itérations en moyenne et pour un α de 1 il est de 300 en moyenne. Concernant le paramètre ω, les nombres d'itération les plus faibles sont obtenus pour un ω de 0,1 (110 pour le LAIT et 160 pour le CCS) ou 0,2 (125 pour le TP, l'AAR et l'ORT). Le nombre d'itération augmente jusqu'à 230 en moyenne avec un ω de 1. Pour l'hyperparamètre τ, on observe une région ($3,5 \leq \tau \leq 5$) où le nombre d'itération est minimal (140 itérations en moyenne). Comme pour les analyses multiraces, les caractères AAR et ORT sont peu sensibles aux variations du paramètre τ avec un nombre d'itération moyen de 210

2.2.3. Effets des hyperparamètres α, ω et τ sur les évaluations génomiques Saanen.

La Figure 43 présente les précisions et les biais des évaluations génomiques en fonction des valeurs des hyperparamètres α, ω et τ. Les tendances observées dans les analyses multirace et Alpine se retrouvent en race Saanen. Les hyperparamètres α et ω affectent faiblement la précision des évaluations ssGBLUP. Sur l'ensemble des 5 caractères, la précision est de 0,52 en moyenne pour un α de 0, elle atteint 0,59 en moyenne pour un α de 1. On observe les mêmes tendances pour ω, les précisions sont en moyenne de 0,59 pour un ω de 1 et de 0,55 pour un ω de 0. Concernant l'hyperparamètre τ, les précisions augmentent de 0,16 à 0.59 en moyenne lorsque τ passe de 0 à 1. Les précisions les plus élevées (0,59 en moyenne) sont obtenues pour τ égal à 1 pour tous les caractères

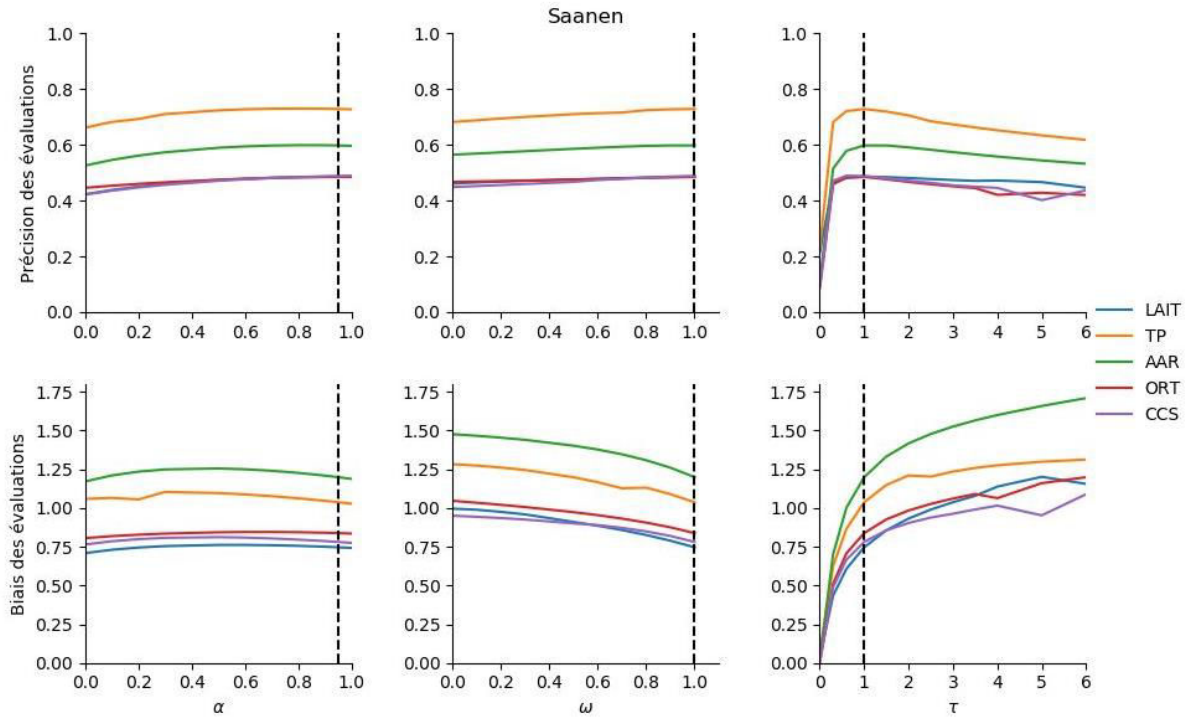


Figure 43. Évolution des précisions génomiques et des biais en fonction des valeurs des hyperparamètres α, ω et τ dans les évaluations génomiques ssGBLUP Saanen pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS)

Comme observé en race Alpine, l'hyperparamètre α a un effet limité sur les pentes des évaluations génomiques. En moyenne, les pentes sont de 0,75 pour le LAIT, 1,07 pour le TP, 1,22 pour AAR, 0,83 pour ORT et 0,79 pour CCS. Pour l'hyperparamètre ω, plus sa valeur tend vers 0, plus les pentes sont élevées. Ainsi, le passage d'un ω de 1 à 0 augmente les pentes de

+25 points pour le LAIT, +25 points pour le TP, +28 points pour AAR, +21 points pour ORT et +17 points pour CCS. Des pentes de 1 (à 0,02 points près) sont obtenues pour les caractères LAIT avec ω=0, et ORT avec ω=0,3. Pour les autres caractères, les pentes sont les plus proches de 1 avec un ω de 1 pour le TP et AAR (avec des pentes de 1,04 et 1,20 respectivement) et pour un ω de 0 pour CCS (pente de 0,95). On observe les mêmes tendances que pour les évaluations multirace et Alpine pour l'hyperparamètre τ. Avec un τ de 0, les pentes sont proches de 0 alors qu'avec un τ de 6, les pentes atteignent 1,16 pour le LAIT, 1,31 pour le TP, 1,20 pour ORT, 1,71 pour AAR et 1,09 pour CCS. Les pentes les plus proches de 1 sont obtenues pour un τ de 2,5 pour le LAIT (0,99), un τ de 1 pour le TP (1,04), un τ de 0,6 pour AAR (1,00), un τ de 2 pour ORT (0,98) et un τ de 3,5 pour les CCS (0,99).

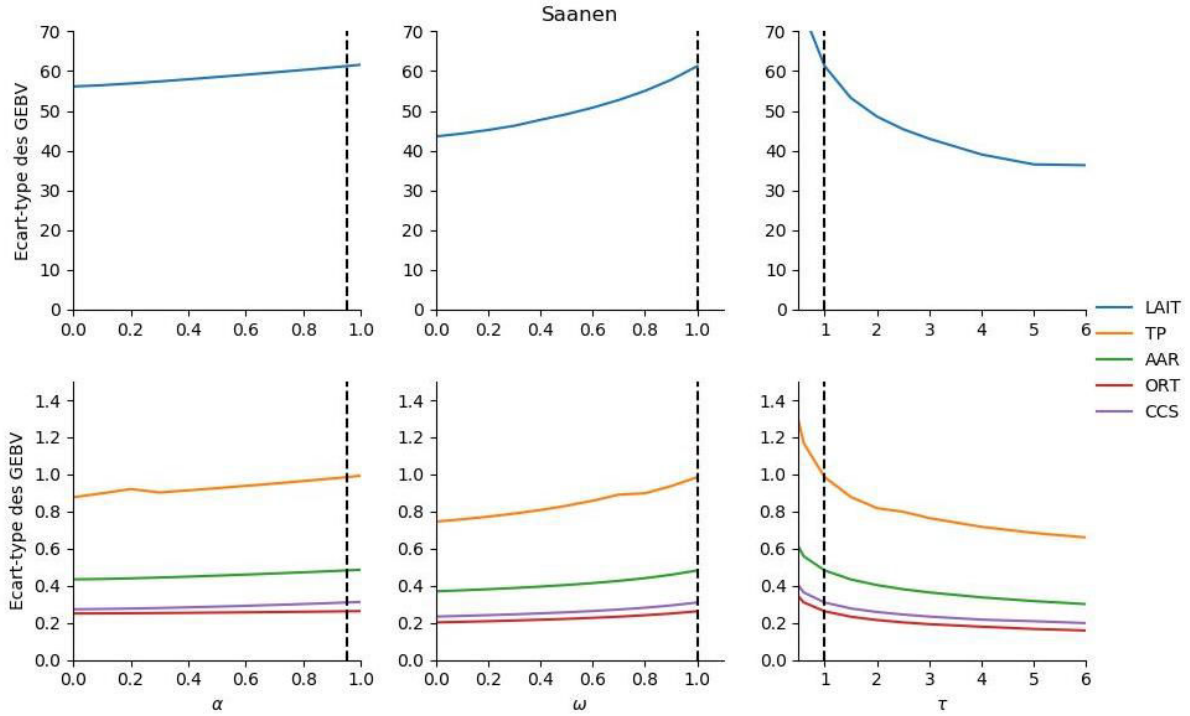


Figure 44. Ecart-type des GEBV de la population de validation en fonction des valeurs des hyperparamètres α, ω et τ pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS) en analyse Saanen

La Figure 44 présente les écarts-types des GEBV pour les 5 caractères (LAIT, TP, AAR, ORT et CCS) en fonction des valeurs des hyperparamètres α, ω et τ pour les évaluations Saanen. Les tendances sont à nouveau similaires aux analyses multirace et Alpine. La baisse de α ou ω par rapport aux valeurs par défaut entraine une diminution des écarts-types des GEBV. Ils diminuent de 5 kg pour le LAIT, de 0,12 g/kg pour le TP, de 0,05 pour AAR, de 0,01 point pour ORT et de 0,04 point pour CCS, lorsque α passe de 1 à 0. La réduction des écarts-types est plus importante pour les valeurs de l'hyperparamètre ω, lorsque ce dernier passe de 1 et 0. Elles sont de 18 kg pour le LAIT, 0,24 g/kg pour le TP, 0,11 point pour AAR, 0,06 point pour ORT et 0,07 point pour CCS. Comme pour les analyses multirace et Alpine, les écarts-types des GEBV sont très élevés pour τ égal à 0 (valeurs non représentées sur la Figure 44) : 547 kg pour le LAIT, 5,40 g/kg pour le TP, 4,92 points pour AAR, 12,91 points pour ORT, 5,20 points pour CCS. Les écarts-types des GEBV diminuent lorsque τ tend vers 6 pour atteindre 36 kg pour le

LAIT, 0,66 g/kg pour le TP, 0,30 point pour AAR, 0,16 point pour ORT et 0,20 point pour CCS.

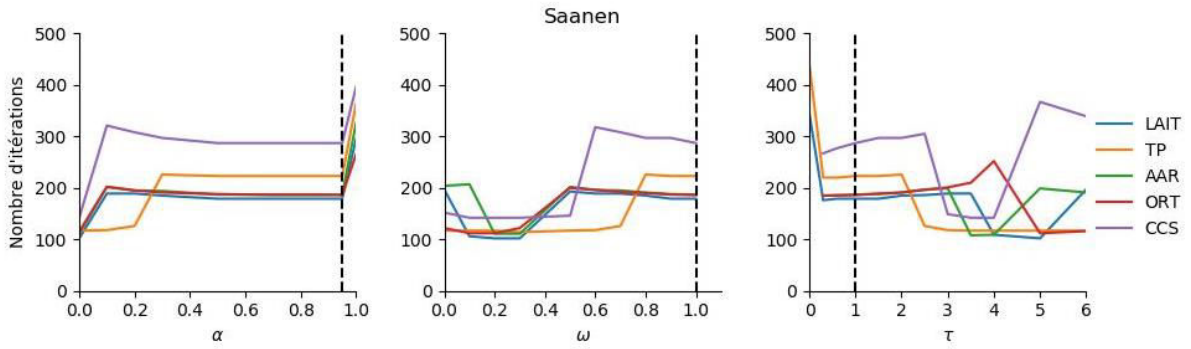


Figure 45. Nombre d'itérations pour atteindre la convergence du ssGBLUP en fonction des valeurs des hyperparamètres α, ω et pour les caractères lait (LAIT), taux protéique (TP), qualité de l'attache arrière (AAR), orientation des trayons (ORT) et comptage de cellule somatique (CCS) en analyse Saanen

L'analyse de la convergence pour les évaluations Saanen présente des profils similaires à ceux obtenus pour les évaluations multirace et Alpine (Figure 45). L'hyperparamètre α a peu d'influence sur le nombre d'itération pour un α compris entre 0,1 et 0,95 (210 itérations en moyenne). Pour un α de 0, le nombre d'itération moyen baisse à 120 et atteint 320 itérations pour un α de 1. Pour l'hyperparamètre ω, le nombre d'itération moyen diminue lorsque ω est inférieur à 0,5, il est minimum pour un ω de 0,2 (115 itérations en moyenne). Pour l'hyperparamètre τ, on observe quelques différences comparées aux évaluations multirace et Alpine, les caractères AAR et ORT présentent des profils différents. Pour ORT, le nombre d'itération augmente lorsque τ passe de 0 à 4 (185 itérations à 252 itérations) puis chute lorsque τ passe de 4 et 5 (112 itérations). Pour l'AAR, le nombre d'itération moyen est de 191 sauf pour un τ de 3,5 et 4 où il est de 108 et 109 respectivement. Pour le LAIT, un minimum d'itération est atteint avec τ compris entre 4 et 5 (105 en moyenne). Pour CCS, ce minimum est atteint avec un τ compris entre 3 et 4 (145 itérations en moyenne). Pour le TP, peu de changement avec un τ supérieur à 2,5 où le nombre d'itération reste constant (autour de 120 itérations).

2.3. Discussion

Les évaluations génomiques ssGBLUP avec les valeurs par défaut des hyperparamètres α, ω, et τ ont majoritairement des pentes inférieures à 1, suggérant des biais dans les évaluations génomiques multirace, Alpine ou Saanen. Ces biais conduisent souvent à une surestimation des GEBV sauf pour le TP. Gowane et al., (2018a) se sont intéressés aux biais des évaluations en fonction de l'héritabilité, l'architecture génétique du caractère, la taille de la population de référence ainsi que de la méthode de reproduction en utilisant les méthodes BLUP, GBLUP et ssGBLUP. Ils ont simulé des populations avec le logiciel QMSim (Sargolzaei and Schenkel, 2009). Une population historique a été créée sur 100 générations. Pour les générations 101 à 110, une sélection a été mise en place. La sélection des animaux a été réalisée avec 3 méthodes différentes : (RR) une sélection des animaux et des accouplements aléatoires, (SR) une sélection des animaux basée sur les EBV, mais des accouplements aléatoires et (SA) une sélection des animaux et des accouplements basés sur EBV. Les phénotypes générés avaient une moyenne de zéro et une variance de 1. Les phénotypes ont été simulés pour avoir une héritabilité de 0,1, 0,3 ou 0,5. Pour les 9 dernières générations (génération 101 à 109), les EBV des animaux ont été estimés avec un BLUP. Le génotypage des animaux mâles est mis en place pour les 4 dernières générations (générations 106 à 109) avec 125, 250 ou 500 génotypages par génération. Ces animaux ont formé les différentes populations de référence comptabilisant 500, 1000 ou 2 000 animaux génotypés respectivement. La génération 110 a été utilisée comme population de validation pour calculer la précision et les biais des évaluations génomiques. Les (G)EBV

des animaux de la génération 110 ont été estimés avec un BLUP, GBLUP et un ssGBLUP. Leurs résultats montrent que le scénario SA est plus biaisé que le scénario RR, quel que soit l'héritabilité du caractère ou le nombre de QTLs. Pour un caractère avec une héritabilité de 0,3, 90 QTLs et une population de référence de 1000 animaux génotypés, les biais avec les évaluations ssGBLUP sont de 0,99 pour la méthode RR et de 0,53 pour la méthode SA. Ils ont également observé que plus l'héritabilité est faible et plus les biais seront grands. Par exemple, pour le scénario SA avec 90 QTLs, les évaluations ssGBLUP avec une population de référence de 1 000 animaux génotypés ont des pentes égales à 0,67 (héritabilité de 0,5), de 0,53 (héritabilité de 0,3) et de 0,25 (héritabilité de 0,1). Les pentes sont plus élevées lorsque les animaux sont évalués avec la méthode BLUP (0,78 pour une héritabilité de 0,5, 0,72 pour une héritabilité de 0,3 et 0,76 pour une héritabilité de 0,1). En caprins, ces biais apparaissent avec la méthode ssGBLUP car les pedigrees sont incomplets et la population de référence ne contient pas les animaux de la population de base. Lorsque les éléments de **G** sont mis à l'échelle de la matrice $\mathbf{A_{22}}$, les éléments de **G** seront plus petits que $\mathbf{A_{22}}$ pour les animaux avec de longs pedigrees (Vitezica et al., 2011 ; Misztal et al., 2013 b) et inversement pour des animaux avec des pedigrees courts. Des biais dans les évaluations génomiques ssGBLUP ont été mis en évidence pour d'autres espèces d'élevage tels que chez les bovins laitiers (Aguilar et al., 2010). Pour des femelles Holstein US génotypées avec la puce Illumina BovineSNP50 BeadChip (Illumina, San Diego, CA) et évaluées pour le score final en Holstein, Aguilar et al., (2010) ont observé des biais de 0,66 pour des évaluations ssGBLUP. Des biais ont également été observés pour les évaluations génomiques ovines (Legarra et al., 2014a). Dans cette étude, les animaux sont génotypés avec la puce OvineSNP50 Bead-Chip (Illumina Inc., San Diego, CA) et phénotypés pour le LAIT (DYD). Les biais pour les évaluations ssGBLUP étaient entre 0,19 et 1,07 pour les races Basco-Béarnaise, Manech Tête Noire, Manech Tête Rousse, Latxa Cara, Negra Euskadi, Latxa Cara Rubia et Latxa Cara Negra Navarre.

Dans notre étude, les hyperparamètres α, ω et τ ont les mêmes effets sur les précisions, les biais et les écarts-types des GEBV dans les analyses multirace, Alpine et Saanen. Les valeurs optimales des hyperparamètres seraient $\alpha = 0.95$, $\omega = [0.1, 0.3]$ et $\tau = [3\text{-}4]$ pour obtenir des évaluations précises, sans biais et une bonne convergence pour la majorité des caractères étudiés. Les précisions varient peu sur la gamme des valeurs de α et ω testées et pour τ compris entre 0.9 et 6. Le TP dans les analyses multirace, Alpine et Saanen et AAR dans les analyses Saanen présentent les précisions les plus fortes et des évaluations non biaisées (pente proche de 1) pour les valeurs par défaut des hyperparamètres α, ω et τ. Ces 3 hyperparamètres agissent également sur la dispersion des GEBV comme nous avons pu l'observer dans notre étude. Martini et al. (2018) ont constaté également que l'augmentation de l'hyperparamètre τ ou la réduction de l'hyperparamètre ω induisaient une baisse de la variabilité des GEBV. En revanche, dans notre étude, l'hyperparamètre α semble avoir un effet très limité sur la variabilité des GEBV. Dans la littérature, les valeurs optimales des hyperparamètres sont variables et généralement spécifiques de chaque population étudiée. Pour l'hyperparamètre α, Misztal et al., (2013) recommande d'utiliser un $\alpha \approx 0{,}8$ chez les bovins laitiers pour réduire les biais des évaluations génomiques. En bovins allaitants, Croué et Ducrocq, (2017) se sont intéressés à l'hyperparamètre α pour la prédiction de caractères de carcasse. Les meilleures pentes sont obtenues avec un α de 0,4 (pente de 1,02 pour l'âge à l'abattage, 0,99 pour la conformation de la carcasse et 0,86 pour le poids de la carcasse). Pour l'hyperparamètre ω, Tsuruta et al., (2011) ont analysé 18 caractères de morphologie en Holstein. Les biais sont en moyenne de 0,74 pour un ω de 1, et de 0,91 pour un ω de 0,7. Ils préconisent donc d'utiliser un ω de 0,7 dans les évaluations génomiques, comme Lourenco et al., (2014). Pour Croué et Ducrocq, (2017), le ω optimal est de 0,9 pour la conformation de la carcasse et l'âge à l'abattage (pente de 0,98 et 0,99 respectivement) et de 0,7 pour le poids de la carcasse (0,94). Pour l'hyperparamètre τ, Misztal et al., (2010), Tsuruta et al., (2011) et Lourenco et al., (2014) recommande l'utilisation

d'un τ de 1 pour des caractères de conformation et pour des caractères de production laitière chez la Holstein.

L'amélioration des biais dans notre étude semble possible soit en diminuant l'hyperparamètre ω, soit en augmentant l'hyperparamètre τ. Cette situation suggère que la compatibilité entre les matrices **A** et **G** puisse être améliorée en les pondérant de façon adéquate dans le ssGBLUP (τ pour la matrice **G** et ω pour la matrice **A**). Cependant dans notre étude, les hyperparamètres α, ω et τ ont été testés individuellement, les résultats ne donnent donc pas d'information sur les précisions et les biais que l'on obtiendrait si ces hyperparamètres étaient modifiés simultanément. Une approche intéressante serait de sélectionner une grille de valeur pour les hyperparamètres ω et τ comme présentée dans l'étude de Martini et al., (2018). Au vu des résultats obtenus et par comparaison avec des résultats de la littérature, nous avons choisi de conserver pour la suite de nos études les paramètres par défaut du ssGBLUP, soit $\alpha = 0.95$, $\omega = 1$ et $\tau = 1$.

Chapitre 4 : Intégration de gènes majeurs, mutations causales ou régions génomiques d'intérêt dans les évaluations génomiques

1. Analyse du gène de la *caséine* α_{s1} et du gène *DGAT1* ayant une influence sur les taux protéique et butyreux

Nous avons testé plusieurs méthodes d'évaluations génomiques qui intègrent l'effet d'un gène majeur dans les modèles d'évaluation génomique. L'objectif était d'étudier l'impact de la prise en compte de l'information d'un gène majeur identifié dans les modèles génomiques sur les précisions. Les méthodes que nous avons comparées utilisent les génotypes de la puce 50K et/ou les génotypes pour une mutation identifiée. Les premiers travaux ont porté sur le gène de la *caséine* α_{s1} sur le TP. Dans un deuxième temps, nous avons étudié l'effet de ces méthodes sur le TB avec la prise en compte de l'information génomique du gène *DGAT1*.

1.1. Intégration du gène de la *caséine* α_{s1} dans les évaluations génomiques du taux protéique

1.1.1. Déséquilibre de liaison entre les génotypes du gène de la caséine α_{s1} et les marqueurs de la puce 50K

Certaines méthodes testées n'utilisent pas les génotypes du gène de la *caséine* α_{s1}. Il est donc intéressant de savoir si les marqueurs de la puce 50K (en particulier ceux du chromosome 6) sont en déséquilibre de liaison avec le gène de la *caséine* α_{s1}. Cette question est d'autant plus importante que le gène de la *caséine* α_{s1} est polymorphe et qu'un seul SNP ne peut pas capter toute l'information de ce gène. Pour tester le déséquilibre de liaison entre les génotypes de la puce 50K et les génotypes de la *caséine* α_{s1}, nous avons réalisé un test du chi-deux sur l'ensemble des 903 animaux à la fois génotypés sur la puce 50K et pour le gène de la *caséine* α_{s1} (paragraphe 3.3.2, Chapitre 1). Certains animaux peuvent avoir un génotype manquant (codé sous la forme de 5), ils ont été éliminés pour réaliser le test du chi-deux. Les résultats de ce test sont présentés dans la Figure 46. On observe un pic avec de fortes valeurs pour les analyses multirace et Alpine et Saanen en fin du chromosome 6. Pour les analyses multirace, les 10 SNPs avec les plus fortes valeurs (142 et 318) sont localisés dans le 82eme Mb du chromosome 6. En Alpine et Saanen, les SNPs sont localisés dans la même zone avec des pics un peu moins haut compris entre164 et 196 en Alpine et 142 et 154 en Saanen.

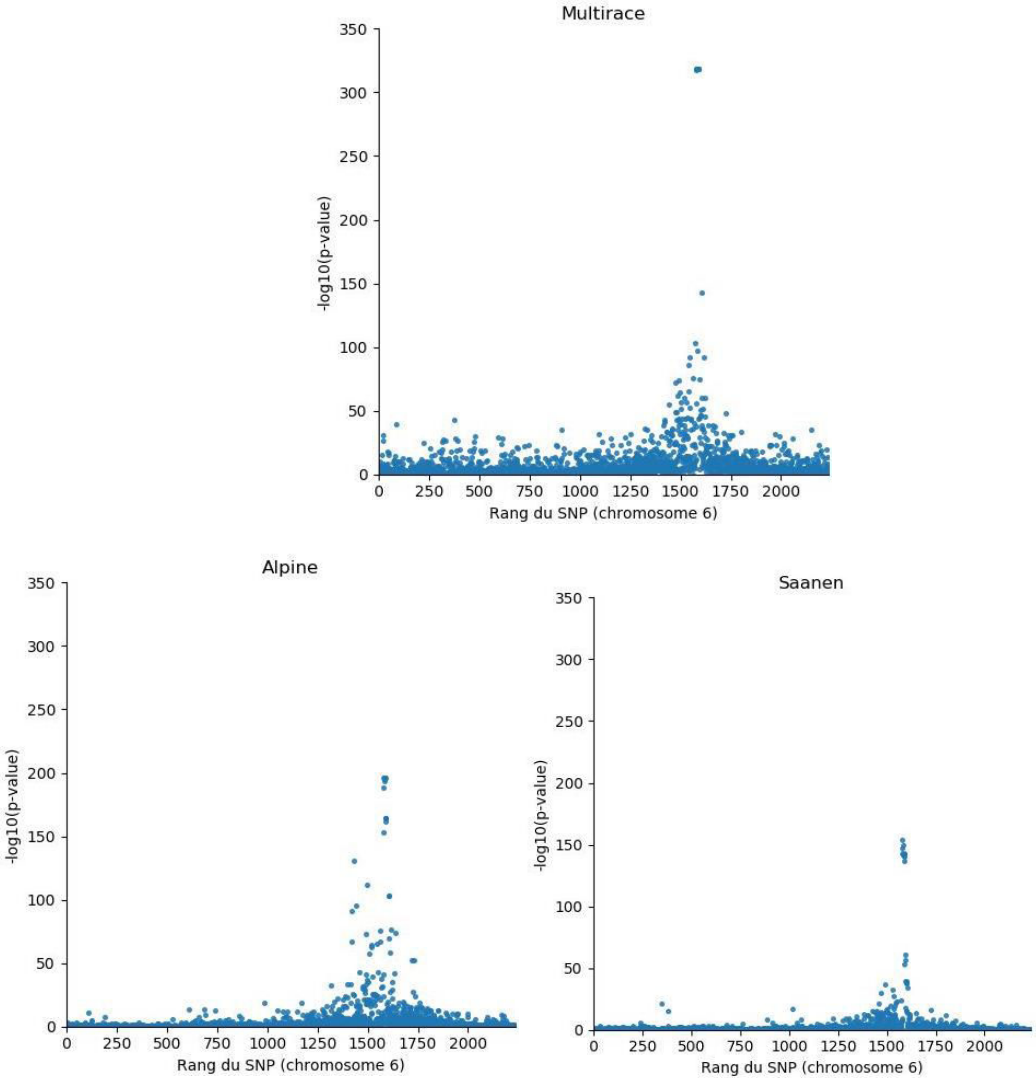


Figure 46. Valeur de –log10(p-valeur) du test du Chi-deux entre les génotypes du gène de la caséine α_{s1} et les génotypes de la puce 50K (sur le chromosome 6) pour les analyses multirace, Alpine et Saanen

La Figure 47 présente le nombre de SNPs communs entre les analyses multirace, Alpine et Saanen parmi les tops 10. La majorité des SNPs (7) sont communs aux trois analyses, seuls 2 SNPs sont communs entre les analyses multirace et Alpine et 1 SNP en commun entre les analyses Alpine et Saanen. Les derniers SNPs sont eux spécifiques à chaque analyse (1 SNP pour les analyses multirace et 2 SNPs pour les analyses Saanen).

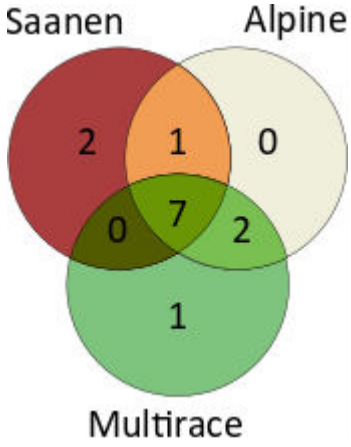


Figure 47. SNPs communs entre les analyses multirace, Alpine et Saanen parmi les tops 10 des SNPs les plus fortement associés entre les génotypes de la caséine α_{s1} et les génotypes de la puce 50K en utilisant un test du chi-deux

Le Tableau 20 présente l'exemple du SNP (snp59416-scaffold980-293987) dont le $-\log_{10}(p-valeur)$ est égal à 318 pour les analyses multirace, c'est également le SNP qui a les plus fortes valeurs de $-\log_{10}(p-valeur)$ pour les analyses Alpine et Saanen. On observe que 91 % des animaux, ayant un des génotypes au gène de la *caséine* α_{s1} dit « effet positif fort sur le TP », sont homozygotes « 0 » au marqueur (snp59416-scaffold980-293987) de la puce 50K. 92 % des animaux ayant un des génotypes au gène de la *caséine* α_{s1} dit «effet intermédiaire sur le TP » sont hétérozygotes au marqueur SNP. Et on note que tous les individus EE au gène de la *caséine* α_{s1} ont le génotype 2 au marqueur SNP (snp59416-scaffold980-293987). Ce marqueur SNP semble donc être un bon indicateur de l'effet du gène de la *caséine* α_{s1} sur le TP.

Tableau 20. Table de contingence entre les génotypes caséine α_{s1} et les génotypes de la puce 50K pour le SNP « snp59416-scaffold980-293987 » pour les analyses multirace

		Génotype 50K (snp59416-scaffold980-293987)		
Effet sur le TP	**Génotype *caséine* α_{s1}**	**Nb homozygote (0)**	**Nb hétérozygote (1)**	**Nb homozygote (2)**
Positif fort	AA	271	0	0
	AB	107	1	0
	AC	30	0	0
	BB	5	0	0
	BC	5	0	0
	CC	1	0	0
Intermédiaire	AE	1	249	0
	AF	31	0	0
	AO	1	0	0
	BE	0	34	0
	BF	4	1	0
	CE	0	4	0
	CF	2	0	0
Faible	EE	0	0	117
	EF	0	24	0
	FF	1	0	0

1.1.2. Résultats des évaluations génomiques intégrant l'effet du gène de la caséine α_{s1}

L'article, publié dans la revue Genetics Selection Evolution, s'intéresse à l'intégration de l'effet du gène de la *caséine* α_{s1} dans les modèles d'évaluations génomiques. Nous avons comparé les méthodes ssGBLUP, WssGBLUP classique et ses alternatives, le TABLUP et le gene content, implémentées dans les logiciels de la famille blupf90 (Misztal et al., 2002). Les phénotypes, pedigrees et génotypes 50K utilisés sont présentés au paragraphe 1.1 du chapitre 3. Les populations d'apprentissage et de validation sont celles du scénario E. Les génotypes du gène de la *caséine* α_{s1} utilisés sont présentés au paragraphe 1.1.1 (chapitre 3). Les analyses ont porté sur une population multirace (les données Alpine et Saanen sont analysées simultanément), la population Alpine et la population Saanen. Les précisions génomiques ont été calculées pour comparer les méthodes testées.

Article I : Le single-step GBLUP pondéré améliore les précisions des évaluations génomiques chez les caprins français pour le TP : un caractère contrôlé par un gène majeur.

Marc Teissier, Hélène Larroque, Christèle Robert-Granié, 2018. Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: A quantitative trait influenced by a major gene. Genetics Selection Evolution, 50:31

Teissier et al. Genet Sel Evol (2018) 50:31
https://doi.org/10.1186/s12711-018-0400-3

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene

Marc Teissier*, Hélène Larroque and Christèle Robert-Granié

Abstract

Background: In 2017, genomic selection was implemented in French dairy goats using the single-step genomic best linear unbiased prediction (ssGBLUP) method, which assumes that all single nucleotide polymorphisms explain the same fraction of genetic variance. However, ssGBLUP is not suitable for protein content, which is controlled by a major gene, i.e. a_{s1} *casein*. This gene explains about 40% of the genetic variation in protein content. In this study, we evaluated the accuracy of genomic prediction using different genomic methods to include the effect of the a_{s1} *casein* gene.

Methods: Genomic evaluation for protein content was performed with data from the official genetic evaluation on 2955 animals genotyped with the Illumina goat SNP50 BeadChip, 7202 animals genotyped at the a_{s1} *casein* gene and 6,767,490 phenotyped females. Pedigree-based BLUP was compared with regular unweighted ssGBLUP and with three weighted ssGBLUP methods (WssGBLUP, WssGBLUP$_{Max}$ and WssGBLUP$_{Sum}$), which give weights to SNPs according to their effect on protein content. Two other methods were also used: trait-specific marker-derived relationship matrix (TABLUP) using pre-selected SNPs associated with protein content and gene content based on a multiple-trait genomic model that includes a_{s1} *casein* genotypes. We estimated accuracies of predicted genomic estimated breeding values (GEBV) in two populations of goats (Alpine and Saanen).

Results: Accuracies of GEBV with ssGBLUP improved by + 5 to + 7 percent points over accuracies from the pedigree-based BLUP model. With the WssGBLUP methods, SNPs that are located close to the a_{s1} *casein* gene had the biggest weights and contributed substantially to the capture of signals from quantitative trait loci. Improvement in accuracy of genomic predictions using the three weighted ssGBLUP methods delivered up to + 6 percent points of accuracy over ssGBLUP. A similar accuracy was obtained for ssGBLUP and TABLUP considering the 20,000 most important SNPs. Incorporating information on the a_{s1} *casein* genotypes based on the gene content method gave similar results as ssGBLUP.

Conclusions: The three weighted ssGBLUP methods were efficient for detecting SNPs associated with protein content and for a better prediction of genomic breeding values than ssGBLUP. They also combined fast computing, simplicity and required ssGBLUP to be run only twice.

*Correspondence: marc.teissier@inra.fr
GenPhySE, INRA, INPT, ENVT, Université de Toulouse, 31326 Castanet-Tolosan, France

Background

The availability of molecular data has enabled the development and commercial application of genomic selection in various livestock species, such as dairy cattle [1, 2], dairy sheep [3, 4], meat sheep [5, 6] and dairy goats [7–9]. Meuwissen et al. [10] proposed genomic prediction of animals based on dense single nucleotide polymorphism (SNP) maps, by deriving the effects of SNPs from a reference population, for which animals are both phenotyped and genotyped. Genomic estimated breeding values (GEBV) of selection candidates (i.e., usually young individuals with genotypes but without phenotypes) can be estimated by summing up the effects of the SNP alleles carried by each animal.

Methods such as genomic best linear unbiased prediction (GBLUP) [11–15], are used to predict GEBV by replacing the pedigree relationship matrix used for pedigree-based BLUP with a realized genomic relationship matrix. The GBLUP method was further improved with single-step GBLUP (ssGBLUP) [12], which uses simultaneously all phenotypic, pedigree and genotypic information, including phenotypic information on non-genotyped individuals. Therefore, in ssGBLUP, the relationship between each pair of animals (genotyped and non-genotyped) is estimated with a relationship matrix that combines pedigree and genotype information. Several studies have reported that the accuracy of genomic prediction obtained with these methods is higher than with genetic evaluation using pedigree-based BLUP [16–18]. However, the accuracy obtained from genomic information depends on several parameters including reference population size [19, 20], extent of linkage disequilibrium (LD), heritability of the trait [20, 21], relationship between training and validation populations [10] and the genetic architecture of the trait, which relates to the relative size of allele substitution effects at quantitative trait loci (QTL) [10, 22].

The GBLUP and ssGBLUP methods usually assume that each SNP follows the same distribution [11, 12, 16, 23–25], thus, all SNPs have the same variance and the same weight for SNP variance. However, different genomic evaluation methods have been developed to allow the variance of the effect of SNPs to differ between SNPs. A priori information can be used to modify the distribution of SNP effects. Giving more variance to some SNPs allows these methods to take the presence of major genes or QTL that affect the trait of interest into account. For instance, various Bayesian methods, which estimate the effect of SNPs from animals that are both genotyped and phenotyped, have been proposed [10, 26–28]. The main difference between these Bayesian methods lies in the definition of an a priori distribution of the effects of SNPs. SNPs can be attributed to different distribution classes, which explain different parts of the total genetic variance, with one class possibly containing the SNPs that have no effect on the trait. Because animals need to be phenotyped and genotyped to apply Bayesian methods, phenotypes from non-genotyped animals cannot be included. In dairy breeding programs, genotypes are mainly determined on the males whereas phenotypes come from the females. Thus, daughter yield deviations (DYD) or de-regressed proofs are calculated to obtain pseudo-phenotypes for the males. However, multi-step methods may create bias in genomic predictions [29].

Other methods based on the ssGBLUP framework such as weighted ssGBLUP (WssGBLUP) or on the trait-specific marker-derived relationship matrix (TABLUP) have been proposed [30]. WssGBLUP is an extension of ssGBLUP in which weights for SNP variances are used when forming the genomic relationship matrix [12]. WssGBLUP can set more weight to SNPs that are in high LD with a causal mutation or associated with QTL with a relatively large effect. These weights are estimated from the variance explained by each SNP as presented by Wang et al. [23]. The weighting of SNP variances was also investigated by Zhang et al. [24] who proposed to use the same weight for SNPs that are within a defined window along the genome. The TABLUP method proposes to construct the genomic relationship matrix based on genotypes from a subset of pre-selected SNPs. Selection of SNPs can be performed after GWAS analysis or based on weights that are estimated with WssGBLUP. The selected SNPs are then equally weighted for the analyses [30]. Furthermore, an alternative to the previous methods is the gene content method proposed by Gengler et al. [31], which is based on a multiple trait model and considers the gene content for specific genotypes as a new trait. This method can combine information from SNPs and genotypes for a causal mutation [31, 32]. The number of alleles carried by each animal is considered as a second trait correlated to the quantitative trait. Then, the causal mutation is integrated directly in the ssGBLUP multiple-trait model. Its advantage is that it can be extended to multi-allelic genes and used when genotypes for a causal mutation are missing [33].

In French dairy goats, the first step towards genomic selection for milk production traits, udder type traits and somatic cell score was taken by Carillier-Jacquin et al. [8, 9] for French Alpine and Saanen dairy goat breeds. Carillier-Jacquin et al. [8, 9] compared ssGBLUP and other methods of genomic evaluation that require several steps (GBLUP or Bayesian methods). GBLUP and Bayesian methods usually use performances based on pseudo-phenotypes (DYD) whereas ssGBLUP is based on female performance. These authors found that ssGBLUP gave more accurate predictions of the genetic merit

of selection candidates than the previous official genetic evaluation that did not use genomic information, or the use of multi-step genomic methods. However, the increase in accuracy due to using genomic information was not expected to be high because the reference population was small.

Currently, the next step in the genomic evaluation of French dairy goats is to investigate better ways to use genotyping information to improve the accuracy of genomic evaluation. One possibility is to take prior knowledge about major genes into account. Several major genes have been identified, such as *DGAT1* for fat content [34] and α_{s1} *casein* for protein content [35]. For protein content, Carillier-Jacquin et al. [33] reported that the genetic variance explained by the α_{s1} *casein* gene reached 38% in the Saanen and 43% in the Alpine breed. The caprine α_{s1} *casein* gene has six alleles (*A*, *B*, *C*, *E*, *F* and *O*) that have been identified in the French dairy goat population. Allele *A* is predominant in the Alpine breed, whereas alleles *A*, *E* and *F* are the most frequent in the Saanen breed [33]. Carillier-Jacquin et al. [33] showed that integrating the α_{s1} *casein* gene for protein content with the gene content method improved the accuracy of genomic evaluation (+8 to 14% for Alpine and Saanen populations) compared with ssGBLUP.

In this study, our aim was to investigate different methods of genomic prediction that estimate and integrate the fact that chromosomal regions are strongly associated with a trait. Protein content in French dairy goats was analyzed by applying WssGBLUP, two alternatives of the WssGBLUP method, the TABLUP method and the gene content method. These methods were compared with pedigree-based BLUP and ssGBLUP based on the accuracies of predicted breeding values.

Methods

Animals, phenotypes and genotypes

The dataset used in this study was provided by the French national milk records system and included animals from the two main French dairy goat breeds, Alpine and Saanen. Phenotypes for protein content, pedigree data, genotypes and environmental fixed effects used in the ssGBLUP method were obtained from the official genetic evaluation of January 2016 [36]. Analyses were performed with a multi-breed dataset (Alpine and Saanen animals combined) and in two separate within-breed analyses.

The trait analyzed was protein content (g/kg) with measurements from 6,767,490 lactations and 2,458,453 females recorded between 1980 and 2010. Descriptive statistics (animal and record numbers, minimum, mean, maximum, coefficient of variation) for each breed are in Table 1.

The pedigree consisted of 2,543,789 animals (1,449,991 Alpine and 1,093,798 Saanen). In addition, it was completed with 36 unknown parent groups. Unknown parent groups were defined for each breed and for animals born before 1975, and then for cohorts born in 2-year windows up to 2010.

Animals that were genotyped with the Illumina goat SNP50 BeadChip (50K SNP) [37] were also used in the analysis. Quality control (QC) for a dataset of 3347 genotyped animals (2020 Alpine and 1278 Saanen) and 53,347 SNPs was performed independently for each breed. SNPs with a minor allele frequency (MAF) lower than 1% and a call rate lower than 95% were removed. Hardy–Weinberg equilibrium was also tested and the associated Chi squared statistic was calculated for each SNP. SNPs with a Chi squared statistic higher than 24 were removed. Finally, animals with a SNP call rate lower than 99% were discarded from the analyses. After QC, 2955 (1749 Alpine and 1206 Saanen) animals and 46,849 SNPs remained for further analyses. Some SNPs within the α_{s1} *casein* gene were present on the 50 K SNP but since they did not pass QC, they were removed [33].

Genotypes for the α_{s1} *casein* gene were available for 3696 Alpine individuals (2154 males and 1542 females), and 3506 Saanen individuals (2049 males and 1457 females) born between 1982 and 2012. The α_{s1} *casein* gene is located on caprine chromosome 6 at 82 Mb and is multi-allelic in the French dairy goat population, with six different alleles (*A*, *B*, *C*, *E*, *F* and *O*) and 19 genotypes detected among the 21 possibilities (*FO* and *OO* genotypes have never been detected in the French dairy goat population) [33]. Genotypes of animals with one missing allele were removed from the analysis. The estimated effects of the 19 α_{s1} *casein* genotypes on protein content were computed and reported previously [33]. Table 2 includes the number of animals (males and females for

Table 1 Summary statistics on protein content (g/kg) in Alpine and Saanen breeds

Breed	Number of lactations	Number of females with phenotypes	Minimum[a] (g/kg)	Mean[a] (g/kg)	Maximum[a] (g/kg)	CV
Alpine	3,844,071	1,392,399	10.47	30.42	54.81	0.11
Saanen	2,923,419	1,066,054	10.00	29.67	54.63	0.09

CV coefficient of variation

[a] Minimum, mean, maximum protein content

Table 2 Number of animals with information on the α_{s1} *casein* genotype and/or 50 K SNP genotypes

Breed	Gender	Animals with 50 K SNP genotype	Animals with α_{s1} *casein* genotype	Animals with both 50K SNP and α_{s1} *casein* genotype
Alpine	Males	512	2154	510
	Females	1237	1542	0
Saanen	Males	393	2049	393
	Females	813	1457	0

Alpine and Saanen breeds) used in this study with information on their α_{s1} *casein* and/or 50 K SNP genotypes.

Genomic prediction with and without considering information on the α_{s1} *casein* genotypes

ssGBLUP was implemented in 2017 in the official genetic evaluations for the two main French dairy goats. This method and pedigree-based BLUP were used as the reference method in our study and compared with WssGBLUP, two alternatives of the WssGBLUP method, TABLUP and the gene content method. Analyses were performed using the blupf90 software [38].

Single-step GBLUP (ssGBLUP) method

For both multi-breed and within-breed scenarios, the following model was applied:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{p} + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is a vector of performances (female phenotypes) for protein content (phenotypes are based on standardized 250-day lactation records). β is a vector of fixed effects including herd within year (32 years from 1980 to 2012) and within parity (1, 2 and ≥ 3) (188,933 levels in total); age at delivery within year and within region (four regions in France depending on goat breeding management) (3224 levels in total); month at delivery within year and region (1448 levels in total); and length of dry period within year and region (1107 levels in total); a fifth fixed effect for breed (two levels) was added for multi-breed analyses. $\mathbf{u}$ is a vector of random additive genetic effects assumed to be normally distributed $N(0, \mathbf{H}\sigma_u^2)$, $\mathbf{p}$ is a vector of random permanent environmental effects assumed to be normally distributed $N\left(0, \mathbf{I}\sigma_p^2\right)$, $\mathbf{e}$ is a vector of random residuals that is normally distributed $N(0, \mathbf{I}\sigma_e^2)$. $\mathbf{X}$ is the incidence matrix relating phenotypes to the fixed effects (β). $\mathbf{Z}$ is the design matrix allocating phenotypes to breeding values ($\mathbf{u}$) and $\mathbf{W}$ is the incidence matrix relating phenotypes to permanent environmental effects ($\mathbf{p}$).

