THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (INP Toulouse)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

Mme CELINE CARILLIER

le vendredi 16 octobre 2015

Titre :

ETUDE DE LA PREDICTION GENOMIQUE CHEZ LES CAPRINS: FAISABILITE ET LIMITES DE LA SELECTION GENOMIQUE DANS LE CADRE D'UNE POPULATION MULTIRACIALE ET A FAIBLE EFFECTIF

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-GRANIÉ

MME HELENE LARROQUE

Rapporteurs :

M. NICOLAS GENGLER, UNIVERSITE DE LIEGE

M. VINCENT DUCROCQ, INRA JOUY EN JOSAS

Membre(s) du jury :

M. LEOPOLDO SANCHEZ RODRIGUEZ, INRA ORLEANS, Président

M. FRANCOIS GUILLAUME, EVOLUTION, Membre

Mme CHRISTELE ROBERT-GRANIÉ, INRA TOULOUSE, Membre

Mme HELENE LARROQUE, INRA TOULOUSE, Membre

Remerciements

Cette thèse a été financée par la région Midi-Pyrénées ainsi que par le métaprogramme SELGEN (Xgen) de l'INRA que je tiens à remercier pour m'avoir permis de réaliser ce travail. Je remercie également les programmes Genovicap et Phénofinlait (ANR, Apis-Gène, CASDAR, FranceAgriMer, France Génétique Élevage et le Ministère de l'Agriculture Française) qui ont participé au financement des génotypages des animaux nécessaires à mes recherches.

Je remercie les membres du jury de cette thèse : Vincent Ducrocq, Nicolas Gengler, Leopoldo Sanchez et François Guillaume. Je tiens également à remercier les membres de mon comité de thèse : Pascal Croiseau, Andrés Legarra, Pascale Leroy et Florence Phocas, pour leurs conseils et leur regard critique sur mon travail de thèse.

Je remercie bien évidemment mes directrices de thèse, Christèle Robert-Granié et Hélène Larroque, qui m'ont encadrée pendant ces trois années de thèse ainsi que pendant les six mois de stage qui ont précédés. Je tiens à leur faire part de ma gratitude pour leur encadrement bien veillant qui a conseillé mon travail tout au long de ces trois ans, tout en me laissant une certaine autonomie. Je tiens à les remercier également pour leur gentillesse. Cet encadrement m'a été très bénéfique et je tiens à les remercier pour sa qualité que je sais être rare.

Je souhaite également remercier particulièrement Hélène Larroque et Virginie Clément pour m'avoir ouvert les portes du monde de la recherche avec mon premier stage de fin d'étude. Je les remercie de m'avoir encouragée à continuer dans cette voie et d'avoir toujours été bienveillantes à mon égard. Je remercie également tout le groupe « caprin », Rachel Rupp, Gwenola-Tosser Klopp, Isabelle Palhière, Christophe Huau, Virginie Clément et bien sûr Hélène Larroque et Christèle Robert-Granié dont les quelques réunions ont su orienter mon travail. Je remercie également ce groupe pour l'intérêt porté à cette thèse qui a su être un moteur de motivation pour moi.

Je remercie également pour leur chaleureux accueil les membres de l'unité GenPhySE, et plus particulièrement les équipes GesPR et MG^2 auxquelles j'ai été intégrée pendant ma thèse, ainsi que les secrétaires toujours disponibles et bienveillantes. Je remercie également pour sa gentillesse Hervé Garreau mon collègue de bureau pendant plus de 3 ans.

Et comment ne pas remercier tous mes amis rencontrés à l'INRA : Charlotte, Chloé, Diane, Marie-Line, Mathilde, Maxime, Mélanie, Morgane, Pauline, Réjane, Samira, Sophie,

Yoannah, et celles avec qui je partage les bons moments « cheval » du lundi soir Claire, Gaëlle, Diane et Réjane bien sûr ! Merci à tous pour tous ces bons moments passés ensemble.

Je remercie également ma famille pour son soutien et ses encouragements durant cette thèse. Merci à mes parents et mes grands-parents qui m'ont permis d'en arriver là et m'ont encouragée et soutenue tout au long de mon parcours dans la réussite comme dans l'échec.

Enfin, je ne saurai terminer sans remercier celui qui est là chaque jour pour moi et qui étant passé par là, a su être présent pendant ces 3 années de thèse. Je ne te remercierai jamais assez d'être à mes côtés chaque jour (ou presque…) et de me soutenir toujours dans les bons et les moins bons moments. Merci aussi, même si je ne te le dis pas assez, d'avoir su faire entrer le bonheur dans ma vie et d'être le moteur de mes projets.

Et pour ceux qui se souviennent de cette phrase qui m'a permis d'avancer : « l'important ce n'est pas ce que vous êtes aujourd'hui mais ce que vous serez demain », je crois que demain … c'est maintenant ! Merci à tous d'avoir fait de moi ce que je suis aujourd'hui.

Carpe Diem

Résumé

La sélection génomique, qui a révolutionné la sélection génétique des bovins laitiers notamment, est désormais envisagée dans d'autres espèces comme l'espèce caprine. La clé du succès de la sélection génomique réside dans la précision des évaluations génomiques. Chez les caprins laitiers français, le gain de précision attendu avec la sélection génomique était un des questionnements de la filière en raison de la petite taille de la population de référence disponible (825 mâles et 1945 femelles génotypés sur une puce SNP 50K). Le but de cette étude est d'évaluer comment augmenter la précision des évaluations génomiques dans l'espèce caprine. Une étude de la structure génétique de la population de référence caprine constituée d'animaux de races Saanen et Alpine, a permis de montrer que la population de référence choisie est représentative de la population élevée sur le territoire français. En revanche, les faibles niveaux de déséquilibre de liaison (0,17 entre deux SNP consécutifs) de consanguinité et de parenté au sein de la population, similaires à ceux trouvés en ovins Lacaune, ne sont pas idéaux pour obtenir une bonne précision des évaluations génomiques. De plus, malgré l'origine commune des races Alpine et Saanen, leurs structures génétiques suggèrent qu'elles se distinguent clairement d'un point de vue génétique. Les méthodes génomiques (GBLUP ou Bayésiennes) « two-step », basées sur des performances pré-corrigées (DYD, EBV dérégressées) n'ont pas permis une amélioration significative des précisions des évaluations génomiques pour les caractères évalués en routine (caractères de production, de morphologie et de comptages de cellules somatiques) chez les caprins laitiers. La prise en compte des phénotypes des mâles non génotypés permet d'augmenter les précisions des évaluations de 3 à 47% selon le caractère. L'ajout des génotypes de femelles issues d'un dispositif de détection de QTL améliore également les précisions (de 2 à 14%) que ce soit pour les évaluations two steps ou les évaluations basées sur les performances propres des femelles (single step). Les précisions sont augmentées de 10 à 74% avec les évaluations single step comparées aux évaluations two steps, ce qui permet d'atteindre des précisions supérieures à celles obtenues sur ascendance. Les précisions obtenues avec les évaluations génomiques multiraciales, bicaractères et uniraciales sont similaires même si la précision des valeurs génomiques estimées des candidats à la sélection est plus élevée avec les évaluations multiraciales. La sélection génomique est donc envisageable chez les caprins laitiers français à l'aide d'un modèle génomique multiracial single step. Les précisions peuvent être légèrement augmentées par l'inclusion de gènes majeurs tels que celui de la caséine α_{s1}

notamment à l’aide d’un modèle « gene content » pour prédire le génotype des animaux non génotypés.

Abstract

Genomic selection which is revolutionizing genetic selection in dairy cattle has been tested in several species like dairy goat. Key point in genomic selection is accuracy of genomic evaluation. In French dairy goats, gain in accuracy using genomic selection was questioning due to the small size of the reference population (825 males and 1 945 females genotyped). The aim of this study was to investigate how to reach adequate genomic evaluation accuracy in French dairy goat population. The study of reference population structure (Alpine and Saanen breeds) showed that reference population is similar to the whole population of French dairy goats. But the weak level of linkage disequilibrium (0.17 between two consecutive SNP), inbreeding and relationship between reference and candidate population were not ideal to maximize genomic evaluation accuracy. Despite their common origin, genetic structure of Alpine and Saanen breeds suggested that they were genetically distant.

Two steps genomic evaluation (GBLUP, Bayesian) based on performances corrected for fixed effect (DYD, deregressed EBV) did not improve genetic evaluation accuracy compared to classical evaluations for milk production traits, udder type traits and somatic cells score classically selected in French dairy goat. Taking into account phenotypes of ungenotyped sires increased genomic evaluation from 3 to 47% depending on the trait considered. Adding female genotypes also improved genomic evaluation accuracies from 2 to 4% depending on the method (two steps or single step) and on the trait. When using gemomic evaluation directly based on female performances (single step), accuracy of genomic evaluation reach the level obtained from ascendance in classic evaluation which was not the case using two steps evaluations. Genomic evaluation accuracies were similar when using multiple-trait model, multi-breed or single breed evaluation. But accuracies derived from prediction error variances were better when using multi-breed genomic evaluations.

Genomic selection is feasible in French dairy goats using single step multi-breed genomic evaluations. Accuracies could be slightly improved integrating major gene as α_{s1} casein especially when using « gene content » approach to predict genotypes of ungenotyped animals.

Table des matières

Liste des abréviations

AAR : qualité de l'attache arrière
ACP : analyse en composantes principales
AVP : forme de l'avant pis
BCP : BayesCπ
BCPest : BayesCπ avec π estimé par le programme
BL : Lasso bayésien
BLUP : meilleur prédicteur linéaire non biaisé (best linear unbiased predictor)
CAEV : arthrite encéphalite caprine à virus
CCS : comptages de cellules somatiques
CD : coefficient de détermination
DL : déséquilibre de liaison
EBV : valeur génétique estimée (estimated breeding value)
EM : Espérance Maximisation (expectation-maximization)
DREBV : valeur génétique dé-régressée
GBLUP : meilleur prédicteur linéaire génomique non biaisé (genomic best linear unbiased predictor)
GEBV : valeur génomique estimée (genomic estimated breeding value)
IA : insémination animale
IGGC : consortium international caprin (International Goat Genome Consortium)
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
PEV : variance d'erreur de prédiction (Prediction Error Variance)
PLA : distance plancher-jarret
PRM : profil de la mamelle
REML estimateur du maximum de vraisemblance restreinte (restricted maximum likelihood)
TB : taux butyreux
TP : taux protéique
SCS : score de cellules somatiques
SNP : polymorphisme d'une seule base nucléotidique (single nucleotide polymorphism)

Introduction

Les avancées dans le domaine du séquençage humain (dans les années 2000) ainsi que dans celui des animaux d'élevage ont permis de mettre en évidence un nouveau type de marqueurs moléculaires : les SNP (single nucleotide polymorphism) dont le polymorphisme consiste en la variation d'une base nucléotidique. L'arrivée d'outils comme les puces à SNP contenant plusieurs milliers de marqueurs (54 609 SNP pour la puce Illumina 50k bovine en 2007) permet de mieux comprendre le génome, ce qui présente un avantage pour la sélection génétique des animaux. Cette nouvelle technologie a révolutionné la sélection génétique des bovins laitiers (Boichard et al., 2012).

Avant la découverte des marqueurs génétiques, les évaluations génétiques étaient basées uniquement sur l'utilisation de l'information des performances (phénotypes) mesurées sur des individus et de leurs apparentés, ainsi que sur la connaissance des liens de parentés entre individus (pédigrée). En l'absence d'identification de l'ensemble des gènes régissant les caractères d'intérêt zootechnique, l'évaluation génétique des animaux repose sur le modèle polygénique infinitésimal : un caractère quantitatif est gouverné par un très grand nombre de loci ayant des effets individuels faibles et indépendants. Les évaluations génétiques conduisent ainsi à l'estimation de la valeur génétique d'un individu qui correspond à son potentiel génétique susceptible d'être transmis, en espérance, par moitié à sa descendance. Les reproducteurs ayant les valeurs génétiques estimées les plus élevées sont retenus pour procréer la génération suivante. La sélection des allèles favorables est donc implicite puisqu'ils sont captés à travers la connaissance des phénotypes sans avoir besoin de connaitre les gènes. Mais il est nécessaire de mesurer les phénotypes, c'est-à-dire en ruminants laitiers, d'obtenir les lactations des filles (testage sur descendance). La connaissance du génome peut cependant permettre d'améliorer la précision de l'estimation des valeurs génétiques des individus en particulier lorsque les phénotypes ne sont pas (ou pas encore) connus. La sélection génétique classique ne permet pas, par exemple, de distinguer deux pleins-frères à la naissance puisque l'espérance de leur valeur génétique correspond à la moyenne des valeurs génétiques estimées de leurs parents et que l'aléa de méiose (écart à cette moyenne propre à chacun) ne peut être évalué.

Un premier type de marqueurs génétiques, les microsatellites, découverts dans les années 1990 (en 1996 en caprins (Vaiman et al., 1996)), ont été utilisés pour la sélection génétique de certaines espèces d'élevage. Chez les bovins laitiers français, ces marqueurs ont été utilisés afin de mettre en place un programme de sélection assistée par marqueurs (SAM)

dès la fin de l'année 2000 pour les trois grandes races laitières. Ce programme consistait en une première étape de détection de loci à effet quantitatif (QTL pour quantitative trait locus). L'information des marqueurs flanquants les QTL, ainsi que les performances des apparentés (demi-sœurs, oncles et grand-oncles), permettaient ensuite d'estimer la valeur génétique des jeunes candidats pour leur pré-sélection (Guillaume et al., 2012). En ovins et caprins laitiers, ces marqueurs ont été utilisés afin de réaliser une sélection assistée par gène. Cette méthode a permis de sélectionner les animaux résistants pour le gène prion protein gene (Prp) afin d'éradiquer la tremblante en ovins laitiers. En caprins, elle permet d'éliminer les mâles candidats à la sélection porteur des allèles défavorables pour le gène de la caséine α_{s1} qui a un effet fort sur le taux protéique.

Les importants progrès technologiques survenus dans l'étude du génome ont permis de développer un nouveau type de marqueurs : les SNP. La densité de marquage des puces SNP étant beaucoup plus élevée que celles des puces microsatellites, le déséquilibre de liaison capté est beaucoup plus important avec les puces SNP. Exploitant ces nouvelles informations, Meuwissen et al. en 2001 ont proposé une première définition de la sélection génomique : la valeur génétique d'un individu correspond dans ce cas à la somme des effets de tous les marqueurs SNP. Les effets des marqueurs sont estimés grâce à une population de référence comprenant des individus génotypés pour l'ensemble des SNP et phénotypés. Sur la base de cette définition, dès 2008, des premières évaluations génomiques ont été réalisées pour les trois grandes races bovines laitières françaises (Holstein, Montbéliarde et Normande) (Guillaume et al., 2012). La sélection génomique connait un franc succès en bovins laitiers car elle permet de sélectionner les candidats dès la naissance, avec une précision suffisante permettant de supprimer le testage sur descendance très coûteux. L'engouement pour la sélection génomique dans les autres espèces d'élevage (bovins allaitants, petits ruminants, porcs, poules, chevaux, poissons) est donc très fort et de nombreux projets de génomique ont été lancés pour l'ensemble de ces espèces dans les années 2010. De récents travaux en ovins laitiers Lacaune (Buisson, 2012; Baloche et al., 2013) ont permis la mise en place de la sélection génomique au début de l'année 2015. Cependant, l'avantage de la sélection génomique en termes de coût et de progrès génétique pour les autres espèces ne semble pas être aussi favorable qu'en bovins laitiers.

En caprins laitiers, la disponibilité d'une puce 50k en 2011 a poussé la filière caprine à s'intéresser à la possible mise en place d'une sélection génomique dans cette espèce. Un projet de détection de QTL débuté en 2011 a permis d'identifier de nombreuses zones d'intérêt dans l'espèce caprine pour les caractères en sélection (Maroteau, 2014). L'intérêt de

la sélection génomique dans les schémas de sélection petits ruminants (ovins et caprins laitiers) a été étudiée en amont (avant la disponibilité de la puce caprine) sur la base d'un modèle déterministe (Shumbusho et al., 2013). Cette étude montre que le gain génétique pourrait être augmenté de 26% avec la sélection génomique en race Alpine, mais cela nécessiterait le génotypage de 500 candidats par an, ce qui paraît peu réaliste étant donné que seuls 40 mâles Alpins sont testés sur descendance chaque année.

Comme en bovins laitiers, le testage sur descendance en caprins est long et coûteux et pourrait être arrêté avec la mise en place de la sélection génomique. Mais cet arrêt suppose que la précision des valeurs génomiques estimées des candidats à leur naissance soit suffisante, c'est-à-dire dans l'idéal, égale à celle obtenue après testage sur descendance et au minimum supérieure à celle obtenue sur ascendance. Il semble donc nécessaire d'étudier le gain potentiel de précision obtenu avec la sélection génomique avant d'envisager de la mettre en place en routine. Une première étude en 2012 (Carillier, 2012) réalisée à partir des génotypes disponibles au moment de l'étude (pour 67 mâles ainsi et environ 2 000 femelles) ainsi que de génotypes simulés (700 mâles) a montré que les précisions des évaluations génomiques n'atteignaient pas celles obtenues sur ascendance. La principale raison évoquée pour expliquer ces résultats était la petite taille de la population de référence (environ 800 mâles avec génotypes simulés considérés dans l'étude).

Une étude supplémentaire basée sur l'ensemble des animaux génotypés était donc nécessaire afin d'étudier l'intérêt de la sélection génomique dans l'espèce caprine. Aux vues des résultats de la précédente étude, il semblait nécessaire de déterminer les conditions permettant d'obtenir des précisions d'évaluations génomiques suffisantes pour envisager la mise en place la sélection génomique dans l'espèce caprine. Le but de ma thèse est donc de répondre à cette question. Etant donné la petite taille de la population caprine, ma thèse s'est concentrée sur l'évaluation de différents modèles et méthodes d'évaluation génomique afin de déterminer la modélisation permettant les plus grands gains de précision.

Le chapitre 1 est consacré à la description du contexte de la sélection génétique caprine en France, à une présentation des principaux facteurs influençant la précision des évaluations génomiques ainsi que des différentes méthodes et modèles d'évaluation génomique.

Le chapitre 2 est consacré à l'étude de la structure de la population de référence caprine afin d'identifier ses points faibles et ses points forts vis-à-vis de la sélection génomique. Ils permettront de proposer par la suite différentes stratégies de génotypage, qui seront discutées dans la partie conclusion.

Les chapitres 3 et 4 présentent les différents modèles d'évaluations génomiques que nous avons testées incluant notamment l'étude de l'intérêt d'une population de référence multiraciale ou encore l'intérêt de la prise en compte de génotypes femelles dans les modèles génomiques.

Le cinquième chapitre est consacré à l'étude de modèles d'évaluations génétiques et/ou génomiques prenant en compte l'effet du gène majeur de la caséine α_{s1}.

Enfin, des perspectives à ce travail sont proposées et discutées dans la conclusion générale, notamment en termes de modélisations génomiques, stratégies de génotypages et d'évolution du schéma de sélection caprins.

Chapitre 1 : Synthèse bibliographique

1.I Contexte de la filière caprine vers la génomique

L'élevage des caprins dans le monde représente environ 840 millions de chèvres. Les effectifs sont essentiellement concentrés en Asie avec 64% des effectifs, en Afrique (29%), en Amérique (5%) et en Europe (2%). Dans la majorité des pays sauf en Amérique du nord et en Europe la filière caprine est orientée vers la production de viande. Seuls quelques pays possèdent des programmes de sélection : l'Australie, l'Afrique du Sud et les Etats-Unis pour la filière viande et l'Europe (France, Norvège et Espagne) et le Canada pour la filière laitière. Des petits programmes de sélection sont en cours de mise en place au Brésil et au Mexique. La France, un des premiers pays producteur de lait de chèvre de l'Union Européenne, bénéficie du programme de sélection le plus organisé au monde. C'est également le premier pays à se tourner vers la génomique caprine avec le génotypage des premiers animaux en 2012, ce qui en fait le leader mondial de la sélection génomique caprine. Pour ces raisons, ce premier chapitre est focalisé sur l'étude de la filière caprine française uniquement.

1.I.1 La filière caprine en France

En France, l'élevage caprin est essentiellement tourné vers la production laitière (de 75 à 93% du produit brut de l'élevage caprin français) mais également vers la production de viande, et de phanères pour la race Angora. La viande est en général un coproduit du lait et du fromage de chèvre (Patier, 2012). La production française de lait de chèvre est majoritairement (plus de 80%) destinée à la transformation fromagère. En 2013, environ 20% de la production était transformée en fromage en ferme, ce qui représente la moitié des éleveurs français.

La production de lait de chèvre en France est la 5ème plus importante production mondiale avec 643 millions de litres produits par an. Cependant, la filière caprine française ne représente que 1% du produit brut de l'agriculture française (Morand-Fehr et al., 2012). La France produit 30% de la production du lait de chèvre de l'Union Européenne, derrière la Grèce et l'Espagne avec seulement 10% des effectifs européens, c'est-à-dire 998 000 chèvres, soit 2% des effectifs mondiaux (FranceAgriMer et France Génétique Élevage, 2015; France Génétique Élevage, 2014). De plus, la France est reconnue comme le leader mondial de la génétique dans cette espèce puisqu'elle exporte des semences de boucs améliorateurs dans plus de 25 pays (FranceAgriMer et France Génétique Élevage, 2015; France Génétique Élevage, 2014).

En France, la production est essentiellement concentrée (82% de la production totale) dans le Poitou-Charentes (35%), les Pays de la Loire (15%), en Midi-Pyrénées (11%), la région Centre (11%) et la région Rhône-Alpes (10%) (France Agrimer, 2014). Les deux races principalement élevées sont l'Alpine avec 55% des effectifs français et la Saanen avec 42% des effectifs. Ces deux races ont pour origine commune une race élevée dans le Nord des Alpes, mais elles sont sélectionnées séparément en race pure depuis plus de 40 ans (Babo, 2000). Les autres races constituant le cheptel français sont la Poitevine, la Pyrénéenne, la Rove, la Corse, la Provençale, la chèvre des Fossés, la chèvre du Massif Central, l'Angora, la créole ainsi que des chèvres croisées. En 2013, le cheptel français comptait 5 300 élevages professionnels ayant plus de 10 chèvres (Institut de l'élevage, 2013). Mais cette filière caprine a été fortement touchée par la crise économique des cinq dernières années, avec plus de 13% des producteurs de lait de chèvre ayant cessé leur activité entre 2012 et 2013. La taille des élevages français est relativement élevée avec en moyenne 70 femelles pour les éleveurs-fromagers et 233 femelles pour les éleveurs laitiers (Institut de l'élevage, 2013) dont 35% possède plus de 350 animaux (Morand-Fehr et al., 2012).

Les chèvres françaises ont un très bon niveau de productivité par rapport aux chèvres européennes, avec 886kg et 946kg de lait produit en moyenne par lactation respectivement pour les races Alpine et Saanen (Douguet et al., 2013) . Les teneurs en matière protéique et en matière grasse du lait des chèvres françaises sont plus élevées que les moyennes européennes avec un taux protéique moyen de 32.5g/kg et un taux butyreux moyen de 36.8g/kg (FranceAgriMer et France Génétique Élevage, 2015). Les résultats assez exceptionnels de la filière caprine française sont essentiellement dus à l'organisation de la filière, à la motivation des éleveurs et à la coopération efficace entre la recherche et la filière (Manfredi et Adnøy, 2012).

1.I.2 La sélection génétique caprine

L'organisation de la sélection caprine a débuté dans les années 70 avec le premier testage sur descendance et les premières indexations des mâles en ferme dès 1969 (Renard, 2008). Aujourd'hui, le programme de sélection français est conduit par un seul organisme et l'entreprise de sélection, Capgènes, qui possède un centre de testage. Ce programme de sélection combine d'une part, un choix raisonné des individus reproducteurs sur ascendance, via des accouplements programmés, et sur descendance via un important testage sur descendance (plus de 100 filles par père et par an) ; et d'autre part l'enregistrement des

performances laitières et de morphologie mammaire en ferme, de la fonction sexuelle des mâles en centre et des généalogies (père et mère) des animaux (Manfredi et Adnøy, 2012).