Matrix $\mathbf{H}$ is the genetic relationship matrix combining SNP information and pedigree data, implemented as in Legarra et al. [12]:

$$\mathbf{H} = \begin{pmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix},$$

where $\mathbf{A}$ is a pedigree-based relationship matrix with indices 1 for ungenotyped animals and 2 for genotyped animals, and $\mathbf{G}$ is the genomic relationship matrix derived as in Christensen and Lund [11]:

$$\mathbf{G} = 0.95\frac{\mathbf{M}'\mathbf{M}}{2\sum_{i=1}^{m} p_i(1-p_i)} + 0.05\mathbf{A}_{22},$$

where m is the number of SNPs, p_i is the estimated allele frequency at locus i and $\mathbf{M}$ is a centered matrix of SNP genotypes.

Variance components were estimated by using the restricted maximum likelihood (REML) method in the remlf90 software [38].

Weighted ssGBLUP (WssGBLUP) method

Model 1 was also used for WssGBLUP but $\mathbf{G}$ was constructed differently. Solutions of genomic breeding values from ssGBLUP (Model 1) can be decomposed into SNP effects as modeled in Wang et al. [23]:

$$\hat{\mathbf{a}} = \mathbf{D}\mathbf{M}'\left[\mathbf{M}\mathbf{D}\mathbf{M}'\right]^{-1}\hat{\mathbf{u}}_g,$$

where $\hat{\mathbf{a}}$ is a vector of SNP effects, $\mathbf{D}$ is a diagonal matrix of weights (initially diagonal of 1 for the ssGBLUP), $\mathbf{M}$ is the centered matrix of SNP genotypes and $\hat{\mathbf{u}}_g$ the vector of GEBV from genotyped animals only. Variances of the effect of SNP i were estimated as:

$$\sigma_{u,i}^2 = 2\hat{a}_i^2 p_i(1-p_i),$$

where p_i is the allele frequency of SNP i. The vector of variances of SNP effects was normalized (the normalization process ensured that the sum of the variances remained constant and was equal to the number of SNPs) and used as weights in matrix $\mathbf{D}$ to construct the weighted matrix $\mathbf{G}$ ($\mathbf{G}^*$) as described in Wang et al. [23]:

$$\mathbf{G}^{*} = 0.95\frac{\mathbf{M}'\mathbf{D}\mathbf{M}}{2\sum_{i=1}^{\mathbf{m}} p_i(1-p_i)} + 0.05\mathbf{A}_{22}.$$

GEBV were estimated again with Model 1 by considering weights for each SNP via the $\mathbf{G}^{*}$ matrix included in the $\mathbf{H}$ matrix. This process was carried out iteratively with weights estimated at each iteration as described in Wang et al. [23]. Wang et al. [23] have shown that WssGLUP with only very few iterations may be sufficient to reach a maximum accuracy of GEBV and SNP effects. In this study, we analyzed the influence of the number of iterations (1–10) on the accuracy of genomic predictions.

As proposed by Zhang et al. [24], other methods can be considered to calculate the weight for SNPs in the $\mathbf{D}$ matrix. These methods assign the same weight to several consecutive SNPs within a chromosomal region. Modifications of the WssGBLUP method were considered in this study and the individual weights were computed as follows: (1) the maximum weight of SNPs included in the chromosomal region, or (2) the sum of the weights of the SNPs included in the chromosomal region. These weights were calculated based on the weights estimated with the WssGBLUP. In the end, the vector of the weights was normalized in such a way that the sum of all weights remained constant and equal to the number of SNPs. Chromosomal regions of various lengths were tested: 2, 5, 10, 20, 40, 80, 100, 150, 200 and 250 consecutive SNPs with non-overlapping windows. Hereafter, these methods are named WssGBLUP$_i$ where i denotes the method used to calculate the weights (Max or Sum).

Trait-specific marker-derived relationship matrix (TABLUP) method

Only a subset of SNPs that are more or less associated with protein content was selected to build the $\mathbf{G}$ matrix. One of our objectives was to investigate how the genetic architecture of protein content could be taken into account in the ssGBLUP method. Thus, TABLUP was applied by selecting a subset of SNPs according to their effect on the trait (estimated from the WssGBLUP method described previously). A total of 5000, 10,000, 15,000, 20,000, 25,000, 30,000, 35,000 or 40,000 SNPs were selected to construct $\mathbf{G}$. Two scenarios were tested in which either the most or the least strongly associated SNPs were selected. GEBV were estimated with Model 1 and the $\mathbf{G}$ matrix that was built based on the selected SNPs without weights ($\mathbf{D} = \mathbf{I}$).

Gene content method

The gene content method estimates the GEBV for each animal by taking information on the α_{s1} *casein* genotype, genotypes from the 50K SNP and pedigree into account through a multiple-trait model. The model used here was the same as in [33]:

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{p} + \mathbf{e}\\
\mathbf{y}_A &= \mu_A + \mathbf{Z}_A\mathbf{u}_A + \mathbf{e}_A\\
\mathbf{y}_B &= \mu_B + \mathbf{Z}_B\mathbf{u}_B + \mathbf{e}_B\\
\mathbf{y}_C &= \mu_C + \mathbf{Z}_C\mathbf{u}_C + \mathbf{e}_C\\
\mathbf{y}_E &= \mu_E + \mathbf{Z}_E\mathbf{u}_E + \mathbf{e}_E\\
\mathbf{y}_F &= \mu_F + \mathbf{Z}_F\mathbf{u}_F + \mathbf{e}_F\\
\mathbf{y}_O &= \mu_O + \mathbf{Z}_O\mathbf{u}_O + \mathbf{e}_O
\end{aligned}, \qquad (2)$$

where $\mathbf{y}$ is a vector of female performances for protein content. Fixed effects ($\boldsymbol{\beta}$), random effects ($\mathbf{u}$, $\mathbf{p}$ and $\mathbf{e}$) and incidence matrices $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ are the same as in Model 1. $\mathbf{y}_A$, $\mathbf{y}_B$, $\mathbf{y}_C$, $\mathbf{y}_E$, $\mathbf{y}_F$, and $\mathbf{y}_O$ are vectors of gene content for alleles *A*, *B*, *C*, *E*, *F* and *O*. This corresponds to the number of copies carried by each animal (i.e., 0, 1 or 2). For ungenotyped animals, the value was set to missing. $\mu_A, \mu_B, \mu_C, \mu_E, \mu_F$, and μ_O are the mean fixed effects for alleles *A*, *B*, *C*, *E*, *F* and *O*, $\mathbf{Z}_A$, $\mathbf{Z}_B$, $\mathbf{Z}_C$, $\mathbf{Z}_E$, $\mathbf{Z}_F$, and $\mathbf{Z}_O$ are the incidence matrices relating observations to the random genetic effect ($\mathbf{u}_A$, $\mathbf{u}_B$, $\mathbf{u}_C$, $\mathbf{u}_E$, $\mathbf{u}_F$ and $\mathbf{u}_O$) of gene content for each allele and $\mathbf{e}_A$, $\mathbf{e}_B$, $\mathbf{e}_C$, $\mathbf{e}_E$ and $\mathbf{e}_O$ are the random residual errors for each of the six alleles. For $i \in \{A,B,C,E,F,O\}$, $\mathbf{u}_i$ are normally distributed such that $Var(\mathbf{u}_i) = \mathbf{H}\sigma^2_{u_i}$ and $\sigma^2_{u_i} = 2p_i(1-p_i)$, where p_i is the frequency of allele i at the α_{s1} *casein* locus. Covariances between genetic values ($\mathbf{u}$) and genetic effects of gene content ($\mathbf{u}_A$, $\mathbf{u}_B$, $\mathbf{u}_C$, $\mathbf{u}_E$, $\mathbf{u}_F$ and $\mathbf{u}_O$) were modeled as in Carillier-Jacquin et al. [33]. Variance and covariance parameters from this model were estimated using the restricted maximum likelihood (REML) algorithm implemented in the remlf90 software.

Accuracy of genomic predictions

Genomic evaluations were performed from all phenotypes recorded until January 2010, but we were also interested in the prediction of genotyped animals that constituted our reference population. This reference population was composed of 905 sires born between 1993 and 2012 and genotyped with the 50K SNP chip (Table 2) and was split into a training population of 554 sires born from 1993 to 2007 (307 Alpine and 247 Saanen) with phenotypes of their daughters recorded until January 2010), and 351 validation sires born from 2008 to 2012 (205 Alpine and 146 Saanen) with no daughters in January 2013 (daughters of these animals were removed from the dataset). Then, GEBV and DYD computed from the official genetic evaluation of January 2016 were compared for the 351 animals in the validation set. DYD were average performance values for the daughters corrected

for environmental effects and merit of the dam, and they were weighted by effective daughter contributions as described in VanRaden and Wiggans [39]. Accuracy of genomic predictions was assessed as the Pearson correlation between GEBV estimated with each model and DYD. Pearson correlations obtained with different methods were tested using the Hotelling-Williams test [40].

Results and discussion

The most frequent α_{s1} *casein* genotypes are *AA* for the males and *AE* for the females in the Alpine breed, and *AE* for the females and *EE* for the males in the Saanen breed (present in more than 50% of the animals). Allele *C* is rather rare (less than 5% of the animals carry this allele) in the two breeds. The largest differences in genotype frequency between Alpine and Saanen populations were observed for genotypes *AA* (49% in Alpine vs. 7% in Saanen), *EE* (3% in Alpine vs. 32% in Saanen) and *AE* (49% in Saanen *vs*. 30% in Alpine). These results were consistent with the previous work of Carillier-Jacquin et al. [33] in which fewer genotypes were available. Protein content was analyzed knowing that this trait is highly heritable in both Alpine and Saanen populations (0.5) [41].

Estimation of weights for SNPs with the WssGBLUP method

We compared different genomic methods. First, we used WssGBLUP because we wanted to identify the weights given to SNPs with this method, in order to determine if the chromosomal region including the α_{s1} *casein* gene was considered in the analyses. WssGBLUP is an iterative method, and 10 iterations were performed for multi-breed analyses and within-breed analyses. Accuracy of genomic predictions was evaluated at each iteration (results not shown). The highest accuracies were obtained at the second iteration as reported by Wang et al. [23] and then decreased slightly. Thus, all the results presented for the WssGBLUP multi-breed and within-breed analyses are those obtained for the second iteration (see Fig. 1). The top 50 SNPs (with the biggest weights) were compared between the three analyses and were all located on chromosome 6 i.e. the multi-breed (between 71 and 86 Mb), Alpine (between 64 and 101 Mb) and Saanen analyses (between 71 and 92 Mb), and their weights ranged from 24 to 115 for multi-breed, from 23 to 45 for

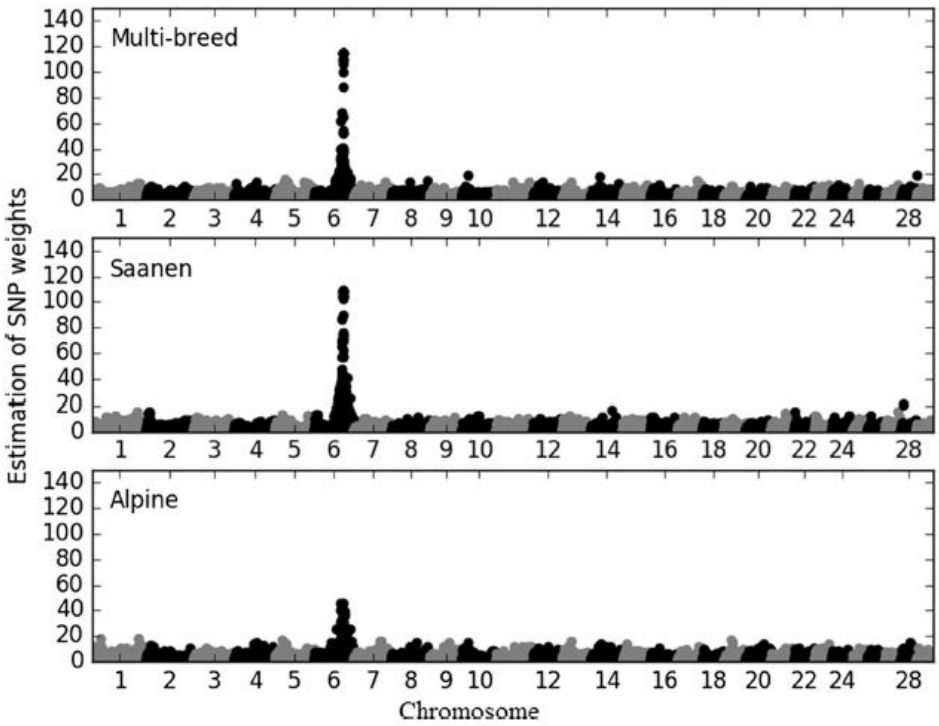


Fig. 1 Estimated weights of SNPs included in the second iteration of the WssGBLUP approach for multi-breed, Alpine and Saanen populations. WssGBLUP used GEBV of genotyped animals and genotypes to estimate weights for each SNP. The estimated weight of each SNP is represented along the genome

Alpine and from 30 to 108 for Saanen analyses. Among these SNPs, 16 were common to the three analyses and located between 78 and 82 Mb; 11 SNPs were common to the Saanen and multi-breed analyses and located between 79 and 83 Mb; 16 SNPs were common to the Alpine and multi-breed analyses and located between 77 and 86 Mb; and only one SNP was common to both the Alpine and Saanen analyses and located at 76 Mb.

WssGBLUP can be used not only for genomic prediction but also for QTL detection as in GWAS [23, 24]. In French dairy goat data, the chromosomal regions detected with WssGBLUP were on caprine chromosome 6, which includes a well-known region that was previously located and described by Martin et al. [34] in a GWAS study. They performed linkage analyses (LA) and linkage disequilibrium (LD) analyses on 1941 dairy goats distributed in 20 half-sib families using all females and their 20 sire genotypes and detected a large QTL between 82.5 and 82.8 Mb on chromosome 6. In our study, SNPs with the biggest weights for SNP variances were located within this region.

The WssGBLUP method developed by Wang et al. [23] has some limitations. Weights for SNP variances are estimated by using a whole-genome regression, which can result in their unstable prediction due to multi-collinearity between SNPs because of LD between SNPs. In our study, we tested common weights for several SNPs instead of individual weights for SNP variances, using $WssGBLUP_{Max}$ or $WssGBLUP_{Sum}$. These methods are expected to limit the large variation in prediction of weights for SNP variances by smoothing weights of SNPs that are in the same window. In our study, $WssGBLUP_{Max}$ and $WssGBLUP_{Sum}$ gave higher accuracies of genomic prediction than the classical WssGBLUP. With $WssGBLUP_{Max}$ or $WssGBLUP_{Sum}$, window sizes were used to allocate the same weights to consecutive SNPs. Another approach would be to use the LD between SNPs, which could limit the multi-collinearity between the SNPs used in the genomic evaluation. Since the weight of SNPs is included through the **D** matrix, this matrix can be replaced by the weights derived from the GWAS approach.

Including the effect of the α_{S1} *casein* gene in WssGBLUP or gene content methods

Figure 2 presents accuracies of genomic evaluation for pedigree-based BLUP, ssGBLUP, gene content and WssGBLUP in a multi-breed population and in the Alpine and Saanen breeds. Accuracies with pedigree-based BLUP (0.72 in multi-breed, 0.71 in Alpine and 0.66

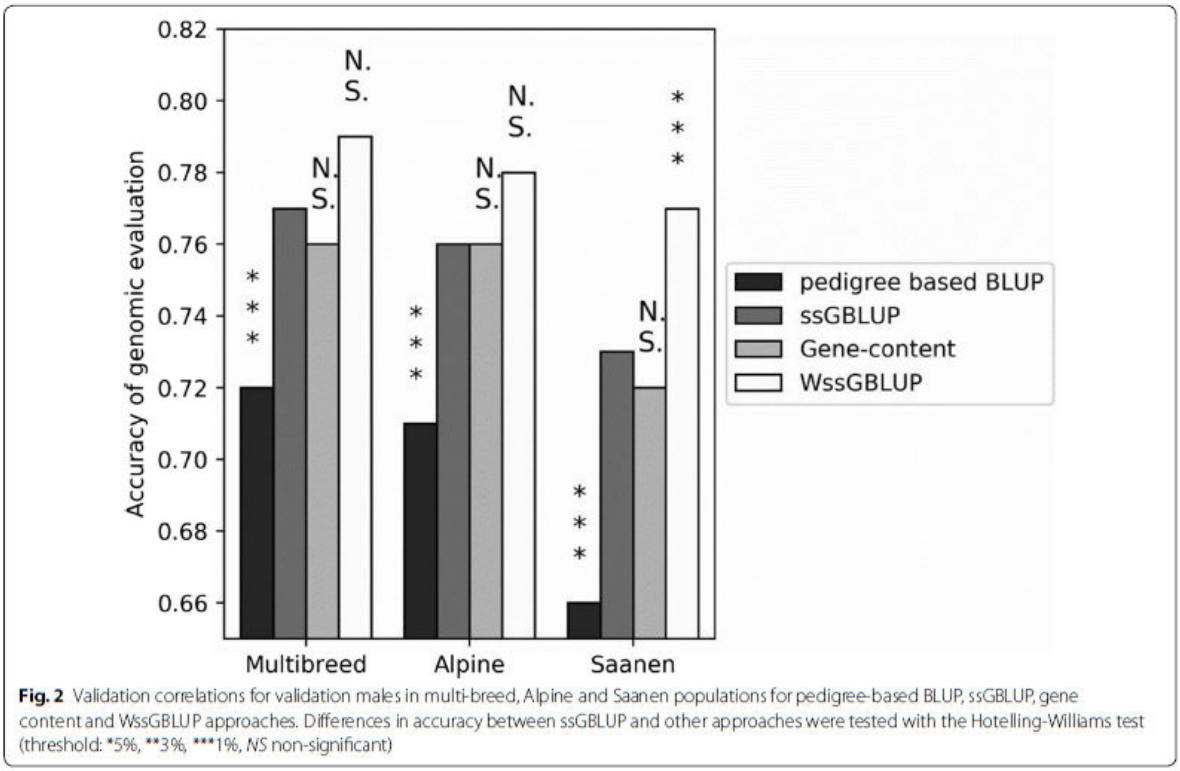


Fig. 2 Validation correlations for validation males in multi-breed, Alpine and Saanen populations for pedigree-based BLUP, ssGBLUP, gene content and WssGBLUP approaches. Differences in accuracy between ssGBLUP and other approaches were tested with the Hotelling-Williams test (threshold: *5%, **3%, ***1%, *NS* non-significant)

in Saanen) were lower than accuracies with ssGBLUP (0.77 in multi-breed, 0.76 in Alpine and 0.73 in Saanen), gene content (0.76 in multi-breed, 0.76 in Alpine and 0.72 in Saanen) or WssGBLUP (0.79 for multi-breed, 0.78 for Alpine and 0.77 for Saanen). The gene content method did not improve accuracy of genomic predictions for the three populations compared to ssGBLUP (accuracy was 1 percent point lower for gene content in the multi-breed and Saanen analyses and identical in the Alpine analysis). In addition, accuracies with WssGBLUP were significantly higher than with ssGBLUP for the Saanen population (+4 percent points). We did not observe any significant difference between ssGBLUP and WssGBLUP for multi-breed and Alpine populations.

Previously, Carillier-Jacquin et al. [33] used the gene content and ssGBLUP methods to analyze protein content in French dairy goats. Accuracies obtained with ssGBLUP were higher in our study than in Carillier-Jacquin et al. [33] for the multi-breed (+5 percent points) and Alpine (+8 percent points) analyses, and slightly lower for the Saanen analysis (−2 percent points). A similar trend was observed with the gene content method, with +1 percent point for multi-breed, +8 percent points for Alpine and −14 percent points for Saanen in our study compared to Carillier-Jacquin et al. [33]. The main difference between our study and that of Carillier-Jacquin et al. [33] was the number of animals genotyped with the 50 K SNP chip, number of α_{s1} *casein* genotypes, and the size and composition of the training and validation sets. In our study, 82 males and 2050 females genotyped with the 50 K SNP chip and 50 females and 878 males genotyped for the α_{s1} *casein* gene were added. In Carillier-Jacquin et al. [33], the reference population consisted of a training set with 677 animals born between 1993 and 2009 (384 Alpine and 293 Saanen), and a validation set with 146 animals born between 2010 and 2011 (86 Alpine and 60 Saanen). In our study, we had 554 animals born between 1993 and 2007 (307 Alpine and 247 Saanen) in the training set and 351 animals born between 2008 and 2012 (205 Alpine and 146 Saanen) in the validation set. The main difference between the Carillier-Jacquin et al. study and that reported here was the size of the validation population (2 versus 5 years in our study). The slightly improved results that we obtained may be explained by the larger reference population (823 animals in Carillier-Jacquin et al. [33] compared to 905 in our study), a well-known factor in the literature on genomic selection. For instance, VanRaden et al. [42] report a gain of +5 percent points between genomic prediction and parent average by adding 1000 animals in the training population. These results were consistent with the higher accuracy obtained in the multi-breed analysis compared to the within-breed analyses, especially if the trait has the same genetic determinism in the two breeds that are combined (which is the case for protein content). Accuracy is expected to improve even more the size of the reference population continues to grow over the years.

Carillier-Jacquin et al. [33] showed that the gene content method was more accurate than ssGBLUP (+3 percent points for multi-breed, +5 percent points for Alpine and +11 percent points for Saanen). However, in our study, accuracies of genomic prediction were the same for the gene content method and ssGBLUP. The goat α_{s1} *casein* gene has six alleles in the two main French dairy goats and genotype frequencies vary considerably with some being rare. Predicting α_{s1} *casein* genotypes with the gene content method for non-genotyped animals remains difficult in this case, especially in French dairy goats, for which the number of non-genotyped animals is large compared with that of genotyped animals (only 0.3% of the population is genotyped for the α_{s1} *casein* gene), and 40% of females have unknown parents. This may explain why the gene content method did not outperform ssGBLUP.

The genetic architecture of protein content is similar between the Alpine and Saanen breeds. However, the gain in accuracy with the genomic evaluation methods (ssGBLUP, gene content and WssGBLUP) compared to pedigree-based BLUP was greater for the Saanen than the Alpine breed. As discussed by Carillier-Jacquin et al. [9], the greater gain observed for the Saanen breed between pedigree-based BLUP and genomic evaluation may be explained by a higher level of inbreeding (2.3% in Saanen and 1.8% in Alpine), and a higher kinship coefficient between the training and validation sets (2.4% in Saanen and 1.1% in Alpine using genomic data).

For prediction of GEBV, WssGBLUP was more efficient than gene content, which may be due to the construction of the 50K SNP chip. The region around the α_{s1} *casein* gene was enriched in SNPs in the 1-Mb region at 82 Mb on chromosome 6 (the region that contains the α_{s1} *casein* gene). Overall, 40 SNPs are present within this 1-Mb region, whereas on average only 20 SNPs per Mb are located outside of this region on chromosome 6 or on other chromosomes. Moreover, the Chi squared test between α_{s1} *casein* genotypes and each SNP on chromosome 6 revealed a very strong correlation between α_{s1} *casein* genotypes and SNPs on the 50K SNP chip in this region (results not shown). Giving more weight to SNPs that are more strongly associated with protein content seems to be more efficient to capture the effect of the α_{s1} *casein* gene than using genotype data for this gene. Vallejo et al. [18] investigated the efficiency of WssGBLUP for bacterial cold water disease resistance, for which several QTL are identified. They observed an improvement of 4

percent points with WssGBLUP compared to ssGBLUP. In our study, we observed similar gains with WssGBLUP. Su et al. [43] also observed a superiority of the WssGBLUP over ssGBLUP in dairy cattle for milk traits.

Use of common weights on consecutive SNPs with WssGBLUP

WssGBLUP was significantly more predictive than other genomic evaluation methods for protein content in the Saanen breed but not in multi-breed or the Alpine breed. Zhang et al. [24] reported that $WssGBLUP_{Max}$ and $WssGBLUP_{Sum}$ increase the accuracy of genomic evaluation more efficiently than WssGBLUP. We evaluated these methods and Tables 3, 4 and 5 show the results on the validation population in the multi-breed, Alpine and Saanen populations, respectively using WssGBLUP and the two modified WssGBLUP methods (Max, Sum) according to the size of SNP windows. If identical results were obtained for different window sizes, they were merged in the same column. For the multi-breed population, accuracies of the analyses with $WssGBLUP_{Max}$ and $WssGBLUP_{Sum}$ were very similar and differed only with non-overlapping SNP windows of 40, 80, 100, 150, 200 and 250 SNPs, the accuracy (0.81) of $WssGBLUP_{Sum}$ being slightly higher than that of $WssGBLUP_{Max}$ (0.80). Otherwise, accuracies were equal to 0.79 with a window size of two SNPs and 0.80 for window sizes of five, 10 and 20 SNPs. Finally, accuracies of $WssGBLUP_{Max}$ and $WssGBLUP_{Sum}$ were slightly higher than that of WssGBLUP (0.79) and higher than that of ssGBLUP (0.77).

For both within-breed analyses, increasing the window size barely influenced accuracies. In the Alpine within-breed analysis, a maximum accuracy of 0.79 was reached with the $WssGBLUP_{Sum}$ method and a window size of 40 SNPs and thus, it outperformed WssGBLUP (0.78). For other window sizes (larger or smaller), accuracies with $WssGBLUP_{Sum}$ were equal to 0.78. With the $WssGBLUP_{Max}$ method, accuracies ranged from 0.77 for a window of two consecutive SNPs to 0.78 for windows of 5, 10, 20, 40, 80, 100, 150, 200 and 250 consecutive SNPs. In comparison, genomic evaluations with $WssGBLUP_{Max}$ and $WssGBLUP_{Sum}$ were more accurate than with ssGBLUP (0.76). In the Saanen within-breed analysis, accuracies of 0.78 were reached with $WssGBLUP_{Sum}$ for windows of 40, 80, 100, 150, 200 and 250 consecutive SNPs, and with $WssGBLUP_{Max}$ for windows of 80 and 100 consecutive SNPs. $WssGBLUP_{Max}$ and $WssGBLUP_{Sum}$ outperformed WssGBLUP (0.77) or even ssGBLUP (0.73). Accuracies of 0.77 were obtained with $WssGBLUP_{Sum}$ for windows of 2, 5, 10 and 20 consecutive SNPs and with $WssGBLUP_{Max}$ for windows of 2, 5, 10, 20, 40, 150, 200 and 250 consecutive SNPs.

$WssGBLUP_{Max}$ and $WssGBLUP_{Sum}$ slightly improved the accuracy of genomic predictions for protein content in French dairy goats compared to WssGBLUP. Similar results were observed by Zhang et al. [24] with $WssGBLUP_{Max}$ and $WssGBLUP_{Sum}$ compared to WssGBLUP on simulated data for five QTL. Zhang et al. [27] presented their results for a window size of 20 consecutive SNPs because when they used windows with more than 20 SNPs, accuracies decreased when many QTL affected a trait. This is due to most of the weight being assigned to the windows with large SNP effects and less weight to those with small SNP effects, which may introduce bias in the estimates. For the populations in our study, accuracies varied little with window size. However, 20 consecutive SNPs were not sufficient to reach the highest accuracies and 40 consecutive SNPs were more appropriate. Thus, for a trait that is influenced by few QTL, $WssGBLUP_{Max}$ or $WssGBLUP_{Sum}$ were more

Table 3 Validation correlations for 351 validation males in the multi-breed population using different WssGBLUP and different window sizes of non-overlapping SNPs

Method	Size of non-overlapping SNP windows			
	1	2	5/10/20	40/80/100/150/200/250
WssGBLUP[a]	0.79			
$WSSGBLUP_{Sum}$		0.79	0.80	0.81
$WSSGBLUP_{Max}$		0.79	0.80	0.80

[a] Each SNP has its own weight (WssGBLUP standard)

Table 4 Validation correlations for 205 validation males in the Alpine breed using different WssGBLUP and different window sizes of non-overlapping SNPs

Method	Size of non-overlapping SNP windows				
	1	2	5/10/20	40	80/100/150/200/250
WssGBLUP[a]	0.78				
$WSSGBLUP_{Sum}$		0.78	0.78	0.79	0.78
$WSSGBLUP_{Max}$		0.77	0.78	0.78	0.78

[a] Each SNP has its own weight (WssGBLUP standard)

Table 5 Validation correlations for 146 validation males in the Saanen breed using different WssGBLUP and different windows size of non-overlapping SNPs

Method	Size of non-overlapping SNP window				
	1	2/5/10/20	40	80/100	150/200/250
WssGBLUP[a]	0.77				
$WSSGBLUP_{Sum}$		0.77	0.78	0.78	0.78
$WSSGBLUP_{Max}$		0.77	0.77	0.78	0.77

[a] Each SNP has its own weight (WssGBLUP standard)

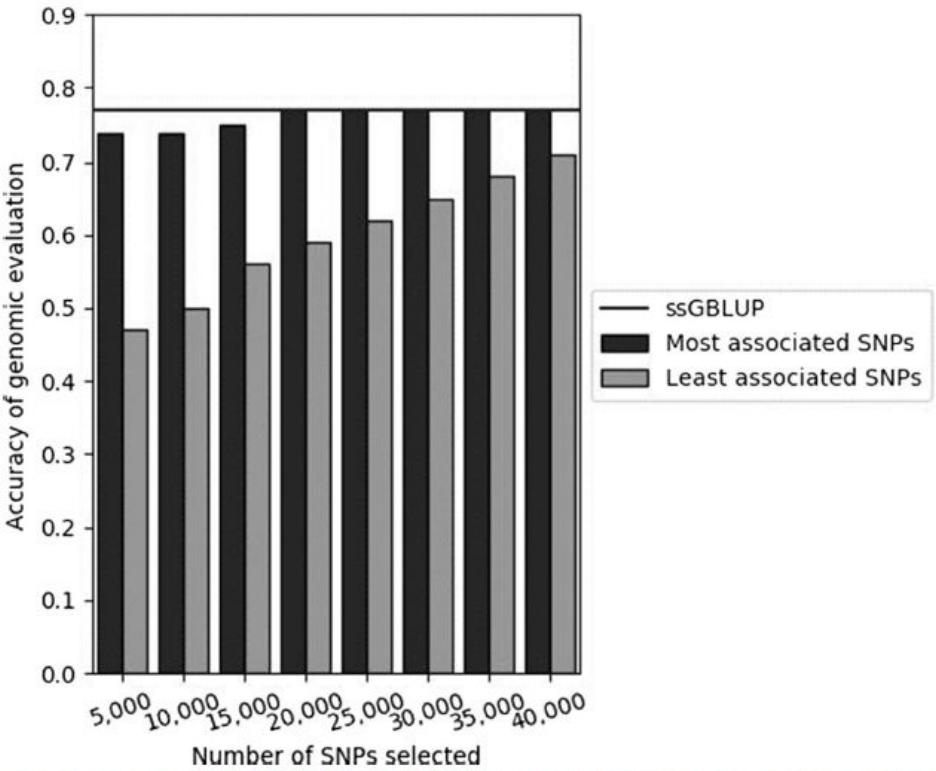


Fig. 3 Validation correlations for 351 validation males in the multi-breed population using the TABLUP approach. TABLUP consists in selecting a subset of the genotypes from the 50K SNPs according to their association with protein content (either the most strongly associated or the least strongly associated with protein content), which is done by selecting SNPs according to their weights estimated with the WssGBLUP approach

efficient to capture clear signals from QTL compared to WssGBLUP with one weight per SNP.

TABLUP method

To validate that ssGBLUP does capture the α_{s1} *casein* gene information, we used TABLUP that consists in selecting a subset of SNPs for constructing the **G** matrix, i.e. we selected the SNPs that were the most or the least strongly associated with protein content. Figure 3 shows the accuracies obtained with ssGBLUP and TABLUP for the multi-breed population according to the number of SNPs conserved (5000 to 40,000 SNPs) to construct the **G** matrix. Since results for both Alpine and Saanen breeds were similar to those for the multi-breed population, they are not shown.

First, for the SNPs that were the most strongly associated with protein content, TABLUP with only 5000 such SNPs led to a high accuracy of genomic prediction (0.74), which is close to that obtained with ssGBLUP (0.77). TABLUP reached the 0.77 accuracy of ssGBLUP with 20,000 such SNPs, which were distributed across the whole genome with on average 42% of the SNPs on each chromosome being retained and 54% on chromosome 6. This indicates that SNPs around the α_{s1} *casein* gene have been more selected than the others. Increasing the number of SNPs from 20,000 to 40,000, did not increase the accuracy furthermore. Conversely, for the SNPs that were the least strongly associated with protein content, TABLUP with 5000 such SNPs led to a very low accuracy (0.47) and increasing their number to 40,000 led to an increase in accuracy of 24 percent points (0.47 with 5000 SNPs and 0.71 with 40,000 SNPs) but accuracy remained significantly lower than that obtained by using the whole 50K SNP BeadChip (0.71 against 0.77).

Using different subsets of SNPs and the BayesA model, VanRaden et al. [44] compared accuracies of genomic predictions in Holstein breed cattle for 33 traits. They used 60K and high-density (HD) SNP panels, and added specific SNPs selected from whole-genome sequence data, which were SNPs based on their annotation (located on exons, splicing sites, indels, 2 kb upstream, 1 kb downstream, untranslated regions, SNPs with large effects). They showed that the highest accuracies were

obtained with the scenario that used 60K SNPs plus the top 1000 SNPs for all 33 traits. Increasing the number of SNPs (using the HD SNP panel for example) did not increase the accuracy of genomic predictions. However, adding selected SNPs from whole-genome sequence to a medium-density SNP BeadChip improved GEBV accuracies. These results agree with those that we obtained with the TABLUP method. In the near future, when whole-genome caprine sequence data become available, it will be possible to select sequence-based variants and add them to the 50K SNP data in the genomic evaluation model, which will improve the accuracy of genomic predictions in these species.

We undertook additional analyses (results not shown) in which SNPs were removed chromosome-wise with the ssGBLUP, WssGBLUP and gene content methods. The same accuracies were observed, regardless of the chromosome from which the SNPs were removed, except for chromosome 6 for ssGBLUP (0.77), WssGBLUP (0.79) and gene content (0.76). When SNPs from chromosome 6 were removed, accuracies dropped to 0.70 for ssGBLUP, 0.66 for WssGBLUP and 0.74 for gene content. However, the loss in accuracy with gene content was smaller than with ssGBLUP and WssGBLUP, i.e. using genotypes for the α_{s1} *casein* gene and SNPs from 28 chromosomes (except chromosome 6) is quite similar to using the 50K SNP chip. The missing genotypes from the 50K SNP chip (i.e. the SNPs on chromosome 6) did not add much information compared to the information contained by the genotypes for the α_{s1} *casein*. Results of TABLUP and chromosome-wise removal of SNPs showed that a part of the effect of the α_{s1} *casein* gene was retained by the ssGBLUP method, which basically does not include information on causal mutations. These results can be explained by the high coverage of SNPs on chromosome 6 around the α_{s1} *casein* gene.

Conclusions

Our aim was to investigate different genomic evaluation methods (using α_{s1} *casein* genotypes and/or 50K SNP information) to integrate information on the α_{s1} *casein* gene in genomic evaluations of dairy goats. Using the trait-specific marker-derived relationship matrix did not improve accuracy of genomic evaluation, which was the same as that obtained by ssGBLUP with a selection of the 20,000 most important SNPs for protein content. With the gene content method, accuracies of genomic evaluation were not improved compared to ssGBLUP, which is probably due to the α_{s1} *casein* gene having many alleles and to the small number of genotyped animals. Putting more weight on SNPs with larger effects improved accuracies of genomic evaluation using WssGBLUP, $\text{WssGBLUP}_{\text{Max}}$ and $\text{WssGBLUP}_{\text{Sum}}$. For $\text{WssGBLUP}_{\text{Max}}$ and $\text{WssGBLUP}_{\text{Sum}}$, accuracies were highest when a common weight was applied to non-overlapping windows of 40 SNPs. Gains in accuracies reached + 12 percent points for the Saanen, + 9 percent points for the multi-breed and + 8 percent points for the Alpine populations compared to a pedigree-based BLUP evaluation. WssGBLUP using common weights for SNPs within non-overlapping windows is efficient if the trait is influenced by few QTL and the true number of QTL is not known. WssGBLUP also combines fast computing and simplicity, and requires ssGBLUP to be run only twice.

Authors' contributions

MT performed the analysis and wrote the paper. MT, CRG and HL interpreted the results. CRG and HL revised and improved the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study would not have been possible without the goat SNP50 BeadChip developed by the International Goat Genome Consortium (IGGC): www.goatgenome.org. The authors thank Ignacy Misztal (University of Georgia, USA) for the blup90iod2 program. We sincerely thank the two anonymous reviewers and Julius Van der Werf (editors for Genetics Selection Evolution) whose comments/suggestions helped improve and clarify this manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available because they were partially produced by private professionals partnerships.

Consent for publication

Not applicable.

Ethics approval and consent to participate

All SNP genotyping was done according to the French National Guidelines for the care and use of animals for research.

Funding

The authors thank the French Genovicap and Phenofinlait programs (ANR, Apis-Gène, CASDAR, FranceAgriMer, France Génétique Elevage, French Ministry of Agriculture Agrifood, and Forestry) and the European 3SR project, which funded part of this work. The first author also received financial support from the Midi-Pyrénées region and the French National Institute for Agricultural Research (INRA) SELGEN program (INCoMINGS).

Publisher's Note

Received: 8 September 2017 Accepted: 30 May 2018
Published online: 15 June 2018

References

1. Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, et al. Genomic selection in French dairy cattle. Anim Prod Sci. 2012;52:115–20.

2. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92:433–43.
3. Baloche G, Legarra A, Sallé G, Larroque H, Astruc J-M, Robert-Granié C, et al. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. J Dairy Sci. 2014;97:1107–16.
4. Duchemin SI, Colombani C, Legarra A, Baloche G, Larroque H, Astruc J-M, et al. Genomic selection in the French Lacaune dairy sheep breed. J Dairy Sci. 2012;95:2723–33.
5. Brito LF, Clarke SM, McEwan JC, Miller SP, Pickering NK, Bain WE, et al. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. BMC Genet. 2017;18:7.
6. Auvray B, McEwan JC, Newman S. a. N, Lee M, Dodds KG. Genomic prediction of breeding values in the New Zealand sheep industry using a 50 K SNP chip. J Anim Sci. 2014;92:4375–89.
7. Mucha S, Mrode R, MacLaren-Lee I, Coffey M, Conington J. Estimation of genomic breeding values for milk yield in UK dairy goats. J Dairy Sci. 2015;98:8201–8.
8. Carillier C, Larroque H, Robert-Granié C. Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. Genet Sel Evol. 2014;46:67.
9. Carillier C, Larroque H, Palhière I, Clément V, Rupp R, Robert-Granié C. A first step toward genomic selection in the multi-breed French dairy goat population. J Dairy Sci. 2013;96:7294–305.
10. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
11. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42:2.
12. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92:4656–63.
13. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. J Dairy Sci. 2010;93:743–52.
14. Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection: predict the accuracy of genomic selection. J Anim Breed Genet. 2011;128:409–21.
15. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
16. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2008;177:2389–97.
17. Daetwyler HD, Swan AA, van der Werf JH, Hayes BJ. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. Genet Sel Evol. 2012;44:33.
18. Vallejo RL, Leeds TD, Gao G, Parsons JE, Martin KE, Evenhuis JP, et al. Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. Genet Sel Evol. 2017;49:17.
19. Andonov S, Lourenco DAL, Fragomeni BO, Masuda Y, Pocrnic I, Tsuruta S, et al. Accuracy of breeding values in small genotyped populations using different sources of external information—a simulation study. J Dairy Sci. 2017;100:395–401.
20. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3:e3395.
21. Viana JMS, Piepho H-P, Silva FF. Quantitative genetics theory for genomic selection and efficiency of genotypic value prediction in open-pollinated populations. Sci Agric. 2017;74:41–50.
22. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245–57.
23. Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res (Camb). 2012;94:73–83.
24. Zhang X, Lourenco D, Aguilar I, Legarra A, Misztal I. Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. Front Genet. 2016;7:151.
25. Strandén I, Garrick DJ. Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. 2009;92:2971–5.
26. Kizilkaya K, Fernando RL, Garrick DJ. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J Anim Sci. 2010;88:544–51.
27. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinform. 2011;12:186.
28. Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 2013;194:573–96.
29. Vitezica ZG, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. Genet Res (Camb). 2011;93:357–66.
30. Zhang Z, Ding X, Liu J, de Koning D-J, Zhang Q. Genomic selection for QTL-MAS data using a trait-specific relationship matrix. BMC Proc. 2011;5:S15.
31. Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 2007;1:21–8.
32. Legarra A, Vitezica ZG. Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using gene content multiple-trait BLUP. Genet Sel Evol. 2015;47:89.
33. Carillier-Jacquin C, Larroque H, Robert-Granié C. Including αs1 casein gene information in genomic evaluations of French dairy goats. Genet Sel Evol. 2016;48:54.
34. Martin P, Palhière I, Maroteau C, Bardou P, Canale-Tabet K, Sarry J, et al. A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content. Sci Rep. 2017;7:1872.
35. Grosclaude F, Mahé M-F, Brignon G, Di Stasio L, Jeunet R. A Mendelian polymorphism underlying quantitative variations of goat αs1-casein. Genet Sel Evol. 1987;19:399–412.
36. Larroque H, Astruc JM, Barbat A, Barillet F, Boichard D, Bonaiti B, et al. National genetic evaluations in dairy sheep and goats in France. In: Proceedings of the 62nd annual meeting of the European Federation of Animal Science: 29 August–2 September 2011; Stavanger; 2011.
37. Tosser-Klopp G, Bardou P, Cabau C, Eggen A, Faraut T, Heuven H, et al. Goat genome assembly, availability of an international 50 K SNP chip and RH panel: an update of the International Goat Genome Consortium projects. In: Proceedings of the International Plant and Animal Genome Conference XX: 14–18 January 2012; San Diego; 2012.
38. Misztal I, Tsuruta S, Strabel T, Auvrey B, Druet T, Lee DH. BLUPF90 and related programs. In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production: 19–23 August 20; Montpellier; 2002.
39. VanRaden PM, Wiggans GR. Derivation, calculation, and use of national animal model information. J Dairy Sci. 1991;74:2737–46.
40. Williams EJ. The comparison of regression variables. J R Stat Soc Ser B Methodol. 1959;21:396–9.
41. Bélichon S, Manfredi E, Piacère A. Genetic parameters of dairy traits in the Alpine and Saanen goat breeds. Genet Sel Evol. 1999;31:529–34.
42. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92:16–24.
43. Su G, Christensen OF, Janss L, Lund MS. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J Dairy Sci. 2014;97:6547–59.
44. VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 2017;49:32.

1.1.3. Analyses complémentaires de l'effet du gène de la caséine α_{s1} dans les évaluations génomiques sur le taux protéique

La Figure 48 présente les précisions des évaluations génomiques obtenues avec le ssGBLUP, le WssGBLUP et le gene content lorsque l'on considère uniquement les marqueurs SNPs d'un seul chromosome dans le cadre d'une analyse multirace. Par exemple, 1 sur l'axe des abscisses de la Figure 48 signifie que les SNPs des chromosomes 2, 3,…, 28 et 29 ont été supprimés de l'analyse. Comme attendu, les précisions les plus grandes sont obtenues lorsque tous les marqueurs sont considérés dans les analyses (indiqué par 50K sur la Figure 48), le WssGBLUP surpasse les méthodes ssGBLUP et gene content. Parmi les situations ne considérant les marqueurs que d'un seul chromosome, la méthode obtenant les meilleures précisions est le gene content (0,58 en moyenne), puis le ssGBLUP (0,51 en moyenne) et enfin le WssGBLUP (0,41 en moyenne). La prise en compte des génotypes du gène de la *caséine α_{s1}* avec le gene content améliore systématiquement les précisions génomiques, exceptées pour le chromosome 6. Pour ce cas, le ssGBLUP et le gene content ont des précisions similaires (0,67 et 0,66 respectivement) et le WssGBLUP a une précision légèrement inférieure (0,62). Il s'agit également de la situation pour laquelle les précisions génomiques sont les plus élevées parmi les situations ne considérant qu'un seul chromosome. Les mêmes résultats ont été observés pour les analyses Alpine et Saanen (résultats non présentés).