L'objectif affiché du programme de sélection caprin commun aux races Alpine et Saanen est d'augmenter la production laitière tout en améliorant la qualité du lait, la morphologie fonctionnelle et les qualités d'élevage (fertilité, précocité, longévité). Les critères de sélection sont d'une part les caractères liés à la production laitière et d'autre part les caractères de morphologie de la mamelle. Les caractères de production sélectionnés sont les quantités de lait, de matière protéique (MP) et de matière grasse (MG) ainsi que les taux butyreux (TB) et protéique (TP). Depuis 2014, le score de cellules somatiques fait partie des caractères sélectionnés. Les critères de sélection de la qualité de la morphologie de la mamelle regroupent le profil de la mamelle (PRM), la distance entre le plancher de la mamelle et le jarret (PLA), la qualité de l'attache arrière (AAR), la forme de l'avant pis (AVP) et l'orientation des trayons (ORT). Ces caractères sont combinés dans un index de synthèse économique appelé index combiné caprin (ICC). Depuis 2012, cet index dépend de l'index de production caprin (IPC) et de l'index morphologique caprin (IMC) : $ICC = IPC + 0{,}6\ IMC$ en race Saanen et $ICC = IPC + 0{,}5\ IMC$ en race Alpine. La formule de l'IPC est la suivante : $IPC = \text{Index MP} + 0{,}4\ \text{Index TP} + 0{,}2\ \text{Index MG} + 0{,}1\ \text{Index TB}$. Il est identique pour les races Saanen et Alpine. L'ensemble des caractères de morphologie est regroupé dans l'IMC avec : $IMC = \text{Index PRM} + \text{Index PLA} + \text{Index AAR} + \text{Index AVP} + \text{Index ORT}$. En septembre 2015, l'index de synthèse inclura aussi le comptage des cellules somatiques (Clément et al., 2014). L'index de synthèse, qui n'était utilisé avant 2012, qu'au niveau du schéma de sélection pour la procréation des boucs d'insémination animale (IA), est actuellement diffusé pour les femelles pointées dont le coefficient de détermination (CD) est supérieur à 0,30 et pour les boucs d'IA (Clément, 2012).

Le programme de sélection caprin ne regroupe qu'une petite partie des éleveurs français (adhérents à Capgènes), puisqu'il n'inclut que 170 000 chèvres ce qui représente 17% du cheptel national. Les performances laitières et les comptages de cellules somatiques de l'ensemble de ces femelles sont enregistrés mensuellement par le biais du contrôle laitier. La totalité des chèvres des éleveurs adhérents au programme de sélection sont également pointées pour les postes de morphologie mammaire et d'aplombs. Le schéma de sélection caprin (Figure 1 .1) est commun aux races Saanen et Alpine bien que la sélection soit conduite séparément dans les deux races. Le programme s'intéresse à la voie mâle de la sélection et est axé principalement autour d'une sélection sévère des boucs reproducteurs. En

effet, entre 1994 et 2004, seulement 4,6% des mâles Saanen et 5,9% des mâles Alpins contre 15% en bovins Holsein (http://www.ciagenesdiffusion.com/genetique-disponible/Holstein-programme-genetique.aspx) issus d'accouplement programmés ont été agréés (Renard, 2008). Cependant, la sélection des mâles caprins est moins drastique que la sélection réalisée en ovins Lacaune où seuls 1,5% des mâles issus d'accouplement programmés sont choisis comme béliers d'IA. Cette différence de pression de sélection avec les ovins laitiers Lacaune est essentiellement due à un tri des béliers avant le testage (sur spermatogenèse) plus sévère qu'en caprins laitiers (18% des mâles conservés contre 35% en caprins). Cette sélection est plus sévère dans cette espèce en raison de l'impossibilité de congeler la semence pour l'IA (Buisson, 2012).

Les mâles candidats au testage sur descendance sont issus de 1 000 accouplements programmés. Ces accouplements sont réalisés entre les meilleurs individus parentaux tout en minimisant l'apparentement afin de limiter la consanguinité à moins de 2% (Colleau et al., 2004). Issus de ces 1 000 accouplements, 200 mâles sont entrés en centre, et seuls 120 seront sélectionnés après les 30 jours de tests sanitaires, de croissance et de conformation.

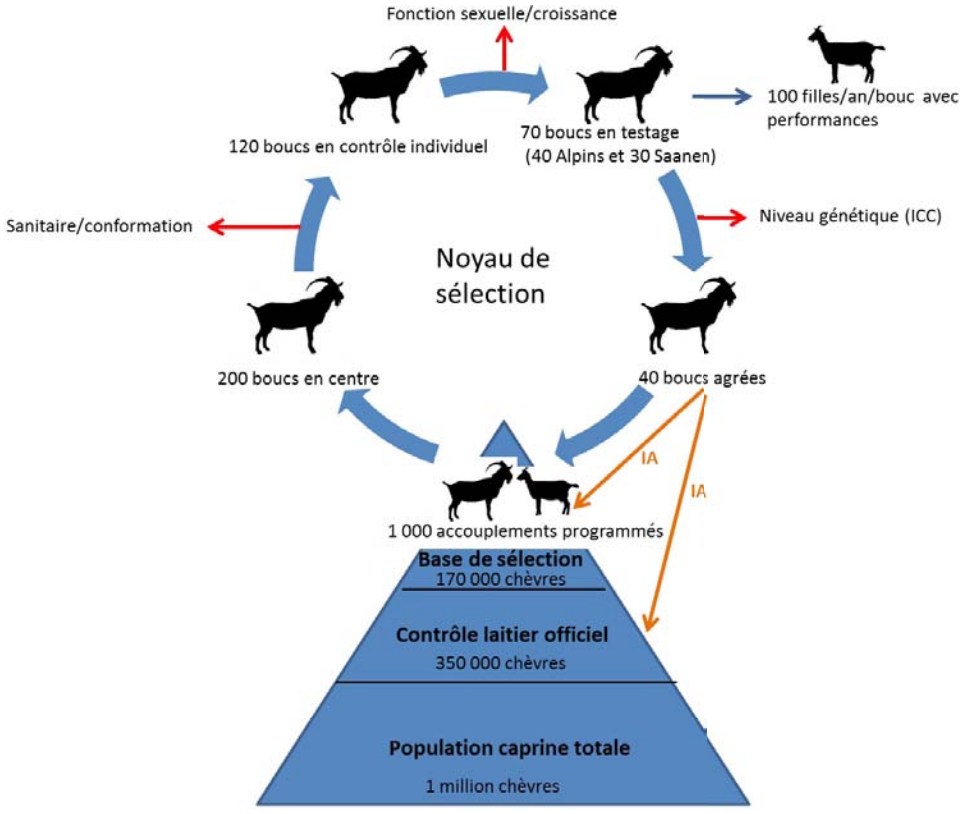


Figure 1.1 : Schémas de sélection caprin des races Alpine et Saanen (source Capgènes)

Les principales causes d'élimination de ces mâles, plus nombreuses en race Alpine qu'en race Saanen, sont des raisons sanitaires (25% pour cause d'arthrite encéphalite caprine à virus (CAEV)), et des problèmes de quantité (volume et concentration en spermatozoïdes de

l'éjaculat), de qualité (motilité et anomalies des spermatozoïdes), et de décongélation de la semence (environ 15%). Enfin, seulement 70 boucs sélectionnés (40 Alpins et 30 Saanens) sur leur comportement sexuel, leur production de semence (qualité et quantité) et l'aptitude de leur semence à être congelée seront testés sur descendance. L'évaluation sur descendance de ces mâles repose sur un testage sur descendance relativement conséquent avec 200 inséminations animales et le contrôle des performances de 80 à 100 filles par bouc et par an. Les 30 à 40 meilleurs boucs sélectionnés sur ICC sont utilisés comme améliorateurs. Sur la série de testage des mâles nés en 2010 (Figure 1 .2), 15 mâles Saanen sur 30 (ICC > 2) et 22 mâles Alpins sur 43 (ICC > 1,8) ont été conservés comme améliorateurs. En fin de carrière, ces boucs améliorateurs auront jusqu'à plus de 2 000 filles réparties sur l'ensemble du territoire. Le renouvellement des boucs améliorateurs est réalisé par accouplement des meilleurs boucs améliorateurs avec les meilleures femelles issues d'IA ayant au minimum 4 lactations enregistrées. Les critères de sélection des meilleurs accouplements (ICC des deux parents) sont propres à chaque race, ce qui peut expliquer des différences d'efficacité de sélection dans les deux races.

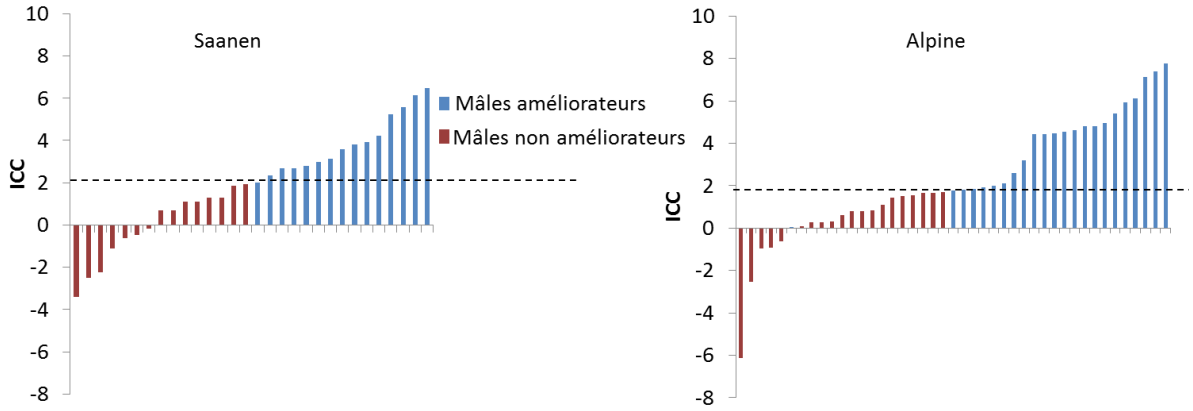


Figure 1.2 : Index Combiné Caprins (ICC) des mâles de testage nés en 2010 par rapport à leur statut (améliorateur ou non)

Le progrès génétique annuel (ΔG), ou augmentation du niveau génétique réalisée en une année est défini par : $\Delta G = \frac{iR\sigma_g}{T}$. L'écart-type génétique (σ_g) est une mesure de la variabilité génétique du caractère, l'intensité de sélection (i) est estimée par la différence entre la valeur génétique moyenne des individus sélectionnés et celle des candidats à la sélection exprimée en écart-type génétique. La précision de l'évaluation génétique (R) définie par la corrélation entre la valeur génétique vraie et la valeur génétique estimée correspond à la racine du CD. L'intervalle de génération (T) correspond à l'âge moyen des parents à la

naissance de leurs descendants. La diffusion du progrès génétique dans la population des femelles est essentiellement réalisée via l'utilisation des boucs améliorateurs en insémination animale. Le taux d'insémination est d'environ 21% sur l'ensemble de la population au contrôle laitier en 2011 et de 40% sur la population des adhérents au schéma de sélection (Clément, 2012). Il est relativement faible comparé aux bovins laitiers et aux ovins laitiers Lacaune (80%), même s'il a augmenté d'environ 15% entre 2002 et 2007 (Renard, 2008). La diffusion du progrès génétique dans la population des femelles est également assurée par l'utilisation des boucs de monte naturelle, fils des boucs améliorateurs. Les intervalles de générations (Tableau 1 .1) qui résultent de ce schéma sont relativement élevés en caprin et limitent la création du progrès génétique. En effet, l'intervalle de génération sur la voie père-fils par exemple est d'environ 5 ans et demi pour les caprins contre environ 4 ans pour les ovins Lacaune (Danchin-Burge, 2011). La diffusion du progrès est également limitée par la production des boucs d'IA ayant une limite physiologique d'environ 10 000 doses par mâle ainsi que par un taux de réussite à l'IA moyen (65%) après synchronisation des chaleurs (Institut de l'élevage, 1997). Ce taux de réussite, meilleur en race Alpine qu'en race Saanen, est cependant plus élevé qu'en bovins laitiers avec des taux de réussite de l'ordre de 50% (Leboeuf et al., 1998).

Tableau 1.1: Intervalles de génération en années pour les quatre voies de transmission des gènes en races Alpine et Saanen

	Alpine	Saanen
Père-fils	5,59	5,61
Mère-fils	3,62	3,42
Père-fille	3,86	3,79
Mère-fille	3,27	3,26

Le progrès génétique ne repose pas seulement sur l'organisation du schéma de sélection, il est également basé sur l'évaluation précise de la valeur génétique des reproducteurs pour les caractères en sélection. Ces évaluations sont réalisées par l'INRA pour les races Alpine, Saanen et Angora en utilisant la méthode BLUP modèle animal. Elles utilisent l'information des performances de toutes les femelles enregistrées au contrôle laitier ce qui représente 350 000 lactations par an, soit environ 40% du cheptel national (Manfredi et Adnøy, 2012). L'ensemble des lactations des femelles depuis le 1er septembre 1979 d'une durée minimum de 47 jours pour les lactations terminées et de 76 jours pour les lactations en cours au moment de l'indexation sont utilisées (7 614 194 lactations en 2012 (Clément, 2012)). Cependant environ 10% des lactations enregistrées sont invalidées car non conformes au règlement technique du contrôle laitier qui impose des règles strictes notamment sur

l'intervalle entre deux contrôles dans un élevage. Ces lactations ne sont donc pas utilisées pour les évaluations génétiques. Selon leur durée, les lactations en cours ou terminées sont extrapolées, tronquées ou corrigées pour que leur durée soit équivalente à 250 jours. Les enregistrements des caractères de morphologie mammaire ne sont réalisés que pour les femelles du schéma de sélection nées depuis 1999. Les femelles ne sont pointées pour ces caractères qu'une seule fois dans leur carrière en principe lors de la première lactation sauf pour quelques bonnes femelles non pointées en première lactation et qui le sont donc en deuxième lactation. En 2013, l'évaluation pour les caractères de production laitière comptait 2 875 004 femelles. En plus des données de performances, l'indexation nécessite l'enregistrement précis des pédigrées de tous les animaux. Ainsi l'indexation de janvier 2013 comprenait 2 915 165 femelles et 66 644 mâles en 2012. Cependant, un nombre conséquent de femelles n'ont pas de généalogie connue (52% en 2011) et cette proportion est en constante augmentation depuis 2005 (40%) (Clément, 2012). Ce phénomène, qui n'est pas optimal pour l'évaluation génétique, s'explique d'une part par un faible intérêt des éleveurs pour l'enregistrement des généalogies et d'autre part par l'augmentation de la conduite en lot par rapport à la monte en main entre autre pour les femelles dont l'IA a échoué (Virginie Clément, Idele, communication personnelle).

Les indexations pour les races Alpine et Saanen sont réalisées trois fois par an : en janvier et septembre pour les indexations officielles et en juin pour une indexation pour la gestion du schéma réservée à Capgènes (Hélène Larroque, INRA, communication personnelle). Chaque indexation consiste en trois évaluations : une pour les caractères de production laitière, une pour le comptage de cellules somatiques et une pour les caractères de morphologie mammaire. L'évaluation des caractères de production laitière est multiraciale étant donné que les paramètres génétiques estimés pour ces caractères sont très proches dans les deux races (Boichard et al., 1989). Ces caractères de production sont évalués simultanément par le biais d'une seule évaluation unicaractère sans qu'aucune corrélation entre les caractères ne soit considérée. La deuxième évaluation concernant la sélection pour la résistance aux mammites est basée sur l'analyse du score de cellules somatiques corrigé (LSCS). Le score non corrigé (SCS) est estimé par transformation logarithmique du comptage de cellules somatiques (CCS) $SCS = \mathrm{Log}_2\left(\frac{CCS}{100000}\right) + 3$ afin d'obtenir une performance normalement distribuée. Les SCS sont ensuite corrigés pour le rang et le stade de lactation. Ils sont ensuite pondérés et moyennés pour obtenir une performance à la lactation (LSCS). Cette

évaluation est réalisée séparément dans les deux races en raison des différences observées pour ce caractère (Clément et al., 2008). Les effets fixes considérés pour ces deux évaluations sont le troupeau, le mois de mise bas, l'âge à la mise bas et la durée de tarissement précédent, tous ces effets étant définis intra campagne de lactation (année), rang de lactation et région (excepté pour l'effet troupeau). Ces évaluations prennent également en compte un effet aléatoire d'environnement permanent, car les données sont répétées. La variance d'erreur est modélisée selon 2 facteurs (variance hétérogène) : un effet fixe campagne-région-numéro de lactation et un effet aléatoire élevage-campagne-numéro de lactation sans autocorrélation. Enfin, la troisième évaluation concerne les caractères de morphologie mammaire. En raison du plus faible nombre d'enregistrements (pas de répétitions et moins de femelles) considérés pour les caractères de morphologie, que pour les caractères de productions ainsi qu'en raison de fortes corrélations génétiques entre certains caractères (PLA et AAR, Tableau 1 .2) cette troisième évaluation est multicaractère. Elle est également uniraciale c'est à dire réalisée séparément dans les deux races, en raison des différences observées entre les deux races (Manfredi et al., 2001). Le modèle d'analyse pour les caractères de morphologie prend en compte les effets fixes d'âge et de stade de lactation au moment du pointage ainsi que l'effet du troupeau intra campagne de lactation.

Tableau 1.2 : Corrélations génétiques entre les caractères de morphologie mammaire pour la race Alpine au-dessus de la diagonale et en dessous pour la race Saanen (source (Manfredi et al., 2001))

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

Les corrélations génétiques entre ces caractères de morphologie sont présentées dans le Tableau 1 .2. Ces corrélations sont fortes entre PLA et AAR (0,77 et 0,72 en races Alpine et Saanen respectivement), PRM et ORT, AAR et AVP et entre PLA et AVP (0,51 et 0,58 respectivement en races Alpine et Saanen). Elles sont faibles entre AVP et PRM (0,03 et -0,09 respectivement en race Alpine et Saanen), entre AAR et PRM, entre AVP et ORT et entre PLA et ORT. Ces corrélations sont proches, bien que plus faibles, de celles observées en bovins laitiers Holstein (Berry et al., 2004). Les corrélations génétiques entre les caractères de production laitière (Figure 1 .3) étant également fortes (entre -0,29 et 0,92), une évaluation

multicaractère pourrait être envisagée. En effet, ces corrélations sont fortement positives entre la quantité de lait et les matières, entre les matières et entre les taux. Elles sont négatives entre le lait et les taux butyreux et protéique.

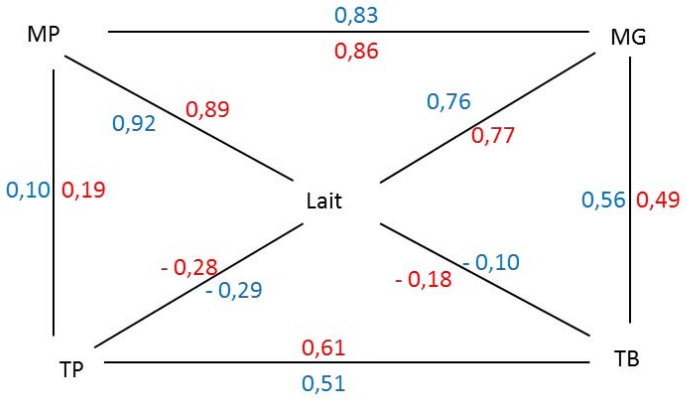


Figure 1.3 : Corrélations génétiques entre les caractères laitiers en race Saanen (bleu) et en race Alpine (rouge) (source Manfredi et Adnøy,2012))

Les héritabilités des caractères (Tableau 1 .3) sont relativement moyennes, (entre 0,19 et 0,36) excepté pour les taux butyreux et protéique pour lesquels les héritabilités sont fortes (0,5) (Manfredi et Adnøy, 2012). Ces héritabilités sont proches de celles estimées en ovins Lacaune (h^2 = 0,28 pour la quantité de lait (Rupp et al., 2003)) et en bovins Holstein (h^2 = 0,31 pour la quantité de lait (Short et Lawlor, 1992)). Les héritabilités les plus faibles sont obtenues pour le LSCS (0,19 et 0,21), la race Saanen ayant l'héritabilité la plus forte. Elles sont comparables à celles estimées en ovins Lacaune avec des héritabilités pour les LSCS de 0,12 à 0,13 dépendant du rang de lactation (Rupp et al., 2003). Les héritabilités pour les caractères de morphologies sont légèrement plus élevées entre 0,25 et 0,36 que pour les LSCS comme en ovins Lacaune (0,19 < h^2 > 0,33 ; (Etancelin et al., 2005)). Ces héritabilités ainsi que la variabilité génétique des caractères de morphologie permettent une sélection légèrement plus efficace pour ces caractères que pour les LSCS.

Les légères différences de paramètres génétiques en race Alpine et en race Saanen suggèrent un progrès génétique potentiel pour la MP et le TP plus élevé en race Alpine (h^2=0,58 pour le TP) qu'en race Saanen (h^2=0,50 pour le TP). En revanche, l'amélioration simultanée de la quantité de lait et du TB est plus facile en race Saanen du fait des corrélations génétiques plus élevées entre ces deux caractères dans cette race (Bélichon et al., 1999). Les corrélations génétiques entre les caractères de morphologie et de production laitière, y compris les LSCS, sont très modérées. De même, les corrélations entre les caractères de production laitière et les LSCS sont modérées à l'exception de la corrélation entre le lait et distance plancher jarret (- 0,5 en moyenne) qui indique qu'augmenter la production laitière

dégrade la distance plancher-jarret. En effet, une forte production laitière nécessite un volume plus important de la mamelle.

Tableau 1.3 : Paramètres génétiques utilisés pour les évaluations génétiques officielles en race Alpine et Saanen

	Alpine	Saanen
	héritabilité	héritabilité
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
LSCS	0,19	0,21
PLA	0,34	0,33
PRM	0,36	0,27
AAR	0,26	0,33
AVP	0,26	0,28
ORT	0,33	0,25

La sélection caprine est efficace. En effet, les cheptels ayant plus de 70% de chèvres issues d'IA ont une production laitière supérieure de 200kg/an, un TB supérieur de 1,7g/kg/an, un TP supérieur de 1,3g/kg/an par rapport aux cheptels n'utilisant pas de semences testées (France Génétique Élevage, 2014). De plus, les femelles utilisées comme mères à boucs ont une production laitière supérieure de plus de 40% par rapport à la moyenne du cheptel français. En moyenne chaque année, le progrès génétique dans l'ensemble du cheptel français permet d'augmenter la quantité de lait de 12kg par lactation, et le TB et TP de 0.1g/kg. La production laitière a ainsi augmenté de 125 kg en 10 ans (FranceAgriMer et France Génétique Élevage, 2015; France Génétique Élevage, 2014).

Le progrès génétique global sur l'ensemble des caractères en race Saanen est plus important que celui obtenu en race Alpine dû à un progrès génétique plus important sur le lait en Saanen, à une opposition génétique plus marquée entre les taux (surtout le TB) et le lait en race Alpine ainsi qu'à une dégradation de l'IMC en Alpine ces dernières années (Renard, 2008). Cependant la totalité du progrès réalisé sur la production laitière ne peut pas être attribuée uniquement à la sélection génétique mais aussi à l'amélioration des conditions d'élevage par exemple (Clément et al., 2002). Cette sélection efficace a cependant un coût relativement élevé pour l'éleveur : l'adhésion à Capgènes est de 28€ par chèvre et par an, le coût du contrôle laitier est de 10 à 25€ par chèvre et par an, le coût de l'insémination animale de 22 à 24€ par chèvre. L'adhésion au schéma de sélection obligeant les éleveurs à inséminer au minimum 30% des chèvres du cheptel, le coût de l'amélioration génétique s'avère assez lourd (Vandiest, 2006).

Enfin, malgré l'efficacité de la sélection génétique en caprins laitiers en France, les résultats n'atteignent pas le même gain qu'en bovins laitiers en raison notamment du faible taux d'insémination animale dans cette espèce. La sélection génétique caprine en France est d'autre part actuellement limitée car l'augmentation des séries de testage (intensité de sélection) n'est pas envisageable dans la mesure où le schéma de sélection ne génère pas les fonds nécessaires à l'augmentation de la capacité de testage (Manfredi and Adnøy, 2012). Ces limitations pourraient être revues avec l'arrivée de la sélection génomique dans cette espèce.

1.I.3 Vers la génomique

Depuis la fin des années 80, les données moléculaires sont disponibles et utilisables pour la sélection des animaux. Une sélection que l'on pourra qualifiée de « sélection par gène » (SAG), qui permet de sélectionner les individus sur des mutations causales connues (Manfredi and Adnøy, 2012), est appliquée depuis les années 90 pour le gène de la caséine α_{s1} dans le schéma de sélection caprin (Barbieri et al., 1995). En effet, tous les mâles candidats au testage sur descendance ainsi que toutes les mères à boucs potentielles sont génotypés pour ce gène. Ce génotypage permet d'éliminer du noyau de sélection tous les individus avec des génotypes défavorables surtout pour le taux protéique (cf. chapitre 5, article III). L'intégration de la sélection du gène de la caséine α_{s1} a donc été réalisée sans modification du schéma de sélection. Le gène de résistance à la tremblante caprine est également connu depuis de nombreuses années, et un outil tel que celui développé pour la caséine α_{s1} pourrait être envisagé afin de sélectionner les animaux résistants (Manfredi and Adnøy, 2012).

Avec l'arrivée de nouvelles technologies d'analyse du génome et de typage à haut débit en bovins laitiers, un consortium international (International Goat Genome Consortium (IGGC)), a été créé en 2010 afin de développer une puce à SNP pour l'espèce caprine. Une puce caprine 50k, l'Illumina goat SNP50k BeadChip (Tosser-Klopp et al., 2014) est désormais disponible et l'essentiel des boucs de testage français nés depuis la fin des années 90 ont été génotypés avec cette puce depuis 2012 (Carillier et al., 2013). Les données de génotypage ont été dans un premier temps utilisées pour détecter des QTL (quantitative trait loci) pour les caractères de production laitière, les LSCS et les caractères de morphologie mammaire (Roldán et al., 2008; Maroteau et al., 2013). En bovins laitiers français, les données de QTL ont été utilisées depuis 2001 pour réaliser une sélection dite assistée par marqueurs (SAM) permettant ainsi un progrès génétique plus important qu'avec une sélection génétique classique (Fritz et al., 2008). La mise en place d'une sélection de type SAM dans cette espèce a nécessité au préalable une détection de QTL d'abord basée sur des marqueurs microsatellites en 1990

(Boichard et al., 2002). Une approche de type SAM n'a pas été envisagée dans l'espèce caprine. Cependant, la disponibilité récente des génotypages 50k, permet d'envisager, à court terme, la sélection génomique dans cette espèce afin d'améliorer le progrès génétique.

Le principe de la sélection génomique est basé sur la prédiction des valeurs génétiques des individus de la population cible ou candidate à partir des informations (SNP et phénotypes) de la population de référence. Les relations établies dans la population de référence entre les génotypes aux marqueurs et les phénotypes permettent de prédire la valeurs génétiques des candidats sur la base de leurs génotypes sans avoir besoin de leurs phénotypes (Robert-Granié et al., 2011). Comme en bovins laitiers, l'intérêt de la sélection génomique chez les caprins laitiers est de pouvoir sélectionner les mâles dès la naissance économisant ainsi le coût et le temps du testage sur descendance. Les intervalles de génération relativement long en caprins notamment sur la voie père-fils pourraient être ainsi diminués. De plus, le coût du testage ainsi économisé, une augmentation du nombre de candidats à la sélection pourrait être envisagé afin d'augmenter l'intensité de sélection (nombre de sélectionnés / nombre de candidats) si le coût du génotypage est relativement faible. Le gain génétique pourrait donc être augmenté par le biais de la sélection génomique. Une simulation par rapport aux gains observés en bovins laitiers a montré que pour un caractère avec un héritabilité de 0,2 dont les marqueurs expliqueraient 50% de la variance phénotypique, le progrès génétique annuel pourrait passer de 0,25 σ_g en sélection classique à 0,50 σ_g avec la sélection génomique (Colleau et al., 2009). De plus, le progrès ainsi réalisé sur les caractères déjà en sélection pourrait donner des marges de manœuvre pour intégrer de nouveaux caractères (comme des caractères de santé avec par exemple la résistance à la tremblante) aux objectifs de sélection. L'ensemble de ces décisions (i.e. diminution de l'intervalle de génération, augmentation de l'intensité de sélection, intégration de nouveaux caractères aux objectifs de sélection, etc.) nécessite quelques études pour quantifier l'intérêt et le gain génétique apporté par la sélection génomique. En effet, il existe de nombreux scenarii possibles afin d'optimiser le schéma de sélection caprin avec l'arrivée de la sélection génomique qui ne prennent pas seulement en compte la disparition du testage sur descendance (Shumbusho et al., 2013). Les réflexions envisagées devront tenir compte du coût du génotypage (environ 130 € en 2012) bien qu'inférieur au coût du testage reste supérieur à la valeur des reproducteurs (Manfredi and Adnøy, 2012). Les coûts de génotypages évoluent cependant rapidement avec un prix aujourd'hui proposé de moins de 70€ incluant le prix de la puce et du génotypage (Gwenola Tosser-Klopp, INRA, communication personnelle).

Le gain génétique envisagé avec la sélection génomique suppose une précision suffisante des index génomiques des mâles à la naissance. La précision des évaluations génomiques est définie comme la corrélation entre les valeurs génomiques prédites et les valeurs génétiques « vraies » des candidats. En bovins laitiers, des précisions d'évaluations génomiques de l'ordre de 0,5 à 0,7 ont permis d'envisager la suppression du testage (Fritz et al., 2010). Les gains de précision avec la sélection génomique obtenus en ovins Lacaune sont de l'ordre de 13% (sur la quantité de lait et les LSCS) à 22% sur le TB (Baloche et al., 2013). En bovins laitiers Hostein australiens, les gains ont été estimés entre 12,5% pour un caractère de fertilité et 61% pour la matière protéique (Hayes et al., 2009b). Les gains observés dans la population Holstein française sont légèrement plus élevés, entre 21% pour la fertilité et 64% pour le TB (Boichard et al., 2012b). Dans une population de bovins allaitants Angus américains, le gain de précisions observé maximal observé est de 350% pour le poids de carcasse (Saatchi et al., 2011).

Nous avons vu précédemment que la mise en place de la sélection génomique dans l'espèce caprine pourrait permettre d'augmenter le progrès génétique par le biais de la diminution des intervalles de générations. Les gains de précision obtenus dans les autres espèces avec la sélection génomique, permettent d'espérer pour l'espèce caprine un gain de progrès génétique via l'augmentation des précisions. Cependant, en caprins laitiers, les précisions des index sur ascendance des candidats dans le schéma de sélection actuel sont déjà élevées (CD de 0,38 en moyenne). En effet, les boucs améliorateurs, pères des candidats ont en moyenne plus de 1 000 filles (CD approchant les 0,99) et les mères choisies ont plus de quatre lactations (CD > 0,50). Il sera donc probablement difficile d'obtenir un gain de précision aussi conséquent que celui observé en bovins laitiers.

D'autre part, la précision des index génomiques dépend de l'héritabilité du caractère, de la taille et de la structure de la population de référence, de l'apparentement entre la population référence et la population candidate, du modèle et de la méthode d'évaluation génétique utilisés (Hayes et al., 2009c; Meuwissen et al., 2001; Daetwyler et al., 2008) (cf. 1.II.5). La structure de la population sera étudiée dans ce manuscrit afin d'évaluer les facteurs influençant l'efficacité de la sélection génomique dans l'espèce caprine. Cependant, la taille limitée de la population caprine comprenant seulement 70 boucs testés sur descendance par an laisserait à penser que les précisions obtenues avec la sélection génomique seront moins élevées que celles trouvées en bovins laitiers.

1.II Structure de la population et facteurs influençant la précision de l'évaluation génomique

De nombreux auteurs ont montré que les structures de la population de référence et de la population candidate ont une influence sur la précision des évaluations génomiques (Goddard and Hayes, 2007; Hayes et al., 2009c; Meuwissen et al., 2001; Daetwyler et al., 2008). La taille, l'apparentement, la consanguinité, le déséquilibre de liaison de la population sont des éléments clés des populations qui peuvent influencer cette précision. Nous verrons donc ici comment sont définis et calculés ces différents éléments et quelle est leur influence sur la précision des évaluations génomiques.

1.II.1 Déséquilibre de liaison

Le déséquilibre de liaison (DL) est défini comme l'association préférentielle d'allèles à deux positions (loci) différentes sur le génome. En équilibre de liaison, les fréquences alléliques constatées dans la population pour les différents haplotypes (groupes d'allèles à différents loci) seront les mêmes (25% dans chaque cas pour deux loci bialléliques).

Tableau 1.4 : Fréquences allèliques constatées dans une population en équilibre de liaison pour un haplotype de 2 loci bialléliques

		Allèles au locus 1	
		A	a
Allèles au locus 2	B	0,25	0,25
	b	0,25	0,25

Le Tableau 1.4 illustre ce cas d'équilibre de liaison dans le cas de deux marqueurs bialléliques ayant respectivement pour allèles A/a et B/b. Tout écart à cet équilibre de liaison est le signe d'un déséquilibre de liaison. Les causes les plus fréquentes du DL sont :

- les mutations ou créations d'un nouvel allèle dans une population idéale sans sélection. Elles apparaissent chez un individu donné et par transmission à la descendance, crée du déséquilibre de liaison;
- les recombinaisons entre 2 loci sur un chromosome. Elles font apparaitre de nouvelles associations entre allèles qui n'étaient pas observables chez les individus parentaux;
- la sélection des individus. Qu'elle soit naturelle ou orientée, la sélection de certaines combinaisons d'allèles favorables à l'espèce crée une prépondérance de cette combinaison allélique dans la population et donc par définition du déséquilibre de liaison;

- le mélange des populations. Même si deux populations sont en équilibre de liaison, le mélange de deux populations crée du déséquilibre de liaison quand les fréquences alléliques dans les deux populations sont différentes.

Le DL étant l'écart des fréquences alléliques par rapport à l'équilibre de liaison, il peut être calculé à partir de la déviation (D) des fréquences constatées par rapport à une association aléatoire entre allèles. Le déséquilibre de liaison évolue en fonction du nombre de générations écoulées et de la distance génétique entre loci en raison des phénomènes de recombinaisons. On peut montrer que le déséquilibre de liaison à la génération n est équivalent à $(1-r)^t D_0$ avec D_0 le déséquilibre de liaison initial et r le taux de recombinaison. Plus le taux de recombinaison est élevé, plus le déséquilibre de liaison est faible. Le DL sera plus élevé pour deux loci proches puisque la probabilité de recombinaisons dans ce cas est très faible. Ce phénomène est d'autant plus marqué que le nombre de générations écoulées est faible (Feingold, 1991).

Considérant l'exemple biallélique précédent, la déviation se définit comme suit (Lewontin and Kojima, 1960) : $D = p(\mathrm{AB}) \times p(\mathrm{ab}) - p(\mathrm{Ab}) \times p(\mathrm{aB})$, où p(AB), p(ab), p(Ab), p(aB) sont les fréquences des haplotypes AB, ab, Ab et aB dans la population. Cette déviation comprise entre - 0,25 et 0,25, vaut 0 s'il y a équilibre de liaison. Cependant cette déviation D n'est pas la meilleure mesure pour comparer le DL entre plusieurs populations car elle dépend des fréquences alléliques observées dans une population qui peuvent différer d'une population à l'autre (Feingold, 1991).

Une autre mesure du DL souvent utilisée en génétique des populations est la mesure du D' (Lewontin, 1964) : $\mathrm{D}' = \frac{|D|}{D_{\max}}$ où $D_{\max} = \begin{cases} \min(p_A \times (1-p_b), p_a \times (1-p_B)) \text{ si } D > 0 \\ \min(p_A \times (1-p_B), p_a \times (1-p_b)) \text{ si } D < 0 \end{cases}$ avec p_A la fréquence de l'allèle A dans la population. D' varie de -1 à + 1.

La Figure 1 .4 représente un calcul de D et D' dans un cas biallélique entre deux loci. D' n'est pas souvent utilisée comme mesure du déséquilibre de liaison entre populations car pour des petites populations ou des populations à faibles fréquences alléliques, le DL est surestimé (Weiss and Clark, 2002).

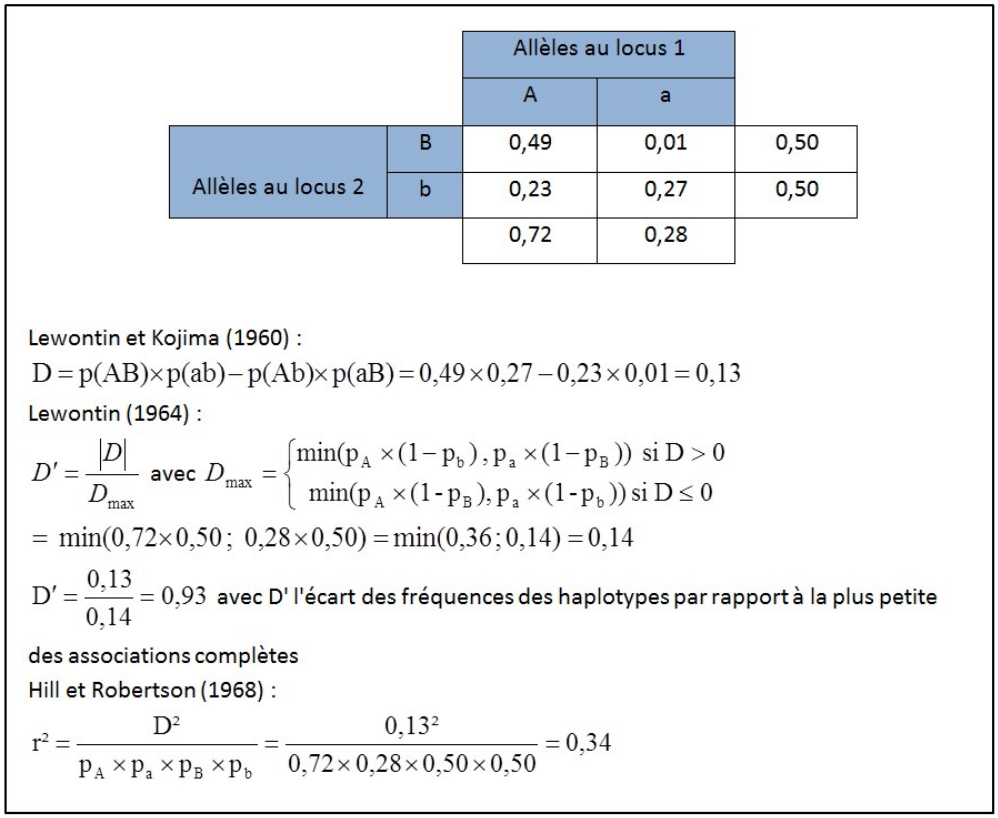


Figure 1.4 : Exemple du calcul de déséquilibre de liaison par mesure du D, D' et r²

La mesure du déséquilibre de liaison standardisé ou r² (Hill and Roberson, 1968) est moins dépendante des fréquences alléliques que les mesures précédentes. Elle est la mesure du DL la plus utilisée dans la littérature. Le r² est calculé selon la formule :

$$r^2 = \frac{D^2}{p_A \times p_a \times p_B \times p_b} = \frac{D^2}{p_A \times (1-p_A) \times p_B \times (1-p_B)}$$ où p_A est la fréquence de l'allèle A dans la population. Un exemple de ce calcul est présenté en Figure 1 .4. Le r² peut prendre des valeurs entre 0, lorsqu'il n'y a aucun DL et 1 lorsque le DL est complet c'est-à-dire lorsqu'un des allèles du premier marqueur est associé à un seul des allèles du second marqueur.

L'ensemble des mesures citées précédemment suppose que les fréquences alléliques ainsi que les phases alléliques (i.e. relations entre les allèles de deux marqueurs liés) sont connues. Or dans le cas des populations animales, la phase allélique est peu souvent connue, c'est pourquoi Rogers et Huff (2009) ont développé une mesure plus adaptée à ce cas. Elle correspond à une corrélation entre génotypes pour tous les individus à 2 loci donnés :

$r^2 = \frac{[\mathrm{cov}(g_i, g_j)]^2}{\mathrm{var}(g_i) \times \mathrm{var}(g_j)}$ (Rogers and Huff, 2009) avec g_i et g_j les génotypes aux SNP i et j. La Figure 1 .5 représente un schéma explicatif de ce calcul du DL.

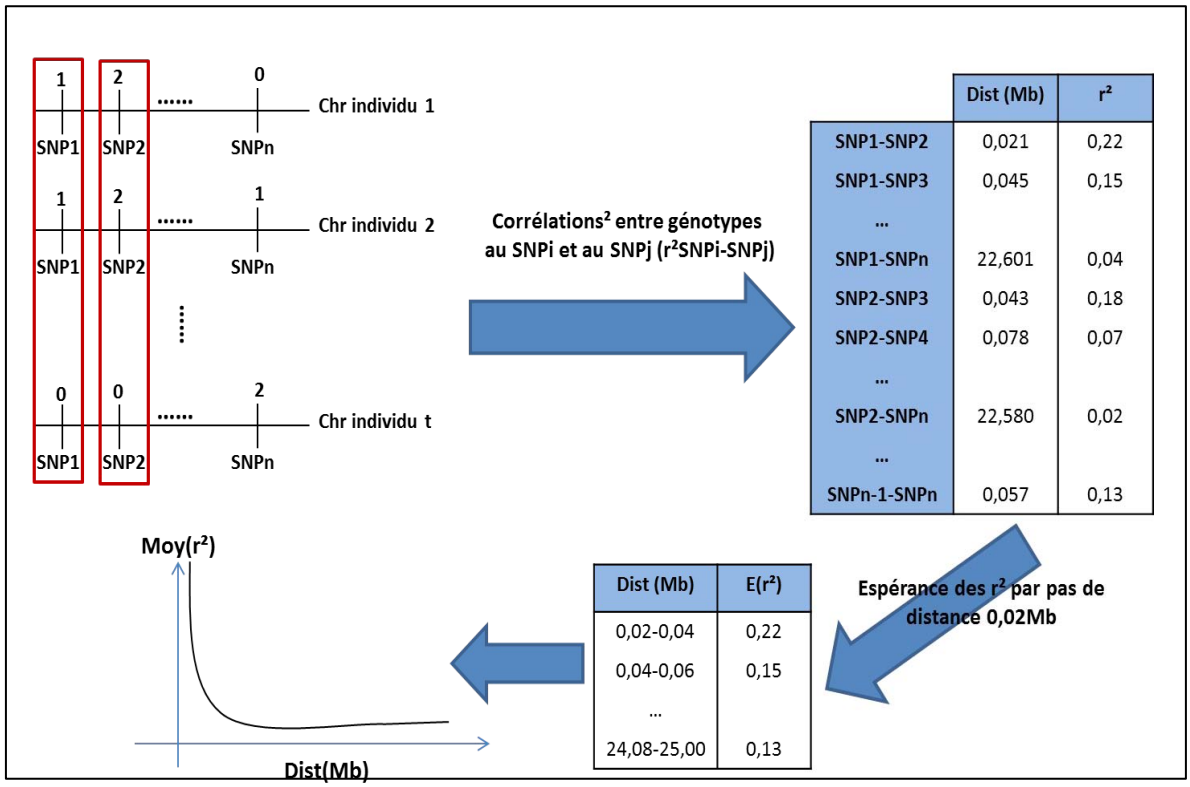


Figure 1.5 : Schéma du calcul du déséquilibre de liaison avec la méthode de Roger et Huff (2009)
Les valeurs 0, 1 ou 2 indiquées à chaque SNP correspond au codage utilisé pour les SNP, 1 pour les hétérozygotes, 0 pour un des homozygotes et 2 pour l'autre homozygote

La persistance des phases du DL entre populations (ou races) permet d'estimer si l'association préférentielle entre allèles de deux loci différents observée dans une population est également observée dans l'autre. Cette persistance est estimée par un coefficient de corrélation de Pearson entre les différentes valeurs de DL estimées dans chaque population pour chaque paire de SNP. Les valeurs de DL utilisées dans ce cas sont les r $\left(r = \frac{\mathrm{cov}(r_{pop1}, r_{pop2})}{\sqrt{\mathrm{var}(pop1) \times \mathrm{var}(pop2)}}\right)$ et non les r², car les r² peuvent avoir la même valeur dans deux populations alors que l'association entre les marqueurs est inversée. Ce niveau de persistance du DL entre deux populations peut être modélisé par $r_0^2(1-c)^{2T}$ avec r_0^2 une mesure du DL dans la population ancestrale commune, c le taux de recombinaison fonction de la distance et T le nombre de générations de divergence entre les deux populations (Sved et al., 2008). Plus simplement, (De Roos et al., 2008) ont montré que cette persistance peut être

estimée par e^{-2cT} . Ces deux équations impliquent que la persistance du déséquilibre de liaison est fortement corrélée négativement avec le temps de divergence entre sous-groupes d'une population.

Le principe de la sélection génomique repose sur l'existence de déséquilibre de liaison entre marqueurs et QTL. En effet, le génome d'un individu peut être considéré comme un ensemble de loci ayant un effet sur un caractère quantitatif (QTL ou quantitatif trait locus en anglais). Les marqueurs en fort déséquilibre de liaison avec ces QTL permettent donc de capter l'information des QTL pour prédire la valeur génétique d'un individu. Le DL a donc une influence sur les précisions des valeurs génomiques estimées (GEBV pour genomic estimated breeding value) (Meuwissen et al., 2001). Plus le déséquilibre de liaison est élevé, plus les précisions des GEBV sont élevées (Calus et al., 2008; Solberg et al., 2008). La précision des GEBV due au DL (ρ_j^{DL} à la génération j) peut être estimée par la formule suivante : $\rho_i = x_{1i} d_j + x_{2i} \rho_j^{DL} + e_i$ avec ρ_i la précision des GEBV à la génération i, $x_{1i} = \frac{r(\text{EBV})_{\text{génération}(i)}}{r(\text{EBV})_{\text{generation}(j)}}$ où r(EBV) est la précision des valeur génétiques (EBV pour estimated breeding value), d_j la différence entre la précision des GEBV et la précision due au DL à la génération j, x_{2i} la diminution du DL par génération et e_i un terme de résidus (Habier et al., 2007, 2010).

Le niveau de déséquilibre de liaison obtenu dans une population est lié positivement au nombre de marqueurs présents sur la puce. Le niveau de DL dans un population est également lié au degré d'apparentement au sein de la population (Robert-Granié et al., 2011; Wientjes et al., 2013). Il est de ce fait difficile de distinguer la part de précision des GEBV due au DL de celle due à l'apparentement (Daetwyler et al., 2012). Cette distinction est pourtant intéressante car la part de précision due au DL est plus conservée dans le temps contrairement à la part due à l'apparentement qui évolue à chaque génération (Habier et al., 2007).

1.II.2 Diversité génétique de la population

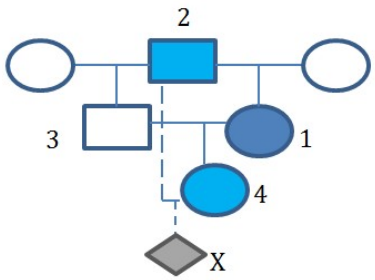
1.II.2.1 Consanguinité et parenté

Un individu est dit consanguin si à un locus donné, il possède deux allèles identiques supposés issus d'un ancêtre commun. Le coefficient de consanguinité correspond donc à la

probabilité qu'un individu reçoive les mêmes allèles de ses deux parents à un locus donné. Il est calculé comme suit: $F_j = \left(\frac{1}{2}\right)^n \times \left(1 + F_A\right)$ où n est le nombre d'individus dans la ligne parentale (entre j et l'ancêtre commun) et F_A le coefficient de consanguinité de l'ancêtre commun (Wright, 1969; Malécot, 1948). La Figure 1 .6 donne un exemple du calcul de ce coefficient. Le calcul du coefficient de consanguinité à partir de données génomiques pour l'individu j correspond à $F_j = \frac{1}{n}\sum_{L=1}^{n}\left(1 - \frac{H_{0L}}{H_{eL}}\right)$ où n est le nombre de loci, H_{eL} est le nombre de génotypes hétérozygotes attendus au locus L estimé à partir des fréquences alléliques observées dans la population de base et H_{0L} le nombre de génotypes hétérozygotes observés au locus L (Baumung and Sölkner, 2003).

Le coefficient de parenté entre deux individus I et J (Φ_{IJ}) est la probabilité qu'à un locus donné, l'allèle pris au hasard chez un premier individu soit identique à un allèle pris au hasard chez un deuxième individu (Minvielle, 1990). Il est fortement lié au coefficient de consanguinité puisqu'il correspond au coefficient de consanguinité qu'aurait la descendance issue des deux individus. La Figure 1 .6 présente un calcul du coefficient de parenté à partir des données de pédigrée. Ce coefficient peut être estimé à partir des données génomiques selon la formule suivante : $\hat{\Phi}_{IJ} = \frac{N_{\text{I et J hétérozygotes}} - 2N_{\text{I et J homozygotes}}}{N_{\text{I hétérozygote}} + N_{\text{J hétérozygote}}}$ avec $N_{\text{I et J hétérozygotes}}$ le nombre total de SNP pour lesquels les 2 individus I et J sont hétérozygotes, $N_{\text{I et J homozygotes}}$ le nombre total de SNP pour lesquels les individus I et J sont homozygotes, $N_{\text{I hétérozygote}}$ (respectivement $N_{\text{J hétérozygote}}$) le nombre total de SNP pour lesquels I (respectivement J) est hétérozygote (Manichaikul et al., 2010). Deux types de parenté ont une influence sur les précisions des valeurs génomiques estimées. Le premier correspond à la parenté au sein de la population de référence. Plus il est élevé, plus il est facile de faire le lien entre les marqueurs et les phénotypes (Habier et al., 2010; Meuwissen, 2009). Le deuxième est la parenté entre la population référence et la population des candidats. Une parenté élevée entre ces deux

populations permet une meilleure précision des GEBV (Moser et al., 2010; Lund et al., 2009; Habier et al., 2010; Pryce et al., 2011).