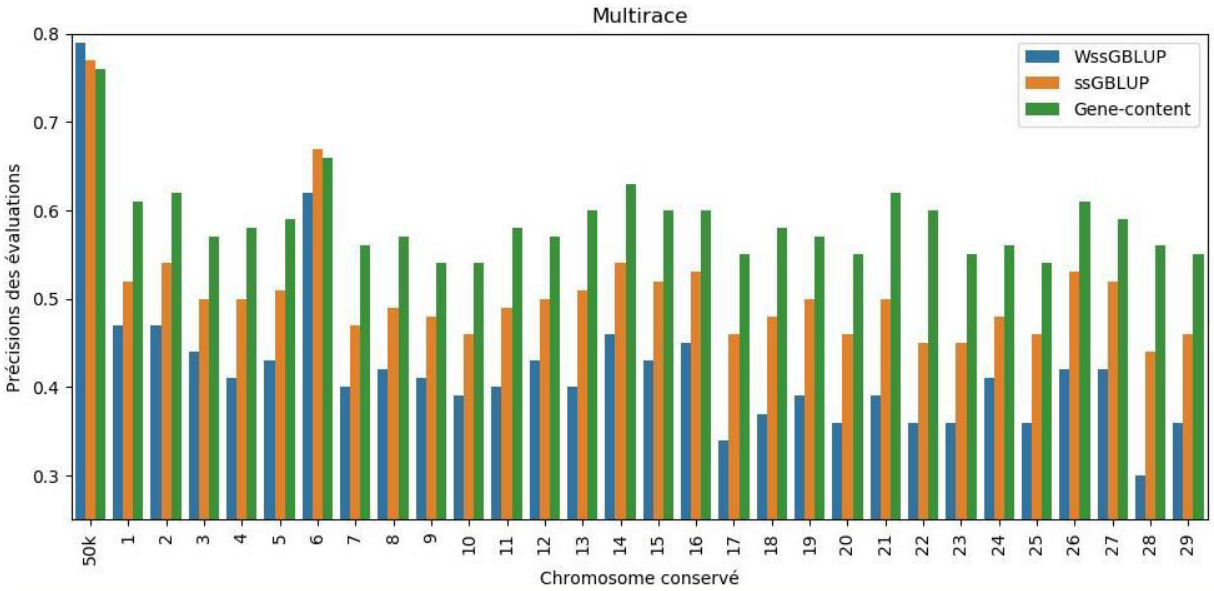


Figure 48. Précisions des évaluations génomiques multirace avec le ssGBLUP, le WssGBLUP et le gene content en utilisant les SNPs d'un seul chromosome pour les analyses multirace. 50K représente la situation en utilisant les SNPs de tous les chromosomes

Nous avons étudié les estimations des poids des SNPs obtenus avec les évaluations WssGBLUP multirace, Alpine et Saanen (Figure 49). Ces figures ne représentent que les SNPs du chromosome 6. Entre les analyses Alpine et Saanen, on observe que les SNPs avec les poids les plus forts en Alpine et en Saanen sont différents. Nous observons des SNPs avec des poids de 45 pour la race Alpine qui ont des poids proches de 0 pour la race Saanen. Pour la Saanen, les SNPs avec les poids les plus élevés (au-dessus de 100) ont des poids moyens pour la race Alpine (autour de 20). Les mêmes tendances sont retrouvées entre les analyses multirace et Alpine. Les SNPs avec un poids supérieur à 100 pour les analyses multirace sont autour de 20 pour les analyses Alpine. Pour la race Saanen, cet effet est moins important, les SNPs avec des poids supérieurs à 100 pour les analyses Saanen sont aussi les SNPs avec un poids supérieur à 100 pour les analyses multirace. Ces analyses montrent qu'il y a des différences entre les deux races. Ainsi, les SNPs les plus fortement associés aux génotypes du gène de la caséine α_{s1} pour

la race Alpine ne sont pas les SNPs les plus fortement associés aux génotypes du gène de la caséine α_{s1} pour la race Saanen. Les estimations des poids des SNPs pour les évaluations multirace et Saanen sont plus proches qu'entre les évaluations multirace et Alpine.

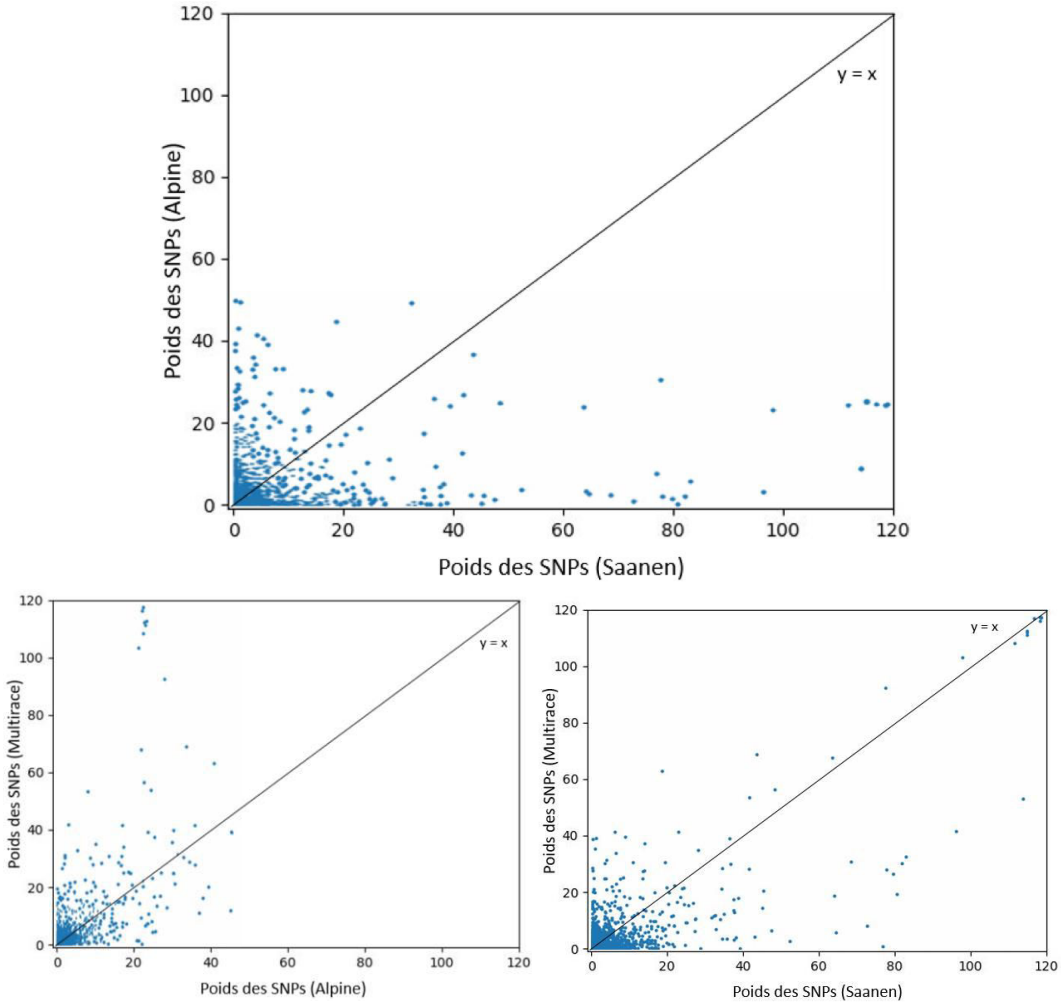


Figure 49. Comparaison des poids des SNPs obtenus pour les analyses multirace, Alpine et Saanen avec le WssGBLUP

1.1.4. Précisions des évaluations génomiques avec la méthode BayesR

De nombreuses méthodes bayésiennes ont été développées pour prendre en compte l'architecture génétique des caractères et intégrer des QTL ou des mutations causales dans les modèles d'évaluation génomique. L'approche BayesR est une de ces méthodes que nous avons étudiée sur les données caprines françaises.

1.1.4.1. Données utilisées pour les évaluations avec la méthode BayesR

La méthode BayesR fait partie de la famille des méthodes estimant directement les effets des marqueurs SNPs, les phénotypes utilisés ici sont les DYD. Le nombre de DYD, le nombre d'animaux dans le pedigree ainsi que le nombre de génotypes sont présentés dans le Tableau 21. La moyenne des DYD pour la race Saanen (0,41) est plus faible que celle de la race Alpine (0,85) (Tableau 22). La moyenne des DYD pour les analyses multirace est de 0,66.

Tableau 21. Nombres de phénotypes, d'animaux dans le pedigree et de génotypes pour les évaluations multirace, Alpine et Saanen utilisé dans l'approche BayesR

			Génotypes	
	Phénotypes	**Pedigree**	**Apprentissage**	**Validation**
Multirace	554	20 141	554	351
Alpine	307	10 700	307	205
Saanen	247	7 565	247	146

Les GEBV des animaux de la population de validation sont estimés comme :

$$GEBV_i = \sum_{j=1}^{m} x_{ij} g_j$$

où m correspond aux nombres de SNPs, x_{ij} est le génotype au marqueur j pour l'individu i et $g_j$est l'effet estimé du SNP j. Les 4 classes de SNPs utilisés pour ces analyses sont les mêmes que celles présentées dans le paragraphe 1.4 (Chapitre 2) :

$$\sigma_1^2 = 0, \sigma_2^2 = 0{,}001\sigma_g^2, \sigma_3^2 = 0{,}005\sigma_g^2 \; et \; \sigma_4^2 = 0{,}01\sigma_g^2$$

où σ_i^2 pour i ∈ {1, 2, 3, 4} est la part de variance génétique expliquée par le SNP i et σ_g^2 est la variance génétique totale. Pour chaque analyse (multirace, Alpine et Saanen), la méthode BayesR a été réalisée avec 25 000 itérations, dont 5000 itérations de burn-in, en utilisant le logiciel Bessie (Boerner and Tier, 2016)

Tableau 22. Statistiques descriptives des performances (DYD) utilisées pour les évaluations multirace, Alpine et Saanen avec le BayesR

	Moyenne	**Variance**	**Minimum**	**Maximum**
Multirace	0,66	0,56	-1,27	3,18
Alpine	0,85	0,52	-1,27	3,18
Saanen	0,41	0,50	-1,15	2,90

1.1.4.2. Résultats des évaluations génomiques avec le BayesR pour les analyses multirace, Alpine et Saanen

Le Tableau 23 présente le nombre moyen de SNP dans les 4 classes. On observe peu de différence entre les analyses pour la classe 4 (σ_4^2) avec 133, 130 et 129 SNPs en moyenne pour les analyses multirace, Alpine et Saanen respectivement. Cette classe est celle avec la plus grande part de variance génétique expliquée par un SNP. La classe 2 (σ_2^2) est celle qui contient le moins de SNPs en moyenne (119, 121 et 137 en moyenne pour les évaluations multirace, Alpine et Saanen respectivement). Une majorité de SNPs sont en moyenne placés dans la classe 1 (σ_1^2) et n'ont aucun effet sur le TP.

Tableau 23. Nombre moyen des SNPs dans les différentes classes (écart-type) du BayesR pour les analyses multirace, Alpine et Saanen

	$\boldsymbol{\sigma_1^2}$	$\boldsymbol{\sigma_2^2}$	$\boldsymbol{\sigma_3^2}$	$\boldsymbol{\sigma_4^2}$
Multirace	45 744 (431)	119 (97)	853 (504)	133 (62)
Alpine	46 013 (405)	121 (95)	585 (438)	130 (64)
Saanen	45 927 (351)	137 (104)	656 (402)	129 (57)

La Figure 50 représente le ratio entre l'effet moyen du SNP et l'estimation de son écart-type pour les évaluations multirace, Alpine et Saanen. Cette représentation permet d'identifier les SNPs dont les effets sur le caractère sont les plus importants. Pour les analyses multirace, Alpine ou Saanen, un pic sur le chromosome 6 est détecté. Pour les analyses multirace, 3 SNPs sur le chromosome 6 ont des effets plus importants que les autres. Ces 3 SNPs ont des valeurs de ratio égales à 5,63 ; 2,17 et 1,12 respectivement. Ils sont tous les 3 localisés dans le 82e Mb, région contenant le gène de la *caséine* α_{s1}. Les autres SNPs localisés dans ce Mb ont des valeurs assez faibles pour ce ratio (entre 0,003 et 0,543). Pour les autres chromosomes, la valeur du ratio effet du SNP moyen / écart-type est inférieure à 0,65. Pour la race Alpine, un SNP avec une valeur de ratio de 4,25 est détecté sur le chromosome 6 dans le 82e Mb. Comme pour les analyses multirace, les autres SNPs présents dans ce 82e Mb ont des valeurs assez faibles comprises entre 0,002 et 0,365. Enfin sur le reste du génome, les SNPs ont des valeurs de ratio inférieures à 0,44. Pour la race Saanen, le SNP avec le plus fort ratio (1,22) est localisé aussi dans le 82e Mb. Les autres SNPs localisés dans ce même Mb ont des valeurs faibles pour ce ratio comprises entre 0,001 et 0,576. Pour les autres chromosomes, les valeurs du ratio n'excèdent pas 0,42. Le SNP avec la valeur maximale pour les évaluations multirace, Alpine et Saanen correspond au SNP snp59417-scaffold980-295173. Dans nos analyses précédentes (analyses du déséquilibre de liaison avec le test du chi-deux, cf paragraphe 1.1.1, Chapitre 4), ce SNP avait de fortes valeurs du $-\log_{10}(p-valeur)$ (318 en multirace, 153 en Alpine et 146 en Saanen).

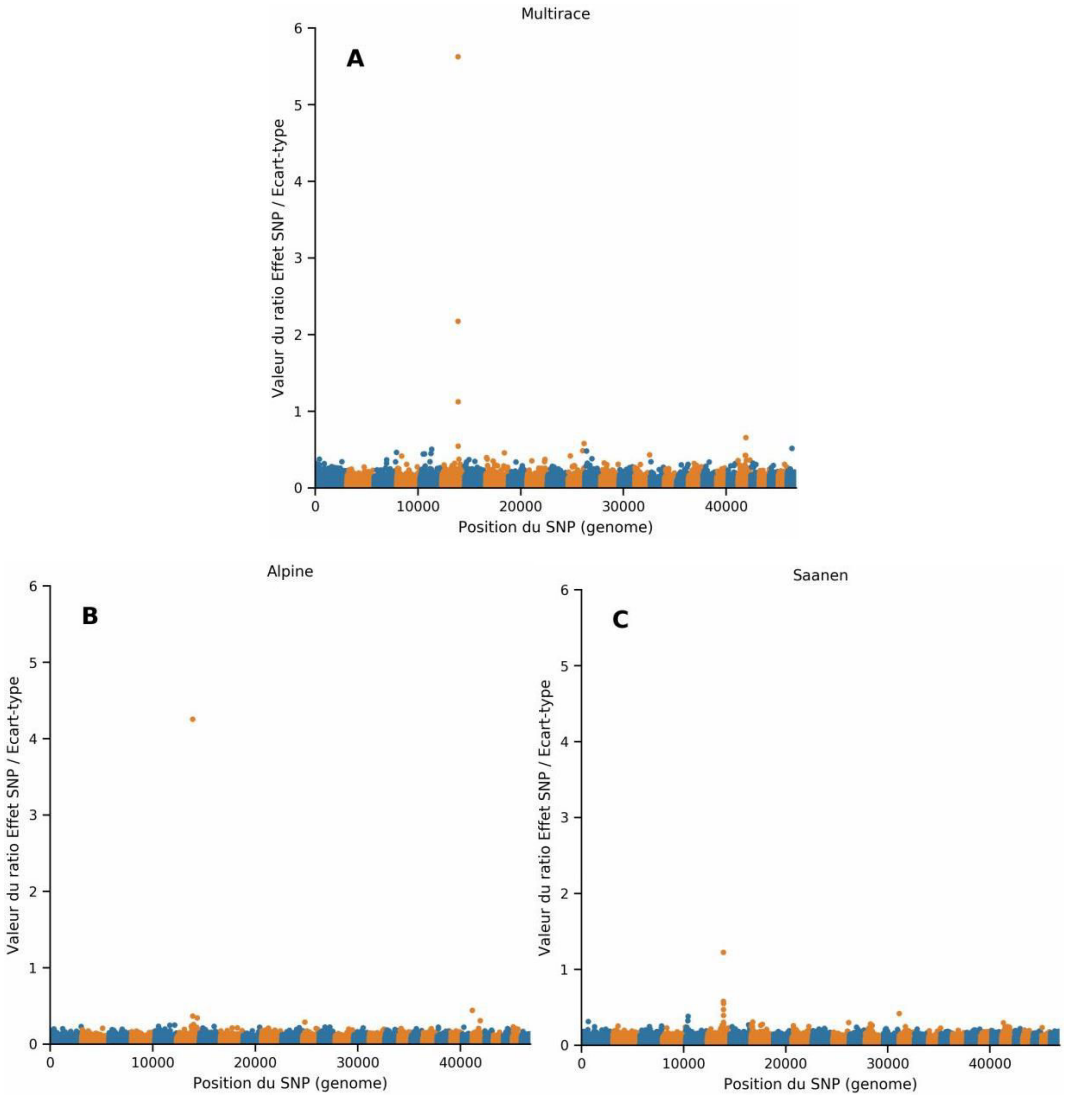


Figure 50. Rapport de l'effet moyen du SNP sur l'écart-type en fonction de la position du SNP le long du génome pour les évaluations génomiques multirace, Alpine et Saanen avec le BayesR

La Figure 51 présente les précisions des évaluations génomiques pour les analyses multirace, Alpine et Saanen avec le ssGBLUP (scénario E présenté dans le paragraphe 1.1, chapitre 3) et le BayesR. Les précisions avec le BayesR sont de 0,68 ; 0,66 et 0,67 pour les analyses multirace, Alpine et Saanen respectivement. Elles sont inférieures aux précisions du ssGBLUP (0,77 ; 0,76 et 0,73 pour les analyses multirace, Alpine et Saanen respectivement).

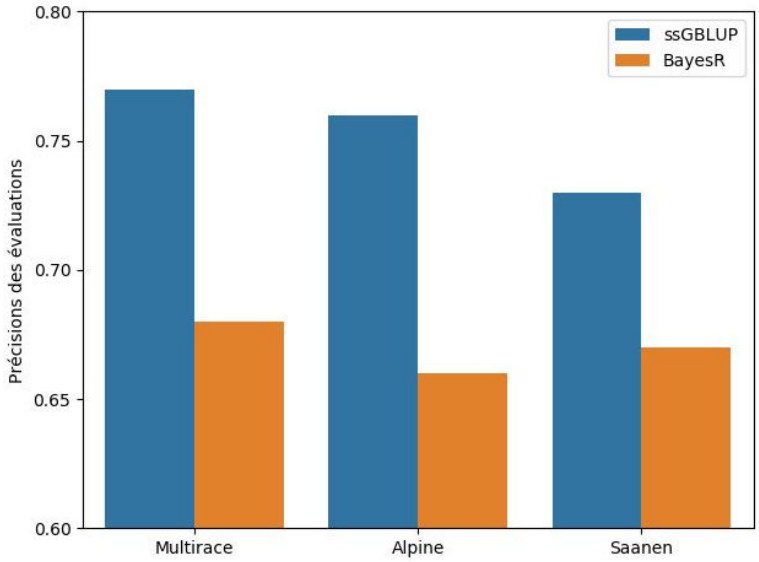


Figure 51. Précisions des évaluations génomiques avec le ssGBLUP et le BayesR pour les analyses multirace, Alpine et Saanen

1.1.5. Discussion

Les résultats du test du chi deux indiquent un LD fort entre le gène de la *caséine* α_{s1} et les SNPs de la puce 50K situés dans la région proche de ce gène. Certains génotypes pour le gène de la *caséine* α_{s1} sont même prédictibles à partir d'un seul SNP (comme le génotype EE avec le SNP « snp59416-scaffold980-293987 »). Ce résultat est cohérent avec la construction de la puce 50K (Tosser-Klopp et al., 2014). En effet, pour construire la puce caprine, la région orthologue du gène de la *caséine* α_{s1} bovine a été identifiée sur le génome caprin et cette région a été densifiée en SNPs par rapport au reste du génome : 56 SNPs ont été sélectionnés dans la région du gène de la *caséine* α_{s1} pour être présents sur la puce 50K. Il est donc assez naturel de retrouver des SNPs fortement liés aux génotypes de la *caséine* α_{s1} et que ces marqueurs captent en partie l'effet du gène de la *caséine* α_{s1}.

Le premier papier intitulé « *Weighted-single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene* » montre l'amélioration de la précision des évaluations génomiques pour le TP avec la méthode Weighted ssGBLUP et ses alternatives. Ces méthodes ont amélioré les précisions des évaluations jusqu'à 4, 3 et 5 points pour les évaluations multirace, Alpine et Saanen par rapport au ssGBLUP. Lorsque l'on analyse les poids des SNPs estimés, on observe que la région du gène de la *caséine* α_{s1} est bien identifiée (SNPs avec les poids les plus importants). Ces résultats sont cohérents avec ceux obtenus par Zhang et al., (2011), Wang et al., (2012), Lourenco et al., (2014) ou encore Vallejo et al., (2017). Le génotypage pour le gène de la *caséine* α_{s1} représente un coût supplémentaire. Des études sont en cours afin de prédire le génotype de la *caséine* α_{s1} à partir des génotypes de la puce 50K (Isabelle Palhière, communication personnelle). De plus, contrairement aux travaux de Carillier-Jacquin et al., (2016), les précisions n'ont pas été améliorées avec l'utilisation du gene content qui utilise à la fois les génotypes de la puce 50K et ceux du gène de la *caséine* α_{s1}. Ces résultats remettent en cause l'intérêt de génotyper les animaux pour ce gène si ces derniers sont génotypés avec la puce 50K. Entre notre étude et celle de Carillier-Jacquin et al., (2016), plusieurs paramètres étaient différents: le nombre de génotypes 50K disponibles (823 animaux génotypés contre 905 dans notre étude), la constitution des populations d'apprentissage et de validation (population de validation constituée d'animaux nés sur 2 millésimes pour Carillier-Jacquin et al., (2016) contre 5 millésimes dans notre étude). Les évaluations génomiques utilisant 28 chromosomes sur les 29 disponibles (on retire un chromosome des analyses) ou 1 chromosome (on conserve

uniquement les SNPs d'un seul chromosome) ont montré que le ssGBLUP capte bien une partie de l'effet du gène de la *caséine* α_{s1} même si cette méthode suppose que chaque SNP contribue de manière homogène et identique à la variabilité génétique. Cet effet est également mis en évidence avec la méthode TABLUP où la sélection des 20 000 SNPs les plus associés au caractère permet d'obtenir une précision aussi bonne que le ssGBLUP. Ce résultat peut s'expliquer par la densification en SNPs sur le chromosome 6 dans la région du gène de la *caséine* α_{s1}. Nos résultats sont cohérents avec une étude similaire réalisée par Rolf et al., (2010). Ils ont prédit la consommation journalière d'aliment et la consommation résiduelle pour la race Angus (825 animaux phénotypés et génotypés sur la puce Illumina BovineSNP50 BeadChip (Illumina, San Diego, CA)). Ils ont construit une matrice de parenté génomique en sélectionnant aléatoirement 100, 500, 1000, 2 500, 5 000, 10 000, 15 000, 20 000, 25 000, 30 000, 35 000 et 40 000 SNPs et ont fait 50 répétitions pour chaque scénario. Contrairement à notre étude, ils n'ont pas comparé les précisions génomiques des évaluations, mais ils ont calculé les corrélations entre les éléments hors-diagonaux des matrices **G** incomplètes (construit avec un sous-ensemble de SNPs) et de la matrice **G** complète (contenant tous les SNPs). Les corrélations sont supérieures à 98 % en sélectionnant au minimum 10 000 SNPs. Au-delà de ce seuil, les corrélations atteignent 99 %. Avec 100 SNPs sélectionnés aléatoirement, cette corrélation n'atteignait que 40%.

La méthode BayesR ne permet pas d'améliorer les précisions des évaluations génomiques caprines pour le TP que ce soit pour les analyses multirace, Alpine ou Saanen. Plusieurs études montrent pourtant que cette méthode est bien adaptée pour les caractères pour lesquels un gène majeur ou un QTL a été identifié. Erbe et al., (2012) ont analysé les précisions des évaluations génomiques avec un GBLUP et un BayesR pour le LAIT, la MG et la MP en Holstein et en Jersiaise. Ils disposaient d'animaux phénotypés (DYD) et génotypés avec la puce Illumina BovineSNP50 BeadChip (Illumina, San Diego, CA) et la puce Illumina High-Density Bovine SNP BeadChip (HD). Tous les animaux ont été imputés sur la puce HD. Les précisions moyennes, en utilisant les données imputées sur la HD, passent de 0,57 à 0,62 entre un GBLUP et un BayesR pour la race Holstein et de 0,43 à 0,45 pour la race Jersiaise. Ils ont également comparé les précisions obtenues avec la méthode BayesR en utilisant la puce HD ou la puce 50K. Dans ce cas, les précisions sont peu modifiées, passant en moyenne de 0,60 à 0,61 entre la puce 50K et la puce HD pour la race Holstein et de 0,46 à 0,45 pour la race Jersiaise. Kemper et al., (2015) ont réalisé une étude similaire en comparant les précisions des évaluations génomiques avec un GBLUP et un BayesR pour les races Holstein et Jersiaise. Ils ont analysé 5 caractères de production laitière (LAIT, MG, MP, TB et TP). Les animaux disposaient de YD (pour les femelles) et de DYD (pour les mâles) et étaient génotypés avec la puce HD ou la puce Illumina BovineSNP50 BeadChip. Tous les animaux étaient ensuite imputés sur la puce HD. Sur l'ensemble des caractères, ils observent des gains de précisions de 0,66 à 0,69 en moyenne pour la race Holstein avec le BayesR comparés au GBLUP et de 0,65 à 0,71 pour la race Jersiaise. Plusieurs raisons peuvent être à l'origine des résultats peu encourageants que nous avons obtenus avec l'approche BayesR sur nos populations caprines. Une de ces raisons est la taille relativement modeste de nos populations de référence. Cette taille réduite peut conduire à des incertitudes sur l'estimation des effets des SNPs (Teo, 2008), en particulier pour les SNPs avec une MAF faible. De plus, la densité de la puce ainsi que le LD dans la population étudiée sont également des paramètres à prendre en compte. Erbe et al., (2012) conclut que le BayesR a peu amélioré les précisions des évaluations intra-race car le LD entre SNPs est suffisamment élevé avec la puce 50K pour capturer les effets des QTLs. Cependant, le LD observé est plus faible dans les espèces caprines (paragraphe 3.4.2, Chapitre 1), on peut donc supposer que la densification en marqueurs permettrait de mieux capturer les effets des QTLs.

1.2. Intégration du gène *DGAT1* dans les évaluations génomiques du taux butyreux

1.1.1. Déséquilibre de liaison entre les génotypes DGAT1 (R251L et R396W) et les marqueurs de la puce 50K

Les Figure 52 et Figure 53 représentent les valeurs – log10 (p-valeur) du test du chi-deux entre les génotypes *DGAT1* (pour les mutations R251L et R396W) et les SNPs de la puce 50k du chromosome 14 pour les analyses multirace, Alpine et Saanen. Les génotypes manquants ont été ignorés pour la réalisation du test comme pour les analyses du gène de la caséine α_{s1}. Les génotypes *DGAT1* pour la mutation R251L (Figure 52) et les marqueurs SNPs du chromosome 14 sont fortement associés pour les analyses multirace et Saanen (plutôt en début et milieu de chromosome), aucune association n'est mise en évidence en race Alpine. Le top 10 des SNPs avec les valeurs de – log10 (p-valeur) les plus élevées sont localisées entre 8 Mb et 35 Mb pour les analyses multirace, entre 9 Mb et 90 Mb pour les analyses Alpine et entre 8 Mb et 45 Mb pour les analyses Saanen.

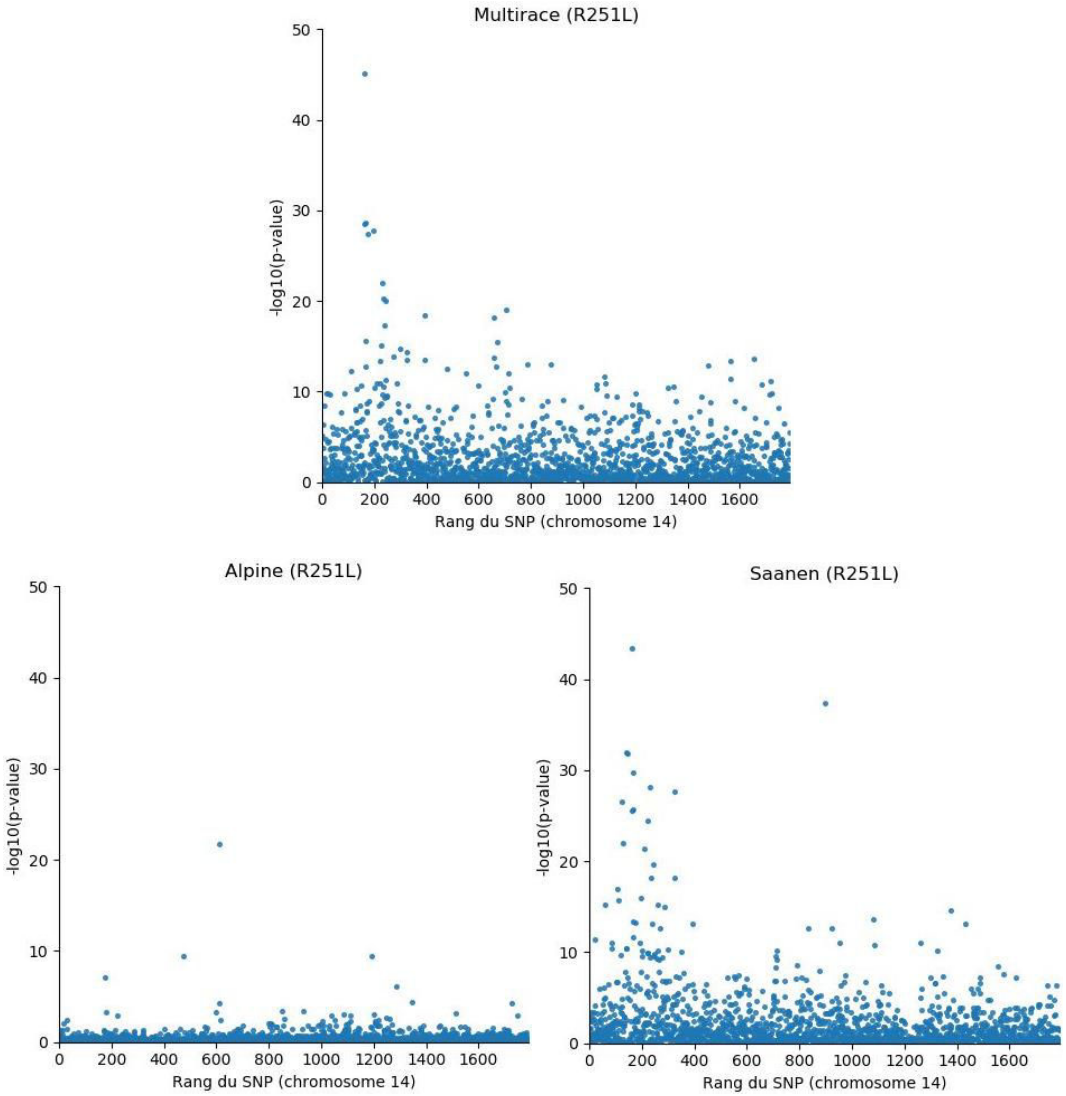


Figure 52. Valeurs de –log10(p-valeur) du test du Chi-Deux entre les génotypes DGAT1 (R251L) et les génotypes de la puce 50 K du chromosome 14 pour les analyses multirace, Alpine et Saanen

Pour la mutation R396W (Figure 53), on observe plusieurs SNPs fortement associés aux génotypes *DGAT1* au début du chromosome 14 pour les analyses multirace, Alpine et Saanen. Le top 10 des SNPs les plus associés sont situés entre 9 Mb et 12 Mb pour les analyses

multirace, entre 8 Mb et 17 Mb pour les analyses Alpine et entre 9 Mb et 20 Mb pour les analyses Saanen.

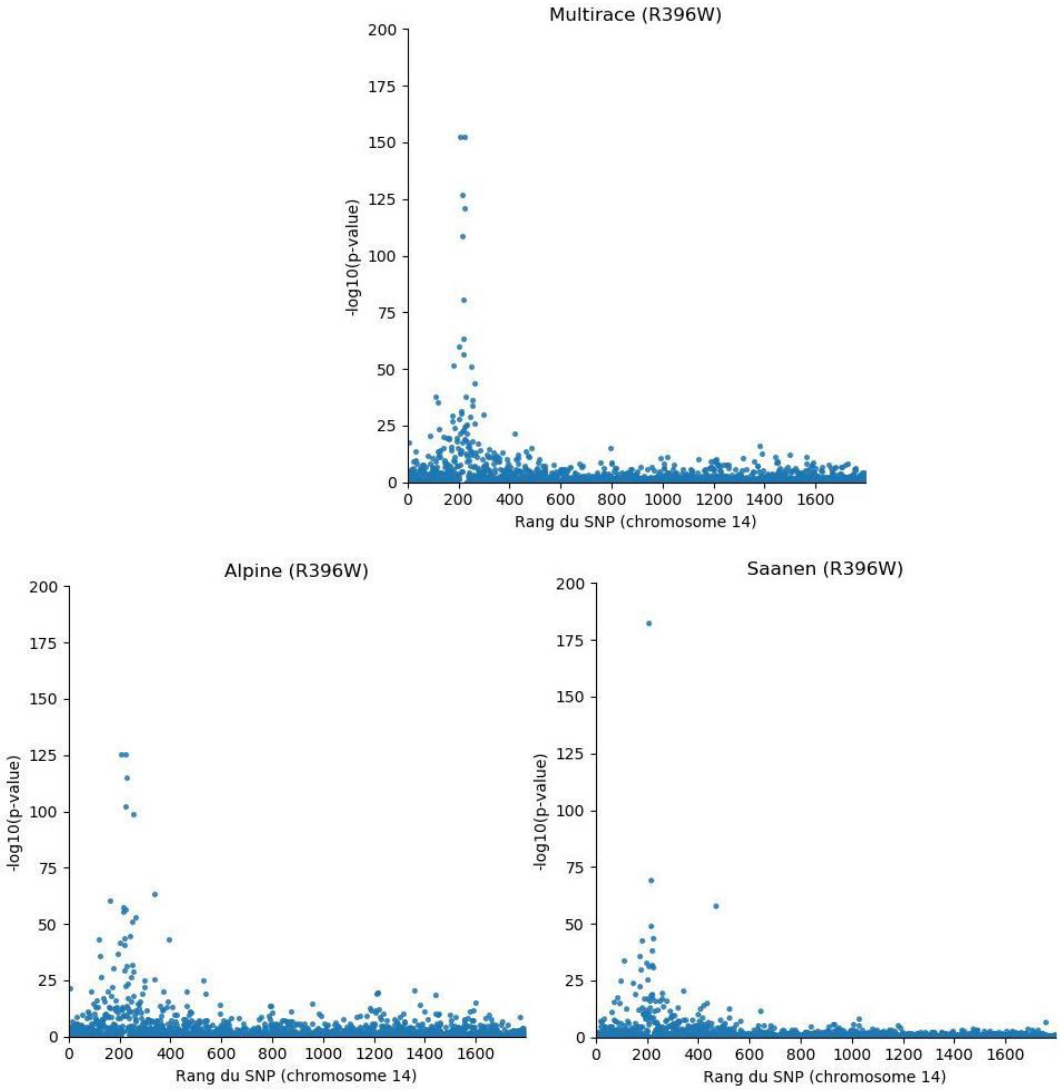


Figure 53. Valeurs de –log10(p-valeur) du test du Chi-deux entre les génotypes DGAT1 (R396W) et les génotypes de la puce 50 K du chromosome 14 pour les analyses multirace, Alpine et Saanen

Les top 10 des SNPs pour la mutation R251L ne sont pas communs entre les analyses multirace, Alpine et Saanen (Figure 54). La plupart des SNPs sont même spécifiques à chaque analyse (7 en multirace, 9 en Alpine et 8 en Saanen). Seuls 2 SNPs sont communs entre les analyses multirace et Saanen et un SNP est commun entre les analyses multirace et Alpine. Pour la mutation R396W, la proportion de SNPs communs (parmi les top 10) entre les analyses est plus élevée. 4 SNPs sont communs aux analyses multirace, Alpine et Saanen ; 3 SNPs en multirace, 5 en Alpine et 4 en Saanen sont spécifiques à chaque analyse.

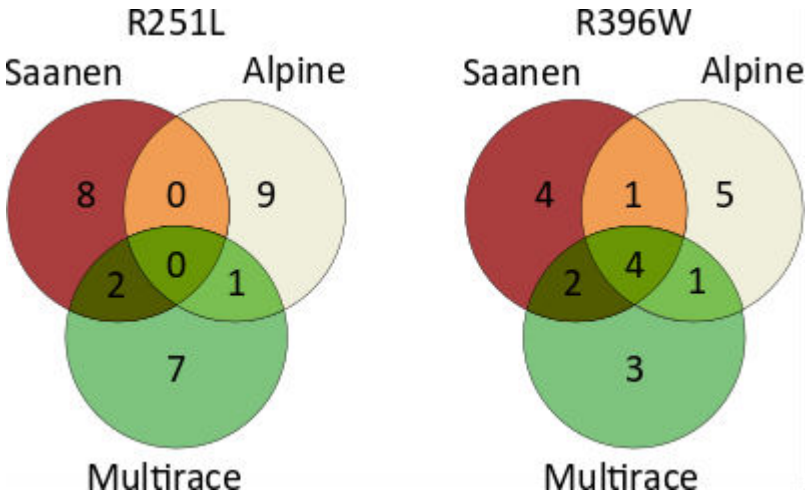


Figure 54. SNPs communs entre les analyses multirace, Alpine et Saanen pour les top 10 des SNPs les plus fortement associés avec un test du chi-deux entre les génotypes DGAT1 et les génotypes de la puce 50 K (chromosome 14) pour les mutations R251L et R396W

1.1.2. Résultats des évaluations génomiques intégrant l'effet du gène DGAT1

Des analyses similaires à celles menées sur le TP et le gène de la *caséine* α_{s1} ont été réalisées pour le TB. Pour ce caractère, des évaluations WssGBLUP en utilisant les génotypages de la puce 50K ont été testés et sont présentés dans le paragraphe 2, chapitre 4 (article 2). En complément des évaluations WssGBLUP, nous avons comparé les précisions génomiques obtenues par 4 méthodes : ssGBLUP (avec uniquement l'information des SNPs 50K), gene content en intégrant les génotypes de la mutation R251L, gene content en intégrant les génotypes de la mutation R396W et WssGBLUP en ajoutant les génotypes des mutations R251L et R396W comme SNPs codés sous la forme : 0, 1, 2 ou 5 (Figure 55). Les précisions génomiques avec le ssGBLUP atteignent 0,70 pour les analyses multirace, 0,66 pour les analyses Alpine et 0,59 pour les analyses Saanen. Le WssGBLUP avec les deux mutations améliore les précisions de 1 point par rapport au ssGBLUP pour les analyses multirace (0,71), de -1 point pour les analyses Alpine (0,65) et de 0 point pour les analyses Saanen (0,59). En incluant une seule des mutations dans l'approche WssGBLUP (résultats non présentés) les précisions restent identiques à celles du WssGBLUP (50K + R251L + R396W). Les précisions avec le gene content incluant la mutation R251L diminuent de 7 points pour les analyses multirace (0,63) et Alpine (0,59) et de 6 points pour les analyses Saanen (0,53) par rapport au ssGBLUP. Cette diminution est moins marquée pour le gene content incluant la mutation R396W : baisse de 5 points, 7 points et 0 point pour les analyses multirace, Alpine et Saanen respectivement par rapport au ssGBLUP.

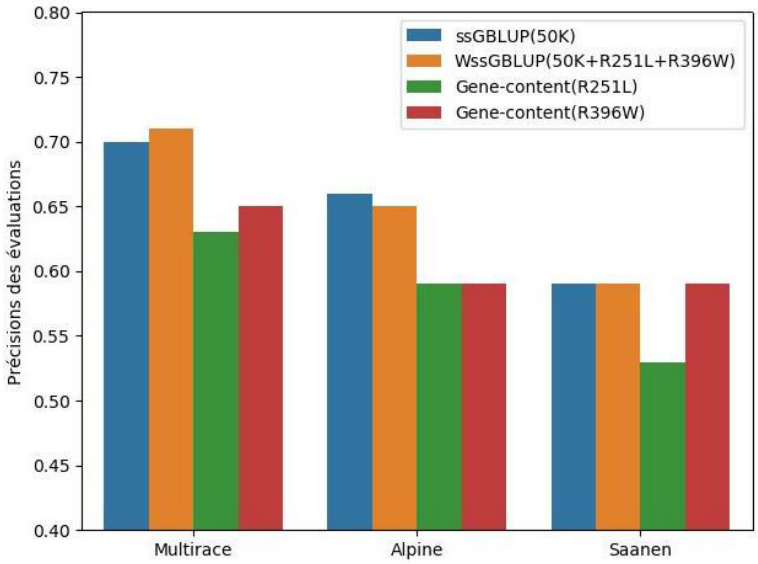


Figure 55. Précisions des évaluations génomiques ssGBLUP, gene content (intégrant les mutations R251L et R396W) et Weighted ssGBLUP (intégrant les génotypes de la puce 50K et les mutations R251L et R396W) pour les analyses multirace, Alpine et Saanen

L'intégration des deux mutations (R251L et R396W) dans les génotypes de la puce 50K a peu d'impact sur l'estimation des poids des SNPs dans les évaluations WssGBLUP (Tableau 24). Les poids du SNPs de la mutation R251L sont faibles et proches de 0 dans les 3 analyses. Ils sont classés à la 46 141ème, 46 849ème et 45 769ème position des poids les plus faibles pour les analyses multirace, Alpine et Saanen respectivement. Les poids sont légèrement plus élevés pour la mutation R396W mais restent inférieurs à 1 (valeur des poids des SNPs par défaut du ssGBLUP). Les poids de ce SNP sont classés à la 36 206ème, 39 546ème et 23 032ème position dans les analyses multirace, Alpine et Saanen respectivement lorsque les SNPs sont triés selon leur poids par ordre décroissant.

Tableau 24. Estimation des poids des SNPs des mutations R251L et R396W pour les évaluations WssGBLUP (50K + R251L + R396W) du TB pour les populations multirace, Alpine et Saanen

	Poids du SNP R251L	**Poids du SNP R296W**
Multirace	$0{,}29 * 10^{-3}$	0,41
Alpine	0,00	0,07
Saanen	$0{,}66 * 10^{-3}$	0,03

1.1.3. Discussion

Les valeurs de – log10 (*p-value)* sont plus faibles pour la mutation R251L que pour la mutation R396W. Cela peut s'expliquer par la fréquence de ces allèles dans les populations génotypées. Que ce soit pour la mutation R251L ou R396W, la fréquence de l'allèle muté (l'allèle T) est rare en Alpine ou en Saanen. Pour la mutation R251L, un seul animal possède le génotype T/T pour la race Alpine et seulement 3 animaux pour la race Saanen (cf paragraphe 3.3.2, Chapitre 1).

Le WssGBLUP incluant les deux mutations causales (R251L et R396W) et les approches Gene content n'ont pas amélioré les précisions des évaluations génomiques par rapport à un

ssGBLUP. Il ne semble donc pas pertinent d'inclure l'information de ces mutations dans les modèles d'évaluation génomique. Des analyses Linkage Analyses (LA) et Linkage Disequilibrium (LD) réalisées par Martin et al., (2017) sur les populations caprines françaises ont identifié un QTL sur le chromosome 14. Leurs travaux ont été utiles pour confirmer l'effet de ces deux mutations (R251L et R396W) sur le TB. Lourenco et al., (2014) ont réalisé des évaluations génomiques avec un GBLUP, BayesCπ, un ssGBLUP et un WssGBLUP sur des caractères de production laitière dans une population bovine Holstein israélienne. Les animaux étaient génotypés avec la puce Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA). Sur le TB, les précisions avec un WssGBLUP étaient améliorées par rapport à un ssGBLUP (0,43 contre 0,40). Elles étaient similaires avec un BayesCπ (0,43) et supérieures à un GBLUP (0,36). Les poids des SNPs intégrés dans les évaluations WssGBLUP ont été réalisés selon l'approche décrite par Wang et al., (2012), cependant d'autres alternatives sont possibles pour modifier la matrice qui contient les poids des SNPs (**D**). En race caprine, il serait possible d'estimer les effets des SNPs à partir d'analyse GWAS. Ces effets de SNPs pourraient être convertis en poids puis intégrer dans les évaluations WssGBLUP. Comme pour le TP, le faible nombre de génotypages disponibles pour les deux mutations *DGAT1* (R251L et R396W) et la présence d'allèle rare pourraient expliquer l'absence de gain en termes de précision génomiques dans les méthodes incluant cette information.