En considérant que l'individu 2 est issu de parents inconnus, son coefficient de consanguinité est nul. Le coefficient de consanguinité de l'individu 4 est :

$F_4 = \left(\frac{1}{2}\right)^n \times (1+F_2) = \left(\frac{1}{2}\right)^3 \times (1+0) = 0{,}125$ avec 3 individus sur la ligne parentale de 4 (1, 2 et 3).

Le coefficient de parenté entre l'individu 4 et l'individu 2 correspond au calcul du coefficient de consanguinité qu'aurait leur descendance même s'il ne peut pas exister. Dans la figure ci-dessus si on considère que l'individu 2 est issu de parents inconnus, le nombre d'individus sur la ligne parentale de X est de 4 (2, 4 et 1).

$$\Phi_{2,4} = F_X = \left(\frac{1}{2}\right)^3 \times (1+F_2) = \frac{1}{8} = 12{,}5\%$$

Figure 1.6 : Exemple d'un calcul du coefficient de consanguinité et de parenté estimés à partir du pédigrée

1.II.2.2 Taille efficace / effectif génétique

Wright (1969) définit la taille efficace (Ne) comme étant le nombre d'individus d'une population idéale (i.e. générations disjointes, taille constante, isolement reproductif, panmixie) pour laquelle le degré de dérive génétique serait équivalent à celui de la population réelle. Dans cette population idéale aucune mutation et aucune sélection ne sont possibles, chaque gène à une génération est donc la copie d'un gène de la génération précédente. La taille efficace est donc un indicateur de la variabilité génétique d'une population.

Il existe, entre autres, deux façons d'estimer la taille efficace de population. La première correspond à une évaluation de l'effectif efficace en termes de variance. Ce qui implique que la population idéale doit avoir la même variance de fréquence allélique que la population réelle. Elle est donc estimée comme suit : $Ne = \frac{p(1-p)}{Var(p')}$ avec p la fréquence allélique d'un allèle donné à une génération t et p' la fréquence du même allèle à la génération t+1. La deuxième façon d'estimer la taille efficace est de considérer une population idéale

avec le même niveau de consanguinité que la population étudiée. Dans ce cas, la taille efficace de la population à la génération t peut être estimée à partir des données du pédigrée par la formule : $Ne_t = \frac{1}{2\left(1-\sqrt[t]{1-F}\right)}$ où Ne_t est la taille efficace de la population à la génération t et F le coefficient de consanguinité dans la population actuelle (Pérez-Enciso, 1995).

Il existe également une méthode permettant d'estimer la taille efficace de la population à partir du DL. Tenesa et al. (2007), reprenant une approximation de (Sved, 1971), ont montré que l'espérance du déséquilibre de liaison pouvait être estimée par la formule : $E(r^2) = \frac{1}{1+4Ne_T c}$ avec c la distance entre deux SNP consécutifs en Morgans et Ne_T la taille efficace à T générations avant l'actuelle avec $T = \frac{1}{2c}$ (Tenesa et al., 2007).

Une autre mesure, le nombre de fondateurs efficaces (Ng) peut permettre d'évaluer la diversité génétique intrinsèque d'une population. Ce nombre Ng correspond à l'inverse de la probabilité que deux gènes tirés au hasard dans la population étudiée proviennent du même fondateur ou du même ancêtre. C'est donc le nombre de fondateurs qui engendrerait la même variabilité observée dans la population s'ils avaient tous contribué au même niveau à la descendance (De Rochambeau et al., 2003).

1.II.3 Taille de la population de référence

La taille de la population de référence correspond au nombre d'individus génotypés et phénotypés d'une population donnée. La précision des valeurs génomiques estimées des individus augmente avec la taille de cette population. Comme en sélection classique, l'augmentation du nombre d'individus considérés dans une évaluation génétique permet de diminuer la variance d'erreur de prédiction et donc d'augmenter la précision de ces prédictions (Liu et al., 2011).

Lorsque les populations étudiées sont considérées comme trop petites pour réaliser une sélection génomique, il existe plusieurs possibilités permettant d'augmenter la taille de la population de référence. La première est de regrouper plusieurs populations européennes de même race. Ainsi, la constitution d'une population européenne de bovins Holstein (Consortium Eurogenomics : 7 pays (Allemagne, Belgique, Danemark, Finlande, France,

Pays-Bas et Suède) comptant 15 996 taureaux testés sur descendance), a permis de créer une population de référence 3 à 4 fois plus importante que les populations nationales séparées. L'utilisation de cette population européenne a engendré des améliorations de précision de 8 à 12% (Lund et al., 2010, 2011). Dans le cas d'un regroupement de populations de référence plus réduit (Finlande, Suède et Danemark), les gains de précision sont légèrement plus limités, allant de 5% à 13% selon le caractère et la population considérée (Brøndum et al., 2011).

Il est également possible de combiner différentes races d'une même espèce afin d'augmenter la taille de la population de référence. Dans ce cas, les effets des marqueurs sont supposés stables entre populations. De plus, ces races doivent être suffisamment apparentées afin d'espérer augmenter les précisions des GEBV (cf.1.II.2). Ces combinaisons de races ont beaucoup été étudiées en bovins laitiers avec le cas Holstein+Jersey (Hayes et al., 2009a; Pryce et al., 2011), ainsi qu'en bovins allaitants à travers le projet « Carcass Merit » (Kizilkaya et al., 2009) et en ovins allaitants (Daetwyler et al., 2012b). Ces différentes études ont montré que le gain de précision est généralement négligeable pour la race ayant le plus fort effectif ainsi que pour les races les moins apparentées à la race principale.

Le génotypage de femelles peut également être envisagé dans le cas d'une taille de population de mâles testés sur descendance trop limitée. Certains projets en bovins laitiers ont évalué l'intérêt de génotyper des femelles avec une puce basse densité (7k) à faible coût (Moser et al., 2010; Dassonneville et al., 2012a). La plupart de ces études ont montré que le gain génétique annuel pouvait être augmenté avec l'ajout de femelles dans la population de référence grâce à l'augmentation des précisions des GEBV des femelles (Schrooten et al., 2005; Buch, 2010; Calus et al., 2011; Mc Hugh et al., 2011; Sonesson et al., 2012). L'apport de femelles dans la population de référence améliore également les précisions des GEBV des mâles (Schaeffer, 2006; Verbyla et al., 2010; Mc Hugh et al., 2011). Le génotypage de femelles à haut potentiel génétique ne donne pas de meilleurs résultats lorsqu'il s'agit de femelles choisies au hasard (Jiménez-Montero et al., 2010).

1.II.4 Optimisation de la structure de population

Afin d'améliorer les précisions des GEBV, il est possible d'envisager une sélection d'individus particuliers, au sein de la population de référence, pour optimiser les caractéristiques de sa structure génétique. Cette optimisation de la structure de la population de référence doit tenir compte de la diminution de la taille de population engendrée par la sélection. Dans le cas de populations de trop petite taille, il n'est pas envisageable de tenter

d'optimiser la structure de cette façon, au risque de diminuer les précisions. Ces méthodes consistent à utiliser différents échantillons de la population comme populations de référence (Rincent et al., 2012), puis à déterminer l'échantillon qui permet d'obtenir les meilleures précisions. Les échantillons sont choisis au hasard (Pszczola et al., 2012a; Asoro et al., 2011) ou selon un critère tel que la variance d'erreur de prédiction (PEV ou prediction error variance en anglais).

A taille de population fixée, il est judicieux de sélectionner les animaux les moins apparentés possibles pour la constitution de la population de référence (Pszczola et al., 2012b). En effet, maximiser la diversité génétique de la population de référence permet de mieux prendre en compte la diversité des allèles présents dans la population.

La précision des valeurs génétiques estimées est plus élevée lorsque l'apparentement entre la population de référence et la population de validation est fort (Clark et al., 2012; Pszczola et al., 2012b). Dans le cas d'une population de taureaux Normands, le passage de 96,5% à 83% des individus de la population de validation ayant au moins un parent dans la population d'apprentissage, diminue les précisions des évaluations génomiques de 25% pour les LSCS à 38% pour le TB (Hozé et al., 2014). Une diminution de l'apparentement entre population de validation et d'apprentissage de 0,5 à 0,125 (et de 0,5 à 0) dans une population de Merinos australiens entraine une diminution de 11% (et respectivement 33%) de la précision des évaluations génomiques.

1.II.5 Équations de prédiction de la précision génomique

Hormis la structure de la population de référence et de la population candidate, il existe d'autres facteurs qui influencent les résultats génomiques comme l'héritabilité des caractères et le nombre de SNP sur la puce. En effet, de même que pour les évaluations génétiques, de nombreux auteurs ont montré que la précision des valeurs génomiques estimées une fonction croissante de l'héritabilité du caractère considéré (Calus et al., 2008; Brard et Ricard, 2014; Pszczola et al., 2012b; Hayes et al., 2009b; Lorenz, 2013; Hayes et al., 2009c). De plus, un nombre de marqueurs élevé ou un marquage très dense (utilisation de puce haute densité 800k par exemple) permet d'augmenter les précisions. Sur données simulées, Solberg et al. (2008) ont montré que la multiplication de la densité du marquage par 8 (de 1Ne/morgan à 8Ne/morgan) permettait d'augmenter les précisions génomiques d'environ 25% pour un caractère avec une héritabilité de 0,5. Cependant, Brard et Ricard (2014) ont montré qu'au-delà de 50 000 marqueurs, la précision n'augmente que faiblement et qu'elle atteint un plateau pour 250 000 marqueurs c'est-à-dire un niveau de DL élevé quelle que soit l'espèce

considérée. En pratique, en bovins laitiers Holstein et Jersey, utiliser une puce 800k au lieu d'une puce 50k ne permet pas d'améliorer les précisions de façon conséquente y compris pour les évaluations multiraciales (Harris et al., 2011). La densité de marquage a également un effet direct sur le DL entre deux marqueurs consécutifs puisque celui-ci dépend de la distance moyenne entre deux marqueurs.

Certains auteurs ont établi des équations de prédictions du niveau de précision des valeurs génétiques estimées en fonction de plusieurs paramètres. La première équation de prédiction des précisions a été donnée par Daetwyler et al. (2008) : $r=\sqrt{\frac{Nh^2}{Nh^2+n_g}}$ où N est le nombre d'animaux dans la population de référence et n_g le nombre de loci indépendants affectant le caractère d'héritabilité h^2. Elle a été établie pour des données de type « cas-contrôle » où tous les effets des QTL estimés ont la même précision. Suite aux travaux de Goddard en 2009, Daetwyler et al. (2010) ont proposé une nouvelle équation : $r=\sqrt{\frac{Nh^2}{Nh^2+Me}}$ où Me représente le nombre de segments chromosomiques indépendants.

Goddard (2009) se basant sur la dérivation proposée par Daetwyler et al. (2008) propose une équation dans laquelle les effets des marqueurs sont estimés à l'aide du BLUP : $r=\sqrt{1-\frac{\lambda}{2N\sqrt{a}}\times\ln\left(\frac{1+a+2\sqrt{a}}{1+a-2\sqrt{a}}\right)}$ avec $a=1+2\frac{Me}{Nh^2\log(2Ne)}$, $\lambda=\frac{Me}{h^2\log(2Ne)}$. Cette équation suppose que la précision est meilleure si les fréquences alléliques au QTL sont intermédiaires car elles permettent d'expliquer une plus grande part de la variance génétique.

Goddard et al. (2011) ont également proposé une équation de prédiction des précisions des évaluations génomiques : $r=\sqrt{b\ \frac{\frac{Nbh^2}{Me}}{1+\frac{Nbh^2}{Me}}}$ où b est la proportion de variance génétique expliquée par les marqueurs. Cette équation est basée sur un modèle où la valeur génomique est estimée à partir des phénotypes et est considérée comme la somme des effets aux

marqueurs. Plus récemment, Meuwissen et al. (2013) ont proposé une prédiction de la précision génomique basée sur les développements précédents qui permet de prendre en compte le fait que la précision des phénotypes utilisés dans les évaluations génomiques (DYD) a une influence sur la variance d'erreur de prédiction. Cette équation est définie par :

$$r=\sqrt{\frac{\theta+1+\sqrt{(1+\theta)^2-4h^2\theta b}}{2h^2}} \quad \text{avec} \quad \theta=\frac{Nbh^2}{Me} \quad .$$

Le nombre de segments chromosomiques indépendants peut être estimé de plusieurs façons (Brard et Ricard, 2014). La première formule pour estimer Me a été proposée par Stam (1980) : $Me=4NeL$ où L est la taille du génome en Morgan. Les autres formules permettant d'estimer Me sont : $Me=\frac{2NeL}{\log(4NeL)}$ (Goddard, 2009), $Me=2NeL$ (Hayes et al., 2009d), $Me=\frac{2NeL}{\log(2NeL)}$ et $Me=\frac{2NeL}{\log(NeL)}$ (Goddard et al., 2011). Il existe de nombreuses autres formules et approximations du nombre de segments chromosomiques dérivées d'autres hypothèses que celles citées précédemment. Cependant toutes ces formules considèrent que l'apparentement entre tous les individus est nul, ce que l'on sait être faux. Ces équations auraient donc tendance à sous-estimer les valeurs de précision qui seront obtenues avec des données réelles.

De nombreux auteurs ont montré que le modèle et les méthodes utilisées pour les évaluations génomiques ont également un impact sur les précisions des valeurs génomiques obtenues (Vitezica et al., 2010; Pryce et al., 2011)..

1.III Méthodologie des évaluations génomiques

1.III.1 Méthodes d'évaluations génétique et génomique

Les modèles d'évaluation génétique reposent sur le postulat qu'un phénotype est la résultante d'effets d'environnement (effets du milieu) et d'effets génétiques. Le modèle général d'évaluation génétique des caractères de production laitière et de morphologie en caprins, est basé sur le modèle linéaire mixte gaussien suivant : $y=\mathbf{X\beta+Zu}+e$ avec **y** le vecteur des observations (phénotypes), **β** le vecteur des différents effets fixes considérés dans le modèle (effets de milieu identifiés), **u** le vecteur des valeurs génétiques aléatoires et **e** le

vecteur aléatoire des erreurs. Les matrices **X** et **Z** sont les matrices d'incidence qui relient les observations respectivement aux effets fixes et aléatoires. Le vecteur des valeurs génétiques polygéniques (**u**) suit une loi normale de moyenne nulle et de variance $A\sigma_u^2$ avec **A** la matrice de parenté construite à partir du pédigrée. De même le vecteur des résidus est supposé suivre une loi normale d'espérance nulle et de variance $I\sigma_e^2$. La méthode d'estimation statistique des différents effets utilisée classiquement pour les évaluations génétiques est le BLUP (Best Linear Unbiaised Predictor) modèle animal. Henderson et al. (1959) ont montré que les estimations des différents effets du modèle peuvent être obtenues par résolution du système des équations du modèle mixte :

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\frac{\sigma_e^2}{\sigma_u^2}A^{-1} \end{bmatrix}\begin{bmatrix} \beta \\ u \end{bmatrix}=\begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}.$$

Cette méthode a l'avantage de permettre l'estimation non biaisée simultanée des effets fixes et aléatoires.

Le principe de l'évaluation génomique est de prédire la valeur génomique estimée (GEBV pour genomic estimated breeding value) des candidats à partir de l'information de leur génotype. Le modèle général pour l'évaluation génomique s'écrit comme suit : $y=\mathbf{X\beta+Mg}+e$ avec **y**, **X**, **β** et **e** définis comme dans le modèle précédent, **M** la matrice contenant une fonction du nombre d'allèles au marqueur et **g** le vecteur aléatoire des effets des SNP dont la distribution est supposée normale. La valeur génétique des individus correspond à la somme des effets de leurs allèles aux marqueurs, on peut donc l'estimer par $\hat{u}_i=m_i'\hat{g}$ où m_i' est un vecteur ligne contenant le génotype de l'individu i considéré pour tous les marqueurs et $\hat{g}$ est un vecteur colonne contenant l'ensemble des effets estimés pour tous les marqueurs. Il existe plusieurs approches permettant d'estimer les effets des SNP (Meuwissen et al., 2001). Nous nous limiterons ici à présenter deux approches : une première basée sur le BLUP et une deuxième basée sur les méthodes bayésiennes.

1.III.1.1 Méthode basée sur la construction d'une matrice génomique de parenté

1.III.1.1.1 Les approches two steps et single step

Afin d'estimer les effets des marqueurs, Meuwissen et al. (2001) ont proposé une première approche basée sur le BLUP. On suppose que tous les SNP ont chacun un faible effet sur le phénotype (modèle infinitésimal). Par analogie avec les équations du modèle mixte, les effets du modèle sont estimés à partir des équations suivantes :

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'M} \\ \mathbf{M'X} & \mathbf{M'M}+\frac{\sigma_e^2}{\sigma_g^2}I \end{bmatrix}\begin{bmatrix} \beta \\ g \end{bmatrix}=\begin{bmatrix} \mathbf{X'y} \\ \mathbf{M'y} \end{bmatrix}.$$

La résolution de ces équations est très gourmande en temps de calcul puisqu'elle nécessite d'inverser la matrice **M'M** très dense (Legarra and Misztal, 2008).

Des stratégies alternatives de résolution plus rapide ont été mises au point pour pallier ce problème tel que le développement d'un modèle équivalent au modèle BLUP animal mais utilisant une matrice génomique de parenté (**G**) au lieu de la matrice de parenté (**A**) estimée à partir du seul pédigrée (VanRaden, 2008; Goddard, 2009). On appellera cette méthode d'évaluation génomique le GBLUP pour BLUP génomique. Dans le cas où le codage des SNP est : 0 pour les hétérozygotes, 1 pour les homozygotes du premier type et 2 pour les homozygotes du deuxième type, la matrice génomique (**G**) peut se construire comme suit :

$$G=\frac{\mathbf{MM'}}{2\sum_j p_j q_j}$$

avec p_j la fréquence d'un des allèles du SNP j et q_j=1-p_j. La matrice génomique de parenté mesure 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'information génomique permettant de capter l'aléa de méiose, l'apparentement mesuré par le biais de la matrice génomique est plus précis que celui estimé à partir du pédigrée.

Dans un premier temps, en raison de l'absence d'outil permettant de gérer les données manquantes, seuls les phénotypes des animaux génotypés étaient pris en compte dans le vecteur des observations (**y**). Pour les espèces laitières, des pseudo-performances étaient construites pour les mâles puisqu'en général ils étaient les seuls à être génotypés. Les phénotypes étaient souvent dans ce cas la moyenne des performances des filles corrigées pour les effets fixes c'est-à-dire les daughter yield deviation (DYD), mais il existe d'autres phénotypes utilisables que l'on détaillera dans la partie suivante (cf. 1.III.2). Cette approche

est appelée two steps car elle nécessite une estimation en deux étapes (cf. Figure 1 .7) : une première étape consiste en une évaluation génétique classique permettant de calculer les phénotypes des mâles, puis une seconde étape comprend l'évaluation génomique elle-même basée sur ces phénotypes estimés. Cependant cette approche n'est pas idéale puisqu'elle nécessite d'utiliser des pseudo-phénotypes (DYD ou autres) qui n'ont pas tous la même précision (i.e. le même nombre de filles pour tous les mâles) et sont parfois biaisés par la sélection (Patry and Ducrocq, 2011). De plus, les phénotypes des animaux non apparentés aux animaux génotypés ne sont pas pris en compte dans une telle approche, ce qui génère une perte d'information.

Afin d'intégrer les phénotypes des individus non génotypés dans les évaluations génomiques, il serait possible de prédire les génotypes de ces individus à l'aide de leur information phénotypique et de l'information génomique des autres individus. Cependant, les précisions de ces prédictions sont en général faibles (Bouwman et al., 2014; Piccoli et al., 2014) surtout dans le cas où un seul des deux parents est génotypé, ce qui est souvent le cas en sélection génomique, puisque généralement seuls les mâles sont génotypés. De plus cette imprécision doit être prise en compte dans la distribution *a posteriori* des effets des marqueurs (Abraham et al., 2007) ce qui n'est pas envisageable dans le cas de gros jeux de données (Legarra et al., 2009).

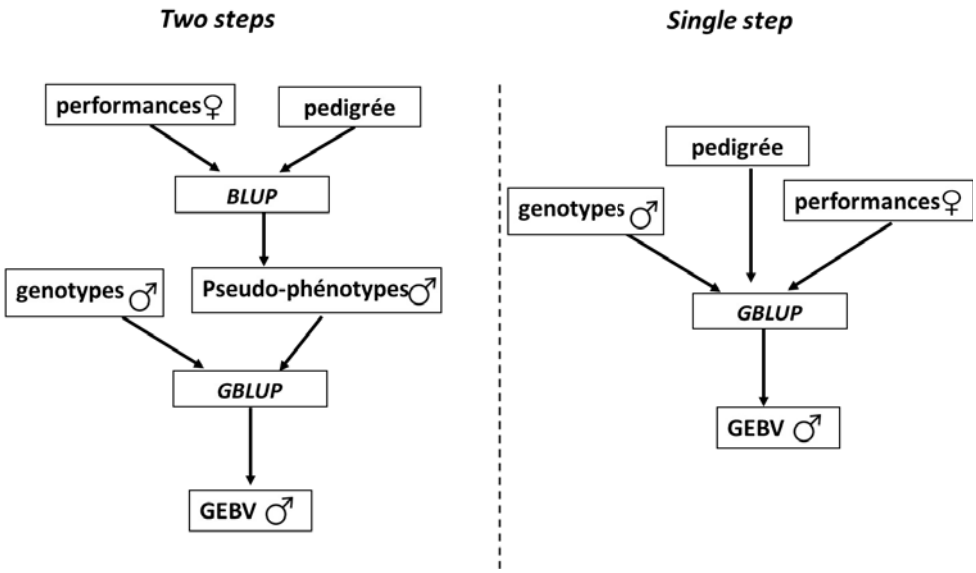


Figure 1.7 : Schémas de réalisation des approches two steps et single step

Une deuxième approche d'évaluation, appelée single step, a été développée plus récemment pour prendre en compte les phénotypes des animaux non génotypés sans avoir à prédire les génotypes manquants (Misztal et al., 2009; Legarra et al., 2009). Les phénotypes

utilisés dans cette approche sont les performances brutes des animaux, c'est-à-dire celles utilisées pour les évaluations classiques (cf. Figure 1 .7). Dans le cas des évaluations des animaux laitiers, les phénotypes considérés sont les performances des femelles. Le modèle général considéré $(y=\mathbf{X\beta+Zu}+e)$ est le même que dans l'approche en deux étapes et la résolution des équations du modèle mixte qui en découle suit le même principe. La matrice de variance-covariance des valeurs génétiques (**H**), définie telle que : $\mathrm{Var}(u)=H\sigma_u^2$, intègre l'information pédigrée et l'information génomique. Elle doit tenir compte du fait que la population de base décrite par la matrice de parenté n'est pas la même que celle décrite par la matrice génomique, c'est-à-dire que les plus vieux animaux génotypés ne sont pas les plus vieux animaux du pédigrée. Dans ce cas, la matrice de parenté doit être ajustée de façon à être compatible avec la matrice génomique (Christensen, 2012). En partitionnant la matrice de parenté en apparentement 1) entre les individus non génotypés ($\mathbf{A}_{11}$), 2) entre les individus génotypés ($\mathbf{A}_{22}$) et 3) entre les individus génotypés et non génotypés ($\mathbf{A}_{12}$ ou $\mathbf{A}_{21}$), la matrice **H** ne peut donc pas s'écrire simplement sous la forme suivante : $H_1=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & G \end{pmatrix}$. Les développements de Christensen et al. (2012) et d'autres auteurs montrent que la matrice **H** peut s'écrire sous la forme suivante : $H=\begin{pmatrix} A_{11}+A_{12}A_{22}^{-1}(G-A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ \mathbf{GA_{22}^{-1}}A_{21} & G \end{pmatrix}$ et son inverse ($\mathbf{H}^{-1}$) est : $H^{-1}=A^{-1}+\begin{pmatrix} 0 & 0 \\ 0 & G^{-1}-A_{22}^{-1} \end{pmatrix}$ (Misztal et al., 2009; Legarra et al., 2009; Christensen, 2012). Cette matrice **H** proposée ici est celle utilisée dans les évaluations génomiques mises en œuvre dans ma thèse. Cette matrice permet d'obtenir de meilleures précisions des évaluations génomiques comparées à la matrice **G** (Gao et al., 2012). Dans ma thèse, nous parlerons également d'une approche appelée « pseudo single step ». Elle consiste à utiliser les phénotypes corrigés de tous les mâles (DYD) ayant plus de 10 filles comme vecteur des observations. Cette approche comme le two steps est basée sur les phénotypes corrigés, en revanche elle permet de prendre en compte tous les phénotypes des mâles qu'ils soient ou non génotypés. Elle ne prend pas en compte les performances de femelles issues de mâles ayant très peu de filles ou les femelles issues de parents inconnus contrairement au single step.

De nombreuses études montrent que l'approche GBLUP single step permet d'augmenter les précisions des évaluations génomiques (GEBV) et d'améliorer les biais par rapport à une approche two steps (Aguilar et al., 2010; Gao et al., 2012; Vitezica et al., 2010). Les précisions sont augmentées de 2 à 34% selon le caractère considéré en ovins laitiers Lacaune (Baloche et al., 2013), de 0,1 à 10% en bovins laitiers Holstein (Gao et al., 2012) et de 3 à 8% en Nordic red cattle (Koivula et al., 2012). Une comparaison de ces approches appliquées à des données caprines françaises sera présentée dans les chapitres 3 et 4. Il existe cependant des cas où l'approche two steps peut être légèrement meilleure, notamment dans le cas de données simulées où aucun traitement préférentiel ni aucune sélection ne sont appliqués (Vitezica et al., 2010; Aguilar et al., 2010).

Même si l'approche single step permet une augmentation sensible des précisions des évaluations génomiques, la taille de la population de référence reste un facteur limitant pour la précision des évaluations génomiques. Certaines études se sont donc intéressées à des évaluations génomiques multiraciales afin d'augmenter la taille de la population de référence et d'améliorer l'efficacité de la sélection.

1.III.1.1.2 Cas particulier des évaluations génomiques multiraciales

Contrairement au regroupement de populations de même race (comme par exemple dans le consortium Eurogenomics en race bovine Holstein), les évaluations génomiques multiraciales nécessitent de prendre en compte les différences entre races. En effet, réaliser une évaluation génomique multiraciale en groupant simplement deux populations de races différentes, implique que les fréquences alléliques sont les mêmes dans les deux races et correspondent à celles estimées dans la population multiraciale. De plus, les effets des marqueurs sont supposés similaires dans les deux races. Afin de prendre en compte les différences de fréquences alléliques, Harris et Johnson (2010) ont proposé de construire une matrice génomique spécifique nécessitant entre autre l'estimation d'une proportion de races pour chaque individu. Un exemple du calcul de cette proportion de race est présenté en Figure 1 .8. Dans cette figure l'individu X est issu d'une mère de race 1 et d'un père de race 2, sa « proportion de race 1 » est donc égale à sa « proportion de race 2 » soit 50%. L'individu Y issu de X et d'un père de race 2 reçoit de son père une proportion de 50% de race 2 et de sa mère (25% de race 2 et 25% de race 1). Cette méthode est coûteuse en temps de calcul et nécessite d'avoir génotypé des individus croisés (Harris and Johnson, 2010; Harris et al., 2014).

Dans le cas d'une population constituée de deux races, Erbe et al. (2012) proposent d'ajuster la matrice génomique pour la différence de fréquences alléliques entre les races, avec $G_c = \frac{M - 2p}{2\sum_i p_i(1-p_i)}$, la matrice **M** étant définie par le modèle BLUP-SNP $(y = \mathbf{X\beta + Mg} + e)$ et p est défini par $p = \alpha p_1 + (1-\alpha)p_2$ avec α la proportion d'animaux génotypés de la race 1, p_1 la fréquence d'un des allèles dans la race 1 et p_2 la fréquence de cet allèle dans la race 2.

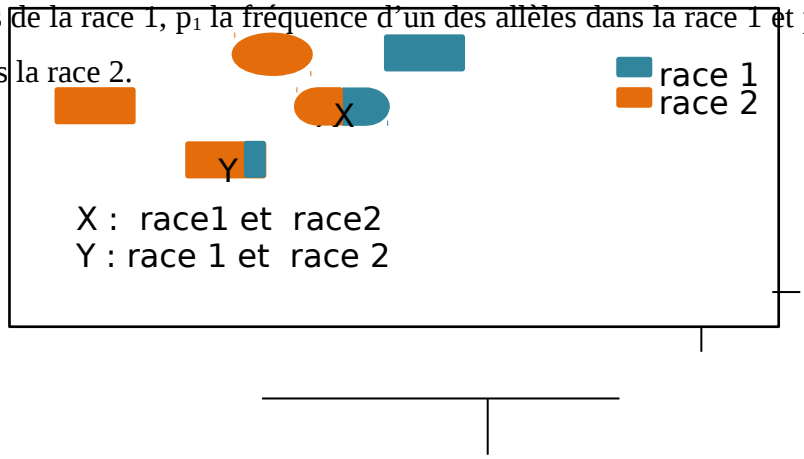


Figure 1.8 : Exemple de calcul d'une proportion de race pour l'individu Y

Les auteurs proposent également de corriger cette matrice $\mathbf{G}_c$ pour la différence de coefficient de consanguinité entre les deux races. Cette correction de la matrice génomique permet d'augmenter les précisions des GEBV de l'ordre de 8% en moyenne pour les caractères de production laitière dans la race Jersiaise en utilisant une population de référence composée de bovins de races Holstein et Jersiaise (Erbe et al., 2012). Dans le cas d'individus de races bien distinctes sans ancêtre commun, la matrice génomique peut être estimée par : $G_{c2} = \frac{\mathbf{LL}'}{m}$ avec m le nombre total de marqueurs et **L** défini par : $L = \left\{ \frac{x - 2p}{\sqrt{p(1-p)}} \right\}_{ij}$ où x_{ij} est le génotype de l'individu i pour le SNP j et p_{ij} est la fréquence allélique pour le SNP j estimée pour la race de l'individu i (Makgahlela et al., 2013b). Les fréquences alléliques utilisées dans cette approche correspondent aux fréquences estimées intra-race. Cette méthode permet d'augmenter les précisions des valeurs génomiques estimées de 3 à 5 % pour la matière protéique, selon la race considérée en bovins laitiers (Harris et al., 2014). D'autres méthodes basées sur une analyse en composante principale (ACP) des génotypes (Price et al.,

2006) ou sur la construction d'une matrice de distances euclidiennes entre les individus i et j de races distinctes (Gianola and Kaam, 2008) permettent d'ajuster la matrice génomique pour les différences observées entre races. La méthode basée sur une ACP ne permet pas d'améliorer les précisions, en revanche la méthode basée sur les distances euclidiennes donne les mêmes résultats que la méthode de Makgahlela et al. (2013) (Harris et al., 2014).

La correction de la matrice de variance-covariance des valeurs génétiques n'est pas la seule façon de prendre en compte les différences raciales dans les évaluations génomiques multiraciales. Olson et al. (2012) et Karoui et al. (2012) ont étudié un modèle multicaractère qui considère qu'un caractère mesuré en race 1 est différent du même caractère mesuré en race 2, les deux caractères sont corrélés génétiquement entre eux. Le modèle général pour n races différentes s'écrit $y = \begin{cases} X_1\beta_1 + T_1u_1 + e_1 \\ \dots \\ X_n\beta_n + T_nu_n + e_n \end{cases}$ avec :

$$\text{var}\begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{pmatrix} = \begin{pmatrix} \sigma^2_{u_1} & \sigma_{u_{1,2}} & \dots & \sigma_{u_{1,n}} \\ \sigma_{u_{2,1}} & \sigma^2_{u_2} & \dots & \sigma_{u_{2,n}} \\ \dots & \dots & \dots & \dots \\ \sigma_{u_{n,1}} & \sigma_{u_{n,2}} & \dots & \sigma^2_{u_n} \end{pmatrix} \otimes \begin{pmatrix} G_1 & G_{1,2} & \dots & G_{1,n} \\ G_{2,1} & G_2 & \dots & G_{2,n} \\ \dots & \dots & \dots & \dots \\ G_{n,1} & G_{n,2} & \dots & G_n \end{pmatrix}$$

, où $\sigma^2_{u_i}$ est la variance génétique pour le caractère considéré de la race i, $\sigma_{u_{i,j}}$ est la covariance génétique entre le caractère de la race i et ce même caractère de la race j. G_i est la matrice génomique d'apparentement entre les animaux de race i et $G_{i,j}$ est la matrice génomique d'apparentement entre les animaux de race i et ceux de race j, toutes deux construites selon VanRaden (2008) où les fréquences alléliques sont estimées dans la population multiraciale. Cette approche permet d'augmenter les précisions des évaluations génomiques de 8 à 13% pour les caractères de production laitière en bovins laitiers américains (Holstein, Jersey ou Brune) par rapport à une approche multiraciale considérant les deux races comme une seule et même race avec ajustement de la matrice génomique (Olson et al., 2012). Cependant les résultats de ce type de modèle dépendent des corrélations génétiques estimées entre les races. En effet, le niveau des corrélations génétiques indique si l'effet d'un QTL est similaire dans les deux races. Si ces corrélations sont élevées, les effets des QTL sont les mêmes dans les

deux races, il est donc pertinent d'utiliser une population multiraciale pour l'évaluation génomique. C'est le contraire si les corrélations obtenues sont faibles (Karoui et al., 2012).

Que ce soit avec ou sans correction pour les différences entre race, utiliser des évaluations génomiques multiraciales plutôt qu'uniraciales ne permet d'augmenter que faiblement les précisions des évaluations génomiques. Elles sont par exemple augmentées de 2% en moyenne en race bovine laitière française Normande avec une population multiraciale composée de vaches de races Holstein, Normande et Montbéliarde (Hozé et al., 2014). L'augmentation des précisions dépend de l'apparentement entre les races considérées. Les bovins laitiers américains de race Holstein et Jersey étant fortement apparentés, l'utilisation d'une population multiraciale entraine des gains de précisions de 13% en moyenne pour les caractères de production laitière dans la race Jersey (Hayes et al., 2009a). En ovins laitiers, considérer une population de référence multiraciale composée de races génétiquement proches (Manech tête Rousse et Laxta Cara Negra Euskadi ou Manech Tête Noire et Laxta Cara Negra Navarra) permet d'augmenter les précisions jusqu'à 18% pour la Manech Tête Noire (Legarra et al., 2014a). De même, en bovins laitiers Hostein rouge (Danish red, Swedish red et Finnish Ayrshire) les précisions des évaluations génomiques sont augmentées de 6 à 10% avec les évaluations multiraciales (Makgahlela et al., 2012). Inversement, Pryce et al. (2011) montrent que la race Fleckvieh n'est pas assez apparentée aux races Hostein et Jersey pour que l'utilisation d'une population multiraciale soit bénéfique.

Les résultats dépendent également de la taille des populations de référence étudiées. Olson et al. (2012) ont montré que les précisions des évaluations génomiques ne sont pas meilleures en Holstein et Jersey lorsque l'on considère une population multiraciale composée d'animaux de race Hostein, Jersey et Brune comparée à des populations de référence uniraciales (Holstein ou Jersey). En revanche, les précisions des GEBV des animaux de race Brune sont augmentées de 30 à 44% pour les caractères de production laitière quand la population de référence passe de 185 mâles à 3 105 mâles en considérant les trois races (185 Brunes, 413 Jersey et 2 507 Holstein) ensemble.

Les méthodes d'estimation des valeurs génomiques utilisant une matrice génomique ont l'avantage d'être similaires à celles utilisées pour les évaluations génétiques classiques. Elles permettent de réaliser des ajustements de la matrice génomique pour prendre en compte des différences de structure de population entre races différentes par exemple. Cependant, la méthode GBLUP suppose que tous les marqueurs ont un petit effet sur le caractère. Or il existe des gènes connus ou régions QTL qui ont un fort effet sur certains caractères.

1.III.1.2 Méthode basée sur une estimation des effets des SNP

Les approches bayésiennes développées dans le cadre des évaluations génomiques permettent de contrôler la proportion de SNP conservés. Le modèle général de l'évaluation génomique ($\mathbf{y = X\beta + Mg + e}$) suppose une distribution normale de l'effet global des effets des SNP ($\mathbf{g} \sim N(0, I\sigma_g^2)$ où la majorité des SNP sont supposés avoir un effet faible. Les méthodes bayésiennes développées (BayesA, BayesB, BayesC, BayesCπ, Bayes R et LASSO Bayésien par exemple) diffèrent par la distribution *a priori* des effets des SNP. Dans la méthode Bayes A, les effets des SNP suivent une loi normale avec une variance spécifique pour chaque SNP. Les variances de ces effets suivent une loi de χ^2 inversée. La méthode Bayes B comme le LASSO Bayésien, et contrairement au Bayes A, ne sélectionne que les marqueurs dont l'effet sur le caractère est significatif. La proportion de SNP ayant un effet sur le caractère est supposée connue dans la méthode BayesB.

La méthode BayesR suppose qu'une certaine proportion de marqueurs a un effet nul tandis que les autres marqueurs ont un effet faible ou modéré. Les effets des SNP sont donc distribués selon un mélange de quatre lois normales de variances différentes (0%, 0,01%, 0,1 % et 1% de la variance génétique totale (σ_g^2). Dans cette méthode, la variance génétique totale est égale à la variance du caractère multipliée par la fiabilité (précision au carré) du caractère. Dans le cas du LASSO Bayésien, les effets des SNP sont supposés suivre une loi de Laplace qui permet de faire l'hypothèse que très peu de marqueurs ont un effet important alors que la majorité ont un effet quasi-nul. La loi *a priori* des effets des marqueurs est la suivante : $p(g|\sigma^2,\lambda) = \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|g|}{\sqrt{\sigma^2}}\right)$ avec $\sigma_g^2 = \frac{2\sigma^2}{\lambda}$ où λ est un paramètre d'échelle permettant de définir l'intensité de sélection des SNP. Le défaut de cette méthode est d'utiliser la même variance (σ^2) pour modéliser les effets des marqueurs et les résidus, ce qui n'est pas optimal car les effets des marqueurs ne sont pas liés aux effets résiduels (Legarra et al., 2011). Legarra et al. (2011) ont développé un LASSO Bayésien qui corrige ce phénomène en séparant la variance résiduelle (σ_e^2) de la variance génétique (σ_g^2). Les lois des effets des résidus et des effets des SNP sont ainsi définies par : $\mathbf{g}|\lambda, \sigma_g^2 \sim$

$\prod_i \frac{\lambda}{2\sigma_g} \exp\left(\frac{-\lambda|g_i|}{\sigma_g}\right)$ et $e|\sigma_e^2 \sim \text{MVN}\left(0, I\sigma_e^2\right)$, les résidus suivant une loi normale multivariée. Les précisions des évaluations génomiques obtenues avec le LASSO Bayésien sont similaires à celles obtenues avec la méthode BayesB, mais le LASSO est beaucoup moins coûteuse en temps de calcul (Campos et al., 2009). Elles sont cependant légèrement meilleures (entre 4% pour la matière protéique et 14% pour le taux butyreux) que celles obtenues avec la méthode BayesA en bovins laitiers Hostein espagnols (Jiménez-Montero et al., 2013).

La méthode BayesC est semblable au BayesB excepté que la variance des effets des SNP avec un effet sur le caractère étudié, est supposée la même pour tous les marqueurs. Cette méthode, comme la méthode BayesB implique de connaitre la proportion de SNP (π) qui ont un effet sur le caractère. Or la valeur de π a une influence sur les précisions des valeurs génomiques obtenues (Su et al., 2010). Une méthode appelée BayesCπ a été développée afin d'estimer ce paramètre (Habier et al., 2011). La loi des effets des marqueurs dans le cas du BayesCπ est définie par :

$$\begin{cases} p\left(g|\pi,\sigma_g^2\right)=0 & \text{avec une probabilité } 1-\pi \\ p\left(g|\pi,\sigma_g^2\right) \sim \text{N}\left(0,\sigma_g^2\right) & \text{avec une probabilité } \pi \end{cases}.$$

Les précisions des évaluations génomiques obtenues en bovins laitiers avec le BayesCπ sont similaires voire meilleures (+17% pour le taux butyreux) que celles obtenues avec le LASSO Bayésien (Colombani et al., 2013).

Les méthodes Bayésiennes, permettant de prendre en compte des effets différents par marqueur, apparaissent particulièrement judicieuses dans le cas de caractères gouvernés par des gènes majeurs à effet fort, comme le gène DGAT1 pour le taux butyreux en bovins laitiers. Il est cependant difficile de conclure sur l'intérêt des méthodes bayésiennes par rapport au GBLUP pour augmenter les précisions des évaluations génomiques. Certaines études, basées sur des pseudo-performances (GBLUP two steps) montrent une légère amélioration des précisions avec les approches bayésiennes (Calus and Veerkamp, 2011; Hayes et al., 2009b). En ovins laitiers Lacaune, l'augmentation des précisions obtenues avec l'utilisation du BayesCπ est d'environ 2 à 5% (Duchemin et al., 2012). En bovins laitiers Hostein et Montbéliarde, l'utilisation du LASSO Bayésien ou du BayesCπ a permis d'augmenter les précisions de 2 à 19% pour des caractères d'héritabilité moyenne à forte (h^2 de 0,3 et 0,5 (Colombani et al., 2013). Les précisions sont cependant meilleures avec une méthode GBLUP pour le caractère de fertilité (+9%) dont l'héritabilité est beaucoup plus

faible. Pryce et al. (2011) ainsi que Makgahelela et al (2011) ont montré que les précisions obtenues avec le BayesA et le BayesC sont similaires ou moins bonnes que celles obtenues avec le GBLUP. Dans une population de bovins laitiers Norvégiens, les précisions des évaluations génomiques obtenues avec le BayesB sont moins élevées de 2,5 à 8% que celles obtenues avec le GBLUP (Luan et al., 2009). L'utilisation de méthodes Bayésiennes pour réaliser des évaluations génomiques en caprins laitiers pourra donc être testée au même titre que l'utilisation du GBLUP.

1.III.2 Phénotypes utilisés

Nous avons vu précédemment qu'il est possible de réaliser des évaluations génomiques sur les performances brutes des animaux via la méthode GBLUP single step. Dans ce cas, la question du choix du phénoytpe à utiliser ne se pose pas. En revanche, pour les évaluations génomiques basées sur les pseudo-performances (approche two steps), plusieurs phénotypes peuvent être envisagés. Dans la plupart des cas les phénotypes utilisés sont les daughter yield deviation (DYD) pour les mâles et les yield deviation (YD) pour les femelles. Les YD sont les performances brutes corrigées des effets fixes, des effets aléatoires autres que génétiques, ainsi que de l'effet génétique de la mère. On peut poser le modèle suivant :

$$y = \beta + p + \frac{1}{2} g_{\text{père}} + \frac{1}{2} g_{\text{mère}} + \mathbf{ms} + e$$

où **y** est le vecteur des performances, **β** le vecteur des effets de milieu, $g_{\text{père}}$ et $g_{\text{mère}}$ les effets polygéniques du père et de la mère respectivement, **p** l'effet aléatoire d'environnement permanent, **ms** l'aléa de méiose et **e** le vecteur des résidus. Dans ce cas, les yield deviations sont définies par : $\text{YD} = y - \left(\hat{\beta} + \hat{p}\right)$ (Szyda et al., 2008). Les DYD représentent la moyenne des performances des filles corrigées des effets fixes et aléatoires non génétiques et du niveau génétiques des mères de ces filles. Les DYD sont donc calculés comme la moyenne pondérées des YD de toutes les filles corrigés du niveau génétique de leur mère :

$$\text{DYD} = \frac{\sum_i \left(Y\hat{D}_i - \frac{1}{2} g_{\text{mère}_i} \right) \hat{w}_i}{\sum_i \hat{w}_i}$$

où YD_i est le YD de la femelle i, $g_{\text{mère}_i}$ est l'effet polygénique de la mère de la femelle i et w_i le poids associé dépendant du nombre de lactations de la femelle. Dans le cas de performances répétées pour une femelle, YDi est égal à la somme pondérée par le poids des lactations de tous les YD de

cette femelle. Les DYD sont analogues à une performance de mâle avec une héritabilité égale à la précision de l'index sur descendance. Fikse et al. (2001) ont montré que lorsque les DYD sont utilisés dans les évaluations génomiques, le poids optimal à leur associer est la contribution des filles ou effective daughter contribution (EDC) en anglais. Les EDC peuvent être calculées selon la formule suivante : $\mathrm{EDC}_i = \sum_k \frac{\frac{4-\mathrm{h}^2}{\mathrm{h}^2} R_k}{4 - R_k\left(1+R_m\right)}$ où EDC_i est l'EDC pour l'animal i, R_m le coefficient de détermination (CD) de la valeur génétique de la mère et $R_k = \frac{n_k \mathrm{h}^2}{1+\left(n_k-1\right)r}$ où r est la répétabilité de la performance et n_k est défini par $n_k = \sum_l 1 - \frac{1}{n_{\mathrm{jkl}}}$, n_{jkl} étant la taille du groupe de contemporaines k dans lequel la fille j a effectué sa l[ème] lactation (Fikse and Banos, 2001).

Lorsque que les DYD ne sont pas disponibles, par exemple dans le cas d'évaluations internationales où seules les valeurs génétiques estimées (EBV) sont disponibles, les index dérégressés ou deregressed proofs (DRP) en anglais peuvent être utilisés comme phénotype (Robert-Granié et al., 2011). Dans le cas d'un modèle génétique $y = \mathbf{X\beta + Zu} + e$ où **y** est le phénotypes, **β** le vecteur des effets fixes, **u** le vecteur des valeurs génétiques et **e** le vecteur des résidus (cf 1.III) avec $\mathrm{Var}\left(u\right) = A\sigma_u^2$, A étant la matrice de parenté estimée à partir du pédigrée, les index dérégressés peuvent s'écrire : $\mathrm{DRP} = \left[I + \left[Z'Z - Z'X\left(X'X\right)^{-}X'Z\right]^{-1} A^{-1}\lambda\right]\mathrm{EBV}$ où λ est le ratio entre les variances résiduelle et génétique (Thomsen et al., 2001). Les DRP sont équivalents aux DYD car par construction si on les utilise comme phénotypes dans une évaluation de type BLUP, on retrouve les mêmes estimations que celles obtenues à partir du modèle complet. Un index dérégressé correspond à la valeur génétique estimée divisée par sa précision. Garrick et al. (2009) ont montré qu'il est préférable de corriger les DRP pour l'information parentale, puisque par construction ils sont régressés sur l'information des parents, ce qui peut poser problème lorsque les parents n'ont pas de performance. La formule permettant le calcul de l'index dérégressé dans le cas où seuls les EBV des animaux sont disponibles est complexe :

$$DRP_i = \frac{\hat{y}_i}{Z_i' Z_i} \quad \text{avec} \quad \hat{y}_i = -2\left(\frac{1-h^2}{h^2}\right) g_{PA} + \left(Z_i' Z_i + 2\frac{1-h^2}{h^2}\right) g_i \ ,$$

$$Z_i' Z_i = \left(\frac{0{,}5 - r_{PA}^2}{1 - r_i^2}\right) Z_{PA}' Z_{PA} + 2\left(\frac{1-h^2}{h^2}\right) \times \left(2\left(\frac{0{,}5 - r_{PA}^2}{1 - r_i^2}\right) - 1\right) \quad \text{et}$$

$$Z_{PA}' Z_{PA} = \left(\frac{1-h^2}{h^2}\right) \times \left(0{,}5\left(\frac{1}{0{,}5 - r_{PA}^2}\right) - 4\right) + 0{,}5 \times \frac{1-h^2}{h^2} \times \sqrt{\left(\frac{1}{0{,}5 - r_{PA}^2}\right)^2 + \frac{16}{\frac{0{,}5 - r_{PA}^2}{1 - r_i^2}}} \ .$$

Ce calcul dépend de la valeur génétique (g_i) de l'individu i, de la moyenne de la valeur génétique de ses parents ($g_{PA} = \frac{g_{mère} + g_{père}}{2}$), du CD de l'individu i (r_i) et de la moyenne du CD des parents $\left(r_{PA}^2 = \frac{r_{mère}^2 + r_{père}^2}{2}\right)$. Le poids associé à ce phénotype (w_i) est estimé à l'aide de la formule suivante : $w_i = \frac{1-h^2}{\left(c + \frac{1-r_i^2}{r_i^2}\right) h^2}$ où c correspond à la part de variance génétique non captée par les marqueurs, elle est considérée comme égale à 0,5 dans l'article de Garrick et al., (2009).