2. Précisions des évaluations génomiques avec le WssGBLUP pour des caractères de productions laitières, de morphologie de la mamelle et pour le comptage de cellules somatiques

Le deuxième article de ma thèse, accepté dans la revue Journal of Dairy Science (décembre 2018), présente les précisions génomiques obtenues sur l'ensemble des caractères évalués en routine (excepté pour le TP) en utilisant la méthode WssGBLUP et ses alternatives ($WssGBLUP_{Max}$, $WssGBLUP_{Sum}$). Ces méthodes utilisent uniquement les marqueurs de la puce 50K et sont comparées au ssGBLUP. Les données utilisées dans cet article sont celles du scénario E (paragraphe 1.1, Chapitre 3). Pour les analyses ssGBLUP et WssGBLUP, nous avons utilisé les logiciels de la famille blupf90 pour estimer les effets des SNPs ainsi que les GEBV des animaux. Pour les alternatives du WssGBLUP (Max et Sum), des fenêtres de 2, 5, 10, 15, 20, 25, 30, 35, 40, 45 et 50 SNPs consécutifs ont été testées pour estimer les poids communs attribués aux SNPs d'une même fenêtre.

Article II : Précisions des évaluations génomiques avec le single-step GBLUP pondéré pour des caractères de productions laitières, de morphologie de la mamelle et pour le comptage de cellules somatiques pour les chèvres laitières françaises.

Teissier, M., H. Larroque, and C. Robert-Granie. 2019. Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats. J. Dairy Sci.. doi:10.3168/jds.2018-15650.

J. Dairy Sci. 102:1–13
https://doi.org/10.3168/jds.2018-15650

Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats

M. Teissier,* H. Larroque, and C. Robert-Granie
GenPhySE, Université de Toulouse, INRA, INPT, ENVT, 31326 Castanet-Tolosan, France

ABSTRACT

Genomic evaluation of French dairy goats is routinely conducted using the single-step genomic BLUP (ssGBLUP) method. This method has the advantage of simultaneously using all phenotypes, pedigrees, and genotypes. However, ssGBLUP assumes that all SNP explain the same amount of genetic variance, which is unlikely in the case of traits whose major genes or QTL are segregating. In this study, we investigated the effect of weighted ssGBLUP and its alternatives, which give more weight to SNP associated with the trait, on the accuracy of genomic evaluation of milk production, udder type traits, and somatic cell scores. The data set included 2,955 genotyped animals and 2,543,680 pedigree animals. The number of phenotypes varied with the trait. The accuracy of genomic evaluation was assessed on 205 genotyped Alpine and 146 genotyped Saanen goats born between 2009 and 2012. For traits with unknown QTL, weighted ssGBLUP was less accurate than, or as accurate as, ssGBLUP. For traits with identified QTL (i.e., QTL only present in the Saanen breed), weighted ssGBLUP outperformed ssGBLUP by between 2 and 14%.
Key words: genomic evaluation, quantitative trait loci, weighted single-step genomic best linear unbiased predictor, French dairy goat

INTRODUCTION

Genomic evaluation is routinely used in an increasing number of species including dairy cattle (Hayes et al., 2009; Boichard et al., 2012), poultry (Wolc et al., 2016), dairy sheep (Duchemin et al., 2012; Baloche et al., 2014), meat sheep (Auvray et al., 2014; Brito et al., 2017), pigs (Christensen et al., 2012), and dairy goats (Carillier et al., 2013, 2014; Mucha et al., 2015). Several genomic methods have been tested and implemented but the most widely used is single-step genomic BLUP (**ssGBLUP**; Legarra et al., 2009). Single-step genomic BLUP has the advantage of simultaneously using the phenotypes of genotyped and nongenotyped animals, pedigrees, and genotypes. The method constructs a relationship matrix based on the numerator relationship matrix (**A**) and the genomic relationship matrix (**G**) called the hybrid relationship matrix (**H**). The use of ssGBLUP increases the accuracy of genomic evaluation in many contexts and species compared with pedigree-based BLUP or genomic BLUP (**GBLUP**; Chen et al., 2011; Carillier et al., 2014; Onogi et al., 2015; Matilainen et al., 2016). However, the expected increase in the accuracy of genomic evaluation depends on several parameters including the size of the reference population (Lourenco et al., 2014; Andonov et al., 2017), the relationship between the training and validation population (Meuwissen et al., 2001), the extent of linkage disequilibrium (**LD**; Zhou et al., 2018), or the genetic architecture of the trait concerned (Goddard, 2009; Carillier-Jacquin et al., 2016; Zhou et al., 2018).

The main French dairy goats breeds are Alpine and Saanen, and their standard evaluation is based on milk production traits, udder type traits, and SCS. Carillier et al. (2013, 2014) investigated the feasibility of genomic evaluation of French dairy goats using the goat SNP50 BeadChip (Illumina Inc., San Diego, CA). These authors showed that LD is less extensive than in dairy cattle (Carillier et al., 2013), and that the reference population is limited in the Alpine and Saanen breeds. In 2016, the reference population consisted of 2,955 genotyped animals (2,050 females and 905 males; Teissier et al., 2018). Carillier et al. (2013, 2014) concluded that ssGBLUP was more accurate than either the pedigree-based BLUP or GBLUP. Using a multibreed approach, the authors reported a −4 to 39% change in accuracy for milk production traits using ssGBLUP compared with a pedigree-based BLUP, a 61 to 96% gain in accuracy for udder type traits, and a 54% gain in accuracy for SCS.

Received September 5, 2018.
Accepted December 5, 2018.
*Corresponding author: marc.teissier@inra.fr

In another study, GWAS analyses were performed in dairy goats (French Alpine and Saanen, or mixed-breed goats) to identify QTL that affect traits under selection. Martin et al. (2018) and Mucha et al. (2018) investigated the genetic architecture of different traits. These authors observed a large QTL associated with milk yield, fat yield, protein yield, udder floor position, rear udder attachment, and SCS on chromosome 19. For the standard traits, they also identified important genomic regions on different chromosomes in all breeds or in only one breed. In both French dairy goat breeds, the α_{S1} casein gene associated with protein content was found to be located on chromosome 6 (Grosclaude et al., 1987) and the *DGAT1* gene associated with fat yield on chromosome 14 (Martin et al., 2017). Using these results, our aim was to investigate whether using information on the location of the detected QTL would improve the accuracy of genomic evaluation in an appropriate ssGBLUP. To this end, we tested the incorporation of previous analyses of the effect of the α_{S1} casein gene in the genomic evaluation method (Carillier-Jacquin et al., 2016; Teissier et al., 2018): gene content (Gengler et al., 2007; Legarra and Vitezica, 2015), weighted ssGBLUP (**WssGBLUP**; Wang et al., 2012), WssGBLUP alternatives (Zhang et al., 2016), and TABLUP (Zhang et al., 2015). Gene content is a multiple-trait ssGBLUP model in which the genotype for a specific causal mutation is considered as a new trait, thus enabling the combination of information from SNP and genotypes for a causal mutation. It can be extended to multi-allelic genes and used when a causal mutation is missing. The WssGBLUP and alternatives are based on the ssGBLUP framework in which weights for SNP variances are used to form the genomic relationship matrix **G**. The WssGBLUP can give more weight to SNP that are in high LD with a causal mutation or associated with QTL with a relatively large effect. The weights were estimated from the variance explained by each SNP as described by Wang et al. (2012). **G** is trait specific and depends on the genetic architecture of the trait (traits with QTL or polygenic traits). With WssGBLUP, one weight is allocated to each SNP, whereas alternative WssGBLUP use the same weight for SNP that are located within a defined window along the genome (Zhang et al., 2016). With alternatives WssGBLUP, the weight in a defined window is calculated as the sum of all SNP weights of the window (**WssGBLUP$_{Sum}$**) or as the maximum of the SNP weights of the window (**WssGBLUP$_{Max}$**). Finally, TABLUP is ssGBLUP with a genomic relationship matrix based on genotypes from a subset of pre-selected SNP. The SNP can be selected after GWAS analysis or based on weights estimated with WssGBLUP. The selected SNP are then given equal weights for the analyses (Zhang et al., 2011). Carillier et al. (2016) and Teissier et al. (2018) showed that only WssGBLUP and their alternatives are able to outperform pedigree-based BLUP and ssGBLUP (Teissier et al., 2018). Compared with ssGBLUP, neither the gene content method nor TABLUP on protein content improved the accuracy of genomic evaluations in either breed. On the other hand, with WssGBLUP and their alternatives (WssGBLUP$_{Sum}$ or WssGBLUP$_{Max}$), improvements were observed, with +6 percentage points of accuracy over ssGBLUP. The advantage of WssGBLUP and alternative WssGBLUP over the gene content method is that only genotypes from SNP50 BeadChip are required.

The aim of this study was to investigate the use of WssGBLUP and WssGBLUP alternatives for other widely selected traits with different genetic architectures and, in some cases, with QTL identified as having a relatively large effect. The accuracy of WssGBLUP methods and ssGBLUP were compared. The weights of SNP and their effect on the genomic relationship matrix were investigated for all the traits. The effect of this weighting on the accuracy of genomic evaluation was also investigated and compared with that obtained with the ssGBLUP method.

MATERIALS AND METHODS

Pedigree, Genotyped, and Phenotyped Animals

The data sets included phenotypes, pedigree, genotypes (Illumina goat SNP50 BeadChip), and environmental effects of the 2 main French dairy goat breeds (Alpine and Saanen) obtained from the French National Milk Recording System (http://fr.france-genetique-elevage.org). Data from the official genetic evaluation in January 2016 were used in this study. All analyses were within breed.

The standard traits selected in French dairy goats include 4 milk production traits, milk yield (**MY** in kg), fat and protein yields (**FY** and **PY**, respectively, in kg), and fat content (**FC** in g/kg), 5 udder type traits, teat angle (**TA**: scored from 1 to 9), udder floor position (**UFP**: scored from 1 to 9), rear udder attachment (**RUA**: scored from 1 to 9), fore udder (**FU**: scored from 1 to 9), and udder shape (**US**: scored from 1 to 9), and SCS (log-transformed SCC) were analyzed. Descriptive statistics on the number of records, and the mean and the heritability of each trait and each breed studied are presented in Table 1. Milk production traits were expressed as 250-d yields. Almost 4 million phe-

Table 1. Number of records, mean, and heritability (h^2) for the traits studied in the Alpine and Saanen breeds

	Alpine			Saanen		
Item	Performances (no.)	Mean	h^2	Performances (no.)	Mean	h^2
Milk yield[1] (kg)	3,844,314	802.12	0.31	2,923,531	823.08	0.26
Fat yield[1] (kg)	3,742,129	28.4	0.28	2,887,051	27.44	0.25
Protein yield[1] (kg)	3,844,071	24.36	0.31	2,923,419	24.32	0.25
Fat content[1] (g/kg)	3,742,129	35.33	0.48	2,887,051	33.39	0.51
Teat angle (score)	150,676	3.63	0.42	102,967	4.05	0.45
Udder floor position (score)	150,676	6.37	0.51	102,967	6.16	0.57
Rear udder attachment (score)	150,676	4.57	0.47	102,967	4.96	0.52
Fore udder (score)	150,676	3.19	0.44	102,967	3.38	0.42
Udder shape (score)	150,676	5.76	0.40	102,967	6.22	0.47
SCS	1,262,187	165.03	0.20	1,031,450	158.99	0.16

[1]Expressed as 250-d yields.

notypes in Alpine and 3 million phenotypes in Saanen were recorded for milk production traits. More than 150,000 phenotypes were available in the Alpine breed and more than 100,000 phenotypes in the Saanen breed for udder type traits, and 1.2 million phenotypes in the Alpine breed and 1 million phenotypes in the Saanen breed for SCS.

The pedigree file contained animals born between 1936 and 2012. For milk production traits, 1,446,296 Alpine animals and 1,097,384 Saanen animals were used. For udder type traits, the pedigree file included 290,656 Alpine animals and 206,154 Saanen animals. For SCS, the pedigree file contained 788,576 Alpine animals and 648,461 Saanen animals. The pedigree file was then completed with unknown parent groups: one group was created for animals born before 1975 and then pooled groups (sires and dams) were defined every 2 yr. Males and females were pooled together in unknown parent groups because few animals had unknown dams.

French dairy goats were genotyped with the Illumina goat SNP50 BeadChip (50K SNP; Tosser-Klopp et al., 2014). Quality control (**QC**) was applied to 2,056 genotyped Alpine and 1,349 genotyped Saanen animals (born between 1983 and 2012) for 53,347 SNP, independently for each breed. During the QC, SNP with a minor allele frequency (i.e., less than 1% and a call rate of less than 95%) were removed. The Hardy-Weinberg equilibrium for each SNP was tested by calculating the associated chi-squared statistic. The SNP with a *P*-value lower than 1.10^{-6} were removed (threshold of 5% corrected for multiple testing). Animals with a SNP call rate of less than 90% were discarded from the analyses. Finally, parent-progeny Mendelian conflicts were checked. After the QC, 1,749 genotyped Alpine (512 males and 1,237 females) and 1,206 genotyped Saanen (393 males and 813 females) for 46,849 SNP remained for analyses. These animals were born between 1993 and 2012.

ssGBLUP

The ssGBLUP is a routinely used method for genomic evaluation of 11 traits selected in the 2 main French dairy goats (Carillier et al., 2014; Venot et al., 2017). It can simultaneously combine information on female phenotypes, pedigrees, and genotypes. Each trait was analyzed with a single trait model. For milk production traits (MY, FY, PY, FC) and SCS, the same model as in the routine genetic evaluation was used (Clément et al., 2002):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{p} + \mathbf{e}, \qquad \text{[model 1]}$$

where $\mathbf{y}$ is a vector of phenotypes, $\boldsymbol{\beta}$ is a vector of fixed effects including 4 combined effects: herd, age and month at kidding, and length of the dry period. The herd effect was estimated within year (32 yr from 1980 to 2012) and parity (1, 2, and ≥3); age and month were within year and region estimations (4 regions in France depending on goat breeding management). The length of the dry period was an estimation within a year and region. $\mathbf{u}$ is a vector of genomic breeding values (**GEBV**) assumed to be normally distributed $N\left(\mathbf{0}, \mathbf{H}\sigma_u^2\right)$, where $\mathbf{H}$ represents the relationship matrix and σ_u is the variance of the random additive genetic effect, $\mathbf{p}$ is a vector of random permanent environmental effects assumed to be normally distributed $N\left(\mathbf{0}, \mathbf{I}\sigma_p^2\right)$, with σ_p the variance of the permanent environmental effect, and $\mathbf{e}$ is a vector of random residual normally distributed $N\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$. with σ_e the variance of residuals. $\mathbf{X}$ is the incidence matrix relating phenotypes to fixed effects ($\boldsymbol{\beta}$); $\mathbf{Z}$ is the design matrix which allocates phenotypes to genomic breeding values ($\mathbf{u}$) and $\mathbf{W}$ is the incidence matrix that links phenotypes to permanent environmental effects ($\mathbf{p}$). Solutions of $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{p}$ were obtained by solving the following system:

$$\begin{Bmatrix} \mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'W} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\frac{\sigma^2_e}{\sigma^2_u}\mathbf{H}^{-1} & \mathbf{Z'W} \\ \mathbf{W'X} & \mathbf{W'Z} & \mathbf{Z'W}+\frac{\sigma^2_e}{\sigma^2_p}\mathbf{I} \end{Bmatrix} \begin{Bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \\ \mathbf{p} \end{Bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \\ \mathbf{Z'y} \end{bmatrix}.$$

A different model was used to analyze udder type traits. The only difference is that no permanent environmental effect was estimated because the animals were scored only once in their life (during their first parity). The model was the following:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \qquad \text{[model 2]}$$

where $\mathbf{y}$, $\mathbf{u}$, and $\mathbf{e}$ are the same vectors previously described in model 1 and $\boldsymbol{\beta}$ is the vector of 3 combined fixed effects: herd, age at scoring, and stage at scoring. Herd effect and parity, age at scoring, and stage at scoring were within year estimations. The matrix $\mathbf{H}^{-1}$ is expressed as

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

which depend on the inverse of the relationship matrix estimated from the pedigree $\mathbf{A}$ (subscript 22 refers to genotyped animals) and on the inverse of the genomic relationship matrix $\mathbf{G}$. The $\mathbf{G}$ matrix was estimated using genotypes as in Legarra et al. (2009) or Misztal et al. (2013):

$$\mathbf{G} = 0.95\frac{\mathbf{M'M}}{2\sum_{i=1}^{m} p_i\left(1-p_i\right)} + 0.05\mathbf{A}_{22},$$

where m is the number of SNP, p_i is the estimated allele frequency at the locus i, and $\mathbf{M}$ is a centered matrix of SNP genotypes.

Variance components were estimated using the REML method in the remlf90 software and ssGBLUP analyses were performed with the blup90iod2 software (Misztal et al., 2002).

Weighted ssGBLUP

The construction of the $\mathbf{G}$ matrix presented above assumes that each SNP explains the same amount of genetic variance. Consequently, this assumption is not valid for traits with a major gene or QTL. Wang et al. (2012) proposed another genomic approach called WssGBLUP based on a model similar to ssGBLUP, to include major genes or QTL with a relatively large effect using a weighted $\mathbf{G}$ ($\mathbf{G}^*$). This genomic relationship matrix $\mathbf{G}^*$ is constructed as follows:

$$\mathbf{G}^* = 0.95\frac{\mathbf{M'DM}}{2\sum_{i=1}^{m} p_i\left(1-p_i\right)} + 0.05\mathbf{A}_{22},$$

where $\mathbf{A}_{22}$, $\mathbf{M}$, p_i, and m are the same as in $\mathbf{G}$ and $\mathbf{D}$ is a diagonal matrix of size $m \times m$, where each element of the diagonal corresponds to SNP weights.

The WssGBLUP approach is based on an iterative algorithm with different steps: (1) run ssGBLUP with the $\mathbf{G}^*$ matrix (at iteration 1, SNP weights in the $\mathbf{D}$ matrix are equal to 1 and are equivalent to a ssGBLUP), (2) estimate SNP effects from solutions of genomic breeding values in the previous step, (3) estimate variances of the effect of each SNP, (4) normalize the vector of variances of SNP effects to get the SNP weights (this normalization process ensures that the sum of the variances remain constant and equal to the number of SNP), (5) use SNP weights to construct the $\mathbf{D}$ matrix, and (6) loop to step (1).

The WssGBLUP was applied to each trait studied with model 1 and 2, respectively, using blupf90 family software (blup90iod2, Misztal et al., 2002). The SNP effects and SNP weights were estimated using postGSf90 software. In this study, 3 WssGBLUP approaches were investigated, each one using a specific $\mathbf{G}^*$: WssGBLUP, $\text{WssGBLUP}_{\text{Sum}}$, and $\text{WssGBLUP}_{\text{Max}}$. The WssGBLUP is the method presented by Wang et al. (2012), which consists in attributing one weight to each SNP. With $\text{WssGBLUP}_{\text{Sum}}$ and $\text{WssGBLUP}_{\text{Max}}$ (Zhang et al., 2016), SNP on the whole genome are split into different-sized nonoverlapping windows, and the same weight is given to each SNP of the window. Windows of 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 consecutive SNP were tested. To compute these weights, the sum of all SNP weights present in the window was given to those SNP ($\text{WssGBLUP}_{\text{Sum}}$), or the SNP with the highest weight in the window was given to all SNP in the same window ($\text{WssGBLUP}_{\text{Max}}$). The final step consists in normalizing the new vector of SNP weights to ensure that the sum of the variances remains constant and equal to the number of SNP. A previous study has shown that the 2nd iteration of the WssGBLUP with 40 SNP was the most accurate for $\text{WssGBLUP}_{\text{Max}}$ and $\text{WssGBLUP}_{\text{Sum}}$ (Teissier et al., 2018), and results were presented for this scenario.

Accuracy of the Genomic Evaluation

The ssGBLUP used female phenotypes, pedigrees, and genotypes. In the French Alpine and Saanen dairy goat breeding scheme, genetic selection is performed on progeny-tested bucks and all these bucks born after 1993 were genotyped. The reference population used to assess the accuracy of genomic evaluation comprised only genotyped males even if genotypes of females were also used in ssGBLUP and WssGBLUP evaluations. This reference population was split into 2 subsets: a training set and a validation set. The training population included 307 Alpine bucks and 247 Saanen bucks born between 1993 and 2007, all the information on these animals (genotype, the pedigree of their ancestry and their progeny, and the phenotypes of their progeny) was kept in the data sets to estimate GEBV. The validation set included 205 Alpine bucks and 146 Saanen bucks born between 2008 and 2012. For these animals, the phenotypes of their progeny were removed from the analysis, and only the genotypes and pedigree of their ancestry were retained. The accuracy of genomic evaluation was calculated as the Pearson correlation between GEBV estimated using the validation set and daughter yield deviations (**DYD**) calculated using the official genetic evaluation of January 2016. The number of daughters used to calculate these DYD was between 46 and 2,509 (with a median of 177 daughters), indicating that the DYD were relatively accurate. Accuracies of genomic evaluations were compared between ssGBLUP and WssGBLUP and its alternatives with the Hotelling-Williams test (Van Sickle, 2003).

Relationship Coefficients Estimated Using Pedigree and Genomic Information

Elements of the off-diagonal of the numerator relationship matrix for genotyped animals ($\mathbf{A}_{22}$) and the weighted genomic relationship matrix ($\mathbf{G}^*$) were compared. To this end, the Pearson correlation between the 2 vectors was calculated.

RESULTS AND DISCUSSION

Accuracy with WssGBLUP Over Iterations

WssGBLUP is based on an iterative process. The first iteration corresponds to ssGBLUP; SNP weights are all equal to 1. The accuracy of the genomic evaluations from iteration 2 to iteration 4 for each breed and each trait are presented in Figure 1. The average accuracies for the 10 traits in the Alpine and Saanen breeds were, respectively, 0.42 and 0.51 at iteration 2, 0.39 and 0.49 at iteration 3, and 0.35 and 0.44 at iteration 4. For all traits in the Alpine breed, accuracy at iteration 2 was higher than the accuracy at iteration 3, which in turn was higher than accuracy at iteration 4. In the Saanen breed, MY accuracy increased at iteration 3 (0.58) compared with iteration 2 (0.56) and then decreased at iteration 4 (0.52). For (FY), accuracy at iteration 2 and 3 was 0.47, then decreased to 0.43 at iteration 4. For all the other traits (i.e., PY, FC, TA, UFP, RUA, FU, US, and SCS), accuracy decreased over the 3 iterations. In both breeds, the decrease in accuracy between iterations 3 and 4 was bigger than the decrease in accuracy between iterations 2 and 3.

In a previous study, we investigated a similar approach (WssGBLUP) to the analysis of protein content in the same 2 French dairy goat populations (Teissier et al., 2018). We concluded that WssGBLUP at iteration 2 provided the most accurate genomic evaluation. In the present study, we obtained the same results for all the standard traits selected in French national genomic evaluations. Our results are also in agreement with those of Wang et al. (2012), who reported WssGBLUP produced the most accurate genomic evaluation at iteration 2. However, after iteration 2, loss of accuracy in our study was much greater than that observed by Wang et al. (2012). These differences could be due to the fact that Wang et al. (2012) used simulation in their study to mimic a trait with a phenotypic mean of 5, variance of 1, and heritability of 0.5. They simulated 2 chromosomes each with 15 QTL sampled from a gamma distribution with a shape factor of 0.4 and a scale factor of 1. They repeated the simulation 10 times. Overall, the average effect of the QTL was 0.16 (0.04). In our study, the situation was probably more complex because we analyzed real data concerning traits with different genetic architectures. According to Wang et al. (2012) and to our previous study (Teissier et al., 2018), the decrease in accuracy could be due to over/underweighting of some SNP across iterations. In the present study, we consequently investigated this point to check if our SNP weights across the WssGBLUP iterations were inflated.

Effects of Iterations on the WssGBLUP Method

Figure 2 presents the SNP weights for RUA in the Saanen breed between iteration 2 and iteration 4 using WssGBLUP. At iteration 2, the highest SNP weights located on chromosome 19 reached 46. On the whole genome, 95% of SNP had weights under 4. At the 4th iteration with WssGBLUP, SNP weights reached 8,621 on chromosome 19 and 95% of the SNP had weights under 0.14. The results at iteration 3 are not shown but were intermediate between those at iteration 2 and it-

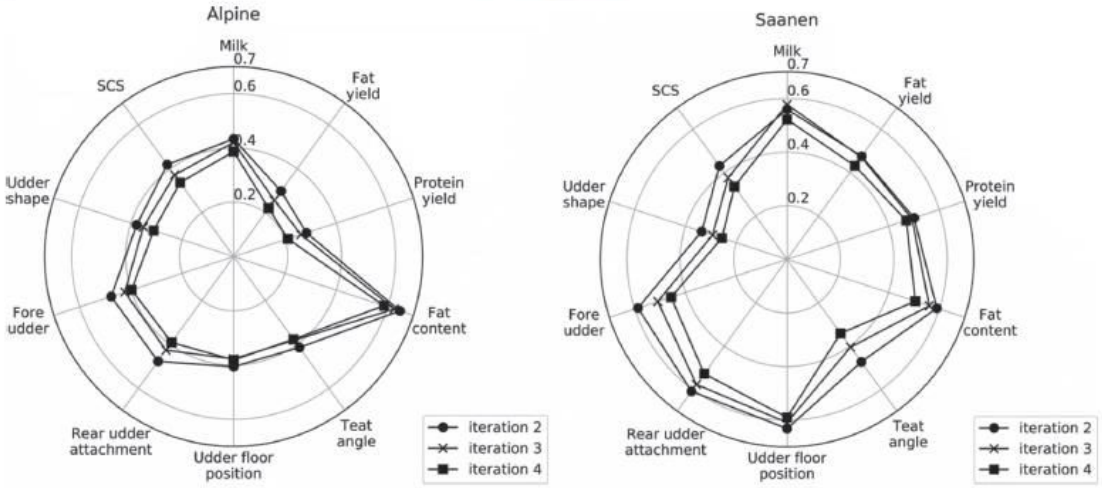


Figure 1. Validation correlations for 205 and 146 validation Alpine and Saanen males, respectively, for 4 milk production traits, 5 udder type traits, and SCS using the weighted single-step genomic BLUP (WssGBLUP) approach at iterations 2, 3, and 4.

eration 4 with a maximum of 1,409 for SNP weights on chromosome 19 with 95% of SNP weights under 2.93. We also observed this huge inflation of SNP weights in the Saanen breed on chromosome 19 for MY, PY, UFP, and SCS. For the other traits (FY, FC, TA, FU, and US), high SNP weights were attributed to some SNP at iteration 4 in both breeds. However, these high SNP weights were not located on a specific chromosome and their maxima were much lower than those observed for MY, PY, UFP, and SCS on chromosome 19 in the Saanen breed. For instance, in the Alpine breed, the highest SNP weights for RUA reached 2,000 at iteration 4 with WssGBLUP whereas they reached 8,621 in the Saanen breed.

The SNP weights were very highly inflated between 2 iterations; at iteration 4 for RUA in the Saanen breed, 18% of SNP weights were allocated to only one SNP on chromosome 19. For all traits and the 2 breeds, we observed that some SNP strongly associated with the traits considered had very high weights and that SNP

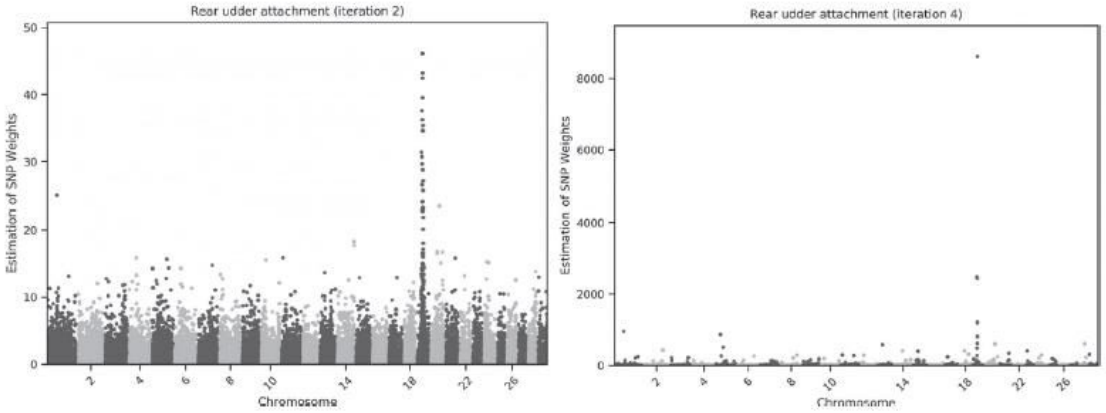


Figure 2. Estimated weights of SNP for rear udder attachment in the Saanen breed at iterations 2 and 4 with the weighted single-step genomic BLUP (WssGBLUP) approach.

weights increased markedly from one iteration to the next, whereas the other SNP weights decreased toward zero. The SNP weights estimated with WssGBLUP were used as weights in matrix $\mathbf{D}$ to construct the weighted matrix $\mathbf{G}^*$. This matrix, which is included in the $\mathbf{H}$ matrix, could affect the structure of the relationship matrix. We compared off-diagonal elements between $\mathbf{G}^*$ and $\mathbf{A}_{22}$ to observe how much effect SNP weights have on elements of the genomic relationship matrix. Figure 3 shows the correlation between the off-diagonal elements of the $\mathbf{G}^*$ matrix and those of the $\mathbf{A}_{22}$ matrix for each breed and trait. In the Alpine breed, the average correlation for the 10 traits was 0.87 at iteration 2. Few variations were observed among the traits with correlations ranging from 0.84 to 0.91. The best correlations were obtained for milk production traits (MY, FY, PY, and FC). At iteration 3, the average correlation was lower (0.76), range: 0.72 to 0.80. Finally, at iteration 4, we observed a low average correlation (0.47), range: 0.41 to 0.51. In the Saanen breed, similar conclusions were drawn. The average correlation for the 10 traits was 0.80, 0.45, and 0.28 at iterations 2, 3, and 4, respectively, with values ranging from 0.70 to 0.90 at iteration 2, from 0.17 to 0.74 at iteration 3, and from 0.12 to 0.48 at iteration 4.

The profiles of the average correlations in the Alpine and Saanen breeds were similar at iteration 2 but differed markedly at iterations 3 and 4. The loss of correlation between iteration 2 and 3 in the Alpine breed was roughly equal to 10 percentage points and 30 percentage points between iteration 3 and 4 for any trait included in this study. In the Saanen breed, the loss of correlation was trait dependent and reached 62 percentage points for MY and 7.5 percentage points for US between iteration 2 and 3.

From iteration 3, the off-diagonal elements of $\mathbf{G}^*$ and $\mathbf{A}_{22}$ differ significantly in the Saanen breed. Martin et al. (2018) found QTL for MY, FY, PY, UFP, RUA, and SCS in the French Saanen breed. These traits are those for which we observed the biggest decrease in the correlation between elements of $\mathbf{A}_{22}$ and $\mathbf{G}^*$. In the Alpine breed, no QTL were detected, suggesting that these traits have polygenic architecture (Martin et al., 2018). We conclude that for those traits for which QTL have been detected, the weights assigned to the SNP most strongly associated with the trait are exacerbated from one iteration to another in the iterative process of WssGBLUP. These results suggest the iterative process of WssGBLUP should be stopped at iteration 2.

Estimation of Weights with WssGBLUP

We analyzed the estimation of SNP weights with WssGBLUP for the 10 traits in both the Alpine and Saanen breeds. These analyses were performed to highlight important chromosomal regions associated with

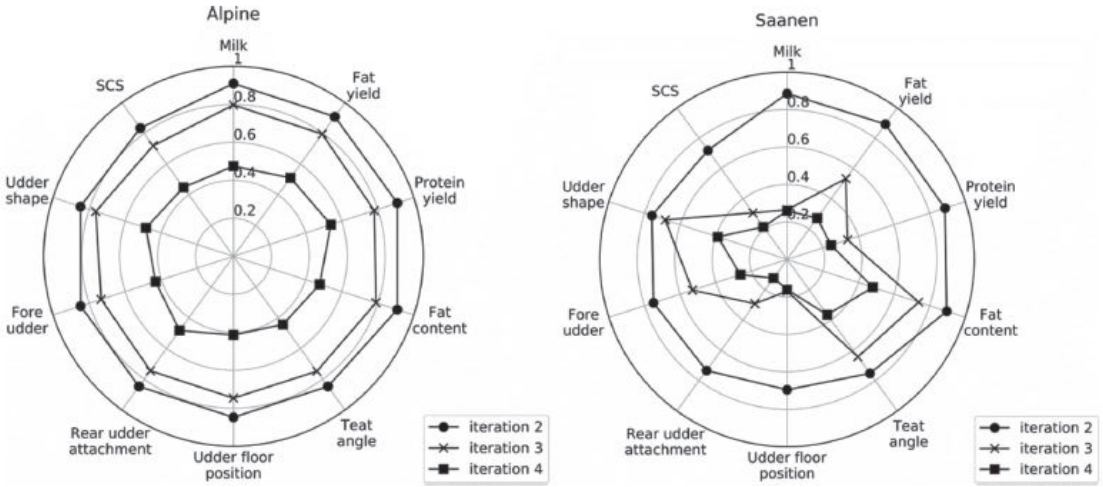


Figure 3. Correlation between off-diagonal elements of the genomic relationship matrix ($\mathbf{G}^*$) and off-diagonal elements of $\mathbf{A}_{22}$ (pedigree $\mathbf{A}$ with subscript 22 referring to genotyped animals) on 205 and 146 validation Alpine and Saanen males, respectively, with the different iterations of the weighted single-step genomic BLUP (WssGBLUP) for 4 milk production traits, 5 udder type traits, and SCS in Alpine and Saanen populations.

selected traits in French dairy goats. We identified 2 different groups: (1) traits with high SNP weights detected in the Saanen breed on one chromosome but not in the Alpine breed, and (2) traits with SNP weights homogeneously distributed along the chromosomes in the 2 breeds. Figure 4 illustrates these 2 different groups of SNP weights for UFP and US. For UFP (included in the first group), SNP weights were below 30 in the Alpine breed, whereas in the Saanen breed, SNP weights reached 68 for some SNP on chromosome 19. The SNP on other chromosomes had SNP weights below 30. The top 10 SNP with the highest SNP weights on chromosome 19 were located between 26 and 28 Mb. The MY, PY, RUA, and SCS were included in the first group. Except for chromosome 19, the SNP weights for all chromosomes were below 30. On chromosome 19, the maximum weights observed were 48 for MY (top 10 SNP were located between 26 and 29 Mb), 42 for PY (top 10 SNP between 26 and 29 Mb), 46 for RUA (top 10 SNP between 20 and 28 Mb), and 37 for SCS (top 10 SNP between 23 and 28 Mb). The top 10 SNP covered a chromosomal region between 20 and 29 Mb for all these traits. For US (included in group 2), we observed SNP weights below 30 for all SNP in both Alpine and Saanen breeds. The same profile was observed for FY, FC, TA, FU, and US. Martin et al. (2018) performed LD and linkage analysis in French dairy goats. They showed that chromosome 19 underlies a pleiotropic QTL located between 24.5 and 26.9 Mb (5% CI) affecting MY, FY, PY, UFP, and RUA. In our study with WssGBLUP, the highest SNP weights were located on the same chromosome 19 and in the same region but with a slightly larger interval for MY, PY, UFP, and RUA. For FY on chromosome 19, we did not find any SNP with significant weights like those found by Martin et al. (2017). It is possible that our training population was too small, and that with more genotyped animals, we would have identified SNP with higher weights on chromosome 19. Surprisingly, the chromosomal region of DGAT1 on chromosome 14, which is known to be associated with FC, was not identified with WssGBLUP, whereas it was detected by Martin et al. (2017) with LD and linkage analysis. This result shows that SNP effects estimation with WssGBLUP had some limitations and could be improved in the future. This limitation may be due to the whole-genome regression performed to estimate SNP effects, resulting in unstable prediction of SNP effects because of LD between SNP. Martin et al. (2018) found a QTL for SCS in the Saanen breed located between 33 and 42 Mb on chromosome 19. In our study, the highest SNP weights were located on the same chromosome 19 but in a neighboring chromosomal region (between 23 and 28 Mb). In a previous study Teissier et al. (2018), we conducted WssGBLUP analysis for protein content in which a major gene (α_{S1} casein gene) was identified, but no SNP in the α_{S1} casein gene was on the 50K SNP after QC. We identified some SNP with high weights (between 90 to 101) in the α_{S1} casein gene region on chromosome 6, and this method provides a more accurate genomic evaluation than ssGBLUP in the 2 breeds. This shows that WssGBLUP is able to capture the complexity of this gene.

Accuracy of Genomic Evaluation Using WssGBLUP

Figure 5 presents the accuracy of genomic evaluation using the validation set for the 10 traits. We compared the ssGBLUP method used as a reference method and the WssGBLUP at iteration 2. In the Alpine breed, accuracy was on average slightly lower with WssGBLUP (0.42) than with ssGBLUP (0.44). The loss of accuracy ranged between +0 percentage points (TA or US) to −3 percentage points (SCS); however, these differences were not significant. For this breed, no QTL was identified for these traits (Martin et al., 2018) and no large SNP weight was identified with WssGBLUP. In the Saanen breed, the accuracy of the genomic evaluation was on average slightly higher with WssGBLUP (0.52) than with ssGBLUP (0.51). However, we observed an increase or a decrease in accuracy depending on the trait. With WssGBLUP, accuracy was the same or lower than with ssGBLUP for FC (+0 percentage points), TA (−1 percentage point), US (−2 percentage points); these differences were not significant. However, for FU (−3 percentage points) and SCS (−3 percentage points) significant decrease of accuracies at 0.05 threshold were observed. Among these traits, all SNP weights were low, below 30, except for SCS where a QTL was identified on chromosome 19 (Martin et al., 2018) and SNP weights were higher on a chromosomal region of chromosome 19. Increased accuracy was obtained for the remaining traits with WssGBLUP: +1 percentage point for RUA (not significant at the 0.05 threshold), +4 percentage points for FY and UFP ($P < 0.05$), +5 percentage points for PY ($P < 0.05$) and +7 percentage points for MY ($P < 0.001$). For these traits, QTL were identified on chromosome 19 (Martin et al., 2017, 2018), and except for FY, high SNP weights were also identified in the same chromosomal region with WssGBLUP.

With WssGBLUP, the accuracy of genomic evaluation was improved for traits with segregation of QTL in French dairy goats. In our study, the Saanen breed was mostly concerned with the large QTL on chromosome 19 for MY, PY, UFP, and RUA. These results are consistent with those in our previous study (Teissier et al.,

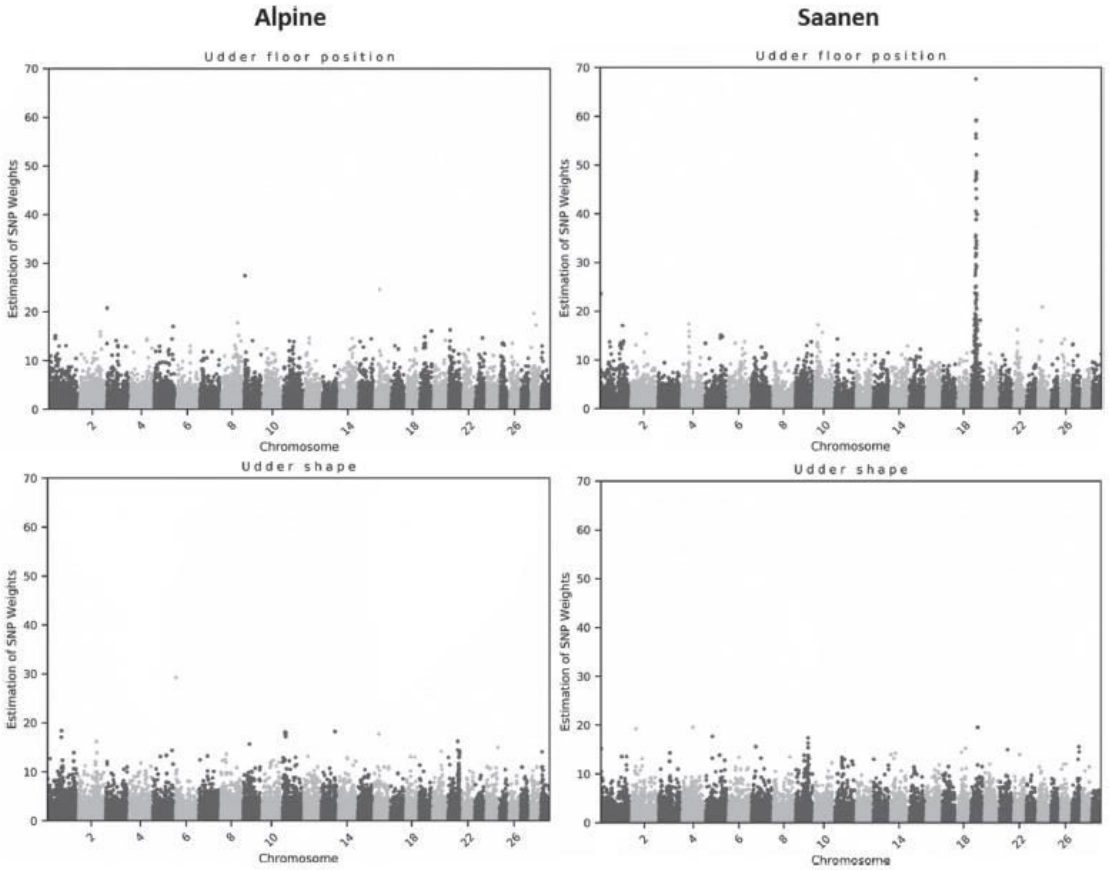


Figure 4. Estimated weights of SNP for udder floor position and udder shape at iteration 2 with weighted single-step genomic BLUP (WssGBLUP) for 4 milk production traits, 5 udder type traits, and SCS in Alpine and Saanen populations.

2018) on protein content, as the α_{S1} casein gene is well known to segregate in these populations. The results of that study showed that with WssGBLUP, accuracy was better than with ssGBLUP in both breeds (+2 and +4 percentage points in the Alpine and Saanen breeds, respectively). Only in Saanen was the WssGBLUP significantly more accurate than ssGBLUP. In another study on dairy cattle using WssGBLUP with a relatively small reference population (1,500 genotyped animals), Lourenco et al. (2014) showed that WssGBLUP could outperform GBLUP or BayesC for traits with QTL with large or moderate effects. On the other side, traits with a polygenic determinism did not benefit from the use of the WssGBLUP method, as a slightly decrease in accuracy was observed compared with ssGBLUP.

The average higher accuracy in Saanen than Alpine breed may be explained by the structure of the population. The level of inbreeding in Saanen (2.3%) is higher than in Alpine (1.8%). There is also a higher kinship coefficient between the training and validation population in the Saanen breed (2.4%) than in Alpine breed (1.1%; Carillier et al., 2013).

Fine Tuning of Weights in the WssGBLUP Method

We observed that with WssGBLUP, SNP weights increased considerably from iteration 2 to iteration 4. Zhang et al. (2016) reported that WssGBLUP alternatives ($\text{WssGBLUP}_{\text{Sum}}$ and $\text{WssGBLUP}_{\text{Max}}$) increased the accuracy of genomic evaluation more efficiently

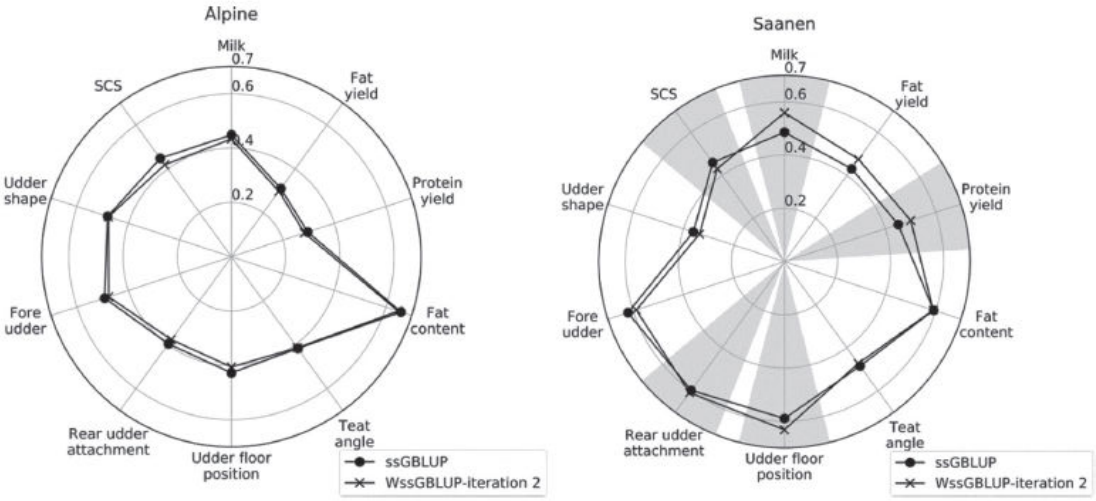


Figure 5. Validation correlations for, respectively, 205 and 146 validation Alpine and Saanen males for 4 milk production traits, 5 udder type traits, and SCS using the single-step genomic BLUP (ssGBLUP) and weighted single-step genomic BLUP (WssGBLUP) at iteration 2. Gray area: traits with high SNP weights detected using WssGBLUP.

than WssGBLUP and limited the increase in SNP weights from one iteration to another. In our previous study (Teissier et al., 2018), we applied $WssGBLUP_{Sum}$ and $WssGBLUP_{Max}$ to protein content and showed that the optimal length of the window was 40 SNP. Even though $WssGBLUP_{Sum}$ and $WssGBLUP_{Max}$ limited the increase in SNP weights over iterations, the best accuracies were obtained at iteration 2. In the present study, we obtained the same results. Table 2 compares the results obtained with ssGBLUP, WssGBLUP, $WssGBLUP_{Sum}$, and $WssGBLUP_{Max}$ with a window size of 40 SNP for the 10 traits at iteration 2 for the Alpine breed and Table 3 compares the same results for the Saanen breed.