Peu d'études comparent les précisions des évaluations génomiques obtenues avec les DRP ou les DYD. Dans la pratique, les index dérégresssés ne sont utilisés que dans les cas où les DYD ne sont pas disponibles. En revanche quelques études ont comparé l'utilisation des DYD par rapport aux valeurs génétiques estimées (EBV) comme phénotypes, principalement pour la détection de QTL, donnant tantôt l'avantage aux DYD (VanRaden and Wiggans, 1991; Hoeschele and VanRaden, 1993; Israel and Weller, 1998), tantôt l'avantage aux EBV (Israel and Weller, 2002). Sur données simulées, les précisions des évaluations génomiques obtenues avec les EVB sont légèrement plus élevées (entre 0,3 et 3,6%) que celles obtenues avec les DYD. Cette augmentation de précision semble être due au fait que les EBV ont des meilleures précisions que les DYD (Guo et al., 2010). Cependant, l'EBV d'un individu prend en compte l'information de tous ses apparentés contrairement aux DYD qui ne prennent en compte que l'information des descendants. Utiliser les EBV dans un modèle d'évaluation génomique

signifie que l'information des apparentés sera comptabilisée deux fois : une première fois dans les estimations des EBV et une deuxième fois lors du calcul des GEBV. Même si ce problème peut être résolu en corrigeant les GEBV obtenues par la précision des EBV des animaux de la population de référence, l'utilisation des EBV comme phénotypes pour les évaluations génomiques ne doit pas être envisagée (Garrick et al., 2009). Dans le cadre de ma thèse, nous étudierons l'impact de l'utilisation de phénotypes différents (DYD, index dérégressés corrigés ou non) sur les précisions des évaluations génomiques chez les caprins français.

1.III.3 Évaluation de la qualité et de l'intérêt des évaluations génomiques

1.III.3.1 Validation des évaluations génomiques

La qualité des évaluations génomiques est évaluée généralement par le biais de ce que l'on appelle la précision de l'évaluation génomique. Elle correspond à la corrélation entre la « vraie » valeur génétique et la valeur génomique estimée. En pratique, la « vraie » valeur génétique est inconnue, elle est donc approchée par les DYD ou YD des animaux. Ces phénotypes corrigés pour les effets fixes, bien que plus précis qu'une performance ne sont qu'une approximation imparfaite de la valeur génétique vraie. Ces précisions sont évaluées par un procédé que l'on appelle validation croisée (Figure 1 .9) correspondant à une étude rétrospective. Elle nécessite de diviser la population de référence en deux sous populations : la population d'apprentissage à partir de laquelle on estime les effets des SNP et la population de validation, plus jeune, pour laquelle on prédit les GEBV. On se place par exemple en 2008, année pour laquelle les individus de la population de validation nés après 2006 n'ont pas encore de performances. On prédit ainsi la valeur génomique des individus de la population candidate à partir de l'information de leurs génotypes. Cinq ans plus tard, en 2013 les individus de la population candidate ont à leur tour des filles avec performances propres, il est donc possible d'obtenir leurs phénotypes (les DYD). La précision de l'évaluation génomique est ensuite calculée comme le coefficient de corrélation de Pearson entre les valeurs génomiques estimées (GEBV) prédites en 2008 et les DYD obtenues en 2013 pour la population de validation. Il est également possible de calculer les pentes de régression des GEBV en fonction des DYD, l'optimum étant une pente égale à 1. Cette mesure permet d'évaluer deux phénomènes la surestimation (pente >1) ou la sous-estimation (pente <1) ainsi que la sur-dispersion (pente >1) ou la sous dispersion (pente <1) des valeurs génétiques estimées des animaux de la population de validation. Certains auteurs appellent ce coefficient

le biais des évaluations génomiques, cependant ce terme est critiqué car les DYD ne sont pas les valeurs génétiques vraies des individus et peuvent être biaisées.

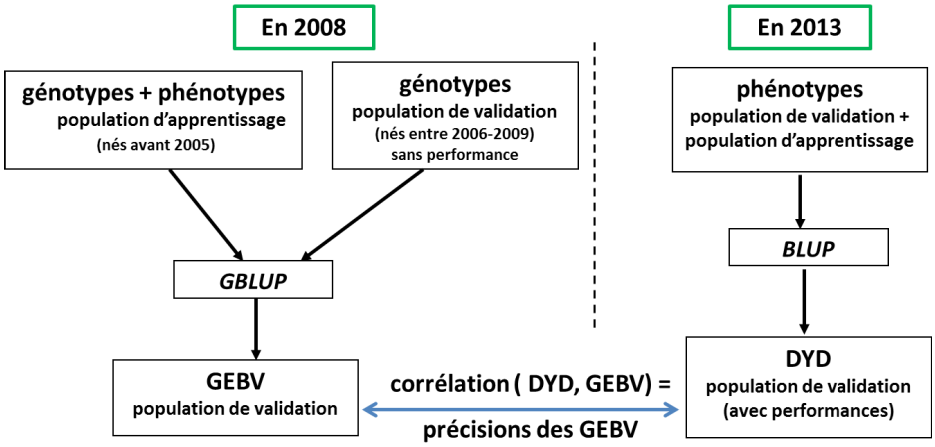


Figure 1.9 : Schéma d'un exemple de validation croisée permettant le calcul de la précision des évaluations génomiques

Le schéma de la validation croisée, c'est-à-dire le choix du partitionnement de la population de référence en population d'apprentissage et de validation est en général défini par la taille de la population de référence. Les auteurs considèrent en général que la population d'apprentissage comprend 75% des individus de la population de référence, ce qui correspond aux individus les plus âgés. Ce partitionnement peut toutefois s'avérer délicat dans le cas de très petites populations. C'était le cas de la race ovine Latxa Cara Negra Navarre (Legarra et al., 2014a) pour laquelle il était difficile d'envisager une évaluation génomique à partir de seulement 67 individus dans la population de référence (trop peu d'animaux étant génotypés).

La qualité des évaluations génomiques peut aussi être étudiée en comparant la précision des évaluations génomiques à la précision des évaluations génétiques par validation croisée. Mais les précisions obtenues par validation croisée ne correspondent pas aux précisions habituellement considérées en sélection classique qui sont les coefficients de déterminations (CD). Ce CD (reliability en anglais) est comparable au carré de la corrélation de validation entre les GEBV prédits et les DYD, appelée accuracy en anglais.

1.III.3.2 Précision génomique théorique des candidats

Dans cette thèse, nous évaluons l'intérêt de la sélection génomique chez les caprins laitiers français en comparant la précision théorique des valeurs génétiques des individus de la population candidate obtenu avec les évaluations génomiques à la précision obtenue sur

ascendance ($\text{précision}_{\text{ascendance}} = \sqrt{\frac{CD_{père} + CD_{mère}}{4}}$) obtenu avec les évaluations génétiques classiques. Dans le cas où la précision génomique théorique moyenne des candidats est supérieure à la précision obtenue sur ascendance, nous considérons que la sélection génomique est envisageable puisque la précision génomique théorique est supérieure à celle obtenue à l'aide des évaluations classiques.

Les précisions génomiques théoriques ($\sqrt{CD}$) sont calculés à partir des variances d'erreur de prédiction (PEV) et de la variance génétique (σ_u^2) estimée à partir des GEBV selon la formule : $\sqrt{CD} = \sqrt{1 - \frac{PEV^2}{\sigma_u^2}}$ (Bijma, 2012). Les variances d'erreur de prédictions sont estimées à partir de l'inverse de la matrice des équations du modèle mixte. Cependant cette inversion est impossible lorsque le système d'équations devient trop grand, ce qui est le cas pour les évaluations génomiques single step (Mizstal et al., 2013). Il existe plusieurs méthodes d'approximation du calcul des précisions théoriques, elles sont basées sur le nombre de performances d'un animal ainsi que sur les contributions des parents et des descendants à la valeur génétique de l'individu (Wiggans et al., 1988). Cependant ces méthodes ne prennent pas en compte les particularités des évaluations génomiques. Ces précisions doivent être proportionnelles au nombre d'animaux génotypés. De plus, les précisions théoriques des GEBV des candidats doivent être indépendantes de l'apparentement génomique entre les candidats. Misztal et al. (2013) ont proposé une approximation pour calculer les précisions théoriques des GEBV basées sur l'inversion de la matrice de variance-covariance des valeurs génétiques des seuls animaux génotypés. Une précision téhorique approchée est défini par la formule suivante : $\sqrt{CD} = \sqrt{1 - \frac{\alpha}{\alpha + d_i}}$ avec α le ratio de variance résiduelle par rapport à la variance génétique, $d_i = \frac{1}{\text{LHS}_i} - \alpha$ avec $\text{LHS}_i \approx \left\{ \left[D_i^r + D_i^p + \left(I + G^{-1} - A_{22}^{-1} \right) \alpha \right]^{-1} \right\}_i$, D_i^r la contribution des phénotypes et D_i^p la contribution du pédigrée.

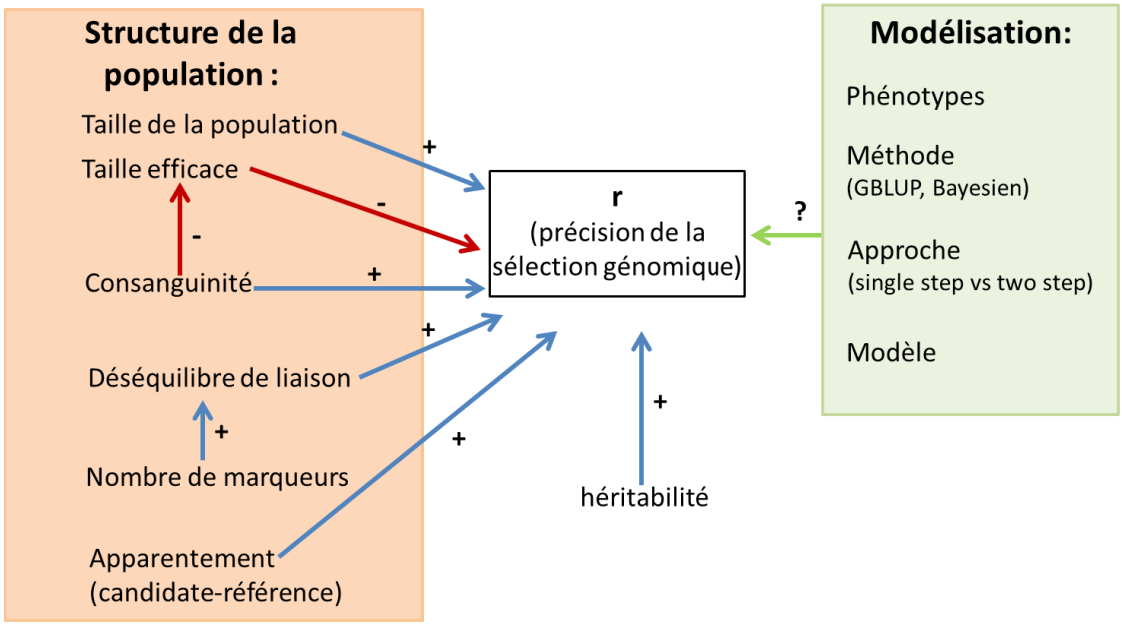


Figure 1.10 : Schéma des différents paramètres influençant la précision des évaluations génomiques

En conclusion de ce chapitre, la précision des évaluations génomiques est un des points clés permettant ou non d'envisager la sélection génomique dans l'espèce caprine. De nombreux paramètres influencent la précision des évaluations génomiques comme le montre la Figure 1 .10. La structure génétique de la population impacte les précisions des valeurs génomiques estimées. En effet, plus le déséquilibre de liaison entre marqueurs est élevé, plus la précision des évaluations génomiques est importante. De même, plus l'apparentement entre la population candidate et la population de référence est élevé, plus les valeurs génétiques estimées des candidats seront précises. En revanche, la diversité génétique (évaluée via la taille efficace et la consanguinité) n'est pas un avantage pour la prédiction génomique car elle rend plus difficile l'établissement du lien entre les marqueurs et le phénotype. Il est cependant important de noter que l'effet négatif de la diversité génétique sur les précisions peut être compensé par la taille de la population de référence. Enfin, la modélisation choisie (modèles et méthodes) a également un impact sur les précisions des valeurs génomiques estimées.

Chapitre 2 : Structure de la population caprine française

2.I Puce et génotypages

Le design de la puce Illumina goat SNP50 BeadChip est basé sur le séquençage de 6 races : Alpine, Saanen et Créole pour les races laitières et Boer, Katjang et Savanna pour les races à viande. Les marqueurs (SNP) ont été choisis comme les plus polymorphes pour le maximum de ces races. Le nombre total de SNP sur la puce est de 52 295 (Tosser-Klopp et al., 2014). La distance moyenne entre 2 SNP consécutifs est de 51 kb alors que la distance médiane est de 43 kb (Figure 2 .11). Certaines zones du génome caprin ayant un effet connu ont été densifiées, c'est le cas pour la composition du lait avec le cluster des caséines sur le chromosome 6 et la région du gène diacylglycerol O-acyltransferase homolog 1 (DGAT1) sur le chromosome 14, ou encore pour la résistance à la tremblante avec le gène protéine prion (PRP) sur le chromosome 13.

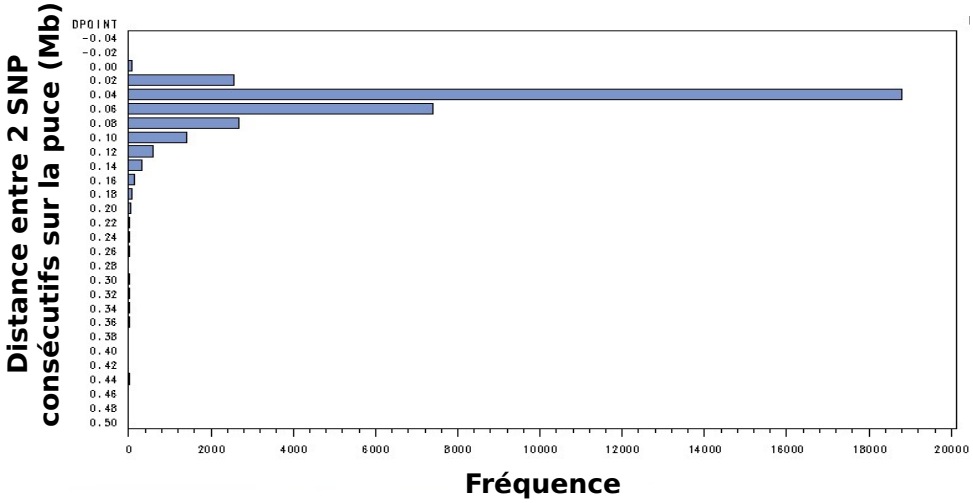


Figure 2.11 : Distribution de la distance entre deux SNP consécutifs sur la puce Illumina goat SNP BeadChip

Le choix des animaux à génotyper a été réalisé au regard de deux principaux objectifs. Le premier objectif était une primo-détection de QTL, le choix des animaux à génotyper s'est donc porté sur 20 familles de pères (11 Alpins et 9 Saanens) comprenant au minimum 100 filles par père (112 en moyenne) nées entre 2008 et 2009. Pour le deuxième objectif, le choix s'est porté sur les mâles d'IA afin d'étudier l'intérêt d'une sélection génomique. Tous les mâles testés sur descendance nés à partir de 1998 et jusqu'en 2011 ont été génotypés, soit un total de 884 mâles (Tableau 2 .5). Les millésimes les plus anciens n'ont pas été retenus, cependant quatre mâles nés entre 1993 et 1997 avec un grand nombre de filles ont été génotypés.

Tableau 2.5 : Tableau des effectifs des animaux génotypés avant et après le contrôle qualité de l'information moléculaire par sexe

	Mâles						Femelles		
	Avant contrôle qualité			Après contrôle qualité			Après contrôle qualité		
Années	Totaux	Alpins	Saanen	Totaux	Alpins	Saanens	Totaux	Alpine	Saanen
< 1997	*4*	4	0	*4*	4	0			
1998	*40*	20	20	*39*	20	19			
1999	*46*	23	23	*44*	23	21			
2000	*38*	19	19	*37*	19	18			
2001	*56*	29	27	*48*	23	25			
2002	*60*	28	32	*57*	28	29			
2003	*73*	42	31	*64*	36	28			
2004	*68*	40	28	*66*	41	25			
2005	*71*	42	29	*71*	42	29			
2006	*70*	40	30	*65*	40	25			
2007	*70*	41	29	*69*	40	29			
2008	*71*	41	30	*45*	28	17			
2009	*70*	40	30	*68*	40	28	420	239	181
2010	*76*	43	33	*76*	43	33	1565	936	629
2011	*75*	45	30	*72*	43	29			
Totaux	***884***	***498***	***386***	***825***	***470***	***355***	***1985***	***1175***	***810***

Un test de contrôle qualité a été effectué afin d'éliminer les SNP et génotypages pouvant être source d'erreurs (Figure 2 .12). Ce contrôle qualité est réalisé à l'aide d'une chaine de programmes écrits en awk et Fortran, validés en ovins Lacaune. La sélection des génotypages suit plusieurs étapes dont la première est focalisée sur les marqueurs. L'apparition de génotypes manquants est régulièrement due à une mauvaise séparation entre les "clusters" prédéfinis de la puce, qui permettent d'affilier chaque individu à un génotype. En général, un premier filtrage consiste à supprimer les SNP dont le pourcentage d'individus génotypés avec succès, appelé call freq, est inférieur à un seuil (ici pris égal à 99%). Puis les marqueurs ne respectant pas l'équilibre de Hardy-Weinberg sont éliminés. L'équilibre de Hardy-Weinberg correspond à une population fermée, d'effectif limité, soumise ni à la sélection ni aux mutations et où les accouplements se réalisent au hasard (panmixie). Cet équilibre induit des fréquences alléliques constantes de génération en génération. On vérifie cet équilibre à l'aide d'un test de χ^2 avec un degré de liberté égales au nombre d'allèles moins un, soit 1. La statistique de ce test est définie par : ($\sum \frac{(fréquence(génotype\ observé) - fréquence(génotype\ attendu))^2}{fréquence(génotype\ attendu)}$), où le génotype attendu correspond au génotype obtenu dans le cas d'une population sous l'équilibre de

Hardy-Weinberg. Enfin, les SNP pour lesquels les erreurs d'incompatibilités parent-descendant (filiation) sont trop fréquentes ainsi que ceux qui n'ont pas été positionnés sur un chromosome ou ceux dont la fréquence de l'allèle mineur (MAF pour minor allele frequency) est inférieure à 1% n'ont pas été retenus pour l'étude. La deuxième étape se focalise sur les individus et comprend deux phases (Figure 2 .12) : la première correspond à une sélection des individus ayant au moins 99% des SNP typés avec succès (call rate), la deuxième correspond à une élimination des animaux dont les erreurs de filiation avec les parents ou descendants concernent plus de 2% SNP. Ce dernier cas correspond majoritairement à des erreurs d'identification d'échantillons, et a abouti à l'élimination de 7,5% des animaux de notre étude.

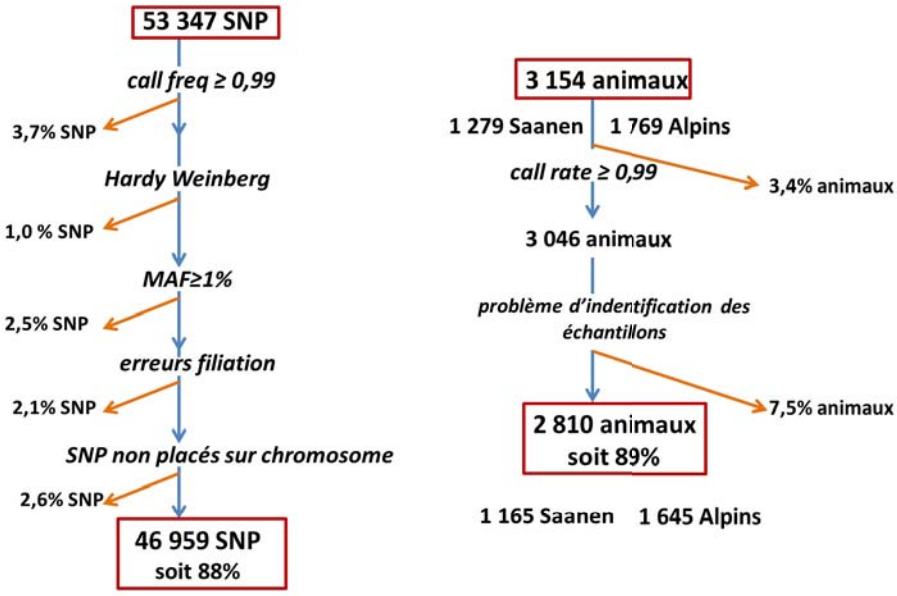


Figure 2.12 : Schéma des différentes étapes du contrôle qualité des génotypages

Une fois ce contrôle qualité réalisé, 46 959 SNP ont été gardés pour 825 et 1 985 femelles (Figure 2 .12). En 2012, au début de notre étude, les 677 mâles génotypés nés avant 2010 ont des filles phénotypées et enregistrées dans les bases de données nationales, ils ont donc une valeur génétique connue sur descendance pour différents caractères. Les 148 jeunes mâles nés entre 2010 et 2011 sont considérés comme des candidats pour toute l'analyse, car n'ayant aucune descendance en 2012.

2.II Déséquilibre de liaison

Le niveau de déséquilibre de liaison (DL) dans la population caprine française a été estimé selon la méthode de Roger and Huff (2009) décrite à la Figure 1 .5 (cf. 1.I.1) par pas de 20 kb à l'aide d'un programme en Fortran développé au sein de l'unité GenPhySE. Cette méthode a été préférée à la méthode de Hill and Robertson (1968) plus couramment utilisée en raison

d'une impossibilité de phasage des génotypes. En effet, une étude préalable (non présentée ici) a montré que le phasage des génotypes à l'aide de méthodes telles que Beagle (Browning and Browning, 2007) n'est pas satisfaisant en raison du peu de mères génotypées.

Le niveau de DL a tout d'abord été estimé dans la population totale (cf. Figure 2 .13) comprenant tous les animaux génotypés mâles et femelles de race Alpine et Saanen sans distinction, soit 2 810 animaux. Comme attendu le niveau de DL estimé décroit avec la distance entre SNP de façon conséquente jusqu'à une distance de 150 kb entre marqueurs. Au-delà de 1 200 kb le niveau de DL moyen converge vers une valeur de 0,03. Le niveau de DL estimé dans la population des 825 mâles (cf. Figure 2 .13) est très proche de celui de la population totale : 0,17 pour une distance entre SNP de 50 kb correspondant à la distance moyenne entre deux SNP consécutifs sur la puce 50K. En revanche le niveau de DL obtenu dans la population des 1 985 femelles est plus important (0,22 pour une distance de 50 kb). Cette différence peut s'expliquer par les liens de parenté plus élevés entre les femelles de cette population (groupes de 100 demi-sœurs).

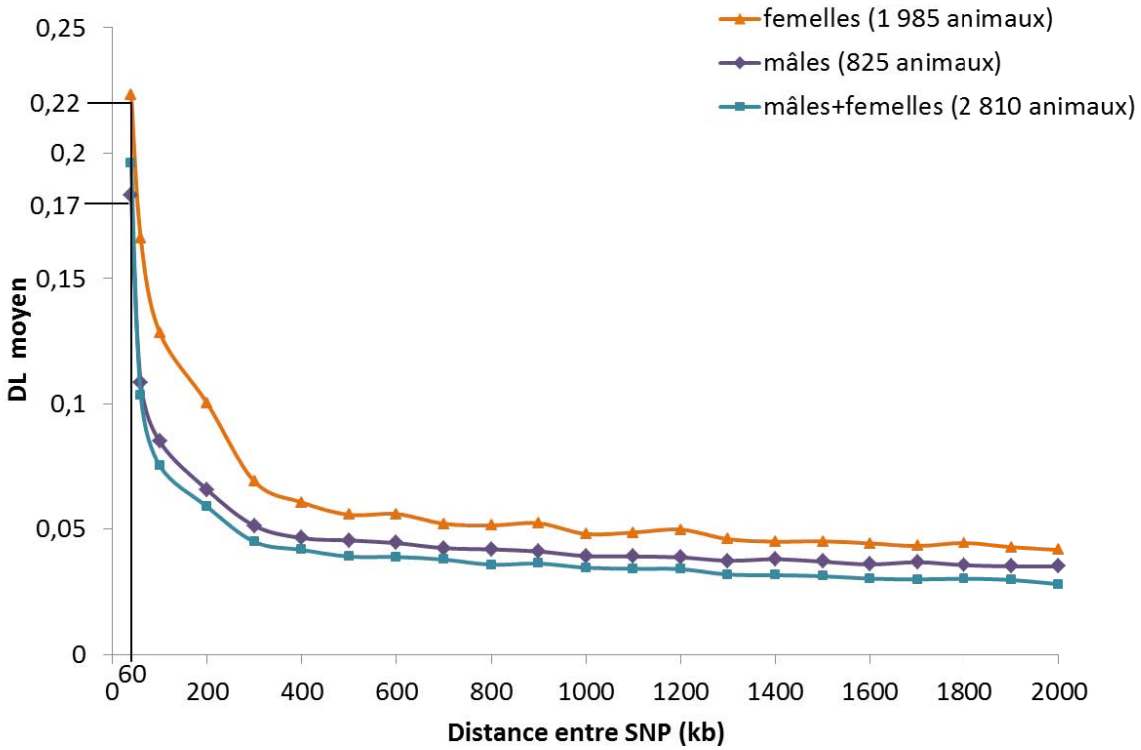


Figure 2.13 : Estimation du niveau de déséquilibre de liaison dans les différentes sous populations étudiées constituées de l'ensemble des animaux (mâles+femelles), des mâles ou des femelles uniquement

Le niveau de DL observé à 50kb dans la population caprine française est similaire au DL estimé dans les populations caprines canadiennes (0,14 en Alpine et 0,15 en Saanen (Brito et al., 2014)). Il est proche de celui obtenu en bovins allaitants (entre 0,13 pour la race

Brahman et 0,23 pour la race Hereford (Porto-Neto et al., 2014)) et laitiers (entre 0,18 et 0,23 pour les Holsteins à 50 kb (De Roos et al., 2008)), ainsi qu'en ovins laitiers (entre 0,13 et 0,14 à 50 kb pour les Lacaune (Baloche et al., 2013), et Manech Tête Noire et Rousse (Legarra et al., 2014a)). Cependant ce niveau est inférieur au DL estimé dans les populations caprines australiennes Toggenburg, Nubian et Boer (entre 0,25 et 0,29 (Brito et al., 2014)) ainsi que dans l'espèce porcine (entre 0,36 pour les Landrace et 0,46 pour les Duroc à 50 kb (Badke et al., 2012)), et chez les volailles (entre 0,24 pour la race White Leghorn et 0,64 pour la race Rhode Island Red entre 50 et 75 kb (Qanbari et al., 2010)).