We first compared $WssGBLUP_{Sum}$ with $WssGBLUP_{Max}$. In the Alpine breed, for MY, FY, FC, TA, and UFP, $WssGBLUP_{Sum}$ was as accurate (no significant difference at 0.05 threshold was observed) as $WssGBLUP_{Max}$. For FU and RUA, $WssGBLUP_{Sum}$ was slightly less accurate (−1 percentage point) than $WssGBLUP_{Max}$. For PY, US, and SCS, $WssGBLUP_{Sum}$ was slightly more accurate (+1 percentage point) than $WssGBLUP_{Max}$. In the Saanen breed, for MY, FY, PY, TA, UFP, and SCS, $WssGBLUP_{Sum}$ was as accurate

Table 2. Pearson correlation between genomic breeding values and daughter yield deviations for the traits studied in the Alpine breed[1]

Trait	ssGBLUP	WssGBLUP	$WssGBLUP_{Max}$	$WssGBLUP_{Sum}$
Milk yield (kg)	0.45	0.43	0.44	0.44
Fat yield (kg)	0.31	0.30	0.30	0.30
Protein yield (kg)	0.30	0.28	0.28	0.29
Fat content (g/kg)	0.66	0.65	0.66	0.66
Teat angle (score)	0.42	0.41	0.41	0.41
Udder floor position (score)	0.43	0.41	0.44	0.44
Rear udder attachment (score)	0.40	0.38	0.42	0.41
Fore udder (score)	0.49	0.48	0.50	0.49
Udder shape (score)	0.48	0.48	0.48	0.49
SCS	0.45	0.42	0.44	0.45
Mean	0.44	0.42	0.44	0.44

[1]ssGBLUP = single-step genomic BLUP; WssGBLUP = weighted single-step genomic BLUP. Accuracies for the maximum of the SNP weights of the window ($WssGBLUP_{Max}$) and the sum of all SNP weights of the window ($WssGBLUP_{Sum}$) are presented for a window size of 40 consecutive SNP.

Table 3. Pearson correlation between genomic breeding values and daughter yield deviations for the traits studied in the Saanen breed[1]

Trait	ssGBLUP	WssGBLUP	WssGBLUP$_{Max}$	WssGBLUP$_{Sum}$
Milk yield (kg)	0.49	0.56	0.56	0.56
Fat yield (kg)	0.43	0.48	0.49	0.49
Protein yield (kg)	0.45	0.50	0.51	0.51
Fat content (g/kg)	0.59	0.59	0.60	0.61
Teat angle (score)	0.49	0.47	0.48	0.48
Udder floor position (score)	0.59	0.63	0.65	0.65
Rear udder attachment (score)	0.60	0.61	0.62	0.63
Fore udder (score)	0.62	0.59	0.62	0.63
Udder shape (score)	0.36	0.34	0.36	0.37
SCS	0.46	0.43	0.46	0.46
Mean	0.51	0.52	0.54	0.54

[1]ssGBLUP = single-step genomic BLUP; WssGBLUP = weighted single-step genomic BLUP. Accuracies for the maximum of the SNP weights of the window (WssGBLUP$_{Max}$) and the sum of all SNP weights of the window (WssGBLUP$_{Sum}$) are presented for a window size of 40 consecutive SNP.

as WssGBLUP$_{Max}$. For FC, FU, RUA, and US, WssGBLUP$_{Sum}$ was slightly more accurate (+1 percentage point) than WssGBLUP$_{Max}$. With WssGBLUP$_{Sum}$, accuracy was very similar to that obtained with WssGBLUP$_{Max}$ whatever the breed and the trait.

In both the Alpine and Saanen breeds, WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$ were at least as accurate as, or more accurate than WssGBLUP, with differences ranging from +0 to +4 percentage points for RUA [WssGBLUP$_{Max}$ compared with WssGBLUP in the Alpine breed ($P < 0.05$)] or FU [WssGBLUP$_{Sum}$ compared with WssGBLUP in the Saanen breed ($P < 0.01$)].

Finally, on average for all traits, WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$ were significantly more accurate (+3 percentage points) than ssGBLUP in the Saanen breed, whereas WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$ were as accurate as ssGBLUP in the Alpine breed. In the Saanen breed, for traits with a QTL on chromosome 19, WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$ were significantly more accurate than ssGBLUP, from +3 percentage points to +7 percentage points for MY, PY, and UFP ($P < 0.01$). For RUA, improvement of accuracy with WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$ (+3 percentage points) was not significant. For the other traits, WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$ did not significantly outperform ssGBLUP (+0 to +2 percentage points). In the Alpine breed, no QTL was detected and the accuracy of all the methods was similar.

Our results are consistent with those reported by Zhang et al. (2016). We conclude that for polygenic traits, the same accuracy can be obtained with ssGBLUP, WssGBLUP$_{Sum}$, and WssGBLUP$_{Max}$, but genomic evaluations made with WssGBLUP are less accurate than with ssGBLUP. However, when QTL are detected for a trait, a slight gain in accuracy can be obtained with WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$ in addition to that obtained with WssGBLUP compared with ssGBLUP.

CONCLUSIONS

Our aim was to investigate different genomic evaluation methods that allow information on the genetic architecture of traits (with or without QTL identified) to be incorporated. We compared ssGBLUP with weighted ssGBLUP and its alternatives (WssBLUP$_{Sum}$ and WssGBLUP$_{Max}$) for the standard traits selected in the 2 main French dairy goat breeds. Weighted ssGBLUP is an iterative algorithm, and we confirmed that the highest accuracies were obtained at the second iteration. The weighted ssGBLUP and its alternatives were able to improve accuracy of genomic evaluations compared with ssGBLUP for traits with a QTL previously identified in a GWAS (MY, FY, PY, UFP, and RUA in the Saanen breed). Compared with ssGBLUP, the gain in accuracy was between 2 and 14% for weighted ssGBLUP and between 3 and 14% for WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$ in the Saanen breed. For traits with no identified QTL (FC, TA, FU, and US in the Saanen breed and all traits studied in the Alpine breed), Weighted ssGBLUP was less accurate than ssGBLUP (between −5 and 0%). With WssGBLUP$_{Sum}$ and WssGBLUP$_{Max}$, the accuracy of the genomic evaluation was close to the accuracy achieved with ssGBLUP (between −2 and 4%). We will recommend the use of the WssGBLUP at the second iteration to predict GEBV of animals in French dairy goat breeding program.

ACKNOWLEDGMENTS

This study would not have been possible without the goat SNP50 BeadChip developed by the International

Goat Genome Consortium (IGGC; www.goatgenome.org). The authors thank Ignacy Misztal (University of Georgia, Athens) for the blup90iod2 program. The authors thank the French Genovicap and Phenofinlait programs [The French National Research Agency (ANR), Apis-Gene, specialized fund for agricultural and rural development (CASDAR), FranceAgriMer, France Genetique Elevage, and the French Ministry of Agriculture Agrifood, and Forestry] and the European 3SR project, which funded part of this work. The first author also received financial support from the Occitanie region and the French National Institute for Agricultural Research (INRA, France) Sélection Génomique (SELGEN) program (INCoMINGS).

REFERENCES

Andonov, S., D. A. L. Lourenco, B. O. Fragomeni, Y. Masuda, I. Pocrnic, S. Tsuruta, and I. Misztal. 2017. Accuracy of breeding values in small genotyped populations using different sources of external information—A simulation study. J. Dairy Sci. 100:395–401. https://doi.org/10.3168/jds.2016-11335.

Auvray, B., J. C. McEwan, S.-N. Newman, M. Lee, and K. G. Dodds. 2014. Genomic prediction of breeding values in the New Zealand sheep industry using a 50K SNP chip. J. Anim. Sci. 92:4375–4389. https://doi.org/10.2527/jas.2014-7801.

Baloche, G., A. Legarra, G. Sallé, H. Larroque, J.-M. Astruc, C. Robert-Granié, and F. Barillet. 2014. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. J. Dairy Sci. 97:1107–1116. https://doi.org/10.3168/jds.2013-7135.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, J. J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. Anim. Prod. Sci. 52:115–120. https://doi.org/10.1071/AN11119.

Brito, L. F., S. M. Clarke, J. C. McEwan, S. P. Miller, N. K. Pickering, W. E. Bain, K. G. Dodds, M. Sargolzaei, and F. S. Schenkel. 2017. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. BMC Genet. 18:7. https://doi.org/10.1186/s12863-017-0476-8.

Carillier, C., H. Larroque, I. Palhière, V. Clément, R. Rupp, and C. Robert-Granié. 2013. A first step toward genomic selection in the multi-breed French dairy goat population. J. Dairy Sci. 96:7294–7305. https://doi.org/10.3168/jds.2013-6789.

Carillier, C., H. Larroque, and C. Robert-Granié. 2014. Comparison of joint versus purebred genomic evaluation in the French multibreed dairy goat population. Genet. Sel. Evol. 46:67. https://doi.org/10.1186/s12711-014-0067-3.

Carillier-Jacquin, C., H. Larroque, and C. Robert-Granié. 2016. Including α s1 casein gene information in genomic evaluations of French dairy goats. Genet. Sel. Evol. 48:54. https://doi.org/10.1186/s12711-016-0233-x.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. E. Meuwissen, S. E. Aggrey, T. Wing, and W. M. Muir. 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. J. Anim. Sci. 89:23–28. https://doi.org/10.2527/jas.2010-3071.

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6:1565–1571. https://doi.org/10.1017/S1751731112000742.

Clément, V., D. Boichard, A. Piacère, A. Barbat, and E. Manfredi. 2002. Genetic evaluation of French goats for dairy and type traits. Pages 235–238 in Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France.

Duchemin, S. I., C. Colombani, A. Legarra, G. Baloche, H. Larroque, J.-M. Astruc, F. Barillet, C. Robert-Granié, and E. Manfredi. 2012. Genomic selection in the French Lacaune dairy sheep breed. J. Dairy Sci. 95:2723–2733. https://doi.org/10.3168/jds.2011-4980.

Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1:21–28. https://doi.org/10.1017/S1751731107392628.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257. https://doi.org/10.1007/s10709-008-9308-0.

Grosclaude, F., M.-F. Mahé, G. Brignon, L. Di Stasio, and R. Jeunet. 1987. A Mendelian polymorphism underlying quantitative variations of goat α_{S1}-casein. Genet. Sel. Evol. 19:399–412. https://doi.org/10.1186/1297-9686-19-4-399.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443. https://doi.org/10.3168/jds.2008-1646.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663. https://doi.org/10.3168/jds.2009-2061.

Legarra, A., and Z. G. Vitezica. 2015. Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using gene content multiple-trait BLUP. Genet. Sel. Evol. 47:89. https://doi.org/10.1186/s12711-015-0165-x.

Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J. I. Weller. 2014. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. J. Dairy Sci. 97:1742–1752. https://doi.org/10.3168/jds.2013-6916.

Martin, P., I. Palhière, C. Maroteau, P. Bardou, K. Canale-Tabet, J. Sarry, F. Woloszyn, J. Bertrand-Michel, I. Racke, H. Besir, R. Rupp, and G. Tosser-Klopp. 2017. A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content. Sci. Rep. 7. https://doi.org/10.1038/s41598-017-02052-0.

Martin, P., I. Palhière, C. Maroteau, V. Clément, I. David, G. T. Klopp, and R. Rupp. 2018. Genome-wide association mapping for type and mammary health traits in French dairy goats identifies a pleiotropic region on chromosome 19 in the Saanen breed. J. Dairy Sci. 101:5214–5226. https://doi.org/10.3168/jds.2017-13625.

Matilainen, K., M. Koivula, I. Strandén, G. P. Aamand, and E. A. Mäntysaari. 2016. Managing genetic groups in single-step genomic evaluations applied on female fertility traits in Nordic Red Dairy cattle. Interbull Bull. 50:71–75.

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

Misztal, I., S. E. Aggrey, and W. M. Muir. 2013. Experiences with a single-step genome evaluation. Poult. Sci. 92:2530–2534. https://doi.org/10.3382/ps.2012-02739.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUPF90 and related programs. Commun. No. 28–07. Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France.

Mucha, S., R. Mrode, M. Coffey, M. Kizilaslan, S. Desire, and J. Conington. 2018. Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. J. Dairy Sci. 101:2213–2225. https://doi.org/10.3168/jds.2017-12919.

Mucha, S., R. Mrode, I. MacLaren-Lee, M. Coffey, and J. Conington. 2015. Estimation of genomic breeding values for milk yield in UK dairy goats. J. Dairy Sci. 98:8201–8208. https://doi.org/10.3168/jds.2015-9682.

Onogi, A., A. Ogino, T. Komatsu, N. Shoji, K. Shimizu, K. Kurogi, T. Yasumori, K. Togashi, and H. Iwata. 2015. Whole-genome prediction of fatty acid composition in meat of Japanese Black cattle. Anim. Genet. 46:557–559. https://doi.org/10.1111/age.12300.

Teissier, M., H. Larroque, and C. Robert-Granié. 2018. Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: A quantitative

trait influenced by a major gene. Genet. Sel. Evol. 50:31. https://doi.org/10.1186/s12711-018-0400-3.

Tosser-Klopp, G., P. Bardou, O. Bouchez, C. Cabau, R. Crooijmans, Y. Dong, C. Donnadieu-Tonon, A. Eggen, H. C. M. Heuven, S. Jamli, A. J. Jiken, C. Klopp, C. T. Lawley, J. McEwan, P. Martin, C. R. Moreno, P. Mulsant, I. Nabihoudine, E. Pailhoux, I. Palhière, R. Rupp, J. Sarry, B. L. Sayre, A. Tircazes, J. Wang, W. Wang, and W. Zhang.International Goat Genome Consortium. 2014. Design and Characterization of a 52K SNP chip for goats. PLoS One 9:e86227. https://doi.org/10.1371/journal.pone.0086227.

Van Sickle, J. 2003. Analyzing correlations between stream and watershed attributes. J. Am. Water Resour. Assoc. 39:717–726. https://doi.org/10.1111/j.1752-1688.2003.tb03687.x.

Venot, E., D. Boichard, V. Ducrocq, S. Fritz, H. Larroque, F. Tortereau, J.-M. Astruc, A. Barbat, M. Barbat, A. Baur, P. Boulesteix, C. Carillier-Jacquin, P. Croiseau, M.-N. Fouilloux, A. Gion, C. Hoze, A. Launay, R. Lefebvre, A. Legarra, V. Loywyck, I. Palhiere, F. Phocas, J. Promp, C. Robert-Granie, R. Rupp, R. Saintilan, M.-P. Sanchez, T. Tribout, A. Vinet, and S. Mattalia. 2017. French genomic experience: Genomics for all ruminant species. In 40th Biennal Session of ICAR.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94:73–83. https://doi.org/10.1017/S0016672312000274.

Wolc, A., A. Kranis, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, A. Avendano, K. A. Watson, J. M. Hickey, G. de los Campos, R. L. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2016. Implementation of genomic selection in the poultry industry. Anim. Front. 6:23–31. https://doi.org/10.2527/af.2016-0004.

Zhang, Z., X. Ding, J. Liu, D.-J. de Koning, and Q. Zhang. 2011. Genomic selection for QTL-MAS data using a trait-specific relationship matrix. BMC Proc. 5:S15. https://doi.org/10.1186/1753-6561-5-S3-S15.

Zhang, Z., M. Erbe, J. He, U. Ober, N. Gao, H. Zhang, H. Simianer, and J. Li. 2015. Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3 Genes Genomes Genet. 5:615–627. https://doi.org/10.1534/g3.114.016261.

Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS. Front. Genet. 7:151. https://doi.org/10.3389/fgene.2016.00151.

Zhou, L., R. Mrode, S. Zhang, Q. Zhang, B. Li, and J.-F. Liu. 2018. Factors affecting GEBV accuracy with single-step Bayesian models. Heredity 120:100–109. https://doi.org/10.1038/s41437-017-0010-9.

Chapitre 5 : Utilisation d'haplotypes ou de pseudo-SNPs dans les modèles et méthodes d'évaluations génomiques

1. Description des haplotypes (ou pseudo-SNPs) construits avec les méthodes DW et LD pour les races caprines Alpine et Saanen

1.1. Construction des haplotypes (ou pseudo-SNPs)

Les phases des génotypes étant inconnues lorsque les animaux sont génotypés avec la puce 50K, nous avons utilisé le logiciel FImpute (avec les paramètres par défaut) pour réaliser ce phasage. Le pédigrée a été utilisé en complément des génotypages 50K et les SNPs manquants (codé 5) ont été imputés par FImpute.

Pour créer les haplotypes selon la méthode DW (paragraphe 2.1, Chapitre 2), nous avons testé des fenêtres de 2, 5, 10, 15, 20, 25, 30, 35, 40, 45 et 50 SNPs de longueur, intra-chromosome. Pour la méthode LD (paragraphe 2.1, Chapitre 2), une étape supplémentaire a été nécessaire pour calculer le LD, via le paramètre r^2 (présenté dans le paragraphe 3.4.2, chapitre 1), entre SNPs. Nous avons utilisé le logiciel PLINK pour cette étape en testant plusieurs seuils de LD : 0,01 ; 0,02 ; 0,03 ; 0,04 ; 0,05 ; 0,06 ; 0,07 ; 0,08 ; 0,09 ; 0,1 ; 0,2 ; 0,3 ; 0,4 ; 0,5 ; 0,6 ; 0,7 ; 0,8 ; 0,9 et 1.

1.2. Analyse du LD pour les populations caprines

La Figure 56 présente le LD moyen pour les analyses multirace, Alpine et Saanen selon la distance entre 2 SNPs. Les LD ont été calculés comme dans le paragraphe 3.4.2 (Chapitre 1). Ils sont très proches pour les 2 races avec une valeur de 0,36 en Alpine et 0,39 en Saanen pour des SNPs distants de 10 kb. Ils atteignent 0,13 pour des SNPs distants de 50 kb (distance moyenne entre 2 SNPs de la puce 50K), et se stabilisent à 0,06 pour des SNPs distants de plus de 900 kb. En analyse multirace, le LD est plus faible : 0,36 pour des SNPs distants de 10 kb, 0,11 pour des SNPs distants de 50 kb et 0,03 pour des SNPs distants de plus de 900 kb.

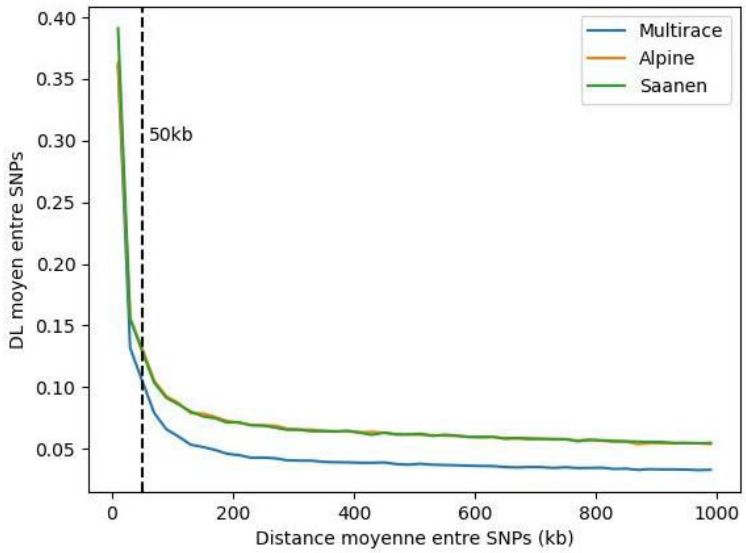


Figure 56. Estimation du LD moyen pour les analyses multirace, Alpine et Saanen selon la distance entre 2 SNPs

1.3. Analyse de la diversité des haplotypes construits avec les méthodes DW et LD

La taille moyenne des haplotypes diminue avec l'augmentation du seuil de LD (Figure 57A). Avec un seuil de LD de 0,01, la taille moyenne des haplotypes est de 2,38 SNPs pour les analyses multirace, 2,69 SNPs pour les analyses Alpine et 2,66 SNPs pour les analyses Saanen. Cette taille moyenne chute rapidement pour finalement ne contenir qu'un SNP avec un seuil de

LD de 1 Les haplotypes sont en moyenne de même taille entre les analyses Alpine et Saanen et légèrement plus courts pour les analyses multirace.

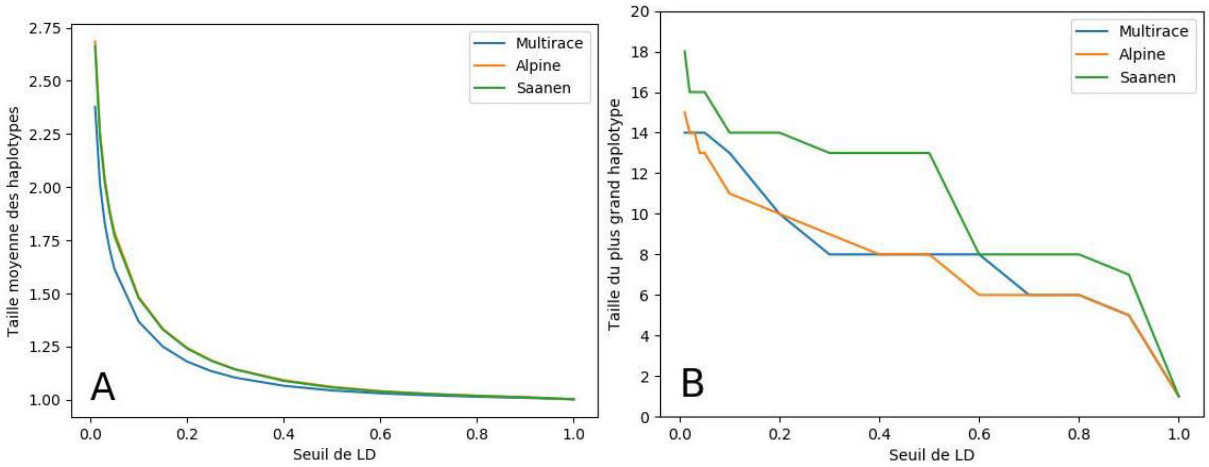


Figure 57 : Taille moyenne des haplotypes (A) et taille maximale des haplotypes (B) selon le seuil de LD choisi pour construire les haplotypes (LD) pour les analyses multirace, Alpine et Saanen

La Figure 57B présente la taille de l'haplotype le plus long selon les différents scénarios testés avec le seuil de LD. Pour un LD de 0,01, l'haplotype le plus long est de 14 SNPs pour les analyses multirace, de 15 SNPs pour les analyses Alpine et de 18 SNPs pour les analyses Saanen. Avec un seuil de LD de 1, l'haplotype le plus long est constitué d'un seul SNP pour les 3 analyses. Les tailles maximales des haplotypes sont plus grandes pour la race Saanen alors qu'elles sont proches ou identiques pour les analyses multirace et Alpine. L'haplotype le plus long pour un LD de 0,01 n'est pas nécessairement le même que l'haplotype le plus long pour un LD de 0,02, etc. Ces haplotypes (les plus longs) peuvent donc se retrouver sur des chromosomes différents selon le seuil de LD. Le Tableau 25 indique sur quel chromosome l'haplotype le plus long se trouve pour chaque analyse. Pour la race Saanen, l'haplotype le plus long pour la majorité des seuils testés se situe sur le chromosome 6 (chromosome contenant le gène de la *caséine* α_{s1}) sauf pour des seuils de LD supérieurs à 0,5 où l'haplotype le plus long est localisé sur le chromosome 11. Pour les analyses multirace et Alpine, l'haplotype le plus long sur le chromosome 6 est retrouvé pour des seuils de LD faibles (entre 0,01 et 0,05). Pour des seuils de LD plus élevés (supérieur à 0,1), les haplotypes les plus longs se situent principalement sur les chromosomes 8, 10, 11, 12, 13, 18 et 22 et 23.

Tableau 25. Localisation des haplotypes les plus longs selon le seuil de LD utilisé pour construire les haplotypes pour les analyses multirace, Alpine et Saanen

Seuil LD	Chromosome Multirace	Chromosome Alpine	Chromosome Saanen
0,01	6;13	6;11	6
0,02	6;12	6	6;12
0,03	6;12	6	6
0,04	6	6	6
0,05	6	6	6
0,1	6	12	6
0,2	8;11	8	6
0,3	12;18	1;8;11;18	6
0,4	11;18	10;11;12	6
0,5	11	11;12	6
0,6	11	1;18;23	11
0,7	11	1;18;23	11
0,8	1;18	18	11
0,9	6;13;18;22	6;13;18;22	11

La méthode LD regroupe des SNPs dans des haplotypes si le LD entre eux est suffisamment élevé. Dans le cas contraire, les SNPs ne sont dans aucun haplotype et sont considérés isolés pour la suite des analyses. Sachant que le LD moyen entre deux SNP (50kb) de la puce est de 0,13 en Alpine et en Saanen, plus le seuil de LD va être élevé, et plus le nombre de SNPs « individuels » (c'est à dire haplotype composé d'un seul SNP) dans les analyses va être important (Figure 58). Avec un seuil de LD de 0,15, 50% des haplotypes ne contiennent qu'un seul SNP pour les analyses Alpine et Saanen. Cette proportion atteint 95% pour un seuil de LD à 0,7. Pour les analyses multirace, la proportion d'haplotype composé d'un seul SNP augmente plus rapidement que pour les analyses intra-race. Ainsi, 50% des haplotypes contiennent un seul SNP lorsque le seuil de LD est de 0,1. On atteint 95% pour un seuil de LD entre 0,6 et 0,7.

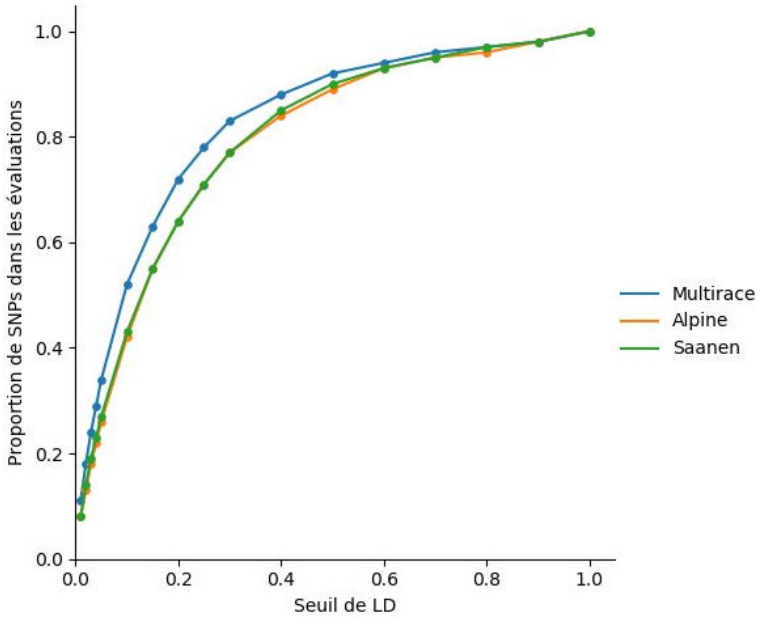


Figure 58. Proportion de SNPs dans les haplotypes selon le seuil de LD pour les analyses multirace, Alpine et Saanen

La taille des haplotypes avec la méthode DW et le seuil de LD pour la méthode LD sont deux paramètres qui vont modifier le nombre d'allèles observé pour chaque haplotype. La Figure 59 présente le nombre moyen d'allèle par haplotype pour les analyses multirace, Alpine et Saanen selon la méthode DW ou LD. Pour la méthode DW, le nombre d'allèle moyen augmente avec la taille de l'haplotype. Cette augmentation est plus forte pour les analyses multirace que pour les analyses Alpine et Saanen. Le nombre d'allèle moyen passe de 4 (2 SNPs) à 630 (50 SNPs) en multirace, 4 (2 SNPs) à 370 (50 SNPs) en Alpine et 4 (2 SNPs) à 300 (50 SNPs) en Saanen. Pour des haplotypes de longueur supérieure à 20 SNPs, on observe un nombre plus faible d'allèle pour la race Saanen, que pour la race Alpine ou l'analyse multirace. Cet écart s'accentue avec la taille des haplotypes. Pour la méthode LD, le nombre d'allèle moyen est largement plus faible que pour la méthode DW. Avec un seuil de LD de 0,01, on observe en moyenne 6,60 allèles par haplotype pour les analyses multirace, 7,90 allèles pour les analyses Alpine et Saanen. Le nombre moyen d'allèle chute rapidement pour atteindre 2,81 pour les analyses multirace et 3,08 pour les analyses Alpine et Saanen avec un seuil de LD de 0,1. Finalement et comme attendu, une moyenne de 2 allèles est obtenue pour un seuil de LD de 1.

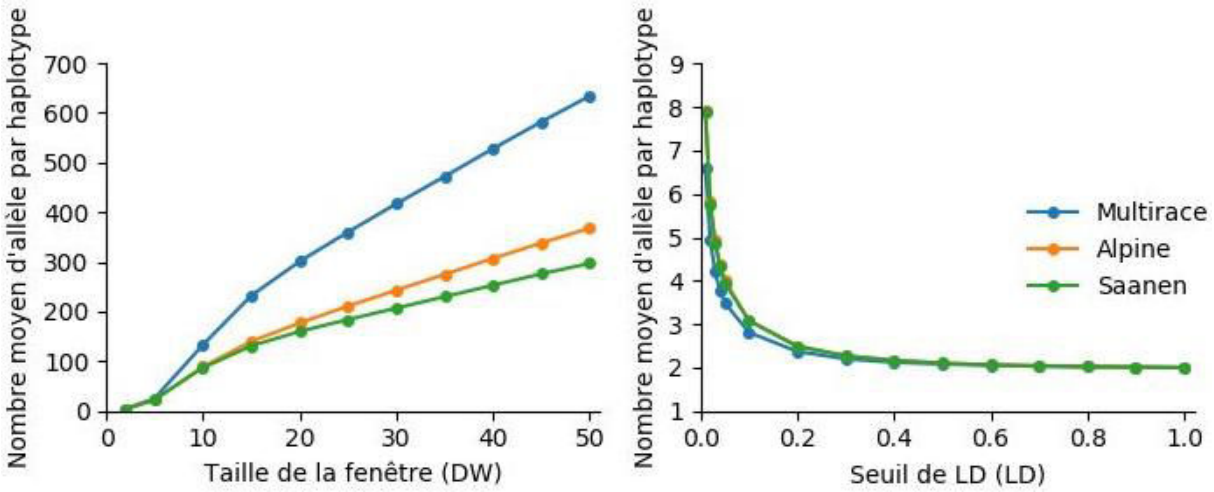


Figure 59. Nombre moyen d'allèles observé pour chaque haplotype selon la méthode DW et LD pour les analyses multirace, Alpine et Saanen

2. Précisions des évaluations génomiques avec des pseudo-SNPs pour des caractères de productions laitières, de morphologie de la mamelle et de comptage de cellules somatiques pour les races caprines

Le troisième article de ma thèse, actuellement en préparation, s'intéresse aux prédictions des évaluations génomiques chez les caprins en utilisant des haplotypes convertis en pseudo-SNPs. Ces pseudo-SNPs ont été intégrés dans des évaluations ssGBLUP et WssGBLUP. Comme pour l'article I et II, les données utilisées dans cet article sont celles du scénario E (paragraphe 1.1, Chapitre 3).

Article III : Prédictions génomiques basés sur les haplotypes pour des caractères de production laitières, des caractères de morphologie de la mamelle et le comptage de cellule somatiques chez les caprins français.

Marc Teissier, Hélène Larroque, Flavio Schenkel, Luiz Brito, Christèle Robert-Granié, 2018. Genomic prediction based on haplotypes for milk production traits, udder type traits and somatic cell scores in French dairy goats.

Genomic prediction based on haplotypes for milk production traits, udder type traits and somatic cell scores in French dairy goats

Marc Teissier[*1], Hélène Larroque[1], Flavio Schenkel[2], Luiz Brito[2], Christèle Robert-Granié[1]

[1]GenPhySE, Université de Toulouse, INRA, INPT, ENVT, 31326 Castanet-Tolosan, France

[2]Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, Ontario, Canada

[*]Corresponding author

Email addresses:

MT: marc.teissier@inra.fr

HL: helene.larroque@inra.fr

FS: schenkel@uoguelph.ca

LB: lbrito@uoguelph.ca

CRG: christele.robert-granie@inra.fr

ABSTRACT

The use of haplotypes could improve accuracy of genomic evaluations by better capturing causal variants than from LD with SNPs. The haplotypes and the length of haplotypes can be defined in many ways, we emphasised the use of pseudo-SNPs (haplotypes converted into SNPs) straightforward used with ssGBLUP software. The aim of this study was to compare ssGBLUP and Weighted ssGBLUP based on haplotypes or individual SNP in terms of accuracy and pseudo-SNPs weights.

Genomic evaluations were performed on 11 traits under selection (milk production traits, udder type traits and somatic cell scores) in French dairy goats (Alpine and Saanen). The training population was constituted of 307 Alpine bucks and 247 Saanen bucks genotyped with the Illumina goat SNP50 BeadChip. The validation population included 205 Alpine bucks and 146 Saanen bucks. Accuracy was evaluated in the validation population as the Pearson correlation between GEBV estimated with all methods and DYD calculated from official genetic evaluation of January 2016.

Results indicated that haplotype-based models could improve accuracy of genomic evaluations for some traits. Gain of accuracy up to +19% (for fat yield) in Alpine and up to +4% (for udder shape) in Saanen were observed with ssGBLUP and up to 22% (for fat yield) in Alpine and 21% (for somatic cell scores) in Saanen with WssGBLUP.

Keywords: genomic evaluation, ssGBLUP, weighted ssGBLUP, haplotype-based models, individual SNP-based models, French dairy goats

INTRODUCTION

A class of genomic evaluation methods uses genotypes to estimate relationships between pairs of animals, such as GBLUP (VanRaden, 2008) or ssGBLUP (Legarra et al., 2009). These

methods are more accurate than pedigree-based BLUP but they assumed that all SNPs explains the same proportion of the genetic variance (Legarra et al., 2009). This is not suitable for traits with important QTLs. To overcome this limitation, the WssGBLUP method was developed to perform GWAS and genomic evaluations (Wang et al., 2012; Zhang et al., 2016). It allocates weights to SNP variances by giving more weight to SNPs that are in high LD with a causal mutation or associated with QTL with a relatively large effect (Wang et al., 2012). Alternatives WssGBLUP, developed by Zhang et al., (2016), propose to use common weights for consecutive SNPs and create the weighed genomic relationship matrix. The common weight in a defined window is calculated as the sum of all SNP weights of the window ($WssGBLUP_{Sum}$) or as the maximum of the SNP weights of the window ($WssGBLUP_{Max}$).

Previous studies were carried out on all traits under selection in French dairy goats (Alpine and Saanen breeds) to evaluate the accuracy of genomic evaluations with ssGBLUP and WssGBLUP and its alternatives ($WssGBLUP_{Sum}$ and $WssGBLUP_{Max}$) (Teissier et al., 2019; 2018a). The accuracies were improved with WssGBLUP and its alternatives compared to ssGBLUP (up to +14%) for traits with a major gene or QTL. This concerned protein content with the α_{s1} *casein* gene (chromosome 6) for both breeds (Alpine and Saanen), and milk yield, fat yield, protein yield, udder floor position and rear udder attachment for the Saanen breed only, with a large QTL on chromosome 19. For the other traits, with a genetic determinism more polygenic, the accuracies with WssGBLUP and its alternatives were similar or slightly lower than accuracies obtained with ssGBLUP (accuracies were 5% to 0% lower). WssGBLUP and its alternatives were suitable methods to take into account the presence of a major gene or QTL in genomic evaluations in French dairy goats.

Another alternative strategy, to improve accuracy of genomic evaluations, is the use of haplotypes. A haplotype could be defined as a group of nearby SNPs on the same chromosome and that are frequently inherited together. The use of haplotypes in genomic evaluations present

some advantages compared to SNPs. Haplotypes are more informative than SNP to describe recent IBD relationship. They may capture LD with multi-allelic QTL better than SNPs which are often bi-allelic (Meuwissen et al., 2014). Then, long haplotypes are better to differentiate IBD and IBS because long shared haplotypes are likely to come from common ancestor (Broman and Weber, 1999). SNPs present on a chip are chosen to have moderate to high minor allele frequency (MAF), therefore, these SNPs are old mutations (new mutations have a low frequency at the beginning (Meuwissen et al., 2014)), so they should be less efficient to trace new mutations compared to haplotypes (Meuwissen et al., 2014).

In practice, accuracy of genomic evaluations with haplotypes varied across studies and traits. Some studies reported no improvement of accuracy for genomic evaluations with haplotypes (Hickey et al., 2013; Meuwissen et al., 2014; Uemoto et al., 2017) whereas others shown better accuracy with haplotypes (Meuwissen et al., 2014a; Jónás et al., 2016; Hess et al., 2017; Karimi et al., 2018). The haplotypes can be defined as grouping together consecutive SNPs (Hickey et al., 2013; Ferdosi et al., 2016), or using linkage disequilibrium (LD) to construct haploblocks (Cuyabano et al., 2014). Jónás et al., (2016) proposed another approach using a preselection of SNPs and optimising allele frequency to construct haplotypes. To integrate haplotypes into genomic models, one possibility is to convert them into pseudo-SNPs. SsGBLUP can easily be implemented using pseudo-SNPs to construct the genomic relationship matrix.

The aim of this study was to investigate genomic evaluation methods using haplotypes converted into pseudo-SNPs. Accuracy of genomic evaluations with ssGBLUP and WssGBLUP were compared with the use of either SNP or pseudo-SNPs.

MATERIALS AND METHODS

Dataset of phenotypes, pedigree and genotypes

90 The datasets were provided by the French national milk records system from official
91 genetic evaluation of January 2016 (Larroque et al., 2011). They contained records from Alpine
92 and Saanen French dairy goat breeds and included phenotypes, pedigrees, genotypes (Illumina
93 goat SNP50 BeadChip) and environmental effects. .

94 Analyses were performed within breed on five milk production traits, five udder type
95 traits and somatic cell scores. Milk production traits were milk yield (MY, kg), fat and protein
96 yield (FY and PY, kg) and fat and protein content (FC and PC, g/kg). The number of phenotypes
97 is about 4 and 3 million in Alpine and Saanen breed respectively (table 1). Udder type traits,
98 scored from 1 to 9, were Udder Floor Position (UFP), Rear Udder Attachment (RUA), Udder
99 Shape (US), Teat Angle (TA), Fore Udder (FU). Around 150,000 phenotypes in Alpine breed
100 and 100,000 phenotypes in Saanen breed were available (table 1) for udder type traits. Animals
101 were scored only once in their career for udder type traits explaining the fewer number of
102 phenotypes than for milk production traits. The last trait studied was somatic cell scores (SCS:
103 log-transformed somatic cell counts). Almost 1.3 million phenotypes were recorded in Alpine
104 and 1 million phenotypes in Saanen.

105 The pedigree file contained animals born between 1936 and 2012. It included 1,446,296
106 Alpine and 1,097,384 Saanen for milk production traits. For SCS, it contained 788,576 Alpine
107 and 648,461 Saanen. Unknown parent groups completed the pedigree with one group for
108 animals born before 1975 and then pooled groups (sires and dams) were defined every two
109 years, because there were few animals with unknown dams.

110 Genotypes with the Illumina goat SNP50 BeadChip (50K SNP) (Tosser-Klopp et al.,
111 2014) were available for animals born between 1980 and 2012 from French dairy goats. Quality
112 control (QC) was applied to 2,056 genotyped Alpine and 1,349 genotyped Saanen animals for
113 53,347 SNPs, independently for each breed. Details of QC can be found in Teissier et al.,

(2018). A total of 46,849 SNPs were remained for 1749 genotyped Alpine (512 males and 1237 females) and 1206 genotyped Saanen (393 males and 813 females). These animals represented our reference population.

Genomic evaluations using individual SNP

The ssGBLUP method. The ssGBLUP uses simultaneously all female phenotypes, pedigrees and genotypes in a single step approach to estimate GEBVs of all animals (Legarra et al., 2009). Genomic evaluations in French dairy goats are based on two different models, one for milk production traits and SCS and a second for udder type traits. For milk production traits and SCS, model used was:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Wp} + \boldsymbol{e}\ [model\ 1]$$

where $\boldsymbol{y}$ is a vector of phenotypes (MY, FY, PY, FC, PC or SCS), $\boldsymbol{\beta}$ is a vector of fixed effects including 4 combined effects (same fixed effects for milk production traits and SCS): herd effect, age and month at delivery effect and length of dry period effect. Herd effect was estimated within year (32 years from 1980 to 2012) and parity (1, 2 and ≥3); age and month were estimated within year and region (four regions in France depending on goat breeding management). Length of dry period was estimated within year and region. $\boldsymbol{u}$ is a vector of genomic breeding values (GEBV) assumed to be normally distributed $N(\mathbf{0}, \boldsymbol{H}\sigma_u^2)$ where $\boldsymbol{H}$ represent the hybrid relationship matrix (Legarra et al., 2009), $\boldsymbol{p}$ is a vector of random permanent environmental effects assumed to be normally distributed $N(\mathbf{0}, \boldsymbol{I}\sigma_p^2)$, $\boldsymbol{e}$ is a vector of random residual normally distributed $N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$. $\boldsymbol{X}$ is the incidence matrix relating phenotypes to fixed effects ($\boldsymbol{\beta}$); $\boldsymbol{Z}$ is the design matrix allocating phenotypes to genomic breeding values ($\boldsymbol{u}$) and $\boldsymbol{W}$ is the incidence matrix relating phenotypes to permanent environmental effects ($\boldsymbol{p}$).

For udder type traits, the model used was:

138 $$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}\ [model\ 2]$$

139 where $\boldsymbol{y}$, $\boldsymbol{u}$ and $\boldsymbol{e}$ were the same vector previously described in model 1 and $\boldsymbol{\beta}$ the vector of 3
140 combined fixed effects: herd, age at scoring and stage at scoring. Herd effect was estimated
141 within year (32 years from 1980 to 2012) and parity (1, 2 and ≥3), age at scoring and stage at
142 scoring was estimated within year.

143 The ssGBLUP used a hybrid relationship matrix ($\boldsymbol{H}$) blending pedigrees and genotypes The
144 inverse of $\boldsymbol{H}$ is expressed as:

145 $$\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G^{-1}} - \boldsymbol{A_{22}^{-1}} \end{bmatrix}$$

146 where $\boldsymbol{A}$ is the numerator relationship matrix estimated from the pedigree, $\boldsymbol{A_{22}}$ is the numerator
147 relationship matrix for genotyped animals and $\boldsymbol{G}$ is the genomic relationship matrix expressed
148 as in VanRaden, (2008):

149 $$\boldsymbol{G} = \frac{\boldsymbol{M'M}}{2\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} p_i(1-p_i)}$$

150 where m is the number of SNPs, p_i the estimated allele frequency at the locus i, $\boldsymbol{M}$ a centered
151 matrix of SNP genotypes. In this study, results from ssGBLUP based on individual SNP will
152 be called ssGBLUP$_{SNP}$ and were performed with the blup90iod2 software (Misztal et al., 2016).

153 ***The Weighted ssGBLUP method (WssGBLUP) using individual SNP.*** The WssGBLUP
154 method allocates SNPs weights according to their effect on the selected trait to estimate
155 relationship between each pair of animals. This method, introduced by Wang et al., (2012) and
156 based on a model similar to ssGBLUP, includes major genes or QTLs with a relatively large
157 effect using a weighted $\boldsymbol{G}$ ($\boldsymbol{G}$*). This genomic relationship matrix $\boldsymbol{G}$* is expressed as:

158 $$\boldsymbol{G^*} = \frac{\boldsymbol{M'DM}}{2\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{p_i(1-p_i)}}$$

where $\boldsymbol{A_{22}}$, $\boldsymbol{M}$, p_i and m was the same as in $\boldsymbol{G}$ and $\boldsymbol{D}$ is a diagonal matrix of size $m * m$ where each element of the diagonal corresponds to SNP weights. SNPs weights are estimated from GEBVs estimated with ssGBLUP. The WssGBLUP approach is based on an iterative algorithm with different steps: (i) run ssGBLUP with the $\boldsymbol{G}^*$ matrix (at iteration 1, SNPs weights in the **D** matrix are equal to one and is equivalent to a ssGBLUP), (ii) estimate SNP effects from solutions of genomic breeding values in the previous step, (iii) estimate variances of the effect of each SNP, (iv) normalize the vector of variances of SNP effects to get the SNP weights (this normalization process ensures that the sum of the variances remain constant and equal to the number of SNPs), (v) use SNP weights to construct the **D** matrix, (vi) loop to step (i). Previous studies have shown that the second iteration of the WssGBLUP was the most accurate (Wang et al., 2012; Teissier et al., 2018a), results were presented for this scenario. Results from WssGBLUP based on individual SNP will be called $\text{WssGBLUP}_{\text{SNP}}$. SNPs effects, SNPs weights and GEBVs were estimated with the blup90iod2 software (Misztal et al., 2016).