La Figure 2 .14 présente le déséquilibre de liaison en fonction de la distance entre SNP estimé sur la population des mâles génotypés pour la race Alpine, la race Saanen ou les deux races combinées (Alpine+Saanen), elles est commentée dans l'article I (cf. 3.I.1 (Carillier et al., 2013)).

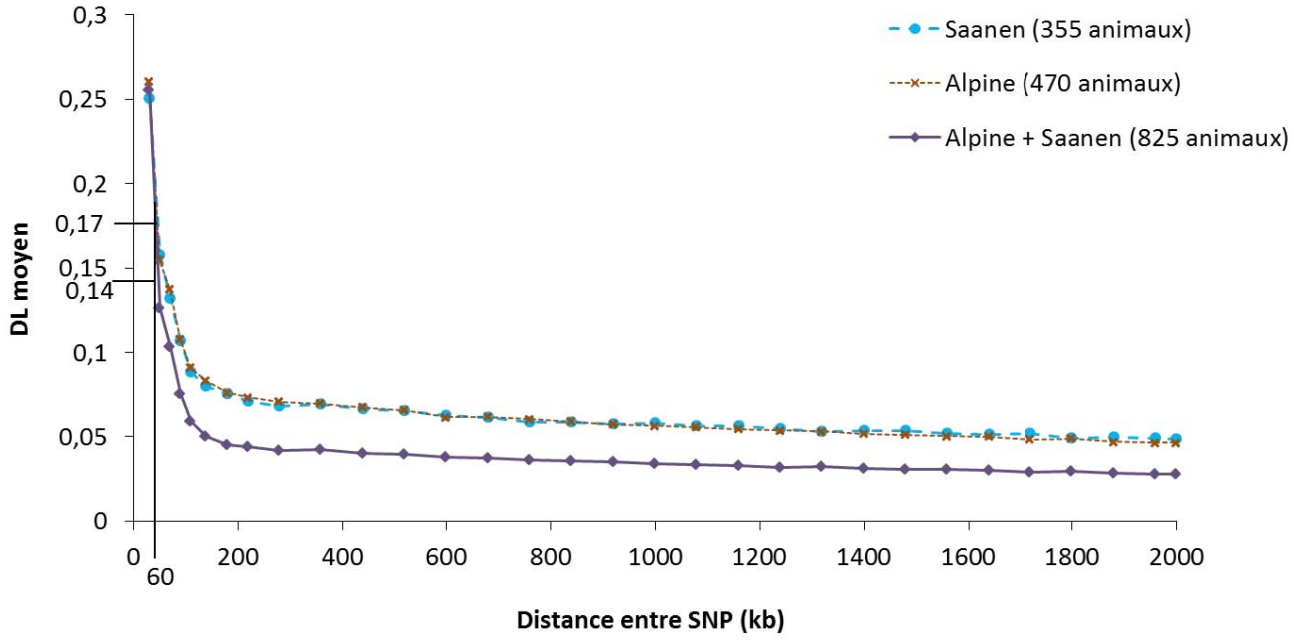


Figure 2.14 : Estimation du niveau de déséquilibre de liaison dans les deux races étudiées (Alpine et Saanen) séparément ou ensemble à partir des génotypages des mâles génotypés

Les mêmes résultats sont obtenus pour les femelles génotypées (Figure 2 .15). Le niveau de DL dans la population multiraciale est moins élevé que celui obtenu en race Alpine ou Saanen (0,14 contre 0,17 à 50 kb pour les mâles). Cette différence de DL entre population multiraciale et populations en race pure a aussi été constatée en bovins laitiers : 0,15 dans la population Montbéliarde et Brune, contre 0,19 et 0,25 pour chacune des deux races respectivement (Hozé et al. 2013). L'écart de DL entre population multiraciale et population uniraciale est d'autant plus élevé que la distance entre marqueurs est grande. Le DL estimé pour de grandes distances entre SNP reflétant l'histoire récente des populations (Hayes et al.,

2003), cette différence pourrait s'expliquer par une sélection intra-race menée depuis plus de 40 ans en Alpine et Saanen. Le DL obtenu à la distance de 60 kb (reflétant l'histoire ancienne des populations) en race Alpine est très proche de celui obtenu en race Saanen (0,172 en Alpine et 0,173 en Saanen pour les mâles). Cette similarité du niveau de DL dans les deux races peut être expliquée par leur origine commune. En effet, la couleur blanche de certains animaux Alpins élevés dans le Nord des Alpes a été sélectionné pour créer la race Saanen, les deux races ayant été ensuite introduites en France dans les années 1910 (Babo, 2000).

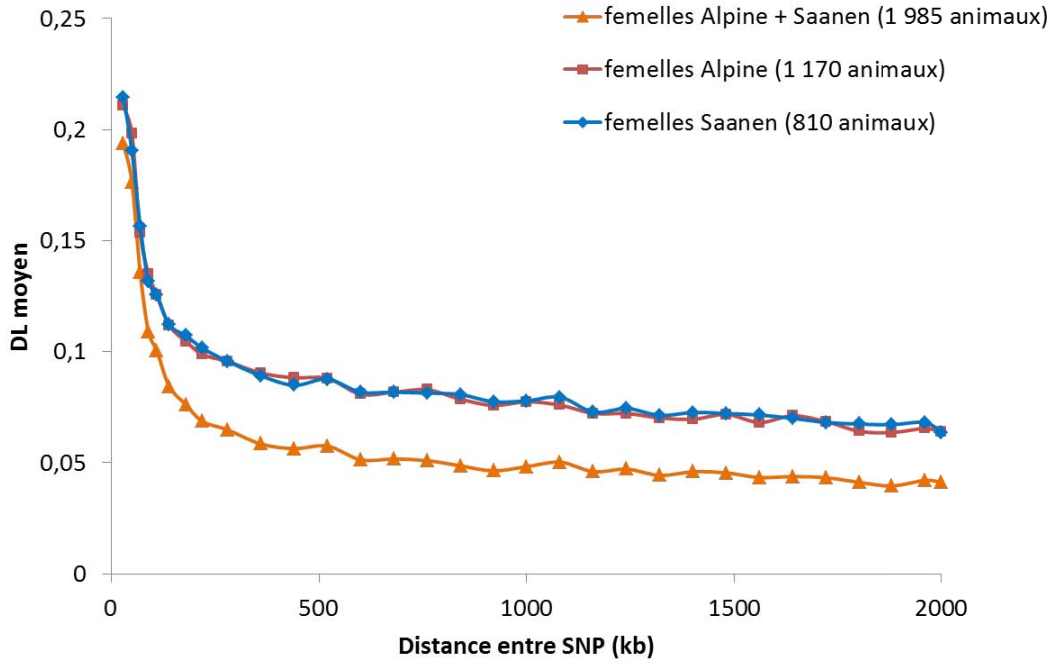


Figure 2.15 : Estimation du niveau de déséquilibre de liaison dans chacune des deux races étudiées (Alpine et Saanen) ainsi que dans la population totale pour l'ensemble des femelles génotypées (Alpine + Saanen)

Le déséquilibre de liaison estimé reflète la moyenne du déséquilibre de liaison par pas de 20 kb. Cependant on peut noter (Figure 2 .16 et Figure 2 .17) qu'une forte proportion de SNP ne présente aucun déséquilibre de liaison quelle que soit la distance entre SNP. Ce phénomène est également observé en ovins laitiers (Andrés Legarra, INRA Genphyse, communication personnelle) ainsi qu'en bovins laitiers (Pascal Croiseau, INRA Gabi, communication personnelle) et allaitants (Florence Phocas, INRA Gabi, communication personnelle).

La précision des évaluations génomiques étant fortement influencée par le niveau de DL (cf. 1.II.1), le faible niveau de déséquilibre de liaison constaté ici en caprins est-il alors rédhibitoire pour une sélection génomique dans cette population ? Cependant, les ovins

laitiers Lacaune avec un niveau de DL équivalent ont pu mettre en place une telle sélection (Baloche et al., 2013).

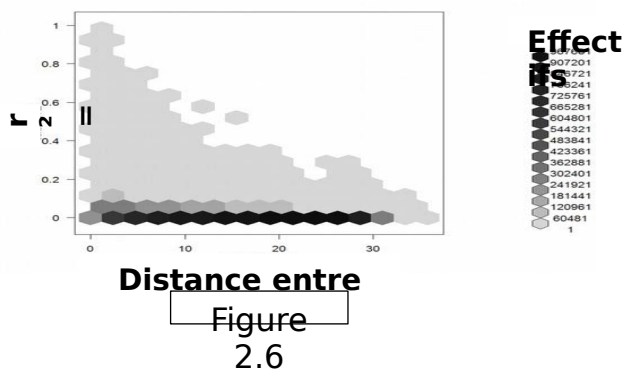


Figure 2.6

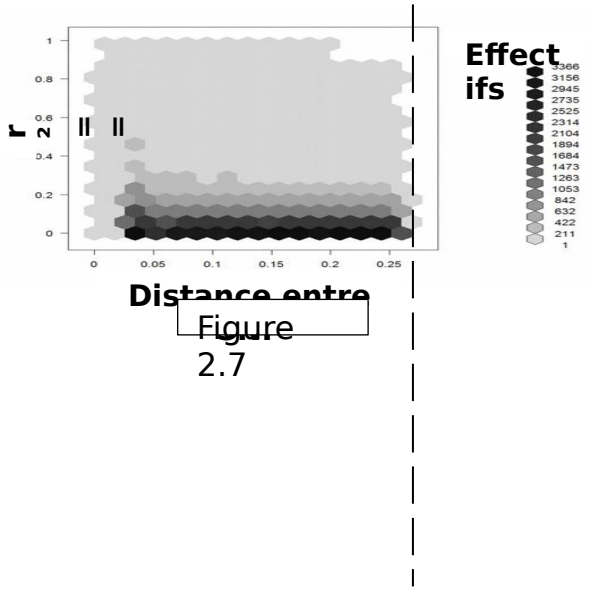


Figure 2.7

Figure 2.16 : Distribution des valeurs du déséquilibre de liaison en fonction de la distance entre SNP entre 0 et 30 Mb

Figure 2.17 : Distribution des valeurs du déséquilibre de liaison en fonction de la distance entre SNP entre 0 et 0,25 Mb

2.III Diversité génétique de la population

2.III.1 Consanguinité

Le coefficient de consanguinité (cf. 1.II.2.1) de la population caprine génotypée a été estimé d'une part à partir du pédigrée selon la méthode de Malécot (1948) à l'aide du logiciel pedig (Boichard, 2006) sans prendre en compte le niveau moyen de consanguinité des parents inconnus, et d'autre part à partir des génotypages selon la méthode de Baumung et Sölkner (2003) à l'aide du logiciel PLINK (Purcell et al., 2007). L'estimation de ces coefficients a d'abord été réalisée sur trois sous-populations multi-raciales différentes (Tableau 2 .6) : 1) la population des mâles de référence (677 boucs génotypés et phénotypés) ; 2) la population de référence totale (677 boucs et 1 985 chèvres) ; 3) la population des candidats comprenant les 148 jeunes mâles qui n'avaient pas encore de filles en janvier 2013. Ces populations correspondent à celles qui sont utilisées dans l'évaluation génomique.

Tableau 2.6: Moyenne des coefficients de consanguinité (µ) estimée selon les populations génotypée à partir du pédigrée (pedig) ou des génotypages (geno)

	Mâles de référence (677) nés entre 1993 et 2009		Référence totale (677 + 1 985) nés entre 1993 et 2009		Candidats (148) nés entre 2010 et 2011	
	pedig	geno	pedig	geno	pedig	geno
µ (%)	2,1	2,2	2,6	2,1	2,1	2,3

Les coefficients de consanguinité moyens ont été estimés entre 2,1 et 2,6% selon la population étudiée et la méthode d'estimation. Ces résultats sont proches de ceux calculés en 2011 (Danchin-Burge, 2011) sur une population de femelles nées en 2009 (2,3% et 2,5% pour les Saanen et les Alpines respectivement). A partir des données du pédigrée, le coefficient le plus fort est obtenu pour la population de référence incluant les femelles (2,6%). La consanguinité la plus forte obtenue à partir des données génomiques est estimée à 2,3% pour la population de jeunes mâles candidats. Les coefficients estimés sur pédigrée à partir des populations génotypées sont assez proches de ceux estimés à partir de la population d'indexation découpée selon les mêmes modalités (Tableau 2 .7), sauf pour les jeunes mâles (3,1% contre 2,1% dans la population génotypée). Les estimations de ces coefficients sur la population totale d'indexation ont été réalisées à partir de 31 217 mâles nés entre 1993 et 2009, 106 823 femelles nées entre 2008 et 2009 et 968 mâles nés entre 2010 et 2011. La consanguinité de la population de référence génotypée semble donc représentative de celle de la population d'indexation.

Tableau 2.7 : Moyenne des coefficients de consanguinité estimés à partir du pédigrée
1) tous les mâles de l'indexation nés entre 1993 et 2009 ; 2) tous les mâles précédents ainsi que les femelles nées entre 2008 et 2009 et 3) tous les mâles nés entre 2010 et 2011

	Mâles (1993-2009)	Mâles (1993-2009) + femelles (2008-2009)	Mâles (2010-2011)
µ (%)	2,3	3,0	3,1

Les moyennes des coefficients de consanguinité estimés à l'aide du pédigrée sont différentes de celles obtenues à l'aide des génotypages. Les coefficients individuels estimés à partir des deux types de données sont en effet mal corrélés (Figure 2 .18). La corrélation entre coefficients de consanguinité pédigrée et génomique est de 0,33 pour la population génotypée multiraciale (mâles + femelles), 0,39 pour les animaux Alpins et 0,35 pour les Saanens génotypés. Ces corrélations sont légèrement moins élevées que celles trouvées dans une population d'ovins finlandais : entre 0,45 et 0,54 selon la population choisie (Li et al., 2011).

Cette différence de corrélation avec l'étude de Li et al. (2011) peut s'expliquer par un nombre important d'animaux issus de parents inconnus dans le fichier des généalogies dans la population caprine française contrairement au cas des ovins finlandais. Les coefficients de consanguinité estimés dans notre population à partir du pédigrée sont donc plus souvent sous-estimés par rapport au coefficient de consanguinité génomique même si le nombre d'équivalent générations connues est conséquent et compris entre 6,1 pour la population de femelles et 8,3 pour la population de mâles.

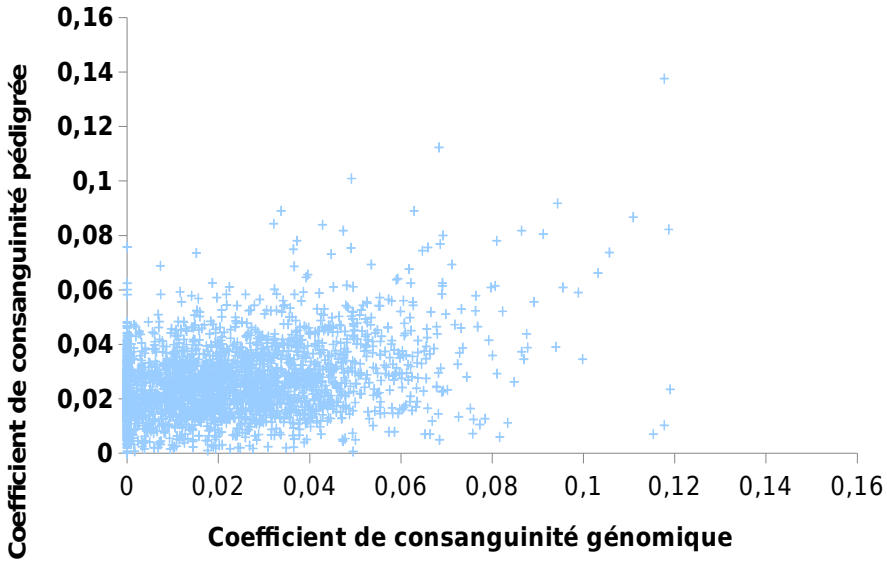


Figure 2.18 : Graphique des coefficients de consanguinités estimés à partir du pédigrée en fonction de ceux estimés à partir des données génomiques sur la population génotypée (825 boucs et 1 985 chèvres)

Les moyennes des coefficients de consanguinités estimées séparément dans chaque race sont présentées dans le Tableau 2 .8. La consanguinité moyenne chez les animaux de race Saanen sur pédigrée est plus élevée que celle des animaux Alpins pour la population de référence. Ce résultat est conforme au résultat d'une précédente étude portant sur les boucs nés entre 1989 et 1998 (Palhière, 2001) : 1,4 % en Saanen et 1% en Alpine. Les différences des moyennes estimées à partir des données pédigrée et des données génomiques sont plus marquées par race qu'en multiracial en raison d'une plus grande variabilité et d'un plus faible effectif intra-race.

Tableau 2.8 : Moyenne des coefficients de consanguinité (en %) estimés à partir du pédigrée (pedig) ou des génotypages (geno) pour les animaux Alpins et Saanens séparement de la population candidate ou de référence

	Alpine		Saanen	
	pedig	geno	pedig	geno
Population de référence (677 + 1 985)	1,8	1,8	2,9	2,3

Population candidate (148)	1,8	1,7	2,4	2,6

La valeur moyenne de la consanguinité estimée dans la population caprine française (entre 1,8% et 2,9%) est proche de celle estimée en ovins laitier : entre 2,3% pour la race Lacaune et 2,6% pour la race Basco Béarnaise. Elle est cependant plus faible celles estimées en bovins laitiers (entre 3,6 en race Normande et 4,4% en Montbélliarde) (Danchin-Burge, 2009).

2.III.2 Parenté

Les coefficients de parenté (cf 1.II.2.1) ont été estimés à partir du pédigrée à l'aide du logiciel pedig (Boichard, 2006) et à partir des données génomiques en utilisant le logiciel king (Manichaikul et al., 2010). Le Tableau 2 .9 présente les moyennes des coefficients de parenté en % estimés entre les individus deux à deux dans la population de référence et dans la population candidate ainsi qu'entre les individus des deux populations. Les coefficients de parenté estimés à partir des génotypages sont nettement plus faibles que ceux estimés à partir du pédigrée. Ce résultat signifie qu'en moyenne les individus partagent moins d'allèles aux SNP en commun que ce qui est attendu par l'examen du pédigrée.

Tableau 2.9 : Coefficient de parenté moyen (en %) estimé à partir des données pédigrée ou génomiques dans la population multiraciale

	Référence mâles (677)	Candidats (148)	Entre référence mâles et candidats	Référence totale (677 + 1 985)	Entre référence totale et candidats
geno	0,9	0,8	0,8	0,9	0,7
pedig	1,8	1,3	1,4	1,7	1,3

Très peu de différences sont observées entre les coefficients de parenté dans les différentes populations lorsqu'ils sont estimés à partir des génotypages. En revanche, les différences entre populations sont plus marquées lorsque les coefficients sont estimés à partir du pédigrée. Les corrélations entre les coefficients de parenté génomique et pédigrée estimées pour la population de référence et la population de candidats sont relativement faibles (68,8%, résultats non montrés). Cette corrélation est plus faible en race Alpine (64,8%) qu'en race Saanen (72,8%).

Le coefficient de parenté est le plus élevé dans les deux populations de référence, incluant ou non les femelles génotypées : 1,8% pour la population de référence mâle et 1,7% pour la population de référence totale estimés à partir du pédigrée. Un coefficient de parenté plus faible entre les candidats peut s'expliquer par l'introduction de nouveaux mâles moins

apparentés dans le schéma depuis l'utilisation de la méthode de J.J. Colleau (Colleau et al., 2004) pour gérer la consanguinité en 2000.

Tableau 2.10 : Coefficient de parenté (en %) moyen estimé à partir des données pédigrée ou génomiques dans la population génotypée totale (mâles et femelles) séparément dans chaque race

	Alpine			Saanen		
	Référence mâles + femelles	Candidats	Entre référence et candidats	Référence mâles + femelles	Candidats	Entre référence -et candidats
geno	1,7	0,7	1,1	2,0	2,3	2,4
pedig	1,9	1,1	1,2	2,1	1,5	1,5

Le coefficient de parenté constaté dans la population des Saanen génotypés est plus élevé que celui estimé chez les Alpines (cf. Tableau 2 .10). Cette différence de parenté entre les deux races était déjà observée pour les mâles nés entre 1989 et 1999 : 2,3% en race Alpine et 3,3% en race Saanen (Palhière, 2001). La différence entre les coefficients de parenté calculés à partir de l'information génomique et pédigrée est plus marquée en race Saanen qu'en race Alpine surtout pour les candidats. La parenté moyenne observée entre les populations de candidats et de référence a été évaluée entre 0,7% (population multiraciale incluant les femelles, Tableau 2 .9) et 2,4% (en race Saanen, Tableau 2 .10). Cette parenté est plus faible que celle habituellement observée en bovins laitiers : entre 3 et 10% (Wientjes et al., 2013; Pszczola et al., 2012b).

Tableau 2.11 : Moyenne des coefficients de parenté (en %) estimés à partir du pédigrée
1) tous les mâles de l'indexation nés entre 1993 et 2009 (mâles anciens); 2) tous les mâles précédents ainsi que les femelles nées entre 2008 et 2009 (population mâle-femelle) et 3) tous les mâles nés entre 2010 et 2011 (jeunes mâles) et entre ces différentes populations

Mâles anciens (1)	Jeunes mâles (3)	Entre mâles anciens et jeunes mâles (entre 1 et 3)	Population mâle-femelle (2)	Entre population mâle-femelle et jeunes mâles (entre 2 et 3)
1,2	1,9	1,4	2,3	2,3

Les coefficients de parenté estimés dans la population génotypée sont inférieurs à ceux estimés dans la population d'indexation correspondante (cf. Tableau 2 .11) sauf pour la population de référence mâles (1,2% pour tous les mâles de l'indexation contre 1,8% pour les mâles génotypés). La parenté de la population génotypée n'est donc pas tout à fait représentative de la population d'indexation totale contrairement à la consanguinité.

2.III.3 Taille efficace

La taille efficace (Ne) de la population génotypée a été estimée dans un premier temps à partir des coefficients de consanguinité selon la méthode de Pérez-Encizo (1995). Dans un deuxième temps, elle a été évaluée à partir des estimations du niveau de DL selon la méthode de Tenesa (2007, cf. 1.II.2).

Tableau 2.12 : Estimations de la taille efficace de population sur les populations d'animaux génotypés Alpins (Alp), Saanen (Saa) et multiraciale (multi) pour un nombre donné de générations avant l'actuelle (nb générations) selon deux méthodes (à partir du pédigree (pédig) ou du déséquilibre de liaison (DL))

		Mâles + Femelles			Mâles			Femelles		
		Alp	Saa	multi	Alp	Saa	multi	Alp	Saa	multi
Ne	pédig	127	99	114	127	92	110	187	184	241
	DL	115	98	194	93	82	127	157	150	220
Nb générations		6	6	5	5	5	5	7	7	6

L'estimation à partir du DL tient compte du nombre de générations avant l'actuelle. Etant donné la distance moyenne entre deux SNP sur la puce caprine, l'estimation de la taille efficace de la population n'a pu être faite qu'à 5 voire 7 générations (minimum imposé par le DL selon la population) avant la génération actuelle ($T=\frac{1}{2c}$ avec c la distance entre deux SNP consécutifs).