Haplotypic genomic evaluations

In this study, haplotypes were constructed with two different methods: considering a fixed number of adjacent SNPs along the chromosome, called Distinct Window (DW) method (Ferdosi et al., 2016) and using a fixed threshold of Linkage Disequilibrium (LD) between each pair of SNP, called here LD method (Cuyabano et al., 2014). The construction of haplotypes requires phased genotypes, so, parental haplotypes were reconstructed using FImpute software (Sargolzaei et al., 2014). Phasing step was performed within breed with all available genotyped animals of the reference population. Figure 1 present, from an example of 2 animals genotyped on five SNPs and knowing parental phases, how to construct haplotypes defined by DW and LD methods (Ferdosi et al., 2016).

183 **Construction of haplotypes**

184 ***Distinct Windows (DW) method.*** The haplotypes with the DW method were defined by
185 considering a fixed number of adjacent SNPs along the chromosome (Hickey et al., 2013;
186 Ferdosi et al., 2016) (Figure 1.B). For the last haplotype of the chromosome, if the number of
187 adjacent SNPs is shorter than those fixed, SNPs from the previous fragment will be used to
188 ensure that haplotype will have the same size. In the example (Figure 1.B), SNP 4 is present in
189 the haplotypes 2 and 3. In this study, the size of haplotypes tested was 2, 5, 10, 15, 20, 25, 30,
190 35, 40, 45 and 50 SNPs.

191 ***Linkage Disequilibrium (LD) method.*** The haplotypes based on LD were constructed as
192 presented in (Cuyabano et al., 2014) and are called haploblocks. Firstly, LD is computed
193 between all pairs of SNPs with the PLINK software (Purcell et al., 2007) using the r² metrics
194 (Rogers and Huff, 2009). This measure ranges from 0 for no LD to 1 for complete LD between
195 two SNPs:

196 $$r^2 = \frac{\left(cov(g_i, g_j)\right)^2}{var(g_i) * var(g_j)}$$

197 where g_i and g_j are genotypes (coded as 0 1 or 2) for SNP i and j. Haploblocks were defined
198 as a group of SNPs where the LD between each pair of SNP was higher or equal to a threshold
199 fixed. In our example (Figure 1.C), LD between two SNPs higher than the threshold was
200 represented in black and in grey otherwise. Haploblocks are presented as a square where all
201 cells are black. It concerned SNPs 1, 2 and 3 that can be gathered in a same haploblock. The
202 LD between SNPs 4 and 5 was not high enough to create a haploblock with 2 SNPs, so 2
203 haploblocks with only 1 SNP were created. Thresholds of LD of 0.01, 0.02, 0.03, 0.04, 0.05,
204 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1 were tested.

205 **The haplotypic genomic relationship matrix**

206 The haplotypes were used in genomic evaluation through the genomic relationship
207 matrix. They were converted into pseudo-SNPs. To achieve this, each allele of each haplotype
208 was considered as a pseudo-SNP, and the number of copies is counted for each animal and each
209 phase (allele count is equal to 0, 1 or 2). The total number of haplotypes was different depending
210 on the size of haplotype used. The Figure 2 shows the results of the transformation of
211 haplotypes to pseudo-SNP from example defined in Figure 1.

212 After converting haplotypes into pseudo-SNPs, pseudo-SNPs with a frequency lower
213 than 1% were excluded from the analyses. As they look like SNPs, the implementation of
214 ssGBLUP and WssGBLUP with pseudo-SNP is straightforward. The pseudo-SNPs were used
215 to construct the genomic relationship matrix, $\boldsymbol{G_{pseudo-SNP}}$ for the ssGBLUP and $\boldsymbol{G^{*}_{pseudo-SNP}}$
216 for the WssGBLUP:

217 $$\boldsymbol{G_{pseudo-SNP}} = \frac{\boldsymbol{M'_{pseudo-SNP} M_{pseudo-SNP}}}{2\sum_{i=1}^{m} p_i(1-p_i)}$$

218 where $\boldsymbol{M_{pseudo-SNP}}$ is a centered matrix of pseudo-SNPs (construct either with the DW or the
219 LD method). The methods using pseudo-SNPs will be called ssGBLUP$_{\text{pseudo-SNPs (DW)}}$ for DW
220 method and ssGBLUP$_{\text{pseudo-SNPs (LD)}}$ for LD method.

221 $$\boldsymbol{G^{*}_{pseudo-SNP}} = \frac{\boldsymbol{M'_{pseudo-SNP} D M_{pseudo-SNP}}}{2\sum_{i=1}^{m} \boldsymbol{p_i(1-p_i)}}$$

222 The methods using pseudo-SNPs will be called WssGBLUP$_{\text{pseudo-SNPs (DW)}}$ for DW
223 method and WssGBLUP$_{\text{pseudo-SNPs (LD)}}$ for LD method.

224 **Accuracy of genomic evaluation**

225 This reference population was split into two subsets: a training set and a validation set.
226 The training population included 307 Alpine bucks and 247 Saanen bucks born between 1993
227 and 2007, all the information on these animals (genotype, the pedigree of their ancestry and

their progeny, and the phenotypes of their progeny) was conserved to estimated GEBV. The validation set included 205 Alpine bucks and 146 Saanen bucks born between 2008 and 2012. For these animals, the phenotypes of their progeny were removed from the analysis and only the genotypes and pedigree of their ancestry were retained. The performance of genomic predictions was measured as the squared Pearson correlation between GEBV and Daughter Yield Deviation (DYD, (VanRaden and Wiggans, 1991)) from the official genetic evaluation of January 2016, in the validation population.

RESULTS

Number of haplotypes and pseudo-SNPs with DW and LD methods

We investigated the number of haplotypes that were created with DW and LD methods, and the total number of pseudo-SNPs. Figure 3 presents the number of haplotypes according to the size of haplotypes (DW) or the threshold of LD (LD). The number of haplotypes with DW method in Alpine and Saanen was the same because SNPs were exactly the same after QC (46,849 SNPs in total). For DW, the number of haplotypes decreased when size of haplotypes increased. It ranged from 23,429 haplotypes with a size of 2 SNPs to 937 haplotypes with a size of 50 SNPs. With LD method, the number of haplotypes increased with a high threshold of LD. The number of haplotypes is equal to 17,445 for Alpine and 17,594 for Saanen with a threshold of LD equal to 0.01. The number of haplotypes reached 46,849 for a threshold of LD equal to 1 for Alpine and Saanen. Between a threshold of 0.01 and 1, number of haplotypes in Alpine and Saanen are almost similar with only a difference of 110 haplotypes on average. With LD method, a part of haplotypes is constituted of individual SNP. Almost 10% of haplotypes were individual SNP for a threshold of LD of 0.01, 50% of haplotypes were individual SNP with a threshold of LD of 0.1 and this proportion reached 90% with a threshold of LD of 0.5 (results not shown).

252 The Figure 4 presents the number of pseudo-SNPs according to the size of haplotypes
253 (DW) or the threshold of LD (LD). With DW, number of pseudo-SNPs increased with a size of
254 haplotypes between 2 and 5 SNPs. A maximum was reached with 5 SNPs with 118,151 pseudo-
255 SNPs in Alpine and 117, 566 pseudo-SNPs in Saanen. Then, the number of pseudo-SNPs
256 dropped to 22,029 in Alpine and 19,674 in Saanen with haplotypes of 50 SNPs. In Alpine and
257 Saanen, 96% of pseudo-SNPs remained after filtering on their allele frequency for haplotypes
258 of 2 SNPs, this proportion decreased with the size of haplotypes and reached only 6% for
259 haplotypes with 50 SNPs. With LD method, the highest number of pseudo-SNPs was observed
260 for a threshold of LD of 0.01 (95,233 pseudo-SNPs in Alpine and 94,980 pseudo-SNPs in
261 Saanen). Then, number of pseudo-SNPs decreased rapidly and was lower than 50,000 (in
262 Alpine and Saanen) for a threshold of LD equal to 0.5. Finally, the number of pseudo-SNPs
263 reached 46,849 pseudo-SNPs in Alpine and Saanen with a threshold of LD equal to one (only
264 individual SNP remained in the analyses). After filtering pseudo-SNPs based on their
265 frequency, 71% of pseudo-SNPs remained with a LD equal to 0 and all the pseudo-SNPs for a
266 LD equal to 1.

267 **SNPs weights with pseudo-SNPs**

268 The Figure 5 presents weights for pseudo-SNPs on chromosome 6 for PC in Alpine and
269 Saanen breeds according to the different size of haplotypes (DW) used in this study. High
270 weights for pseudo-SNPs were located at the end of chromosome 6 for both breeds. In Alpine,
271 when the size of haplotypes increased, the highest weights increased too (from 52 with
272 haplotypes of 2 SNPs to 454 with haplotypes of 50 SNPs). For haplotypes with 2 SNPs, the
273 sum of weights of the 1% pseudo-SNPs with the highest weights explained 15% of the sum of
274 weights of all chromosome 6. It reached 64% for haplotypes with 50 SNPs. In Saanen,
275 maximum weights were equal to 92, 410, 964, 661, 909 and 728 for haplotypes with 2, 5, 10,
276 15, 20, 25 respectively. With longer haplotypes, maximum pseudo-SNP weights were equal to

191 on average. For haplotypes with 5, 10, 15, 20 and 25 SNPs, the sum of weights of the 1% best pseudo-SNPs explained 40% of the sum of weights of all chromosome 6, otherwise it reached 35% on average.

The Figure 6 presents weights for pseudo-SNPs on chromosome 6 for PC according to the threshold of LD in Alpine and Saanen breeds. As for DW method, important weights were observed at the end of the chromosome 6. When the threshold of LD increased, the maximum weights of pseudo-SNPs decreased in both Alpine and Saanen breeds. In Alpine, maximum weights were equal to 82 for a threshold of LD equal to 0.01 and 35 for a threshold of LD equal to 1. In Saanen breed, the maximum weight was equal to 125 for a threshold of LD of 0.01 and equal to 67 for a threshold of LD equal to 1. In Alpine and with a threshold of LD equal to 0.01, the sum of weights of the 1% pseudo-SNPs with the highest weights explained 24% of the sum of weights of all chromosome 6. With a threshold of LD equal to 1, this proportion decreased to reach 13%. The same phenomenon was observed for Saanen breed with 23% of all pseudo-SNPs weights of chromosome 6 explained by the 1% pseudo-SNPs with the highest weights for a threshold of LD equal to 0.01. This proportion is quite constant according to the threshold of LD and was equal to 21% for a threshold of LD equal to 1.

With Saanen breed, similar profile obtained for PC was observed for MY, FY, PY, FU, RUA, UFP and CCS on chromosome 19. SNP weights on chromosome 19 decreased with a higher threshold of LD. A smaller peak was observed on chromosome 13 for FU with DW method. For other traits, no peak (SNPs with high weights) was detected (results not shown). With Alpine breed, no peak was detected for all traits (results not shown) except for PC.

Accuracy of genomic evaluation

The table 2 presents accuracies of genomic evaluations with individual SNP ($ssGBLUP_{SNP}$) and haplotypes ($ssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ and $ssGBLUP_{pseudo\text{-}SNPs\ (LD)}$) for each

301 trait for Alpine breed. Only best accuracies according to the haplotype size (DW) or threshold
302 of LD (LD) are presented. We observed slightly better accuracies for ssGBLUP$_{\text{pseudo-SNPs (DW)}}$
303 or ssGBLUP $_{\text{pseudo-SNPs (LD)}}$ than with ssGBLUP$_{\text{SNP}}$ for MY, FY, PY, FC, FU, UFP, US, TA and
304 SCS. Accuracies were +1 percent point to +3 percent points higher except for PY where
305 accuracies with ssGBLUP$_{\text{pseudo-SNPs (LD)}}$ and ssGBLUP$_{\text{SNP}}$ were the same (0.30). Accuracies
306 were identical between ssGBLUP$_{\text{SNP}}$, ssGBLUP $_{\text{pseudo-SNPs (DW)}}$ and ssGBLUP$_{\text{pseudo-SNPs (LD)}}$ for
307 PC (0.76) and RUA (0.40). Accuracies between ssGBLUP$_{\text{pseudo-SNPs (DW)}}$ and ssGBLUP $_{\text{pseudo-}}$
308 $_{\text{SNPs (LD)}}$ were generally similar between traits except for MY (0.47 and 0.46 respectively), FY
309 (0.37 and 0.33 respectively), PY (0,32 and 0,30 respectively), and US (0.49 and 0.50
310 respectively). For ssGBLUP$_{\text{pseudo-SNPs (DW)}}$, best accuracies were mainly obtained with long
311 haplotypes (25 SNPs to 50 SNPs) for MY, FY, PY, FC, US and RUA. For other traits (PC,
312 TA, UFP and FU), haplotype size was very short and only contained 2 SNPs. For LD, the best
313 accuracies were obtained with the lowest threshold tested (0.01) for MY, FC, UFP, US and
314 SCS, with a threshold of 0.02 for FY, TA, RUA, with a threshold of 0.05 for PC, with a
315 threshold of 0.1 for FU and with a threshold of 0.3 for PY.

316 The table 3 presents accuracies of genomic evaluations with SNPs (ssGBLUP$_{\text{SNP}}$) and
317 haplotypes (ssGBLUP$_{\text{pseudo-SNPs (DW)}}$ and ssGBLUP$_{\text{pseudo-SNPs (LD)}}$ methods) for each trait for
318 Saanen breed. Accuracies were similar between ssGBLUP$_{\text{SNP}}$, ssGBLUP$_{\text{pseudo-SNPs (DW)}}$ and
319 ssGBLUP$_{\text{pseudo-SNPs (LD)}}$ for MY (0.49), FC (0.59), PC (0,73), FU (0.62), UFP (0.59) and RUA
320 (0.60). Genomic evaluation accuracies with ssGBLUP$_{\text{pseudo-SNPs (DW)}}$ were +1 percent point
321 higher than ssGBLUP$_{\text{SNP}}$ for FY, PY, US and TA. Similar results were observed with
322 ssGBLUP$_{\text{pseudo-SNPs (LD)}}$. The highest improvement of accuracy was observed for SCS with a
323 gain of +3 percent points with ssGBLUP$_{\text{pseudo-SNPs (DW)}}$ compared to ssGBLUP$_{\text{SNP}}$. Contrary to
324 Alpine breed, best accuracies with ssGBLUP$_{\text{pseudo-SNPs (DW)}}$ method were obtained with short
325 haplotypes (10 SNPs or less) for all traits except PY (20 SNPs), TA (15 SNPs) and SCS (15

SNPs). For $ssGBLUP_{pseudo\text{-}SNPs\ (LD)}$, best accuracies were observed with a threshold of LD equal to 1 for MY, FY, PC, UFP and RUA, haploblocks were only constituted of individual SNP for these evaluations.

The table 4 presents accuracies of genomic evaluations with $WssGBLUP_{SNP}$, $WssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ and $WssGBLUP_{pseudo\text{-}SNPs\ (LD)}$ for each trait in Alpine and Saanen breeds. To compare results with ssGBLUP, the same window (DW) and the same LD (LD) as in ssGBLUP are presented. For Alpine breed, genomic evaluations were on average more accurate with $WssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ (0.47) or $WssGBLUP_{pseudo\text{-}SNPs\ (LD)}$ (0.47) than with $WssGBLUP_{SNP}$ (0.46). $WssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ and $WssGBLUP_{pseudo\text{-}SNPs\ (LD)}$ were both more accurate than $WssGBLUP_{SNP}$ for FY, TA, UFP, FU and SCS (+1 percent point to +7 percent points). $WssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ were slightly more accurate than $WssGBLUP_{SNP}$ for MY, RUA (+3 to +4 percent points) and $WssGBLUP_{pseudo\text{-}SNPs\ (LD)}$ were more accurate than $WssGBLUP_{SNP}$ for PY (+1 percent point). For other traits, accuracies for $WssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ and $WssGBLUP_{pseudo\text{-}SNPs\ (LD)}$ were equal or lower to accuracies with $WssGBLUP_{SNP}$. For Saanen, results are more method-dependent. The $ssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ slightly outperformed $ssGBLUP_{SNP}$ for MY (+1 percent point), FY (+1 percent point), FC (+1 percent point), RUA (+1 percent point), FU (+2 percent point), US (+1 percent point) and SCS (+6 percent point). On the contrary, $ssGBLUP_{pseudo\text{-}SNPs\ (LD)}$ were as accurate as or less accurate than $ssGBLUP_{SNP}$ for all the traits (0 percent point to – 1 percent point) except for FU (+1 percent point) and SCS (+4 percent points).

DISCUSSION

The number of haplotypes was the same between Alpine and Saanen with the DW method, however, the number of pseudo-SNPs was higher in the Alpine breed for haplotypes longer than 10 SNPs. Inbreeding in Saanen (2.8%) is higher than inbreeding in Alpine (1.8%)

(Carillier et al., 2014). It could result in a higher genetic variability for Alpine and a higher variability of haplotype alleles, especially for long haplotypes. The difference of number of pseudo-SNPs was not present for LD method. (Carillier et al., 2013) have shown that LD is equal to 0.17 in Alpine and Saanen breed for SNPs spaced 50 kb apart (average distance between 2 SNPs on the chip). The LD in French dairy goats is smaller than LD observed in dairy cattle (between 0.18 and 0.23) (de Roos et al., 2008). The LD between SNPs in Alpine and Saanen were not high enough to create long haplotypes. In our study, haplotypes were mainly shorter than 10 SNPs. As a result, the number of alleles did not differ so much between Alpine and Saanen for this method. Karimi et al., (2018) used pseudo-SNPs into genomic evaluations. They used 21,236 Holstein phenotyped for 57 traits (official domestic and MACE proof), traits with a wide range of heritabilites (going from 0.003 to 0.529). They classified traits according to their heritabilities into 3 classes: low (0-0.15), medium (0.15-0.30) and high (> 0.30). Animals were also genotyped with the Illumina BovineSNP50™ Bead Chip (Illumina Inc, San Diego, USA). They performed GBLUP with individual SNP as presented by VanRaden, (2008) and with pseudo-SNPs ($GBLUP_{pseudo\text{-}SNPs}$). The haplotypes were constructed with the DW method using sizes of 5, 10, 15 and 20 SNPs. In comparison with our study, they observed 75,263 pseudo-SNPs for haplotypes of 5 SNPs, it decreased to 37,270 pseudo-SNPs for haplotypes of 20 SNPs, roughly half of the number of pseudo-SNPs observed in French dairy goats. These results showed that French dairy goats present a high genetic diversity.

The haplotypes constructed with DW method lead to a very high number of pseudo-SNPs. However, many of them were removed from the analyses because they were rare alleles. With LD method, a large proportion of haplotypes was constituted of individual SNP because LD in French dairy goats were not high enough to merge SNPs into hapbloblocks. An alternative would be to create fixed-length haplotypes as described by Hess et al., (2017). They used 58,000 New-Zealand dairy cattle genotyped with the Illumina BovineSNP50™ Bead Chip

375 (Illumina Inc, San Diego, USA). In this study, haplotypes with the same length (in kb) were
376 created with sizes of 125 kb, 250 kb, 500 kb, 1 Mb and 2 Mb. They observed improvement of
377 accuracy for milk yield, fat yield, live weight and SCS with haplotypes of length 250 kb
378 compared to genomic evaluations with individual SNP. This method could be useful to create
379 haplotypes of different lengths as in LD method but limiting the number of haplotypes
380 constituted of individual SNP. If the size of the window is limited, it could also limit the
381 presence of many rare alleles as we observed with the DW method.

382 In French dairy goats, previous GWAS studies using linkage analyses and linkage
383 disequilibrium) (Martin et al., 2017, 2018) or WssGBLUP$_{SNP}$ (Teissier et al., 2018a) were
384 performed. Main QTLs were discovered for yields traits, UFP, RUA and SCS on
385 chromosome 19 in Saanen breed. Other chromosomal regions of interest are located on
386 chromosome 6 for protein content with the α_{S1} casein gene (Grosclaude et al., 1987) and on
387 chromosome 14 for fat content with the DGAT1 gene (Martin et al., 2017) in both breeds
388 (Alpine and Saanen). In this study, the use of pseudo-SNPs into WssGBLUP$_{pseudo\text{-}SNPs\ (DW)}$ and
389 WssGBLUP$_{pseudo\text{-}SNPs\ (LD)}$ were able to detect these regions. Nevertheless, the WssGBLUP$_{pseudo\text{-}}$
390 $_{SNPs\ (DW)}$ and WssGBLUP$_{pseudo\text{-}SNPs\ (LD)}$ did not detect the DGAT1 gene for FY. This phenomenon
391 was also observed in Teissier et al., (2018a) with WssGBLUP$_{SNP}$. This suggests that pseudo-
392 SNPs were not able to capture the effect of DGAT1 gene and further development could be
393 made. The use of pseudo-SNPs allowed the detection of new chromosomal regions of interest.
394 In particular, high pseudo-SNP weights were identified for FU in Saanen breed on
395 chromosomes 13 and 19.

396 Some studies reported improvement of accuracy with the use of haplotypes in genomic
397 evaluations (Calus et al., 2008; Cuyabano et al., 2014; Jónás et al., 2016). In these studies,
398 haplotypes were not converted into pseudo-SNPs. The only use of pseudo-SNPs in genomic
399 evaluations were performed by Karimi et al., (2018). They compared GBLUP$_{SNP}$ and

400 $GBLUP_{pseudo\text{-}SNPs}$ with haplotypes constructed with the DW for 5, 10, 15 and 20 SNPs. They
401 reported similar accuracies between $GBLUP_{SNP}$ and $GBLUP_{pseudo\text{-}SNPs}$ whatever the size of
402 haplotypes for traits with low heritabilities (on average 0.20) and medium heritabilities (on
403 average 0.35). They observed improvement of accuracies for traits with high heritabilities and
404 with $GBLUP_{pseudo\text{-}SNPs}$ with 5 SNPs (0.50) compared to $GBLUP_{SNP}$ (0.49). Accuracies with
405 $GBLUP_{SNP}$ (0.49) and $GBLUP_{pseudo\text{-}SNPs}$ with 10 (0.49), 15 (0.48) and 20 (0.47) SNPs were
406 lower or similar. In our study, accuracies of $ssGBLUP_{pseudo\text{-}SNPs}$ were breed and trait-specific.
407 Higher gain of accuracies were observed for yield traits and small gain for udder type traits in
408 Alpine with the use of $ssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ or $ssGBLUP_{pseudo\text{-}SNPs\ (LD)}$ compared to the
409 $ssGBLUP_{SNP}$. In Saanen breed, $ssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ was the best interesting method for
410 SCS.

411 Our study is the first attempt to fit pseudo-SNPs into WssGBLUP method. In Saanen,
412 $WssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ or $WssGBLUP_{pseudo\text{-}SNPs\ (LD)}$ slightly outperformed or were as
413 accurate as $WssGBLUP_{SNP}$ for yields traits, content traits, UFP, RUA and SCS. Those traits
414 have a QTL/major gene identified on several chromosomes (Martin et al., 2017, 2018). They
415 were also able to outperformed $ssGBLUP_{SNP}$ showing the interest of using pseudo-SNPs in
416 genomic evaluations. In Alpine, the use of $WssGBLUP_{pseudo\text{-}SNPs\ (DW)}$ or $WssGBLUP_{pseudo\text{-}SNPs}$
417 $_{(LD)}$ was the most interesting for FY where it outperformed both $ssGBLUP_{SNP}$ and
418 $WssGBLUP_{SNP}$. For other traits, haplotypes were not so interesting in terms of accuracy with
419 mainly lower accuracy than $ssGBLUP_{SNP}$.

420 Genomic evaluations with haplotypes require several additional steps compared to
421 genomic evaluations with individual SNP:(i) phasing the genotypes, (ii) defining the
422 haplotypes, (iii) converting into pseudo-SNPs and (iv) filtering according their frequency.
423 These steps make genomic evaluations with pseudo-SNPs longer and more time-consuming

than genomic evaluations with individual SNP. With the LD method, it is also necessary to estimate LD between SNPs.

CONCLUSION

We compared accuracy of genomic evaluations with ssGBLUP and WssGBLUP with the use of individual SNP and pseudo-SNPs in French dairy goats for 11 routinely selected traits. Genomic evaluations with pseudo-SNPs improved accuracy of genomic evaluations for some traits. With ssGBLUP using pseudo-SNPs, improvements up to +19% (for fat yield) and +4% (for udder shape) in Alpine and Saanen were observed compared to ssGBLUP with individual SNP. The use of the WssGBLUP method with pseudo-SNPs has made it possible to identify QTLs already known in French dairy goats but also to detect new regions of interest. Accuracy of genomic evaluations was improved by a maximum of 22% (for fat yield) and 21% (for somatic cell scores) in Alpine and Saanen between WssGBLUP with individual SNP and WssGBLUP with pseudo-SNPs. Alpine breed seemed more sensitive to genomic evaluations with pseudo-SNPs maybe because Alpine breed present a higher genetic diversity than Saanen breed. However, the construction of haplotypes with pseudo-SNPs required several steps and was more time-consuming methods.

ACKNOWLEDGMENTS

This study would not have been possible without the goat SNP50 BeadChip developed by the International Goat Genome Consortium (IGGC): www.goatgenome.org. The authors thank Ignacy Misztal (University of Georgia, USA) for the blup90iod2 program. The authors thank the French Genovicap and Phenofinlait programmes (ANR, Apis-Gène, CASDAR, FranceAgriMer, France Génétique Elevage, the French Ministry of Agriculture Agrifood, and Forestry) and the European 3SR project, which funded part of this work. The first author also received financial support from the Occitanie region and the French National Institute for Agricultural Research (INRA) SELGEN programme (INCoMINGS).

REFERENCES

Broman, K.W., and J.L. Weber. 1999. Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain. The American Journal of Human Genetics 65:1493–1500. doi:10.1086/302661.

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp. 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178:553–561. doi:10.1534/genetics.107.080838.

Carillier, C., H. Larroque, I. Palhière, V. Clément, R. Rupp, and C. Robert-Granié. 2013. A first step toward genomic selection in the multi-breed French dairy goat population. J. Dairy Sci. 96:7294–7305. doi:10.3168/jds.2013-6789.

Carillier, C., H. Larroque, and C. Robert-Granié. 2014. Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. Genetics Selection Evolution 46:67. doi:10.1186/s12711-014-0067-3.

Curtis, D., B.V. North, and P.C. Sham. 2001. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. Annals of Human Genetics 65:95–107.

Cuyabano, B.C., G. Su, and M.S. Lund. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1171. doi:10.1186/1471-2164-15-1171.

Ferdosi, M.H., J. Henshall, and B. Tier. 2016. Study of the optimum haplotype length to build genomic relationship matrices. Genetics Selection Evolution 48. doi:10.1186/s12711-016-0253-6.

Grosclaude, F., M.-F. Mahé, G. Brignon, L. Di Stasio, and R. Jeunet. 1987. A Mendelian polymorphism underlying quantitative variations of goat αs1-casein. Genetics Selection Evolution 19:399–412. doi:10.1186/1297-9686-19-4-399.

Hess, M., T. Druet, A. Hess, and D. Garrick. 2017. Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genetics Selection Evolution 49:54. doi:10.1186/s12711-017-0329-y.

Hickey, J.M., B.P. Kinghorn, B. Tier, S.A. Clark, J.H.J. van der Werf, and G. Gorjanc. 2013. Genomic evaluations using similarity between haplotypes. Journal of Animal Breeding and Genetics 130:259–269. doi:10.1111/jbg.12020.

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. doi:10.3168/jds.2015-10433.

Karimi, Z., M. Sargolzaei, J.A.B. Robinson, and F.S. Schenkel. 2018. Assessing haplotype-based models for genomic evaluation in Holstein cattle. Can. J. Anim. Sci.. doi:10.1139/CJAS-2018-0009.

Larroque, H., J.-M. Astruc, A. Barbat, F. Barillet, D. Boichard, B. Bonaiti, B. Bonaiti, V. Clément, I. David, G. Lagriffoul, I. Palhière, A. Piacère, C. Robert-Granié, and R. Rupp. 2011.

National genetic evaluations in dairy sheep and goats in France. Page 62. Annual Meeting of the European Federation of Animal Science (EAAP). 2011-08-29, Stavanger, NOR. Wageningen Academic.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92:4656–4663. doi:10.3168/jds.2009-2061.

Martin, P., I. Palhière, C. Maroteau, P. Bardou, K. Canale-Tabet, J. Sarry, F. Woloszyn, J. Bertrand-Michel, I. Racke, H. Besir, R. Rupp, and G. Tosser-Klopp. 2017. A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content. Scientific Reports 7. doi:10.1038/s41598-017-02052-0.

Martin, P., I. Palhière, C. Maroteau, V. Clément, I. David, G.T. Klopp, and R. Rupp. 2018. Genome-wide association mapping for type and mammary health traits in French dairy goats identifies a pleiotropic region on chromosome 19 in the Saanen breed. Journal of Dairy Science 101:5214–5226. doi:10.3168/jds.2017-13625.

Meuwissen, T.H., J. Odegard, I. Andersen-Ranberg, and E. Grindflek. 2014. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. Genetics Selection Evolution 46:49. doi:10.1186/1297-9686-46-49.

Misztal, I., S. Tsuruta, D.A.L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z.G. Vitezica. 2016. Manual for BLUPF90 Family of Programs. Athens: University of Georgia.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575. doi:10.1086/519795.

Rogers, A.R., and C. Huff. 2009. Linkage Disequilibrium Between Loci With Unknown Phase. Genetics 182:839–844. doi:10.1534/genetics.108.093153.

de Roos, A.P.W., B.J. Hayes, R.J. Spelman, and M.E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179:1503–1512. doi:10.1534/genetics.107.084301.

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. doi:10.1186/1471-2164-15-478.

Teissier, M., H. Larroque, and C. Robert-Granié. 2018. Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene. Genetics Selection Evolution 50:31. doi:10.1186/s12711-018-0400-3.

Tosser-Klopp, G., P. Bardou, O. Bouchez, C. Cabau, R. Crooijmans, Y. Dong, C. Donnadieu-Tonon, A. Eggen, H.C.M. Heuven, S. Jamli, A.J. Jiken, C. Klopp, C.T. Lawley, J. McEwan, P. Martin, C.R. Moreno, P. Mulsant, I. Nabihoudine, E. Pailhoux, I. Palhière, R. Rupp, J. Sarry, B.L. Sayre, A. Tircazes, Jun Wang, W. Wang, W. Zhang, and and the International Goat Genome Consortium. 2014. Design and Characterization of a 52K SNP Chip for Goats. PLoS ONE 9:e86227. doi:10.1371/journal.pone.0086227.

528 Uemoto, Y., S. Sato, T. Kikuchi, S. Egawa, K. Kohira, H. Sakuma, S. Miyashita, S. Arata, T.
529 Kojima, and K. Suzuki. 2017. Genomic evaluation using SNP- and haplotype-based genomic
530 relationship matrices in a closed line of Duroc pigs. Animal Science Journal 88:1465–1474.
531 doi:10.1111/asj.12805.

532 VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy
533 Science 91:4414–4423. doi:10.3168/jds.2007-0980.

534 VanRaden, P.M., and G.R. Wiggans. 1991. Derivation, Calculation, and Use of National
535 Animal Model Information. Journal of Dairy Science 74:2737–2746. doi:10.3168/jds.S0022-
536 0302(91)78453-1.

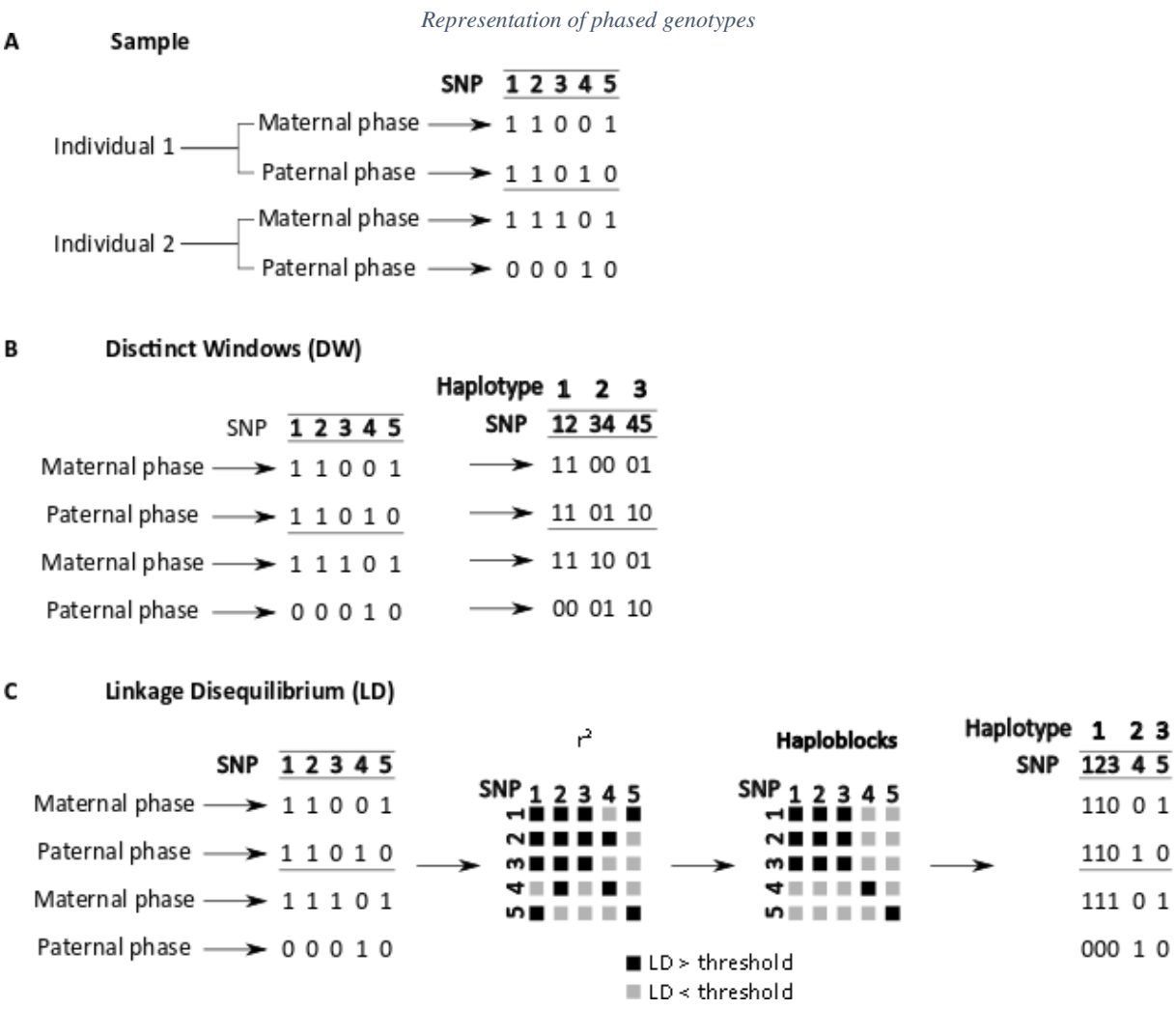
537 Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2012. Genome-wide association
538 mapping including phenotypes from relatives without genotypes. Genetics Research (Camb)
539 94:73–83. doi:10.1017/S0016672312000274.

540 Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting Strategies for
541 Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and
542 GWAS. Frontiers in Genetics 7:151. doi:10.3389/fgene.2016.00151.

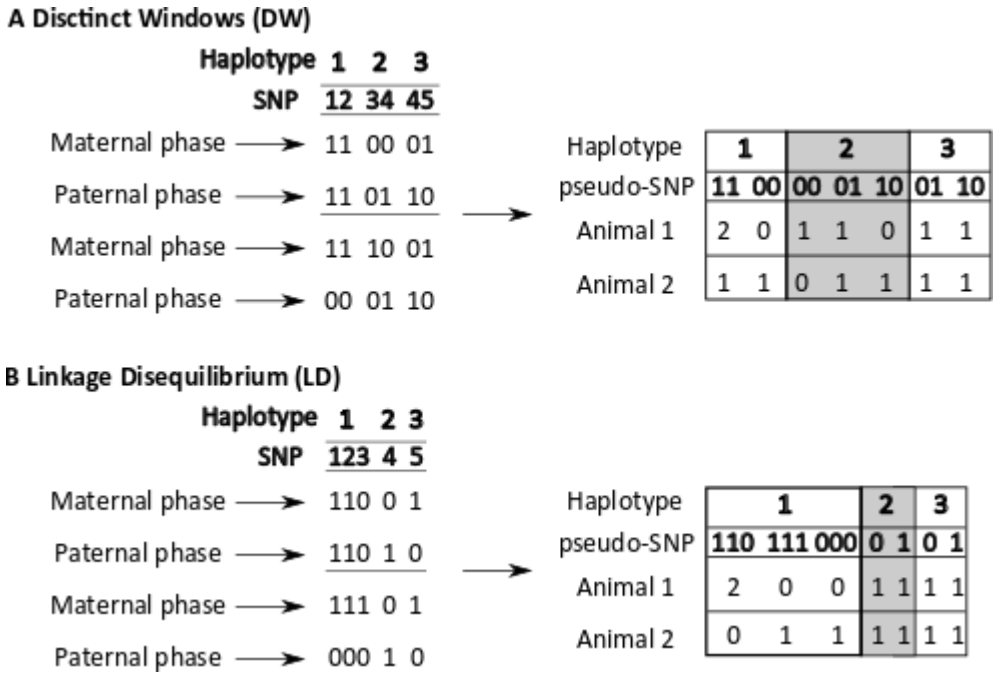
543

FIGURES

Figure 1. Construction of haplotypes using the Distinct Windows (DW) or Linkage Disequilibrium (LD) methods. Previously, genotypes need to be phased (A). With DW (B), the size of the window is required to create haplotypes (here 2 SNPs). With LD (C), it is required to estimate LD between SNP before the construction of the haplotypes.



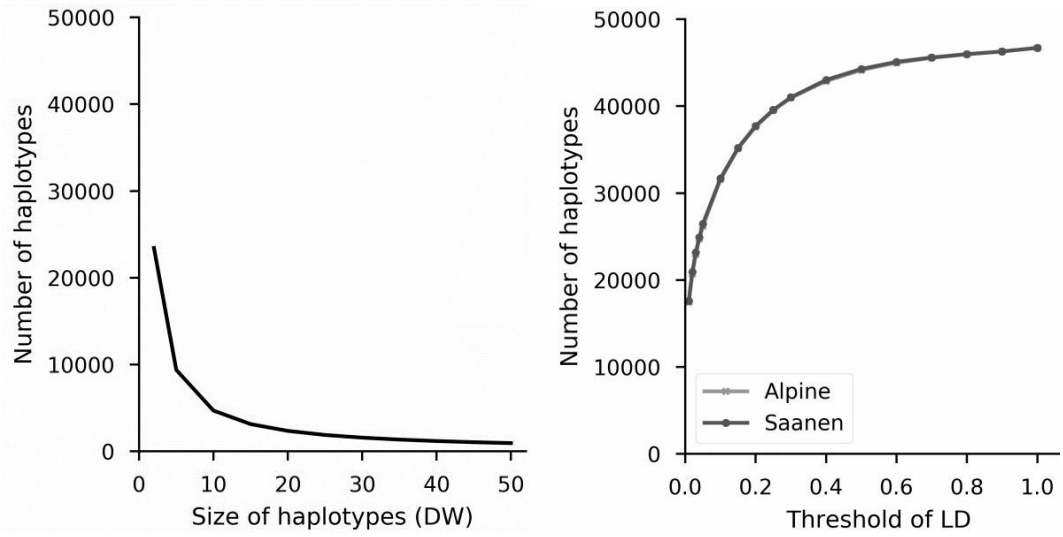
552 *Figure 2. Construction of pseudo-SNPs from haplotypes with DW (A) and LD (B) methods. Pseudo-SNPs are constructed as*
553 *the number of copy of each allele of a haplotype. The number of pseudo-SNPs for one haplotype is equal to the number of*
554 *alleles for this haplotype.*



555

556

557 *Figure 3. Number of haplotypes according to the size of haplotype with the Distinct Windows (DW) method or according the*
558 *threshold of linkage disequilibrium (LD) method in Alpine and Saanen breeds*



559

Figure 4. Number of pseudo-SNPs used in genomic evaluations after filtering on frequency (alleles with a frequency > 0.01 remained) with the DW and LD methods in Alpine and Saanen breeds

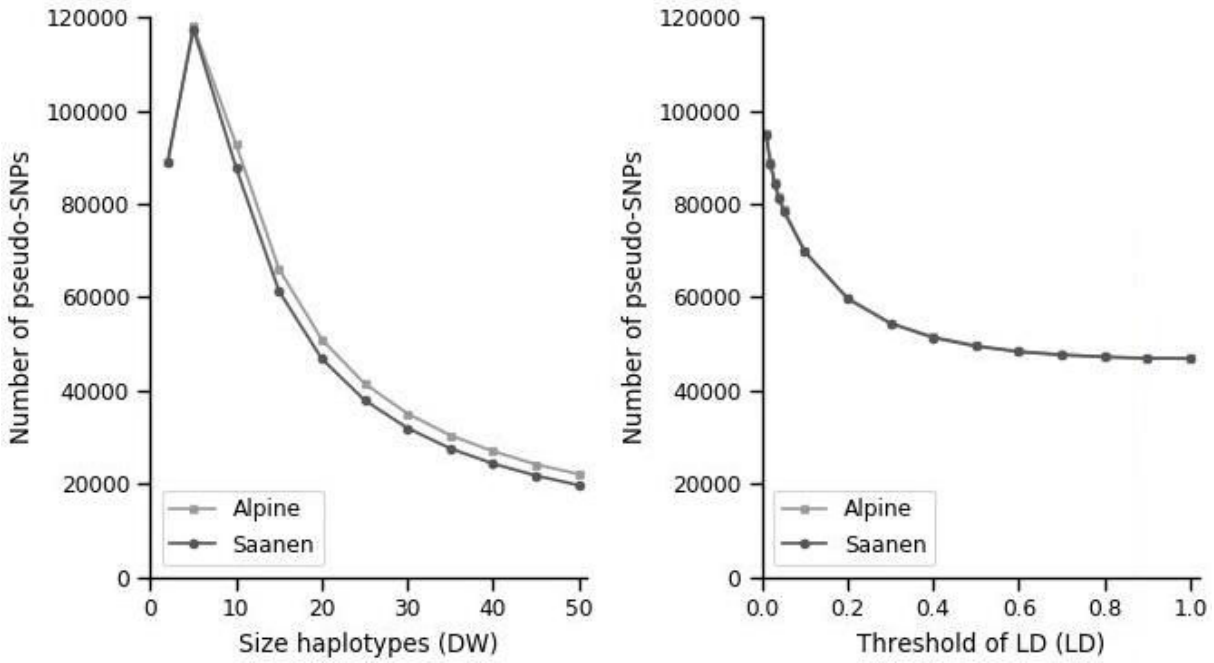


Figure 5. Pseudo-SNPs weights estimation for protein content in Alpine and Saanen for chromosome 6 with the DW method according to the size of haplotypes

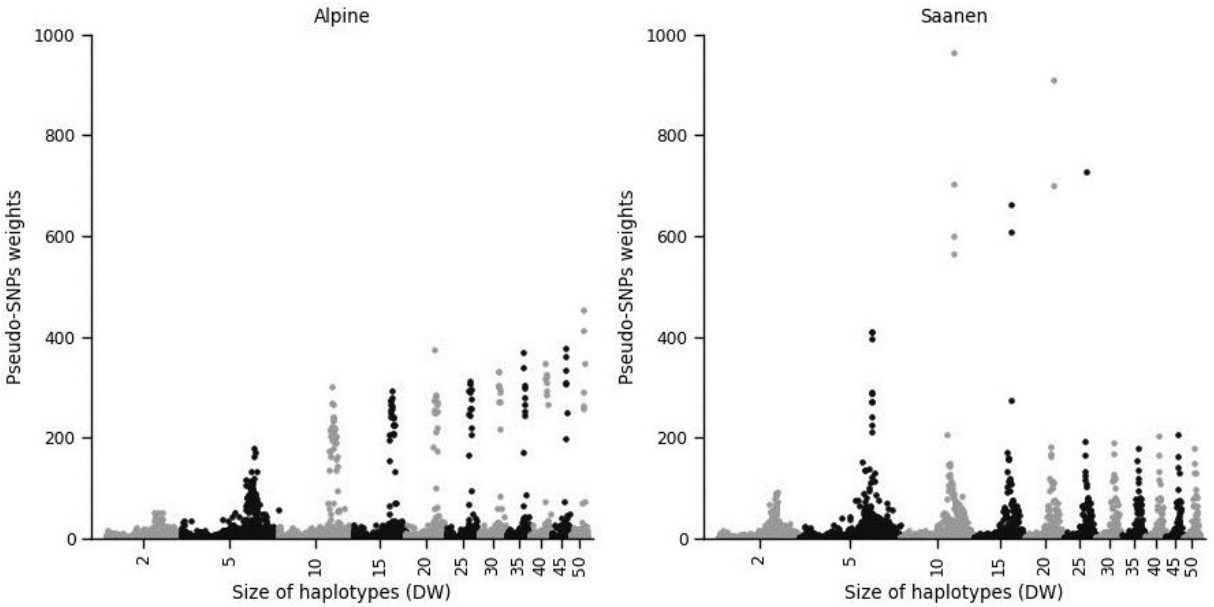
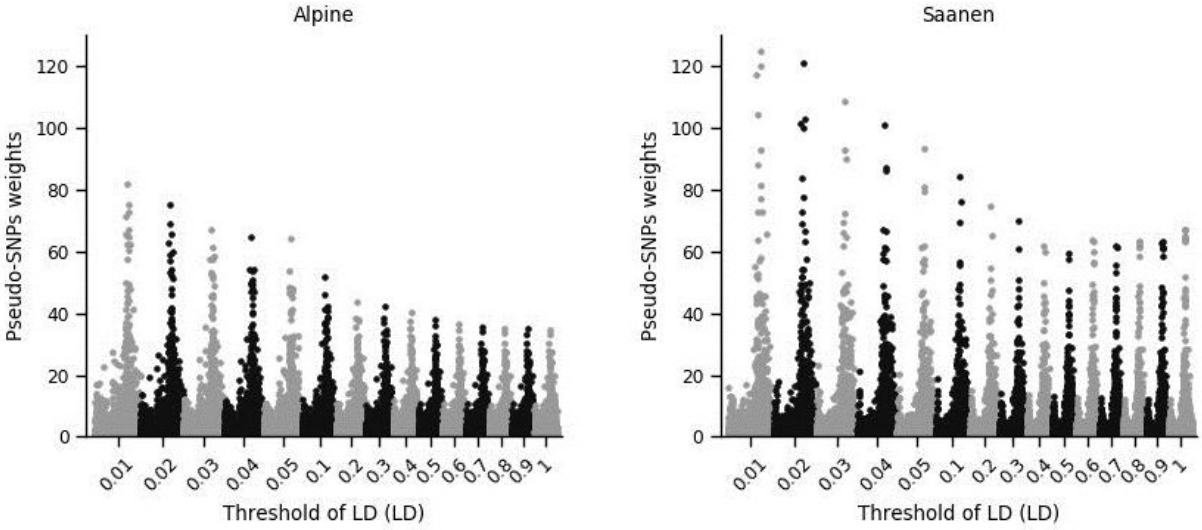


Figure 6. Pseudo-SNPs weights estimation for chromosome 6 for protein content in Alpine and Saanen with the DW method (each size of haplotypes tested is represented)



TABLES

Table 1. Descriptive statistics (number of performances, mean) and heritabilities (h^2) of the 11 traits under selection in French Alpine and Saanen dairy goats

	Alpine			Saanen		
	Number of performances	Mean	h^2	Number of performances	Mean	h^2
Milk yield	3,844,314	802.12	0.31	2,923,531	823.08	0.26
Fat yield	3,742,129	28.4	0.28	2,887,051	27.44	0.25
Protein yield	3,844,071	24.36	0.31	2,923,419	24.32	0.25
Fat content	3,742,129	35.33	0.48	2,887,051	33.39	0.51
Protein content	3,844,071	30.42	0.60	2,923,419	29.68	0.56
Teat angle	150,676	3.63	0.42	102,967	4.05	0.45
Udder floor position	150,676	6.37	0.51	102,967	6.16	0.57
Rear udder attachment	150,676	4.57	0.47	102,967	4.96	0.52
Fore udder	150,676	3.19	0.44	102,967	3.38	0.42
Udder shape	150,676	5.76	0.40	102,967	6.22	0.47
Somatic cell scores	1,262,187	165.03	0.20	1,031,450	158.99	0.16

574 Table 2. Accuracies of genomic evaluation* ssGBLUP based on individual SNP and pseudo-SNPs with DW and LD methods
575 in Alpine breed.

	Alpine				
Trait	**SNP[1]**	**DW[2]**	**LD[3]**	**Window with the best accuracy (DW)**	**Threshold with the best accuracy (LD)**
Milk yield	0.45	0.47	0.46	40	0.01
Fat yield	0.31	0.37	0.33	40	0.02
Protein yield	0.30	0.32	0.30	40	0.30
Fat content	0.66	0.67	0.67	50	0.01
Protein content	0.76	0.76	0.76	2	0.05
Teat angle	0.42	0.43	0.43	2	0.02
Udder floor position	0.43	0.44	0.44	2	0.01
Rear udder attachment	0.40	0.40	0.40	35	0.02
Fore udder	0.49	0.50	0.50	2	0.10
Udder shape	0.48	0.49	0.50	25	0.01
Somatic cell scores	0.45	0.46	0.46	5	0.01

576 *Only scenario with the best accuracies are presented
577 [1] Genomic evaluations with ssGBLUP based on individual SNP ($ssGBLUP_{SNP}$)
578 [2] Genomic evaluations with ssGBLUP based on pseudo-SNPs constructed with the DW method
579 ($ssGBLUP_{pseudo\text{-}SNPs\ (DW)}$)
580 [3] Genomic evaluations with ssGBLUP based on pseudo-SNPs constructed with the LD method
581 ($ssGBLUP_{pseudo\text{-}SNPs\ (LD)}$)

582 *Table 3. Accuracies of genomic evaluation* ssGBLUP based on individual SNP and pseudo-SNPs with DW and LD methods*

583 *in Saanen breed.*

	Saanen				
Trait	**SNP[1]**	**DW[2]**	**LD[3]**	**Window with the best accuracy (DW)**	**Threshold with the best accuracy (LD)**
Milk yield	0.49	0.49	0.49	5	1
Fat yield	0.43	0.44	0.43	5	1
Protein yield	0.45	0.46	0.45	20	0.10
Fat content	0.59	0.59	0.59	5	0.02
Protein content	0.73	0.73	0.73	5	1
Teat angle	0.48	0.49	0.49	15	0.01
Udder floor position	0.59	0.59	0.59	10	1
Rear udder attachment	0.60	0.60	0.60	10	1
Fore udder	0.62	0.62	0.62	2	0.10
Udder shape	0.36	0.37	0.36	5	0.02
Somatic cell scores	0.46	0.49	0.46	15	0.02

584 *Only scenario with the best accuracies are presented

585 [1] Genomic evaluations with ssGBLUP based on individual SNP ($ssGBLUP_{SNP}$)

586 [2] Genomic evaluations with ssGBLUP based on pseudo-SNPs constructed with the DW method

587 ($ssGBLUP_{pseudo\text{-}SNPs\ (DW)}$)

588 [3] Genomic evaluations with ssGBLUP based on pseudo-SNPs constructed with the LD method

589 ($ssGBLUP_{pseudo\text{-}SNPs\ (LD)}$)

590 *Table 4. Accuracies of genomic evaluation WssGBLUP with individual SNP and pseudo-SNPs (DW and LD methods) in Alpine*
591 *and Saanen breeds. Best accuracies obtained with the ssGBLUP scenario are presented (see Table 2 and Table 3 for Alpine*
592 *and Saanen respectively)*

	Alpine			Saanen		
Trait	**SNP[1]**	**DW[2]**	**LD[3]**	**SNP[1]**	**DW[2]**	**LD[3]**
Milk yield	0.43	0.46	0.43	0.56	0.57	0.56
Fat yield	0.30	0.37	0.33	0.48	0.49	0.48
Protein yield	0.28	0.28	0.29	0.50	0.50	0.49
Fat content	0.65	0.60	0.65	0.59	0.60	0.59
Protein content	0.77	0.77	0.77	0.77	0.75	0.77
Teat angle	0.41	0.43	0.42	0.47	0.45	0.47
Udder floor position	0.41	0.44	0.44	0.63	0.61	0.63
Rear udder attachment	0.38	0.42	0.38	0.61	0.62	0.61
Fore udder	0.48	0.49	0.49	0.59	0.61	0.60
Udder shape	0.48	0.44	0.48	0.34	0.35	0.32
Somatic cell scores	0.42	0.45	0.46	0.43	0.49	0.47

593 *Scenario presented for Alpine and Saanen for WssGBLUP is the same than in Table 2 and Table
594 3
595 [1] Genomic evaluations with WssGBLUP based on individual SNP ($WssGBLUP_{SNP}$)
596 [2] Genomic evaluations with WssGBLUP based on pseudo-SNPs constructed with the DW
597 method ($WssGBLUP_{pseudo\text{-}SNPs\ (DW)}$)
598 [3] Genomic evaluations with WssGBLUP based on pseudo-SNPs constructed with the LD method
599 ($WssGBLUP_{pseudo\text{-}SNPs\ (LD)}$)

3. Corrélations entre les éléments de **G** et de $\mathbf{A_{22}}$ dans les évaluations génomiques haplotypiques (ou pseudo-SNPs)

Les Figure 60 à 62 présentent l'évolution des corrélations entre les éléments hors-diagonaux de la matrice **G** et $\mathbf{A_{22}}$ en fonction de la taille des haplotypes ou du seuil de LD pour les analyses multirace, Alpine et Saanen respectivement. Ces corrélations ont été estimées pour les évaluations ssGBLUP basées sur l'information individuelle des SNPs ($ssGBLUP_{SNP}$), les évaluations génomiques haplotypiques ($ssGBLUP_{Ferdosi}$) et pour les évaluations génomiques ssGBLUP ou WssGBLUP basées sur l'information des pseudo-SNPs ($ssGBLUP_{pseudo\text{-}SNPs}$ et $WssGBLUP_{pseudo\text{-}SNPs}$). Pour les analyses $ssGBLUP_{SNP}$, $ssGBLUP_{Ferdosi}$ et $ssGBLUP_{pseudo\text{-}SNPs}$, les matrices **G** sont identiques d'un caractère à l'autre. En revanche, pour le $WssGBLUP_{pseudo\text{-}SNPs}$, la matrice **G** est spécifique à chaque caractère. Les corrélations moyennes ainsi que les écarts-types sont présentés dans les Figure 60 à 62. L'objectif de ces analyses est de voir à quel point les différentes méthodes d'évaluations génomiques considérées affectent la matrice de parenté génomique par rapport à la matrice de parenté **A** basée sur le pedigree.

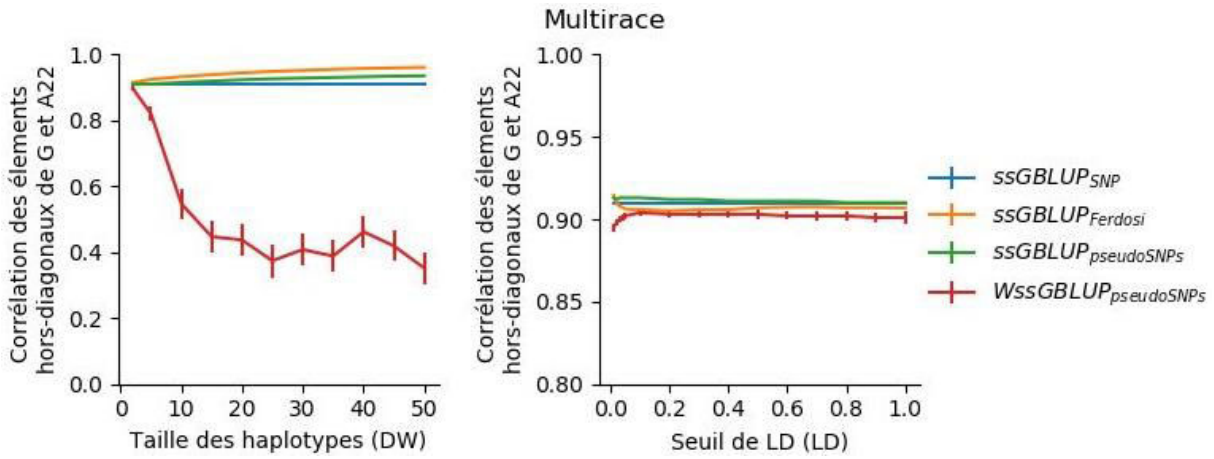


*Figure 60. Evolution des corrélations entre les éléments hors-diagonaux de **G** et de $\mathbf{A_{22}}$ des évaluations multirace $ssGBLUP_{SNP}$, $ssGBLUP_{Ferdosi}$, $ssGBLUP_{pseudo\text{-}SNPs}$ et $WssGBLUP_{pseudo\text{-}SNP}$ en fonction de la taille des haplotypes (DW) et des seuils de LD (LD)*

La Figure 60 présente les résultats des analyses multirace. Pour la méthode DW, les corrélations entre les éléments de **G** et de $\mathbf{A_{22}}$ sont presque identiques, quel que soit la taille des haplotypes pour les évaluations $ssGBLUP_{SNP}$ (moyenne de 0,91), $ssGBLUP_{Ferdosi}$ (moyenne de 0,92) et $ssGBLUP_{pseudo\text{-}SNPs}$ (moyenne de 0,92). Pour les évaluations $WssGBLUP_{pseudo\text{-}SNPs}$, la corrélation moyenne est de 0,73. Cependant, plus la taille des haplotypes est grande et plus cette corrélation est faible, passant de 0,90 (2 SNPs) à 0,35 (50 SNPs). Pour la méthode LD, peu de variation existe entre les corrélations des éléments de **G** et $\mathbf{A_{22}}$. Ces corrélations sont en moyenne identiques pour les évaluations $ssGBLUP_{SNP}$ (0,91), $ssGBLUP_{Ferdosi}$ (0,91), $ssGBLUP_{pseudo\text{-}SNPs}$ (0,92) et $WssGBLUP_{pseudo\text{-}SNPs}$ (0,91).

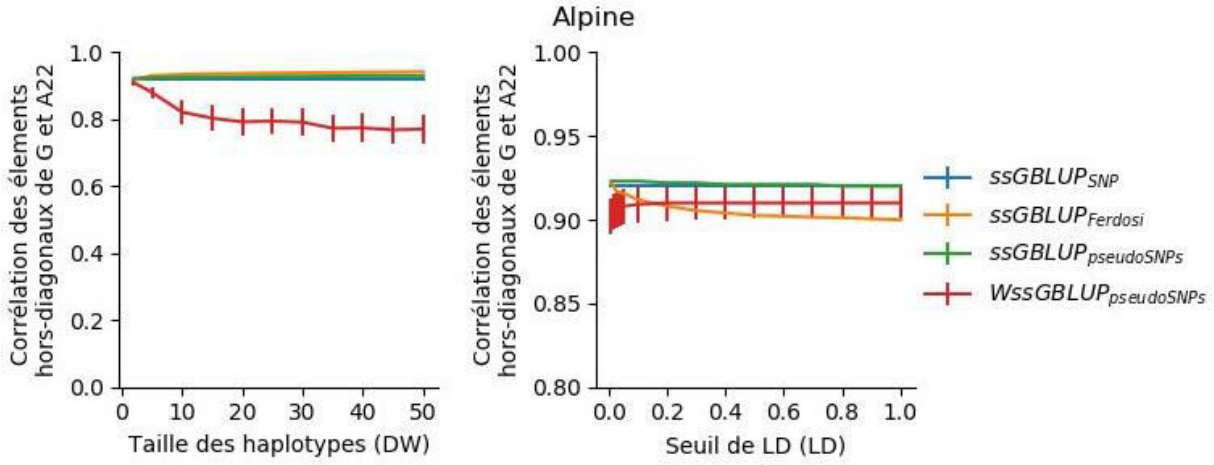


*Figure 61. Evolution des corrélations entre les éléments hors-diagonaux de **G** et de $\mathbf{A_{22}}$ des évaluations Alpine $ssGBLUP_{SNP}$, $ssGBLUP_{Ferdosi}$, $ssGBLUP_{pseudo\text{-}SNPs}$ et $WssGBLUP_{pseudo\text{-}SNP}$ en fonction de la taille des haplotypes (DW) et des seuils de LD (LD)*

La Figure 61 présente les corrélations entre les éléments de **G** et de $\mathbf{A_{22}}$ pour les analyses Alpine selon la méthode DW ou LD. On observe les mêmes tendances que pour les analyses multirace. La baisse des corrélations pour les évaluations $WssGBLUP_{pseudo\text{-}SNP}$ avec la méthode DW est moins prononcée que pour les évaluations multirace. Les corrélations chutent de 0,91 (2 SNPs) à 0,77 (50 SNPs). Les corrélations restent élevées pour les évaluations $ssGBLUP_{SNP}$ (0,92 en moyenne), $ssGBLUP_{Ferdosi}$ (0,94 en moyenne) et $ssGBLUP_{pseudo\text{-}SNPs}$ (0,93 en moyenne). Pour la méthode LD, les corrélations entre $ssGBLUP_{SNP}$ et $ssGBLUP_{pseudo\text{-}SNPs}$ sont très proches (0,92 en moyenne). Pour le WssGBLUP avec la méthode LD, les corrélations sont en moyenne de 0,91. En revanche, pour le $ssGBLUP_{Ferdosi}$, les précisions chutent de 0,93 pour un LD de 0,01 à 0,90 pour un LD de 1.

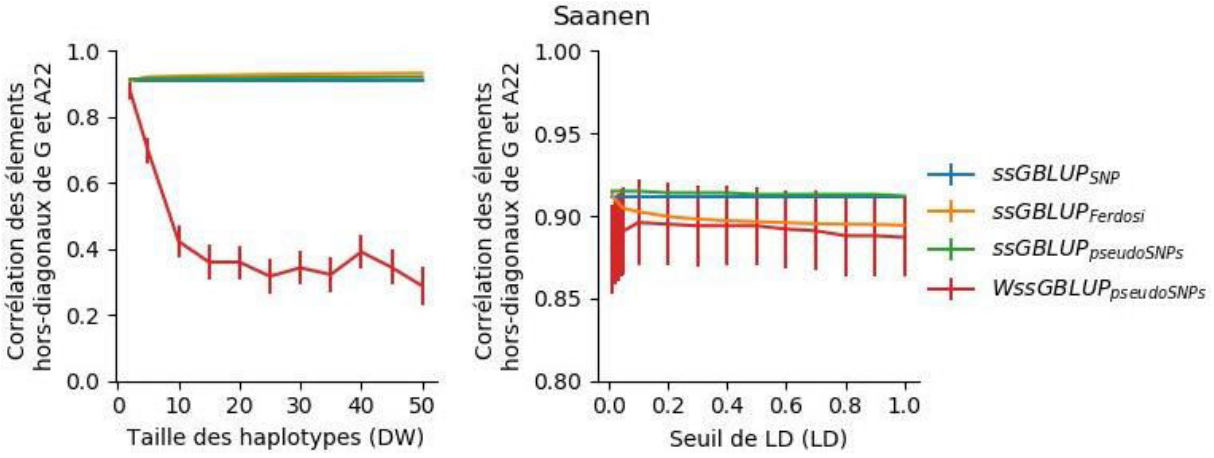


*Figure 62 Evolution des corrélations entre les éléments hors-diagonaux de **G** et de $\mathbf{A_{22}}$ des évaluations Saanen $ssGBLUP_{SNP}$, $ssGBLUP_{Ferdosi}$, $ssGBLUP_{pseudo\text{-}SNPs}$ et $WssGBLUP_{pseudo\text{-}SNP}$ en fonction de la taille des haplotypes (DW) et des seuils de LD (LD)*

Pour les analyses Saanen (Figure 62), on observe les mêmes tendances que pour les analyses multirace pour la méthode DW. La baisse des corrélations est forte pour les évaluations $WssGBLUP_{pseudo\text{-}SNP}$ avec des corrélations chutant de 0,88 (2 SNPs) à 0,29 (50 SNPs). Pour les autres approches, ces corrélations sont en moyenne de 0,91 ($ssGBLUP_{SNP}$), de 0,92 ($ssGBLUP_{Ferdosi}$), de 0,91 ($ssGBLUP_{pseudo\text{-}SNPs}$) et de 0,50 ($WssGBLUP_{pseudo\text{-}SNPs}$). Pour la méthode LD, les précisions pour le $ssGBLUP_{SNP}$ et le $ssGBLUP_{pseudo\text{-}SNPs}$ sont similaires (en

moyenne de 0,91). Ces corrélations sont en moyennes légèrement plus faibles pour le ssGBLUP$_{Ferdosi}$ (0,90) et le WssGBLUP$_{pseudo-SNPs}$ (0,89). On s'aperçoit que les matrices de parenté génomiques construites avec l'ensemble de ces méthodes (ssGBLUP$_{SNP}$, ssGBLUP$_{Ferdosi}$, ssGBLUP$_{pseudo-SNPs}$) sont fortement corrélées aux matrices de parenté basées sur le pedigree, que ce soit en multirace, Alpine ou Saanen. Le WssGBLUP$_{pseudo-SNPs}$, en revanche, affecte beaucoup plus la structure de la matrice de parenté génomique. Cet effet est bien plus fort pour la race Saanen. Un QTL est présent pour la majorité des caractères en Saanen, la présence de ces QTL et l'intégration des poids dans les évaluations ont un impact plus important sur la construction de la matrice de parenté génomique.

4. Précisions des évaluations génomiques haplotypiques pour des caractères de productions laitières, de morphologie de la mamelle et de comptage de cellules somatiques pour les races caprines

Les Tableaux 26 à 28 présentent les précisions des évaluations génomiques haplotypiques multirace, Alpine et Saanen respectivement selon la méthode de Ferdosi et al., (2016) (décrite au paragraphe 2.2, Chapitre 2) pour 5 caractères de production laitière (Lait, MG,MP, TB et TP), les CCS et 5 caractères de morphologie de la mamelle (AVP, PLA, PRM, AAR, ORT et ORT). Seuls les scénarios selon les méthodes DW et LD pour lesquels nous avons obtenu les meilleures précisions sont présentés dans les tableaux. Les précisions des évaluations génomiques haplotypiques sont comparées aux précisions des évaluations ssGBLUP$_{SNP}$.

En moyenne, on observe une légère amélioration (entre +1 et +2 points) des précisions des évaluations génomiques avec le ssGBLUP$_{FerdosiDW}$ (0,50) comparé au ssGBLUP$_{SNP}$ (0,49) dans les analyses multirace pour l'ensemble des caractères (Tableau 26), cependant, ces différences ne sont pas significatives. Seuls le TP et l'AAR ont des précisions identiques entre ssGBLUP$_{SNP}$ et ssGBLUP$_{FerdosiDW}$. Les meilleures précisions sont obtenues avec des haplotypes courts (de 2 à 5 SNPs) pour la majorité des caractères (LAIT, MG, MP, TB, TP, PRM et AAR). Pour les autres caractères (AVP, PLA, ORT et CCS), des haplotypes plus longs (10 à 15 SNPs) ont été nécessaires pour obtenir de meilleures précisions.

Le ssGBLUP$_{FerdosiLD}$ n'améliore pas en moyenne les précisions des évaluations génomiques comparées au ssGBLUP$_{SNP}$ (0,49 pour les deux méthodes). Avec le ssGBLUP$_{FerdosiLD}$, les précisions diminuent légèrement de -1 à -2 points par rapport aux précisions du ssGBLUP$_{SNP}$ pour 5 caractères (TB, TP, PLA, AAR, ORT). A l'inverse, les caractères LAIT, MG et MP voient leurs précisions s'améliorer de +1 à +2 points. Enfin, les précisions sont identiques entre ssGBLUP$_{FerdosiLD}$ et ssGBLUP$_{SNP}$ pour les caractères AVP et PRM. Excepté pour le LAIT, les meilleures précisions sont obtenues lorsque l'on fixe un seuil de LD bas (inférieur à 0,05).

Tableau 26. Précisions des évaluations génomiques haplotypiques (Ferdosi et al., 2016) pour les 11 caractères étudiés avec la méthode Distinct Windows (DW) et la méthode basée sur le déséquilibre de liaison (LD) pour les évaluations multirace

	Multirace				
	ssGBLUP$_{SNP}$	**ssGBLUP$_{FerdosiDW}$**	**ssGBLUP$_{FerdosiLD}$**	**Fenêtre avec la meilleure précisions (DW)**	**Seuil avec la meilleure précision (LD)**
LAIT	0,48	0,50	0,50	2	1
MG	0,39	0,40	0,41	2	0,02
MP	0,36	0,38	0,37	2	0,05
TB	0,70	0,71	0,69	5	0,02
TP	0,77	0,77	0,76	5	0,02
AVP	0,52	0,53	0,52	10	0,02
PLA	0,46	0,47	0,44	10	0,02
PRM	0,41	0,42	0,41	5	0,01
AAR	0,48	0,48	0,47	5	0,04
ORT	0,36	0,37	0,35	15	0,02
CCS	0,45	0,47	0,45	15	0,01
Moyenne	0,49	0,50	0,49	-	-

Pour la race Alpine (Tableau 27), les évaluations génomiques haplotypiques sont en moyenne légèrement plus précises pour le ssGBLUP$_{FerdosiDW}$ (0,48) et le ssGBLUP$_{FerdosiLD}$ (0,48) que le ssGBLUP$_{SNP}$ (0,47). Avec le ssGBLUP$_{FerdosiDW}$, les précisions s'améliorent pour tous les caractères (entre +1 et + 3 points) sauf pour le TP (même précision) par rapport au ssGBLUP$_{SNP}$. Les meilleures précisions sont principalement obtenues avec des haplotypes courts (2 SNPs pour le LAIT, la MP, AVP, PLA, AAR et ORT et de 5 SNPs pour le TP). Pour le ssGBLUP$_{FerdosiLD}$, des améliorations de précisions sont également observées par rapport au ssGBLUP$_{SNP}$ pour tous les caractères (entre +1 et +4 points) sauf le TB (+0 points) et le TP (-1 points). Les précisions les plus élevées sont observées pour de faibles valeurs de LD (entre 0,01 et 0,05) sauf pour le LAIT (0,50) et AAR (0,90).

Tableau 27. Précisions des évaluations génomiques haplotypiques (Ferdosi et al., 2016) pour les 11 caractères étudiés avec la méthode Distinct Windows (DW) et la méthode basée sur le déséquilibre de liaison (LD) pour les évaluations Alpine

	Alpine				
	ssGBLUP$_{SNP}$	**ssGBLUP$_{FerdosiDW}$**	**ssGBLUP$_{FerdosiLD}$**	**Fenêtre avec la meilleure précision (DW)**	**Seuil avec la meilleure précision (LD)**
LAIT	0,45	0,48	0,49	2	0,50
MG	0,31	0,34	0,34	40	0,02
MP	0,30	0,31	0,31	2	0,02
TB	0,66	0,68	0,66	45	0,01
TP	0,76	0,76	0,75	5	0,05
AVP	0,49	0,51	0,50	2	0,01
PLA	0,43	0,44	0,45	2	0,01
PRM	0,48	0,50	0,50	45	0,01
AAR	0,40	0,41	0,42	2	0,90
ORT	0,42	0,43	0,43	2	0,01
CCS	0,45	0,47	0,47	15	0,03
Moyenne	0,47	0,48	0,48	-	-

Pour la race Saanen (Tableau 28), les précisions moyennes des évaluations sont de 0,53 avec les méthodes ssGBLUP$_{SNP}$ et ssGBLUP$_{FerdosiDW}$ et de 0,52 avec le ssGBLUP$_{FerdosiLD}$. Les précisions ssGBLUP$_{SNP}$ et ssGBLUP$_{FerdosiDW}$ sont identiques sauf pour PRM (+2 points pour le ssGBLUP$_{FerdosiDW}$), ORT (+1 point pour le ssGBLUP$_{FerdosiDW}$) et CCS (+2 points pour le ssGBLUP$_{FerdosiDW}$). Les caractères LAIT, MG, MP, AVP et PLA ont des précisions maximales avec le ssGBLUP$_{FerdosiDW}$ pour des haplotypes courts (2 à 5 SNPs). Les autres caractères nécessitent des haplotypes plus longs (entre 10 et 35 SNPs). Pour le ssGBLUP$_{FerdosiLD}$, tous les caractères ont des précisions plus faibles que le ssGBLUP$_{SNP}$ (entre -1 et -2 points) exceptés pour l'ORT où les précisions sont les mêmes. On observe que pour 5 caractères (LAIT, MG, AVP, PLA et AAR) les meilleures précisions sont obtenues pour un LD de 1, c'est-à-dire que les haplotypes ne sont composés que d'un seul SNP. Pour la MP, TB, TP, PRM, ORT et CCS, le seuil de LD pour atteindre des précisions maximales est relativement bas (entre 0,01 et 0,02).

Tableau 28. Précisions des évaluations génomiques haplotypiques (Ferdosi et al., 2016) pour les 11 caractères étudiés avec la méthode Distinct Windows (DW) et la méthode basée sur le déséquilibre de liaison (LD) pour les évaluations Saanen

	Saanen				
	ssGBLUP$_{SNP}$	**ssGBLUP$_{FerdosiDW}$**	**ssGBLUP$_{FerdosiLD}$**	**Fenêtre avec la meilleure précision (DW)**	**Seuil avec la meilleure précision (LD)**
LAIT	0,49	0,49	0,48	5	1
MG	0,43	0,43	0,42	5	1
MP	0,45	0,45	0,44	5	0,05
TB	0,59	0,59	0,57	10	0,02
TP	0,73	0,73	0,71	20	0,02
AVP	0,62	0,62	0,61	2	1
PLA	0,59	0,59	0,58	2	1
PRM	0,36	0,38	0,35	35	0,01
AAR	0,60	0,60	0,59	10	1
ORT	0,48	0,49	0,48	15	0,01
CCS	0,46	0,48	0,48	15	0,01
Moyenne	0,53	0,53	0,52	-	-

5. Discussion

Avec la méthode LD, la proportion d'haplotype composé d'un seul SNP devient vite importante en augmentant le seuil de LD. En caprins, le LD moyen entre 2 SNPs est de 0,17 pour les analyses Alpine et Saanen et de 0,14 pour les analyses multirace Carillier-Jacquin, (2015). L'utilisation d'un seuil de LD trop élevé va limiter le nombre d'haplotype contenant plus de 1 SNP. Cuyabano et al., (2014) ont utilisé la méthode LD pour construire des haplotypes (appellés haploblocks) dans une analyse génomique en race Holstein. Ils disposaient de 5 214 animaux génotypés sur une puce 50K, toutes les données ont été imputées sur la puce HD. Ils ont analysé 3 caractères (le TP, la fertilité et la résistance aux mammites). Dans leurs analyses, en moyenne 9% des SNPs n'appartenaient à aucun haploblock (entre 6 % et 17 % selon le seuil de LD). Ces valeurs sont bien inférieures à celles observées pour les races Alpine et Saanen. Cependant, le LD en bovin laitier (0,18 à 0,23) est supérieur à celui observé en caprin (0,17 pour les races Alpine et Saanen) (paragraphe 3.4.2, Chapitre 1). De plus, pour construire nos haplotypes avec la méthode LD, nous avons utilisé le calcul du r^2. Cuyabano et al., (2014) ont eux préféré utiliser la mesure D' qui présente l'avantage d'être moins sensible lorsque la MAF est faible (McRae et al., 2002) Un LD plus élevé entre SNPs et une densité en SNP plus élevée permettent de créer des haplotypes plus facilement et de limiter le nombre de SNPs « isolés » dans les analyses, ce qui peut expliquer les différences entre l'étude de Cuyabano et al., (2014) et nos résultats.

Avec la méthode DW, plus les haplotypes sont longs et plus le nombre d'allèles moyen est élevé. Le nombre d'allèle par haplotype est en moyenne plus élevé pour la race Alpine que pour la race Saanen, notamment pour des haplotypes de plus de 20 SNPs. Cette diversité plus importante en race Alpine qu'en race Saanen est cohérente avec les résultats de Carillier-Jacquin, (2015). En effet, la consanguinité, calculée à partir de données génomiques, était de 1,8 % en race Alpine et de 2,3 % en race Saanen. Et il est démontré que la consanguinité induit une baisse de la diversité génétique (Charlesworth, 2003). Les différences du nombre d'allèle ne s'observent plus avec la méthode LD. En effet, en raison du LD relativement faible en races Alpine et Saanen, de nombreux haplotypes ne sont constitués que d'un seul SNP. De ce fait, les allèles observés en race Alpine et Saanen sont les mêmes (allèles bi-alléliques). Pour les

haplotypes constitués de plus d'un SNP, ils sont en général relativement courts et, globalement, les mêmes allèles sont observés en Alpine et en Saanen. Pour les analyses multirace, la longueur des haplotypes est en moyenne de 2,7 SNPs pour un seuil de LD de 0,01 et de 2 SNPs pour un seuil de LD de 1. Avec la méthode DW, les SNPs regroupés dans un même haplotype présentent des LD très faibles.

L'analyse des corrélations entre les éléments hors-diagonaux de **G** et de $\mathbf{A_{22}}$ pour les méthodes $ssGBLUP_{SNP}$, $ssGBLUP_{Ferdosi}$, $ssGBLUP_{pseudo\text{-}SNPs}$ et $WssGBLUP_{pseudo\text{-}SNPs}$ permet d'étudier l'influence de ces méthodes dans la construction de la matrice de parenté, qui elle-même influe sur les précisions génomiques. Entre les évaluations ssGBLUP, $ssGBLUP_{Ferdosi}$ (DW ou LD) et $ssGBLUP_{pseudo\text{-}SNP}$ (DW ou LD), les corrélations sont supérieures à 90 % quelle que soit la taille des haplotypes (pour la méthode DW) ou le seuil de LD fixé (pour la méthode LD). Seule l'utilisation de la méthode $WssGBLUP_{pseudo\text{-}SNP}$ avec des haplotypes construits selon la méthode DW induit des baisses de ces corrélations. Dans leurs travaux, Ferdosi et al., (2016) ont comparé les éléments hors-diagonaux de la matrice **G** et les éléments hors diagonaux des matrices haplotypiques (méthode DW). Plus les haplotypes sont grands (méthode DW) et plus les corrélations sont élevées, mais ils n'indiquent pas les niveaux des corrélations obtenus.

Le troisième papier de ma thèse, intitulé « Genomic prediction with pseudo-SNPs for milk production, udder type and somatic cell score in French dairy goats », propose la construction et l'utilisation d'haplotypes, considérés comme des pseudo-SNPs, dans les modèles d'évaluation génomique caprins. L'avantage de ces approches est leur implémentation directe avec le ssGBLUP et le WssGBLUP. Sur la majorité des caractères étudiés, nous n'avons pas montré de supériorité significative de ces approches par rapport aux méthodes ssGBLUP ou WssGBLUP basées sur l'information individuelle des SNPs pour les races Alpine et Saanen. Le seul caractère pour lequel ces approches améliorent les précisions génomiques est la MG en race Alpine (gains de + 19% et + 22% avec la méthode DW par rapport aux ssGBLUP et WssGBLUP respectivement).

En utilisant la méthode de Ferdosi et al., (2016), basée sur la construction d'une matrice de parenté haplotypique (méthode DW), nous avons obtenu des précisions génomiques similaires aux précisions du ssGBLUP (sauf pour la matière grasse en Alpine). Les résultats obtenus avec la méthode de Ferdosi et al., (2016) sont proches des résultats obtenus avec les pseudo-SNPs. Nos résultats différent de ceux observés par Ferdosi et al., (2016). Ils ont analysé des caractères de circonférence du scrotum, âge à la puberté et poids des femelles au moment du premier corps jaune en race bovine Brahman. Ils disposaient de 1007 animaux phénotypés pour la circonférence du scrotum et génotypés avec la puce Illumina BovineSNP50 BeadChip (Illumina, San Diego, CA). Pour le caractère âge à la puberté et poids des femelles au premier corps jaune, 854 animaux étaient phénotypés et génotypés. Ils ont créé des haplotypes selon la méthode DW et les ont intégrés dans les évaluations génomiques à l'aide de la matrice de parenté haplotypique $\mathbf{G_{hap}}$. Ils observent des précisions légèrement plus élevées en utilisant la matrice $\mathbf{G_{hap}}$ qu'en utilisant la matrice **G** (VanRaden, 2008): gain de +5 points (avec des haplotypes de 16 SNPs) pour la circonférence du scrotum , + 2 points pour l'âge à la puberté (avec des haplotypes de 7 SNPs) et + 2 points pour le poids des femelles au premier corps jaune (avec des haplotypes de 11 SNPs). Dans notre étude, les évaluations génomiques haplotypiques basé sur le LD n'ont globalement pas amélioré les précisions des évaluations multirace, Alpine et Saanen. Les haplotypes construits avec cette méthode sont courts (entre 2 et 6 SNPs en moyenne selon le seuil de LD) et souvent ne comporte qu'un seul SNP. Cela pourrait expliquer pourquoi cette méthode n'a pas amélioré les précisions des évaluations.

Les méthodes haplotypiques sont globalement aussi précises que les méthodes d'évaluations basées sur l'information individuelle des SNPs. Cependant, ces analyses nécessitent des étapes supplémentaires (phasage des données de génotypages, définition des tailles des haplotypes (si approche DW) ou du calcul du LD, construction des haplotypes et conversion des haplotypes en pseudo-SNPs) par rapport aux évaluations classiques afin de préparer les haplotypes. De ce fait, l'utilisation des haplotypes dans les modèles d'évaluations génomiques caprines française présente un intérêt limité à ce jour.

Chapitre 6 : Discussions générales et perspectives

1. Bilan des principaux résultats

Suite aux travaux de thèse de Carillier-Jacquin (2015), l'organisme et entreprise de sélection Capgènes a décidé de mettre en œuvre des évaluations génomiques, basées sur le ssGBLUP, pour l'ensemble des caractères en sélection des deux principales races caprines laitières (Alpine et Saanen). Les premiers index génomiques ont été publiés en janvier 2018, aboutissant à une évaluation génétique plus précise, en particulier pour les jeunes boucs sans performance. Ce changement permet de piloter plus finement les schémas de sélection et d'accélérer le progrès génétique en détectant précocement le potentiel des boucs candidats. La précision des index des boucs génotypés à la naissance, exprimée par le CD, augmente d'environ 10 points. La conduite du schéma de sélection, a été revue et prévoit une réduction de l'intervalle de génération, en particulier pour la procréation des filles d'IA. Une meilleure prédiction des index, couplée à une plus forte intensité de sélection et un intervalle de génération réduit contribueront à améliorer le niveau des boucs proposés au catalogue de Capgènes. La génomique entrainera une réorganisation du schéma de sélection à différents niveaux :

- les accouplements programmés : un index génomique plus précoce pour les femelles élites permettra d'abaisser l'âge moyen des mères à boucs ;
- le choix des jeunes mâles d'insémination dans les élevages : la disponibilité d'un index génomique fiable et précoce pour les jeunes mâles nés d'accouplements programmés induira une sélection plus efficace ;
- l'utilisation des jeunes mâles d'insémination : ils seront utilisés dès qu'ils auront la capacité à produire de la semence sans avoir à attendre les résultats de testage sur descendance, ce qui permettra de réduire l'intervalle de génération sur la voie mâle ;
- l'objectif de sélection qui pourra intégrer de nouveaux caractères, notamment ceux qui sont peu héritables et donc difficiles à améliorer dans un schéma classique (caractères fonctionnels, résistance à des maladies) : la construction de populations de référence spécifique pour la prise en compte de nouveaux critères modifiera l'organisation de la collecte des phénotypes.

Les objectifs de ma thèse ont été de comparer les précisions des évaluations génomiques ssGBLUP réalisées par Carillier-Jacquin, (2015) avec des évaluations génomiques ssGBLUP utilisant des données actualisées (nouvelles campagnes de génotypages). Nous nous sommes également intéressés aux biais et à leur amélioration en faisant varier les valeurs des hyperparamètres du ssGBLUP. Dans un deuxième temps, nous avons exploré de nouveaux modèles et méthodes d'évaluations génomiques permettant de prendre en compte des gènes majeurs ou QTL. Ces études ont porté sur des approches basées sur l'utilisation des informations individuelles des SNPs ou la construction d'haplotypes.

Une première étude a consolidé les résultats obtenus par Carillier-Jacquin, (2015), c'est-à-dire que les précisions génomiques ssGBLUP (avec des données actualisées) sont plus précises que celles obtenues avec un BLUP pour l'ensemble des caractères. Nous avons observé des évolutions de précisions entre -10 % et +72 % pour le ssGBLUP comparé au BLUP selon le caractère pour les analyses multirace, Alpine et Saanen. La baisse de précisions concerne uniquement la distance plancher-jarret pour les analyses multirace (-10 %). En revanche, nous avons observé une baisse des précisions des évaluations génomiques pour des caractères importants pour la filière (tel que la matière protéique) en comparant les précisions des évaluations avec une population de validation entre 2006 et 2009 et les précisions des

évaluations avec une population de validation entre 2009 et 2012. L'analyse de la conduite du schéma de sélection ou des données des évaluations (phénotypes, pedigrees ou génotypes) n'ont pas permis d'identifier une origine précise de ces diminutions de précisions. Les évaluations ssGBLUP présentent des biais pour les caractères avec des héritabilités moyennes à faibles. Les hyperparamètres du ssGBLUP (α, ω et τ) réduisent significativement les biais tout en affectant faiblement les précisions des évaluations génomiques. L'ensemble de ces travaux ont été utiles pour fixer les populations d'apprentissage et de validation qui ont servi pour la suite des analyses.

La deuxième partie de ma thèse portait sur l'amélioration des précisions des évaluations en intégrant des gènes majeurs ou des QTLs dans les modèles génomiques. Ces travaux peuvent être classés en 2 catégories : (1) les caractères pour lesquels une mutation causale était connue et pour lesquels des génotypes étaient disponibles, en plus des génotypes de la puce 50K (cas du gène de la *caséine* α_{s1} pour le taux protéique et *DGAT1* pour le taux butyreux), (2) les caractères pour lesquels seuls les SNPs de la puce 50K étaient disponibles. L'utilisation des génotypes pour du gène de la *caséine* α_{s1} et pour *DGAT1* dans les modèles d'évaluations avec la méthode du gene content s'est révélée inefficace pour améliorer les précisions des évaluations pour le taux protéique et le taux butyreux. Pour le taux protéique, les précisions obtenues avec le gene content ont été de -1 % à 0% inférieures aux précisions du ssGBLUP pour les races Alpine et Saanen. Avec le ssGBLUP, la région du gène de la *caséine* α_{s1} joue un rôle particulier. On se rend compte qu'en utilisant uniquement les génotypes de la puce 50K, une partie de l'effet de ce gène sur le taux protéique est pris en compte avec cette méthode. Ce résultat s'explique par une densification plus forte en SNPs dans la région du gène de la caséine. Le gene content a été également utilisé pour prédire les valeurs génétiques des animaux pour le taux butyreux en utilisant les génotypages pour le gène DGAT1. Comme pour le taux protéique, la méthode gene content n'a pas amélioré les précisions par rapport au ssGBLUP avec des baisses comprises entre 0 % et -11%. Pour les méthodes utilisant uniquement les SNPs de la puce 50K, les évaluations WssGBLUP et ses alternatives (poids communs attribués à des SNPs consécutifs) ont amélioré les précisions des évaluations génomiques mais pas de manière équivalente sur tous les caractères. Les précisions ont été améliorées pour le taux protéique dans les deux races (jusqu'à +4 % pour la race Alpine et +7 % pour la race Saanen). Pour la race Saanen, les précisions ont également été améliorées par rapport au ssGBLUP pour le lait, les quantités de matières grasses et protéiques, pour la distance entre le plancher de la mamelle et le jarret ainsi que pour la qualité de l'attache arrière (entre +2 % et +14 %), caractères pour lesquels un QTL est détecté sur le chromosome 19. Il est toutefois important de noter que la méthode WssGBLUP n'a pas permis de mettre en évidence la région du gène *DGAT1*, les précisions pour le taux butyreux n'ont pas été améliorées. Nous avons également étudié l'utilisation d'haplotypes dans les modèles d'évaluations génomiques. Ces approches peuvent améliorer légèrement les précisions par rapport au ssGBLUP mais pas pour tous les caractères. Les précisions des évaluations ssGBLUP haplotypiques (ou pseudo-SNPs) ont été entre -3% et +19% supérieures aux évaluations ssGBLUP selon le caractère pour les races Alpine et Saanen. En utilisant les approches basées sur le WssGBLUP avec des pseudo-SNPs, les précisions ont été entre -9% et +19% supérieures aux évaluations ssGBLUP selon le caractère pour les races Alpine et Saanen. Cependant, la mise en œuvre des modèles haplotypiques nécessite des étapes supplémentaires par rapport aux modèles utilisant individuellement des SNPs : phasage, construction des haplotypes, conversion si besoin des haplotypes en pseudo-SNPs et définition

de la taille optimale des haplotypes. Actuellement et en race Alpine et Saanen, les approches haplotypiques ne concurrencent pas le WssGBLUP, les précisions obtenues sont similaires.

Une amélioration des précisions des évaluations génomiques est possible lorsque le caractère est gouverné par des QTLs ou des gènes majeurs. Le WssGBLUP est la méthode la plus adaptée pour intégrer ces informations. Toutefois, elle est plus intéressante pour la race Saanen qui présente des QTLs pour la majorité des caractères en sélection. Son application en routine ne poserait pas de difficultés majeures puisqu'elle est basée sur une double itération du ssGBLUP, méthode appliquée actuellement en caprin. Pour la race Alpine, en revanche, le ssGBLUP reste la méthode la plus intéressante pour prédire les GEBV des animaux, excepté pour le taux protéique où le WssGBLUP fournit des prédictions plus précises que le ssGBLUP.

2. Perspectives d'amélioration des précisions des évaluations génomiques pour la filière caprine laitière française

La taille et la composition de la population de référence ont un impact sur les précisions des évaluations génomiques (paragraphe 3.4.1, Chapitre 1), les précisions seront plus fortes si la taille de la population de référence est importante. En caprin, cette population est actuellement de taille limitée avec environ 3 000 animaux génotypés. Dans le schéma de sélection caprin actuel, tous les mâles testés sur descendances sont génotypés depuis 1993, ce qui représente environ 70 nouveaux génotypages par an. Les femelles ont, quant à elles, été génotypées dans le cadre d'un projet de détection de QTL et ne sont représentatives que de deux campagnes (2008-2009). Carillier-Jacquin (2015) a proposé quelques perspectives afin d'augmenter plus rapidement la taille de la population de référence en caprins, avec notamment l'acquisition de génotypages femelles ou bien la mutualisation de données avec la création d'une population de référence internationale. Le génotypage 50K reste coûteux comparé au prix de l'animal ce qui rend complexe son utilisation pour génotyper les femelles. Pour contourner ce problème, une alternative serait le développement d'une puce basse densité dont le prix serait suffisamment bas pour rendre le génotypage des femelles intéressant. Ce sont des solutions qui ont pu être développées chez les bovins laitiers ou les ovins laitiers avec une puce 3K SNP et 12K respectivement (Dassonneville et al., 2011; Bolormaa et al., 2015). Ensuite, des premiers travaux exploratoires, réalisés dans le cadre d'une collaboration avec Luiz Brito et Flavio Schenkel en 2017 (Université de Guelph, Canada), ont permis de constituer une population de référence entre la France et le Canada. Cette population était constituée de 721 animaux génotypés pour le Canada et de 3351 animaux génotypés pour la France. Les premiers résultats montrent que les populations françaises et canadiennes (Alpine et Saanen) sont différentes d'un point de vue génétique (travaux en cours, non publiés). Des ACP réalisées sur les génotypes montrent des groupes clairement différents entre les populations françaises et canadiennes mais également entre les races Alpine et Saanen. La persistance des phases gamétiques a également été analysée entre les populations françaises et canadiennes. Pour une distance entre deux SNPs de 50 kb (distance moyenne entre 2 SNPs sur la puce 50k), la persistance des phases gamétiques est de 0,60 entre les Alpines françaises et canadiennes et de 0,52 entre les Saanen françaises et canadiennes. Au-delà de ce seuil, la persistance chute plus fortement pour les races Saanen que pour les races Alpines. La persistance est ainsi inférieure à 0,1 pour la race Saanen pour des SNPs distants d'au moins 500 kb alors qu'une persistance de 0,1 est atteinte pour la race Alpine avec des SNPs distants d'au moins 16 000 kb. Il semble donc complexe de créer une population de référence commune entre le France et le Canada sans travaux supplémentaires. Il pourrait être intéressant d'identifier (grâce à des analyses GWAS par exemple) des SNPs dont les effets

sur un caractère sont similaires dans les populations françaises et canadiennes. Ils pourraient alors constituer une base commune de SNPs à exploiter dans des modèles d'évaluation génomique en utilisant l'ensemble des données françaises et canadiennes.

Les évaluations génomiques ont été mises en place chez les caprins laitiers français en janvier 2018. Elles ont modifié et vont continuer de modifier la conduite du schéma de sélection. Une des principales modifications sera la suppression du testage sur descendance. Pour conserver un coût constant du schéma de sélection, le nombre d'accouplements raisonnés réalisés chaque année va être augmenté de 1000 à 1500, ce qui permet un accroissement du nombre de candidats à la sélection. Tous les animaux candidats seront ainsi génotypés, environ 400 génotypages par an sont prévus contre 70 actuellement. Ces modifications vont permettre d'augmenter la taille de la population de référence plus rapidement et accroitre ainsi les précisions des évaluations génomiques. Il est attendu une diminution de l'intervalle de génération ainsi qu'une augmentation de l'intensité de sélection pour espérer augmenter de +30% le progrès génétique.

Les races Alpine et Saanen présentent une diversité génétique plus importante que d'autres espèces d'élevage. Cela se traduit, par exemple, par une taille efficace des populations plus élevée que les bovins laitiers (Carillier-Jacquin, 2015), une consanguinité plus faible et un déséquilibre de liaison entre marqueurs plus faibles que les bovins laitiers (Chapitre 1). Tous ces paramètres peuvent affecter les précisions des évaluations génomiques (paragraphe 3.4, Chapitre 1). La disponibilité de données de plus haute densité ou de séquence en caprins permettrait de mieux capter les effets des QTLs avec des SNPs en fort DL, voire d'utiliser directement les mutations causales. Dans le cadre du projet international de séquençage VarGoat, coordonné par G. Tosser-Klopp (UMR GenPhySE, INRA), environ 400 séquences viennent d'être disponibles (200 séquences de chèvres internationales et 200 séquences de chèvres françaises dont 80 Alpine et Saanen). L'objectif du travail de thèse de Estelle Talouarn (2017-2020, INRA GenPhySE) est de valoriser l'arrivée de données de séquence en caprins pour cartographier plus précisément les régions d'intérêt (grâce à des analyses GWAS) mais également pour améliorer les précisions des évaluations génomiques. Les travaux en cours visent à tester différentes approches pour imputer les animaux génotypés avec la puce 50K jusqu'à la séquence. L'ensemble des animaux imputés pourront ainsi être utilisé dans le cadre d'analyses GWAS afin d'améliorer la détection de QTL en caprin sur les caractères en sélection et les nouveaux caractères étudiés (comme le chromosome 19 en Saanen), voire découvrir la ou les mutations causales ou encore de nouvelles régions non identifiées avec les analyses GWAS, basées actuellement sur les données de la puce 50K. Pour les évaluations génomiques, des premiers travaux ont été réalisés en bovins laitiers et en volailles pour intégrer des données de séquence dans les évaluations génomiques (van Binsbergen et al., 2014; Heidaritabar et al., 2016; Raymond et al., 2018). Ces études ont toutes utilisé la méthode GBLUP mais elles ont également comparé les précisions avec la méthode Bayesian Stochastic Search Variable Selection (BSSVS) ou avec un BayesC. Ils ont montré que l'intégration des données de séquences avec ces méthodes n'améliorent pas les précisions ou les améliorent très légèrement par rapport à l'utilisation d'une puce 50K ou d'une puce haute densité (HD). Cependant, certaines études soulignent que des amélioration de précisions sont possibles en utilisant des marqueurs présélectionnés sur la séquence (Brøndum et al., 2014; VanRaden et al., 2017) avec les méthodes GBLUP, BayesA ou un Bayesian variable selection model (BVSM). Dans l'étude de VanRaden et al., (2017), des évaluations BayesA en bovin laitier sur des caractères de production laitière et sur des caractères de morphologie ont été réalisées en utilisant la puce 60K complétée de SNPs sélectionnés sur la séquence (+16 k SNPs). Ces SNPs additionnels ont été sélectionnés à partir de leur effet sur le caractère (estimé avec des analyses GWAS). Ils

observent des évaluations en moyenne plus précises de 2,7% comparées aux évaluations utilisant uniquement la puce 60K. Pour l'étude de Brøndum et al., (2014), des évaluations GBLUP et BVSM en bovin laitier pour des caractères de productions laitières, des caractères de morphologie et des caractères de longévité ont été réalisées en utilisant soit les génotypes 54K, soit les génotypes 54K complétés de QTLs sélectionnés à partir de la séquence (sélectionnés sur leur p-valeur après des analyses GWAS). Des gains de 3% à 5% ont été observés pour les caractères de productions laitières en incluant les QTL dans les évaluations GBLUP. Avec la méthode BVSM, les gains étaient plus faibles qu'avec le GBLUP.

La sélection de SNPs à partir de données de séquence serait envisageable en caprin, si les études d'imputation menées actuellement dans le cadre de la thèse d'E. Talouarn sont concluantes. Comme en bovin laitier (Teissier et al., 2018b), à partir des données de séquence imputées, des analyses GWAS pourraient être réalisées sur l'ensemble des caractères en sélection. Les SNPs pourraient être sélectionnés à partir de leur p-valeur à la suite d'analyse GWAS intra-race, ou en fonction de leur effet sur le caractère comme proposé par VanRaden et al., (2017). Pour gagner en puissance, des GWAS multirace pourraient être envisagées à l'aide de méta-analyses afin d'estimer les effets des SNPs et de sélectionner efficacement les SNPs ayant des effets dans les deux races. Les SNPs les plus significatifs pourraient alors être ajoutés aux données de la puce 50K pour construire la matrice d'apparentement génomique **H** dans les modèles d'évaluation génomique ssGBLUP ou WssGBLUP.

Une autre possibilité serait de sélectionner les SNPs sur la séquence à partir d'informations nouvelles telles que les données d'annotation. Ces dernières nous informent sur la localisation des SNPs (régions introniques, exoniques, régulatrices, …). On pourrait ainsi favoriser l'intégration de SNPs présents dans des régions exoniques ou régulatrices dans les modèles d'évaluation génomique. Les données produites dans le cadre du projet Fr-AgEncode, coordonné par des scientifiques du département Génétique Animale et qui a pour objectif d'enrichir les bases de données d'annotation pour plusieurs espèces d'élevage (porcs, poule, chèvre et bovin), pourraient être très utiles. Combiner la sélection de ces SNP, la construction d'haplotypes (convertis en pseudo-SNPs) et l'utilisation des modèles génomiques ssGBLUP ou WssGBLUP méritent d'être explorées pour améliorer les précisions génomiques dans l'espèce caprine.

D'autres approches peuvent être envisagées. MacLeod et al., (2014) propose une méthode, appelée BayesRC, qui permet d'intégrer l'information issue des données d'annotation en se basant sur la méthode BayesR. Les SNPs sont classés dans des groupes spécifiques en fonction de leur annotation. Par exemple, un premier groupe peut regrouper les SNPs proches de gènes candidats (classe I) et un deuxième groupe les SNPs restants (classe II). La méthode est relativement flexible et peut être étendue à plus de 2 classes (comme avec le BayesR). L'avantage de cette méthode est qu'il est possible de choisir des distributions différentes pour chacun des groupes. MacLeod et al., (2014) ont utilisé l'approche BayesRC avec de vraies génotypes en Holstein mais avec des QTL et des phénotypes simulés. Ils ont comparé les précisions des évaluations obtenues avec un BayesR utilisant une puce HD (600K SNPs) ou l'ensemble de la séquence (800K SNPs) et un GBLUP (utilisant une puce HD ou la séquence). Ils ont obtenu des gains allant jusqu'à 16% avec la méthode BayesRC par rapport à la méthode BayesR. Et les évaluations GBLUP (HD ou séquence) étaient moins précises que les évaluations BayesR. De telles approches pourraient être envisagées sur les données caprines françaises grâce à l'acquisition de données d'annotations dans le cadre du projet Fr-AgEncode.

Liste des tableaux

Liste des figures

Références

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. Journal of Dairy Science 93:743–752. doi:10.3168/jds.2009-2730.

Allain, D., and J.M. Roguet. 2003. Genetic and nongenetic factors influencing mohair production traits within the national selection scheme of Angora goats in France. Livestock Production Science 82:129–137. doi:10.1016/S0301-6226(03)00035-6.

Andonov, S., D.A.L. Lourenco, B.O. Fragomeni, Y. Masuda, I. Pocrnic, S. Tsuruta, and I. Misztal. 2017. Accuracy of breeding values in small genotyped populations using different sources of external information—A simulation study. Journal of Dairy Science 100:395–401. doi:10.3168/jds.2016-11335.

Auvray, B., K.G. Dodds, and J.C. McEwan. 2011. Brief communication: Genomic selection in the New Zealand Sheep industry using the Ovine SNP50 Beadchip. Proceedings of the New Zealand Society of Animal Production 11:263–265.

Auvray, B., J.C. McEwan, S. -a. N. Newman, M. Lee, and K.G. Dodds. 2014. Genomic prediction of breeding values in the New Zealand sheep industry using a 50K SNP chip. J. Anim. Sci. 92:4375–4389. doi:10.2527/jas.2014-7801.

Baloche, G., A. Legarra, G. Sallé, H. Larroque, J.-M. Astruc, C. Robert-Granié, and F. Barillet. 2014. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. Journal of Dairy Science 97:1107–1116. doi:10.3168/jds.2013-7135.

Barillet, F., J.-J. Arranz, and A. Carta. 2005. Mapping quantitative trait loci for milk production and genetic polymorphisms of milk proteins in dairy sheep. Genetics Selection Evolution 37:S109–S123. doi:10.1051/gse:2004033.

van Binsbergen, R., M.C. Bink, M.P. Calus, F.A. van Eeuwijk, B.J. Hayes, I. Hulsegge, and R.F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 46:41. doi:10.1186/1297-9686-46-41.

B.O.agri. 2014. Le Règlement Technique du Contrôle de performances lait en espèce caprine - protocoles A, AT, AZ, CZ. https://info.agriculture.gouv.fr/gedei/site/bo-agri/document_administratif-8e8654fe-6c71-478a-9713-a76ab9abd352/telechargement

Boerner, V., and B. Tier. 2016. BESSiE: a software for linear model BLUP and Bayesian MCMC analysis of large-scale genomic data. Genetics Selection Evolution 48:63. doi:10.1186/s12711-016-0241-x.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M.N. Rossignol, M.Y. Boscher, T. Druet, L. Genestout, J.J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. Animal Production Science 52:115–120. doi:10.1071/AN11119.

Bolormaa, S., K. Gore, J.H.J. van der Werf, B.J. Hayes, and H.D. Daetwyler. 2015. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Animal Genetics 46:544–556. doi:10.1111/age.12340.

Brito, L.F., S.M. Clarke, J.C. McEwan, S.P. Miller, N.K. Pickering, W.E. Bain, K.G. Dodds, M. Sargolzaei, and F.S. Schenkel. 2017. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. BMC Genetics 18:7. doi:10.1186/s12863-017-0476-8.

Brito, L.F., M. Jafarikia, D.A. Grossi, J.W. Kijas, L.R. Porto-Neto, R.V. Ventura, M. Salgorzaei, and F.S. Schenkel. 2015. Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats. BMC Genetics 16:67. doi:10.1186/s12863-015-0220-1.

Broman, K.W., and J.L. Weber. 1999. Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain. The American Journal of Human Genetics 65:1493–1500. doi:10.1086/302661.

Brøndum, R., B. Guldbrandtsen, G. Sahana, M. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15:728. doi:10.1186/1471-2164-15-728.

Browning, B.L., Y. Zhou, and S.R. Browning. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. The American Journal of Human Genetics 103:338–348. doi:10.1016/j.ajhg.2018.07.015.

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp. 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178:553–561. doi:10.1534/genetics.107.080838.

Capgènes - Institut de l'élevage. 2017. Bilan de fertilité nationale 2016. http://www.fnec.fr/IMG/pdf/Fertilite_nationale_2016_STAT.pdf

Carillier, C., H. Larroque, I. Palhière, V. Clément, R. Rupp, and C. Robert-Granié. 2013. A first step toward genomic selection in the multi-breed French dairy goat population. J. Dairy Sci. 96:7294–7305. doi:10.3168/jds.2013-6789.

Carillier, C., H. Larroque, and C. Robert-Granié. 2014. Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. Genetics Selection Evolution 46:67. doi:10.1186/s12711-014-0067-3.

Carillier-Jacquin, C. 2015. Etude de la prédiction génomique chez les caprins : faisabilité et limites de la sélection génomique dans le cadre d'une population multiraciale et à faible effectif. Thèse de doctorat en Sciences agronomiques, biotechnologies agro-alimentaires. Toulouse, INPT. 193p.

Carillier-Jacquin, C., H. Larroque, and C. Robert-Granié. 2016. Including α s1 casein gene information in genomic evaluations of French dairy goats. Genetics Selection Evolution 48:54. doi:10.1186/s12711-016-0233-x.

Charlesworth, D. 2003. Effects of inbreeding on the genetic diversity of populations. Philosophical Transactions of the Royal Society of London B: Biological Sciences 358:1051–1070. doi:10.1098/rstb.2003.1296.

Christensen, O.F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genetics Selection Evolution 44:37. doi:10.1186/1297-9686-44-37.

Christensen, O.F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. animal 6:1565–1571. doi:10.1017/S1751731112000742.

Clément, V., D. Boichard, A. Piacère, A. Barbat, and E. Manfredi. 2002. Genetic evaluation of French goats for dairy and type traits. Page, Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France.

Clément, V., H. Caillat, A. Piacère, E. Manfredi, C. Robert-Granie, F. Bouvier, and R. Rupp. 2008. Vers la mise en place d'une sélection pour la résistance aux mammites chez les 185 caprins. 15ème Rencontres des Recherches autour des Ruminants, 3-4 décembre 2008, Paris, Fran.

Clément, V., I. Palhière, and H. Larroque. 2015. Evaluation génétique dans l'espèce caprine : Caractères de production laitière, de comptage de cellules somatiques et de morphologie. Compte-rendu n 00, 14(202), 041.

Clément, V., P. Martin, and F. Barillet. 2006. Elaboration d'un index synthétique caprin combinant les caractères laitiers et des caractères de morphologie mammaire. Renc. Rech. Ruminants.

Clément, V., P. Martin, and R. Rupp. 2017. Prise en compte des concentrations cellulaires dans l'objectif de sélection caprin. Journée de restitution Mamovicap, 7 mars 2017. Paris: Maison du lait

Croué, I., and V. Ducrocq. 2017. Genomic and single-step evaluations of carcass traits of young bulls in dual-purpose cattle. Journal of Animal Breeding and Genetics 134:300–307. doi:10.1111/jbg.12261.

Cuyabano, B.C., G. Su, and M.S. Lund. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1171. doi:10.1186/1471-2164-15-1171.

Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLOS ONE 3:e3395. doi:10.1371/journal.pone.0003395.

Danchin-Burge, C., D. Allain, V. Clément, A. Piacère, P. Martin, and I. Palhière. 2012. Genetic variability and French breeding programs of three goat breeds under selection. Small Ruminant Research 108:36–44. doi:10.1016/j.smallrumres.2012.03.016.

Dassonneville, R., R.F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M.S. Lund, V. Ducrocq, and G. Su. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. Journal of Dairy Science 94:3679–3686. doi:10.3168/jds.2011-4299.

Dong, Y., M. Xie, Y. Jiang, N. Xiao, X. Du, W. Zhang, G. Tosser-Klopp, J. Wang, S. Yang, J. Liang, W. Chen, J. Chen, P. Zeng, Y. Hou, C. Bian, S. Pan, Y. Li, X. Liu, W. Wang, B. Servin, B. Sayre, B. Zhu, D. Sweeney, R. Moore, W. Nie, Y. Shen, R. Zhao, G. Zhang, J. Li, T. Faraut, J. Womack, Y. Zhang, J. Kijas, N. Cockett, X. Xu, S. Zhao, J. Wang, and W. Wang. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). Nature Biotechnology 31:135–141. doi:10.1038/nbt.2478.

Duchemin, S.I., C. Colombani, A. Legarra, G. Baloche, H. Larroque, J.-M. Astruc, F. Barillet, C. Robert-Granié, and E. Manfredi. 2012. Genomic selection in the French Lacaune dairy sheep breed. Journal of Dairy Science 95:2723–2733. doi:10.3168/jds.2011-4980.

Erbe, M., B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, B.A. Mason, and M.E. Goddard. 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. doi:10.3168/jds.2011-5019.

INSEE. 2013. Exploitations agricoles − Tableaux de l'Économie Française. https://www.insee.fr/fr/statistiques/1374189?sommaire=1374192.

Fan, B., Z.-Q. Du, D.M. Gorbach, and M.F. Rothschild. 2010. Development and Application of High-density SNP Arrays in Genomic Studies of Domestic Animals. Asian-Australasian Journal of Animal Sciences 23:833–847. doi:10.5713/ajas.2010.r.03.

Ferdosi, M.H., J. Henshall, and B. Tier. 2016. Study of the optimum haplotype length to build genomic relationship matrices. Genetics Selection Evolution 48. doi:10.1186/s12711-016-0253-6.

Forneris, N.S., A. Legarra, Z.G. Vitezica, S. Tsuruta, I. Aguilar, I. Misztal, and R.J.C. Cantet. 2015. Quality Control of Genotypes Using Heritability Estimates of Gene Content at the Marker. Genetics 199:675–681. doi:10.1534/genetics.114.173559.

Garcia-Baccino, C.A., A. Legarra, O.F. Christensen, I. Misztal, I. Pocrnic, Z.G. Vitezica, and R.J.C. Cantet. 2017. Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. Genetics Selection Evolution 49:34. doi:10.1186/s12711-017-0309-2.

Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. animal 1:21–28. doi:10.1017/S1751731107392628.

Gianola, D. 2013. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194:573–596. doi:10.1534/genetics.113.151753.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response.. Genetica 136:245–257. doi:https://doi.org/10.1007/s10709-008-9308-0.

Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. Journal of Animal Breeding and Genetics 124:323–330. doi:10.1111/j.1439-0388.2007.00702.x.

Gowane, G., S.H. Lee, S. Clark, N. Moghaddar, A.A.-M. Hawlader, and J.H.J. van der Werf. 2018a. Optimising bias and accuracy in genomic prediction of breeding values. Proceedings of the World Congress on Genetics Applied to Livestock Production Electronic Poster Session-Method and Tools-Prediction 1:117.

Gowane, G.R., S.H. Lee, S. Clark, N. Moghaddar, H.A. Al-Mamun, and J.H.J. van der Werf. 2018b. Effect of selection on bias and accuracy in genomic prediction of breeding values. bioRxiv 298042. doi:10.1101/298042.

Grisart, B. 2002. Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Research 12:222–231. doi:10.1101/gr.224202.

Groenen, M.A., H.-J. Megens, Y. Zare, W.C. Warren, L.W. Hillier, R.P. Crooijmans, A. Vereijken, R. Okimoto, W.M. Muir, and H.H. Cheng. 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12:274. doi:10.1186/1471-2164-12-274.

Grosclaude, F. 1988. Le polymorphisme génétique des principales lactoprotéines bovines. Relations avec la quantité, la composition et les aptitudes fromagères du lait. Productions Animales 1 (1), 5-17.(1988).

Grosclaude, F., M.-F. Mahé, G. Brignon, L. Di Stasio, and R. Jeunet. 1987. A Mendelian polymorphism underlying quantitative variations of goat αs1-casein. Genetics Selection Evolution 19:399–412. doi:10.1186/1297-9686-19-4-399.

Grossi, D.A., M. Jafarikia, L.F. Brito, M.E. Buzanskas, M. Sargolzaei, and F.S. Schenkel. 2017. Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in Canadian pigs. BMC Genetics 18:6. doi:10.1186/s12863-017-0473-y.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42:5. doi:10.1186/1297-9686-42-5.

Hayashi, T., and H. Iwata. 2013. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics 14:34. doi:10.1186/1471-2105-14-34.

Hayes, B., and M. Goddard. 2010. Genome-wide association and genomic selection in animal breedingThis article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming".. Genome 53:876–883. doi:10.1139/G10-076.

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41:51. doi:10.1186/1297-9686-41-51.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92:433–443. doi:10.3168/jds.2008-1646.

Hayes, B.J., A.J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M.E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genetical Research 89. doi:10.1017/S0016672307008865.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009c. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research 91:47–60. doi:10.1017/S0016672308009981.

Heidaritabar, M., M.P.L. Calus, H.-J. Megens, A. Vereijken, M. a. M. Groenen, and J.W.M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics 133:167–179. doi:10.1111/jbg.12199.

Henderson, C.R., O. Kempthorne, S.R. Searle, and C.M. von Krosigk. 1959. The Estimation of Environmental and Genetic Trends from Records Subject to Culling. Biometrics 15:192–218. doi:10.2307/2527669.

Hess, M., T. Druet, A. Hess, and D. Garrick. 2017. Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genetics Selection Evolution 49:54. doi:10.1186/s12711-017-0329-y.

Hickey, J.M., B.P. Kinghorn, B. Tier, S.A. Clark, J.H.J. van der Werf, and G. Gorjanc. 2013. Genomic evaluations using similarity between haplotypes. Journal of Animal Breeding and Genetics 130:259–269. doi:10.1111/jbg.12020.

Huau, C., G. Foucras, G. Tabouret, C. Caubet, F. Bouvier, T. Fassier, P. Rainard, P. Martin, G. Tosser-Klopp, and R. Rupp. 2015. L'amélioration génétique sur le comptage de cellules somatiques du lait s'accompagne d'une meilleure qualité hygiénique chez la chèvre.. Renc. Rech. Ruminants.

Huentelman, M.J., D.W. Craig, A.D. Shieh, J.J. Corneveaux, D. Hu-Lince, J.V. Pearson, and D.A. Stephan. 2005. SNiPer: Improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. BMC Genomics 6:149. doi:10.1186/1471-2164-6-149.

Institut de l'élevage. 2017a. Chiffre clés caprins 2017.
http://www.fnec.fr/IMG/pdf/2017_Chiffres_cles_caprins.pdf

Institut de l'élevage. 2017b. Résultats du contrôle laitier - Espèce Caprine - 2016. 30p.
http://www.fnec.fr/IMG/pdf/Resultats_controle_laitier_caprin_-_France_2016.pdf

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. doi:10.3168/jds.2015-10433.

Karimi, Z., M. Sargolzaei, J.A.B. Robinson, and F.S. Schenkel. 2018. Assessing haplotype-based models for genomic evaluation in Holstein cattle. Can. J. Anim. Sci.. doi:10.1139/CJAS-2018-0009.

Kemper, K.E., C.M. Reich, P.J. Bowman, C.J. vander Jagt, A.J. Chamberlain, B.A. Mason, B.J. Hayes, and M.E. Goddard. 2015. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genetics Selection Evolution 47:29. doi:10.1186/s12711-014-0074-4.

Khatkar, M.S., P.C. Thomson, I. Tammen, and H.W. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genetics Selection Evolution 36:163–190. doi:10.1051/gse:2003057.

Koivula, M., I. Strandén, G. Su, and E.A. Mäntysaari. 2012. Different methods to calculate genomic predictions—Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). Journal of Dairy Science 95:4065–4073. doi:10.3168/jds.2011-4874.

Lamaix, B. 2004. Comment devient-on un bouc améliorateur ? Accessed June 4, 2018. https://chevre.reussir.fr/actualites/elevage-caprin-genetique-comment-devient-on-un-bouc-ameliorateur:17391.html.

Larroque, H., J.-M. Astruc, A. Barbat, F. Barillet, D. Boichard, B. Bonaiti, B. Bonaiti, V. Clément, I. David, G. Lagriffoul, I. Palhière, A. Piacère, C. Robert-Granié, and R. Rupp. 2011. National genetic evaluations in dairy sheep and goats in France. Page 62.Annual Meeting of the European Federation of Animal Science (EAAP). 2011-08-29, Stavanger, NOR. Wageningen Academic.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92:4656–4663. doi:10.3168/jds.2009-2061.

Legarra, A., G. Baloche, F. Barillet, J.M. Astruc, C. Soulas, X. Aguerre, F. Arrese, L. Mintegi, M. Lasarte, F. Maeztu, I. Beltrán de Heredia, and E. Ugarte. 2014a. Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. Journal of Dairy Science 97:3200–3212. doi:10.3168/jds.2013-7745.

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014b. Single Step, a general approach for genomic selection. Livestock Science 166:54–65. doi:10.1016/j.livsci.2014.04.029.

Legarra, A., and Z.G. Vitezica. 2015. Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using gene content multiple-trait BLUP. Genetics Selection Evolution 47:89. doi:10.1186/s12711-015-0165-x.

Lourenco, D.A.L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J.I. Weller. 2014. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. Journal of Dairy Science 97:1742–1752. doi:10.3168/jds.2013-6916.

Luan, T., X. Yu, M. Dolezal, A. Bagnato, and T.H. Meuwissen. 2014. Genomic prediction based on runs of homozygosity. Genetics Selection Evolution 46:64. doi:10.1186/s12711-014-0064-6.

Lund, M.S., A.P. de Roos, A.G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genetics Selection Evolution 43:43. doi:10.1186/1297-9686-43-43.

MacLeod, I.M., B.J. Hayes, C.J. Vander Jagt, K.E. Kemper, M. Haile-Mariam, P.J. Bowman, C. Schrooten, and M. Goddard. 2014. A Bayesian Analysis to Exploit Imputed Sequence Variants for QTL discovery. Page, Proceedings, 10th World Congress of Genetics Applied to Livestock Production.

Malécot, G. 1948. Les Mathématiques de l'hérédité. Masson et Cie. Paris. 63p

Manfredi, E., et T. Ådnøy. 2012. Génétique des caprins laitiers. INRA Prod. Anim. 25:233-244.

Manfredi, E., A. Piacere, P. Lahaye, and V. Ducrocq. 2001. Genetic parameters of type appraisal in Saanen and Alpine goats. Livestock Production Science 70:183–189. doi:10.1016/S0301-6226(01)00180-4.

Manichaikul, A., J.C. Mychaleckyj, S.S. Rich, K. Daly, M. Sale, and W.-M. Chen. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867–2873. doi:10.1093/bioinformatics/btq559.

Maroteau, C. 2014. Cartographie fine de QTL pour des caractères d'intérêt pour la filière caprine. Thèse de doctorat en génétique quantitative. Toulouse, INPT.

Martin, P. 2016. Identification et caractérisation fonctionnelle de régions du génome associées à des caractères d'intérêt pour la filière caprine. Thèse de doctorat en génétique moléculaire. Klop. Toulouse III - Paul Sabatier,

Martin, P., I. Palhière, C. Maroteau, P. Bardou, K. Canale-Tabet, J. Sarry, F. Woloszyn, J. Bertrand-Michel, I. Racke, H. Besir, R. Rupp, and G. Tosser-Klopp. 2017. A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content. Scientific Reports 7. doi:10.1038/s41598-017-02052-0.

Martin, P., I. Palhière, C. Maroteau, V. Clément, I. David, G.T. Klopp, and R. Rupp. 2018. Genome-wide association mapping for type and mammary health traits in French dairy goats

identifies a pleiotropic region on chromosome 19 in the Saanen breed. Journal of Dairy Science 101:5214–5226. doi:10.3168/jds.2017-13625.

Martin, P., M. Szymanowska, L. Zwierzchowski, and C. Leroux. 2002. The impact of genetic polymorphisms on the protein composition of ruminant milks. Reprod. Nutr. Dev. 42:433–459. doi:10.1051/rnd:2002036.

Martini, J.W.R., M.F. Schrauf, C.A. Garcia-Baccino, E.C.G. Pimentel, S. Munilla, A. Rogberg-Muñoz, R.J.C. Cantet, C. Reimer, N. Gao, V. Wimmer, and H. Simianer. 2018. The effect of the H−1 scaling factors τ and ω on the structure of H in the single-step procedure. Genetics Selection Evolution 50:16. doi:10.1186/s12711-018-0386-x.

Matukumalli, L.K., C.T. Lawley, R.D. Schnabel, J.F. Taylor, M.F. Allan, M.P. Heaton, J. O'Connell, S.S. Moore, T.P.L. Smith, T.S. Sonstegard, and C.P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4:e5350. doi:10.1371/journal.pone.0005350.

McRae, A.F., J.C. McEwan, K.G. Dodds, T. Wilson, A.M. Crawford, and J. Slate. 2002. Linkage Disequilibrium in Domestic Sheep. Genetics 160:1113–1122.

Meuwissen, T.H. 2009. Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. Genetics Selection Evolution 41:35. doi:10.1186/1297-9686-41-35.

Meuwissen, T.H., J. Odegard, I. Andersen-Ranberg, and E. Grindflek. 2014a. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. Genetics Selection Evolution 46:49. doi:10.1186/1297-9686-46-49.

Meuwissen, T.H.E., 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.E., A. Sonesson, and J. Odegard. 2014b. The basis of genetic relationships in the era of genomic selection. Proceedings, 10 th World Congress of Genetics Applied to Livestock Production.

Misztal, I. 2017. Studies on inflation of GEBV in single-step GBLUP for type. Interbull Bulletin 0.

Misztal, I., S.E. Aggrey, and W.M. Muir. 2013a. Experiences with a single-step genome evaluation1. Poult Sci 92:2530–2534. doi:10.3382/ps.2012-02739.

Misztal, I., I. Aguilar, A. Legarra, and T.J. Lawlor. 2010. Choice of parameters for single-step genomic evaluation for type. Journal of Dairy Science 93:533–533.

Misztal, I., S. Tsuruta, D.A.L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z.G. Vitezica. 2016. Manual for BLUPF90 Family of Programs. Athens: University of Georgia.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D.H. Lee. 2002. BLUPF90 and related programs. Commun. No. 28–07. Page, Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France.

Misztal, I., Z.G. Vitezica, A. Legarra, I. Aguilar, and A.A. Swan. 2013b. Unknown-parent groups in single-step genomic evaluation. Journal of Animal Breeding and Genetics 130:252–258. doi:10.1111/jbg.12025.

Moser, G., M.S. Khatkar, B.J. Hayes, and H.W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genetics Selection Evolution 42:37. doi:10.1186/1297-9686-42-37.

Mucha, S., R. Mrode, M. Coffey, M. Kizilaslan, S. Desire, and J. Conington. 2018. Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. Journal of Dairy Science 101:2213–2225. doi:10.3168/jds.2017-12919.

Mucha, S., R. Mrode, I. MacLaren-Lee, M. Coffey, and J. Conington. 2015. Estimation of genomic breeding values for milk yield in UK dairy goats. Journal of Dairy Science 98:8201–8208. doi:10.3168/jds.2015-9682.

Palhière, I. 2015. L'assignation, un nouveau moyen d'obtenir des parentés. 5eme journées technique caprine. 31 mars et 1er avril 2015. Saint-Jean-de-Sixt.

Palhière, I., V. Clément, P. Martin, and J.J. Colleau. 2014. Bilan de la méthode de Sélection à Parenté Minimum après 6 ans d'application dans le schéma de sélection caprin. Renc. Rech. Ruminants 21:253–256.

Palti, Y., R.L. Vallejo, G. Gao, S. Liu, A.G. Hernandez, C.E.R. Iii, and G.D. Wiens. 2015. Detection and Validation of QTL Affecting Bacterial Cold Water Disease Resistance in Rainbow Trout Using Restriction-Site Associated DNA Sequencing. PLOS ONE 10:e0138435. doi:10.1371/journal.pone.0138435.

Patry, C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. Journal of dairy science 94:1011–20. doi:10.3168/jds.2010-3804.

Piacère, A., N. Bouloc-Duval, J.P. Sigwald, C. Larzul, and E. Manfredi. 1997. Utilisation de l'index combiné caprin et du polymorphisme de la caséine alpha s1 dans le schéma de sélection caprin.Piacère, A., et al. " " Renc. Rech. Rum 4 (1997): 187-190.. Rencontre Recherche Ruminants 4:187–190.

Porto-Neto, L.R., J.W. Kijas, and A. Reverter. 2014. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genetics Selection Evolution 46:22. doi:10.1186/1297-9686-46-22.

Pszczola, M., T. Strabel, H.A. Mulder, and M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of Dairy Science 95:389–400. doi:10.3168/jds.2011-4338.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575. doi:10.1086/519795.

Ramos, A.M., R.P.M.A. Crooijmans, N.A. Affara, A.J. Amaral, A.L. Archibald, J.E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M.S. Hansen, J. Hedegaard, Z.-L. Hu, H.H. Kerstens, A.S. Law, H.-J. Megens, D. Milan, D.J. Nonneman, G.A. Rohrer, M.F. Rothschild, T.P.L. Smith, R.D. Schnabel, C.P. Van Tassell, J.F. Taylor, R.T. Wiedmann, L.B. Schook, and M.A.M. Groenen. 2009. Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. PLoS ONE 4:e6524. doi:10.1371/journal.pone.0006524.

Raymond, B., A.C. Bouwman, C. Schrooten, J. Houwing-Duistermaat, and R.F. Veerkamp. 2018. Utility of whole-genome sequence data for across-breed genomic prediction. Genetics Selection Evolution 50:27. doi:10.1186/s12711-018-0396-8.

Robert-Granie, C., A. Legarra, and V. Ducrocq. 2011. Principes de base de la sélection génomique. INRA Prod. Anim. 24:331–340.

Rogers, A.R., and C. Huff. 2009. Linkage Disequilibrium Between Loci With Unknown Phase. Genetics 182:839–844. doi:10.1534/genetics.108.093153.

Rolf, M.M., J.F. Taylor, R.D. Schnabel, S.D. McKay, M.C. McClure, S.L. Northcutt, M.S. Kerley, and R.L. Weaber. 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genetics 11:24. doi:10.1186/1471-2156-11-24.

de Roos, A.P.W., B.J. Hayes, R.J. Spelman, and M.E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179:1503–1512. doi:10.1534/genetics.107.084301.

Rupp, R., V. Clément, A. Piacere, C. Robert-Granié, and E. Manfredi. 2011. Genetic parameters for milk somatic cell score and relationship with production and udder type traits in dairy Alpine and Saanen primiparous goats. Journal of Dairy Science 94:3629–3634. doi:10.3168/jds.2010-3694.

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. doi:10.1186/1471-2164-15-478.

Sargolzaei, M., and F.S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25:680–681. doi:10.1093/bioinformatics/btp045.

Schrooten, C., H. Bovenhuis, W. Coppieters, and J.A.M. Van Arendonk. 2000. Whole Genome Scan to Detect Quantitative Trait Loci for Conformation and Functional Traits in Dairy Cattle. Journal of Dairy Science 83:795–806. doi:10.3168/jds.S0022-0302(00)74942-3.

Shen, X., M. Alam, F. Fikse, and L. Rönnegård. 2013. A Novel Generalized Ridge Regression Method for Quantitative Genetics. Genetics genetics.112.146720. doi:10.1534/genetics.112.146720.

Skapetas, B., and V. Vampidis. 2016. Goat production in the World: present situation and trends. Livest Res Rural Dev 28:200.

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. doi:10.2527/jas.2007-0010.

Solberg, T.R., A.K. Sonesson, J.A. Woolliams, J. Ødegard, and T.H. Meuwissen. 2009. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. Genetics Selection Evolution 41:53. doi:10.1186/1297-9686-41-53.

Su, G., O.F. Christensen, L. Janss, and M.S. Lund. 2014. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. Journal of Dairy Science 97:6547–6559. doi:10.3168/jds.2014-8210.

Szyda, J., E. Ptak, J. Komisarek, and A. Żarnecki. 2008. Practical application of daughter yield deviations in dairy cattle breeding. Journal of Applied Genetics 49:183–191. doi:10.1007/bf03195611.

Teissier, M., H. Larroque, and C. Robert-Granié. 2019. Accuracy of genomic evaluation with weighted single step GBLUP for milk production traits, udder type traits and somatic cell scores in French dairy goats. Journal of Dairy Science. doi:10.3168/jds.2018-15650.

Teissier, M., H. Larroque, and C. Robert-Granié. 2018a. Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene. Genetics Selection Evolution 50:31. doi:10.1186/s12711-018-0400-3.

Teissier, M., M.P. Sanchez, M. Boussaha, A. Barbat, C. Hoze, C. Robert-Granie, and P. Croiseau. 2018b. 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. doi:10.3168/jds.2017-13587.

Teo, Y.Y. 2008. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Current Opinion in Lipidology 19:133. doi:10.1097/MOL.0b013e3282f5dd77.

Tosser-Klopp, G., P. Bardou, O. Bouchez, C. Cabau, R. Crooijmans, Y. Dong, C. Donnadieu-Tonon, A. Eggen, H.C.M. Heuven, S. Jamli, A.J. Jiken, C. Klopp, C.T. Lawley, J. McEwan, P. Martin, C.R. Moreno, P. Mulsant, I. Nabihoudine, E. Pailhoux, I. Palhière, R. Rupp, J. Sarry, B.L. Sayre, A. Tircazes, Jun Wang, W. Wang, W. Zhang, and and the International Goat Genome Consortium. 2014. Design and Characterization of a 52K SNP Chip for Goats. PLoS ONE 9:e86227. doi:10.1371/journal.pone.0086227.

Tsuruta, S., I. Misztal, I. Aguilar, and T.J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. Journal of Dairy Science 94:4198–4204. doi:10.3168/jds.2011-4256.

Uemoto, Y., S. Sato, T. Kikuchi, S. Egawa, K. Kohira, H. Sakuma, S. Miyashita, S. Arata, T. Kojima, and K. Suzuki. 2017. Genomic evaluation using SNP- and haplotype-based genomic relationship matrices in a closed line of Duroc pigs. Animal Science Journal 88:1465–1474. doi:10.1111/asj.12805.

Vallejo, R.L., T.D. Leeds, G. Gao, J.E. Parsons, K.E. Martin, J.P. Evenhuis, B.O. Fragomeni, G.D. Wiens, and Y. Palti. 2017. Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. Genetics Selection Evolution 49:17. doi:10.1186/s12711-017-0293-6.

Van Sickle, J. 2003. Analyzing Correlations Between Stream and Watershed Attributes1. JAWRA Journal of the American Water Resources Association 39:717–726. doi:10.1111/j.1752-1688.2003.tb03687.x.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91:4414–4423. doi:10.3168/jds.2007-0980.

VanRaden, P.M., M.E. Tooker, J.R. O'Connell, J.B. Cole, and D.M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genetics Selection Evolution 49:32. doi:10.1186/s12711-017-0307-4.

VanRaden, P.M., and G.R. Wiggans. 1991. Derivation, Calculation, and Use of National Animal Model Information. Journal of Dairy Science 74:2737–2746. doi:10.3168/jds.S0022-0302(91)78453-1.

Venot, E., D. Boichard, V. Ducrocq, S. Fritz, H. Larroque, F. Tortereau, J.-M. (Institut de l'Elevage Astruc, A. Barbat, M. Barbat, A. Baur, P. Boulesteix, C. Carillier-Jacquin, P. Croiseau, M.-N. Fouilloux, A. Gion, C. Hoze, A. Launay, R. Lefebvre, A. Legarra, V. Loywyck, I. Palhière, F. Phocas, J. Promp, C. Robert-Granié, R. Rupp, R. Saintilan, M.-P. Sanchez, T. Tribout, A. Vinet, and S. Mattalia. 2017. French genomic experience : genomics for all ruminant species.

Villumsen, T.M., and L. Janss. 2009. Bayesian genomic selection: the effect of haplotype length and priors. BMC Proceedings 3:S11. doi:10.1186/1753-6561-3-S1-S11.

Vitezica, Z.G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genetics Research (Camb) 93:357–366. doi:10.1017/S001667231100022X.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research (Camb) 94:73–83. doi:10.1017/S0016672312000274.

Wang, L., P. Sørensen, L. Janss, T. Ostersen, and D. Edwards. 2013. Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. BMC Genet 14:115. doi:10.1186/1471-2156-14-115.

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus. 2012. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics genetics.112.146290. doi:10.1534/genetics.112.146290.

Wolc, A., J. Arango, P. Settar, J.E. Fulton, N.P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D.J. Garrick, and J.C. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genetics Selection Evolution 43:23. doi:10.1186/1297-9686-43-23.

Wolc, A., A. Kranis, J. Arango, P. Settar, J.E. Fulton, N.P. O'Sullivan, A. Avendano, K.A. Watson, J.M. Hickey, G. de los Campos, R.L. Fernando, D.J. Garrick, and J.C.M. Dekkers. 2016. Implementation of genomic selection in the poultry industry. Animal Frontiers 6:23–31. doi:10.2527/af.2016-0004.

Wray, N.R., J. Yang, B.J. Hayes, A.L. Price, M.E. Goddard, and P.M. Visscher. 2013. Pitfalls of predicting complex traits from SNPs. Nature Reviews Genetics 14:507–515. doi:10.1038/nrg3457.

Zahra, K. 2018. Assessing Haplotype-Based Models for Genomic Evaluation in Holstein Cattle - Canadian Journal of Animal Science. Accessed June 20, 2018. http://www.nrcresearchpress.com/doi/abs/10.1139/CJAS-2018-0009.

Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. Frontiers in Genetics 7:151. doi:10.3389/fgene.2016.00151.

Zhang, Z., X. Ding, J. Liu, D.-J. de Koning, and Q. Zhang. 2011. Genomic selection for QTL-MAS data using a trait-specific relationship matrix. BMC Proceedings 5:S15. doi:10.1186/1753-6561-5-S3-S15.

Zhang, Z., M. Erbe, J. He, U. Ober, N. Gao, H. Zhang, H. Simianer, and J. Li. 2015. Accuracy of Whole-Genome Prediction Using a Genetic Architecture-Enhanced Variance-Covariance Matrix. G3: Genes Genomes Genetics 5:615–627. doi:10.1534/g3.114.016261.

Zhou, L., R. Mrode, S. Zhang, Q. Zhang, B. Li, and J.-F. Liu. 2018. Factors affecting GEBV accuracy with single-step Bayesian models. Heredity 120:100–109. doi:10.1038/s41437-017-0010-9.