La taille efficace dans la population génotypée a été estimée entre 82 pour les mâles Saanen et 241 pour les femelles des deux races (Tableau 2 .12). L'estimation de la taille efficace est plus élevée en race Alpine qu'en race Saanen, ce qui est conforme avec la consanguinité moins élevée observée en 2.III.3 pour cette race. La taille efficace estimée dans la population multiraciale ne correspond pas à l'addition de la taille efficace de chaque race. Ce phénomène peut s'expliquer par la présence d'ancêtres communs aux deux races ainsi que par une détermination de la race des animaux uniquement par appréciation visuelle. En effet, le code race renseigné dans l'indexation correspond à l'observation de la couleur de la robe de l'animal et non à la race sur pédigrée des individus. La race Saanen étant à l'origine une variante albinos de l'Alpine (Babo, 2000), il arrive que certains animaux de parents Alpins aient un type racial Saanen.

L'effectif efficace estimé chez les femelles est plus important que celui estimé chez les mâles. Ce résultat est surprenant car les femelles génotypées sont issues de 20 pères

également génotypés. Cependant, ces femelles ont des mères toutes différentes ce qui peut expliquer une variabilité génétique plus grande que pour les mâles. La taille efficace estimée chez les femelles Alpines est proche de celle évaluée sur les femelles nées entre 2010 et 2011 (191 ancêtres efficaces (Danchin-Burge, 2011)). On peut donc penser que la taille efficace a peu évolué depuis 5 générations dans cette race, ce qui n'est pas le cas chez les Saanens. En effet dans cette race, l'effectif efficace estimé dans cette étude (à partir des femelles nées entre 2008 et 2009) est plus grand que celui estimé pour les femelles nées entre 2010 et 2011 (124, Danchin-Burge, 2011). La taille efficace aurait donc été réduite sur les 5 dernières générations en race Saanen ce qui est en accord avec la tendance à l'augmentation du coefficient de consanguinité observé depuis 2001 (Danchin-Burge, 2011). Ce phénomène peut être expliqué par l'utilisation plus intensive de certains mâles ces dernières années.

Les différences d'estimation de taille efficace entre la méthode basée sur les coefficients de consanguinité et celle basée sur le DL sont assez importantes excepté pour les mâles Saanen. Ces différences ne sont pas observées en ovins Lacaune (Baloche et al., 2013) pour lesquels une étude similaire a été réalisée. En revanche l'évolution du Ne en fonction des générations passées (Figure 2 .19) est similaire à celle observée dans l'espèce ovine (Baloche et al., 2013). On note que la tendance à une taille efficace plus importante chez les femelles que chez les mâles génotypés est inversée à environ 800 générations avant l'actuelle. Ceci s'explique principalement par une évolution légèrement différente du niveau de DL en fonction de la distance entre SNP, estimée chez les mâles ou les femelles.

La taille efficace estimée chez les caprins de race Alpine et Saanen dans cette étude est plus élevée que celle estimée en bovins laitiers (entre 37 pour les Montbéliardes et 62 pour les Holsteins (Danchin-Burge, 2009)) et en races porcines (entre 55 en Landrace et 113 en Large white (Maignel et al., 1998). En revanche, elle est proche de celles estimées en ovins laitiers (entre 112 pour les Basco béarnaises et 227 pour les Lacaunes (Danchin-Burge, 2011) et est plus faible que celle calculée en bovins allaitants Charolais (1255 (Danchin-Burge, 2009)).

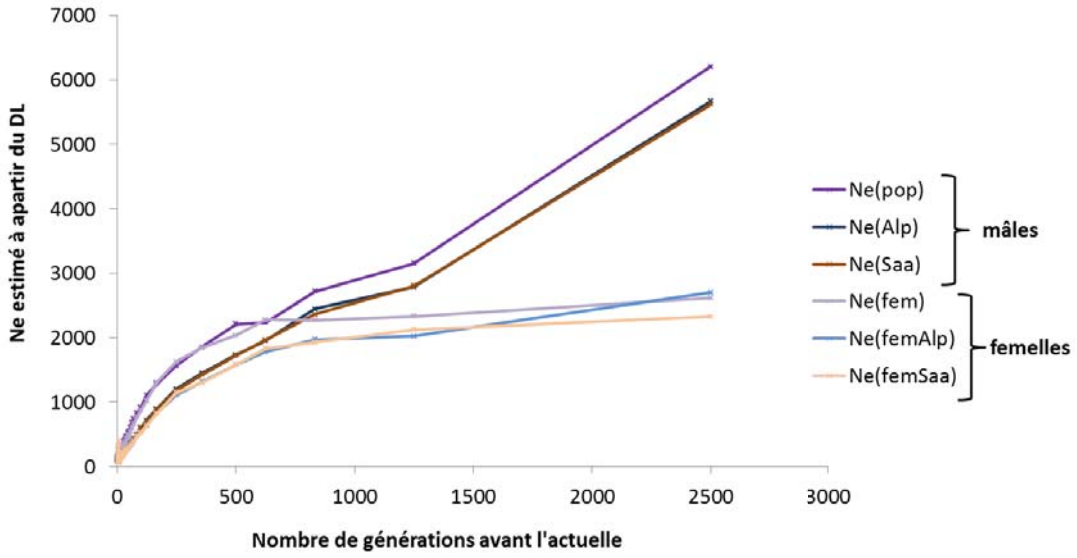


Figure 2.19: Evolution de la taille efficace (Ne) estimée à partir du DL en fonction du nombre de générations antérieures à l'actuelle
dans les populations de mâles génotypés : Saanen (Ne(Saa)), Alpins (Ne(Alp)), des deux races (Ne(pop)) ; et dans celles de femelles génotypées : Saanen (Ne (femSaa)), Alpines (Ne(femAlp)) et des deux races (Ne(fem))

2.IV La différentiation des races Alpine et Saanen

Les races Alpine et les Saanen ont pour origine commune une race Suisse (Babo, 2000). Le coefficient de parenté estimé à partir du pédigrée entre les mâles génotypés de race Saanen et ceux de race Alpine est faible, inférieur à 0,5% (résultat non montré). Cependant, les effectifs de mâles testés sur descendance chaque année dans chaque race étant très faibles, l'augmentation de la population de référence pourrait se faire via une évaluation génomique multiraciale. Afin d'être sûr de l'intérêt d'une telle évaluation, il est nécessaire d'analyser la distance génétique entre ces deux races.

La première étape a consisté à évaluer la persistance du déséquilibre de liaison entre les deux races à l'aide d'un programme Fortran développé au sein de l'unité GenPhySE de l'INRA. Cette persistance du DL est la corrélation entre les valeurs du DL (r) obtenues dans deux races différentes. La Figure 2 .20 représente la persistance du DL en fonction de la distance entre SNP, en ovins (Guillaume Baloche, INRA Genphyse communication personnelle) entre les Manech tête rousse et tête noire, et entre les Manech tête rousse et les Lacaune, ainsi qu'en caprins entre les Alpines et les Saanens. En caprins elle a été estimée uniquement à partir des génotypes des mâles soit 355 boucs Saanens et 470 boucs Alpins. Pour les très courtes distances entre SNP (< 70 kb), la persistance du DL observée entre Alpine et Saanen est similaire à celle observée entre les deux races de Manech. A ces faibles

distances entre SNP, la persistance du DL est représentative de l'histoire ancienne des deux races (Hayes, 2011). On en déduit donc que, comme les Manech tête rousse et tête noire, les Alpines et les Saanen ont une origine commune. En revanche, à partir d'une distance entre SNP de 70 kb, la persistance du DL entre les deux races caprines devient plus proche de celle observée entre deux races ovines bien distinctes (Manech tête rousse et Lacaune). Malgré leur origine commune, les Alpines et Saanens sont donc aujourd'hui deux races bien distinctes, sélectionnées en race pure depuis plus de 40 ans.

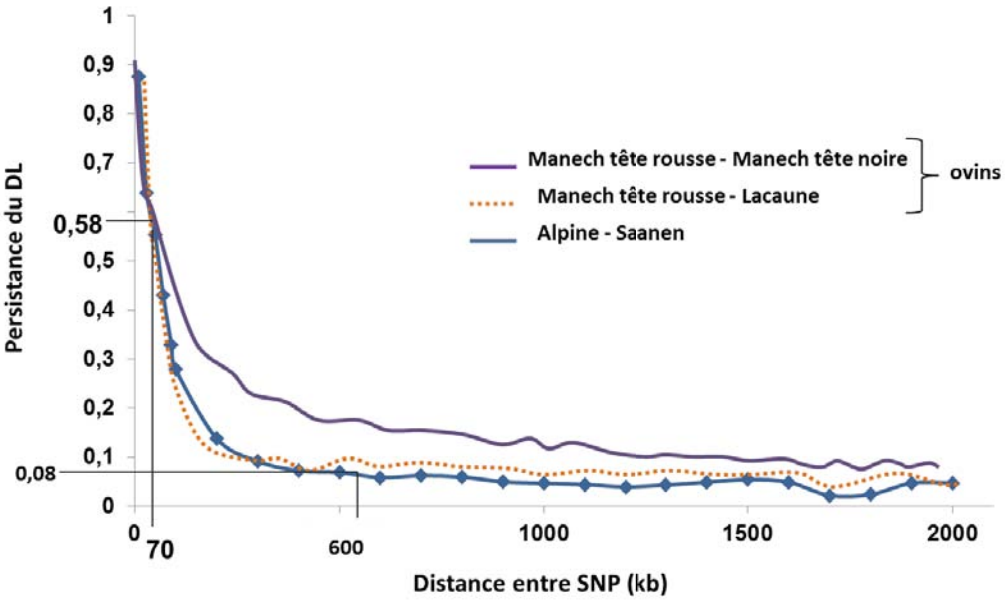


Figure 2.20 : Persistance du DL en fonction de la distance entre SNP en kb entre 1) les ovins laitiers Manech tête rousse et Manech tête noire 2) les ovins laitiers Manech tête rousse et les Lacaune et 3) les caprins de race Alpine et Saanen

Une analyse en composantes principales (ACP) a été réalisée à l'aide du logiciel R, sur les génotypes des mâles en utilisant comme facteur le code race des individus (Figure 2 .21). Cette étude permet de mettre en évidence une différenciation d'un point de vue génomique des races Alpine et Saanen sur les deux premiers axes de l'ACP dont la somme explique plus de 12% de la variance. Le code race visuel semble donc être représentatif de la race génétique des individus (définie par le pédigrée ou les génotypes) pour les mâles.

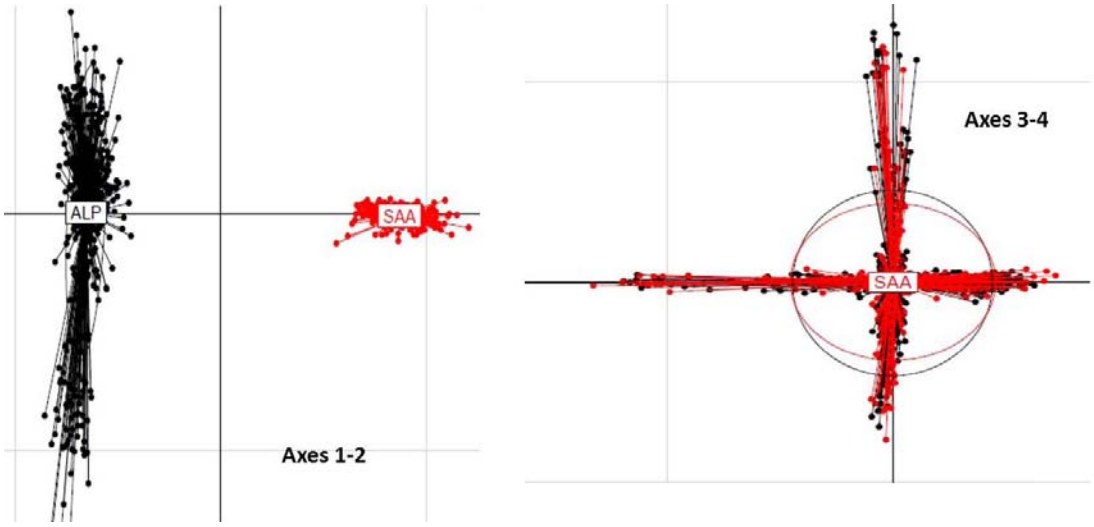


Figure 2.21 : ACP réalisée à partir des génotypes des animaux selon leur code race déterminé visuellement (axes 1 et 2, et axes 3 et 4)

D'autre part, les fréquences alléliques en race Alpine sont moyennement corrélées (0,68 pour les mâles et 0,85 pour les femelles génotypées) à celles obtenues en race Saanen. La plus forte corrélation entre fréquences alléliques observées chez les femelles est cohérente avec la corrélation observée entre leurs 11 pères Alpins et 9 pères Saanen qui est de 0,81. Ces corrélations semblent stables dans le temps (cf. Figure 2 .22) depuis 1998 même si de fortes baisses de corrélations sont constatées en 2008 et 2011. La baisse de corrélations en 2008 peut s'expliquer par un faible effectif de génotypages ayant passés le contrôle qualité.

Les Alpines et les Saanens sont donc deux races bien distinctes d'un point de vue des génotypes, des fréquences alléliques et de la persistance du déséquilibre de liaison, malgré leur origine commune ancienne.

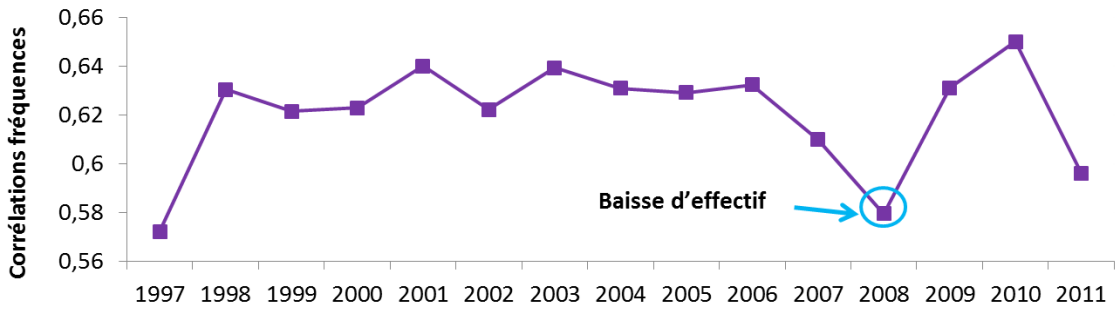


Figure 2.22 : Evolution des corrélations de fréquences alléliques entre race Alpine et Saanen en fonction de l'année de naissance des mâles génotypés

2.V Prédiction du niveau de précision génomique théorique à partir de la structure de population

Les équations de Daetwyler (2010) et de Goddard (2009) décrites dans le paragraphe 1.II.5 ont été utilisées pour prédire le niveau de précision génomique des candidats. Les autres formules (Goddard et al., 2011 ; Meuwissen et al., 2013) n'ont pas été utilisées ici, car elles supposent connue la proportion de variance expliquée par les marqueurs. Deux formules ont été utilisées pour le calcul du nombre de segment indépendants (Me). La première formule utilisée (Me1) est celle qui permet d'obtenir le plus petit Me ($Me = \frac{2NeL}{\log(4NeL)}$) et la deuxième (Me2) est celle qui permet d'obtenir le plus grand nombre de segments indépendants ($Me = 4NeL$) (Brard et Ricard, 2014). La Figure 2 .23 montre ces estimations en fonction de l'héritabilité du caractère considéré, en prenant en compte une taille efficace de 150, une taille de population de 700 (pour 677 mâles génotypés et phénotypés), et une longueur du génome de 30 M.

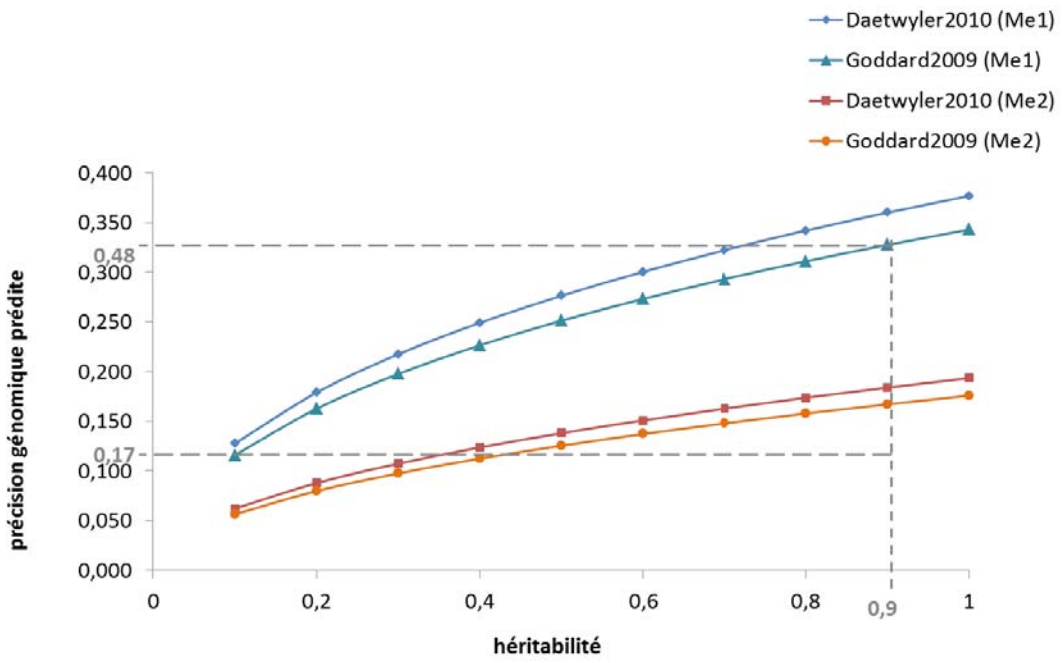


Figure 2.23 : Niveau de précision génomique prédit pour les candidats par les équations de Daetwyler (2010) et Goddard (2009) en fonction de l'héritabilité du caractère

On note d'après la Figure 2 .23 qu'en considérant les paramètres de la population caprine, l'équation la plus optimiste est celle de Daetwyler avec la première formule du Me, la moins optimiste étant celle de Goddard en utilisant la deuxième formule pour calculer le nombre de segments indépendants. Les pseudo-performances (DYD) sont considérées

comme les performances propres des mâles, dont l'héritabilité est égale à la précision de l'index sur descendance, qui est supérieure à 0,9 dans notre étude. Le niveau de précision théorique prédit pour une évaluation génomique basée sur les DYD des mâles chez les caprins se situe entre 0,17 et 0,48 selon les équations, ce qui est plus faible que le niveau de précision obtenue sur ascendance (0,64). Ce niveau est bien plus faible que le niveau de précision théorique des GEBV obtenus en ovins Lacaune (entre 0,47 et 0,71 (Baloche et al., 2013)) et en bovins laitiers sur la production laitière (entre 0,37 en Normande et 0,60 en Holstein (Fritz et al., 2010)). Il a en effet été montré que ces formules sous-estiment globalement les précisions pour de fortes héritabilités (Brard and Ricard, 2014).

En conclusion, la population caprine génotypée, après contrôle qualité des SNP, comprend 677 mâles testés sur descendance, 148 jeunes mâles n'ayant pas encore de filles en 2013 et 1 985 femelles issues de 20 pères génotypés parmi les 677. Dans cette population, le modeste niveau de déséquilibre de liaison, bien que comparable à celui constaté en ovins laitiers Lacaune, ainsi que la bonne diversité génétique (faible consanguinité et grande taille efficace) observés ne sont pas optimaux pour la prédiction des valeurs génomiques des animaux comme nous avons pu le voir au chapitre 1. Cependant, la structure de la population caprine française génotypée est représentative de la population en sélection d'un point de vue de sa consanguinité, de sa parenté et de sa taille efficace. Il n'apparait donc pas nécessaire (ni possible), dans un premier temps de génotyper d'autres animaux afin d'améliorer la composition de la population de référence pour les évaluations génomiques. D'autre part, l'étude de la distance entre les races Alpine et Saanen a permis de montrer que malgré leur origine commune, ces deux races étaient bien distinctes d'un point de vue génomique. Au vu de ces éléments, des évaluations génomiques basées sur une approche en deux étapes seront analysées dans le chapitre suivant afin d'étudier l'intérêt de la sélection génomique chez les caprins laitiers français.

Chapitre 3 : Étude d'évaluations génomiques à partir d'une modélisation en deux étapes

3.I Evaluation génomique basée sur la méthode GBLUP

3.I.1 Evaluation génomique basée sur les DYD

3.I.1.A Evaluation multiraciale

Introduction

Le premier article de ma thèse, publié dans la revue Journal of Dairy Science à l'été 2013 s'inscrit dans un contexte de disponibilité de la puce caprine 50k en 2011 et de l'obtention en septembre 2012 des génotypages des animaux français (825 mâles et 1 945 femelles). Le facteur clé pour la mise en place de la sélection génomique en caprin étant la précision des évaluations, elle-même influencée par la structure de la population (cf. chapitre 1), un des premiers objectifs de cette étude a été de décrire le niveau de déséquilibre de liaison observé dans la population. Ce DL a été estimé dans les deux races séparément ainsi que dans la population multiraciale. La persistance du DL entre les deux races a aussi été étudiée afin de définir si une évaluation génomique multiraciale pouvait être envisagée.

La population des mâles génotypés, comprenant 677 mâles testés sur descendance et 148 mâles nés entre 2010 et 2011 n'ayant pas encore de filles, a été obtenue en deux temps. Une première partie des données génomiques avait été obtenue en janvier 2012 dans le cadre d'un projet de détection de QTL. Elle comprenait 67 mâles et 1 985 femelles génotypés après contrôle qualité (cf. 2.I). L'étude s'est donc concentrée sur quatre populations de référence: 1) une population de 67 mâles, 2) une population de 67 mâles et 1 985 femelles, 3) une population de 677 mâles, et 4) une population de 677 mâles et 1 985 femelles (Figure 3 .24). La population candidate était constituée des 148 jeunes mâles nés entre 2010 et 2011.

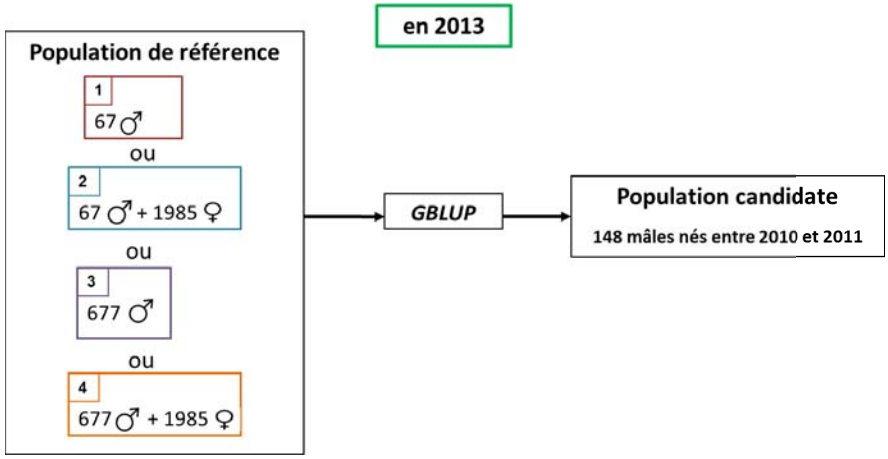


Figure 3.24 : Populations de référence et population candidates utilisées dans l'article I

Le second objectif de cette étude était d'estimer les précisions des GEBV des candidats, calculées à partir des variances d'erreurs de prédiction (1.III.3.2) pour ces quatre populations de référence et de les comparer aux précisions obtenues sur ascendance en évaluation génétique classique. Ces précisions ont été estimées à partir d'évaluation génomique utilisant la méthode GBLUP two steps basée sur les pseudo-phénotypes : les DYD pour les mâles et les YD pour les femelles (calculés à partir de l'évaluation génétique officielle de janvier 2013). Cette stratégie a permis d'étudier l'intérêt de prendre en compte dans une évaluation génomique les génotypes de femelles dans le cas d'une petite ou d'une plus grande population de mâles génotypés. Afin d'appréhender les différences de structure génétique de ces quatre populations, les coefficients de consanguinité ainsi que les coefficients de parenté entre les populations de référence et la population de candidats ont été estimés.

Enfin, la qualité des prédictions génomiques caprines a été évaluée par validation croisée (cf. 1.III.3.1). La validation croisée n'a pu cependant être réalisée que sur la population des 677 mâles génotypés en la divisant en une population d'apprentissage contenant les 425 mâles nés avant 2006 et une population de validation contenant les 252 mâles nés entre 2006 et 2009 (Figure 3 .25).

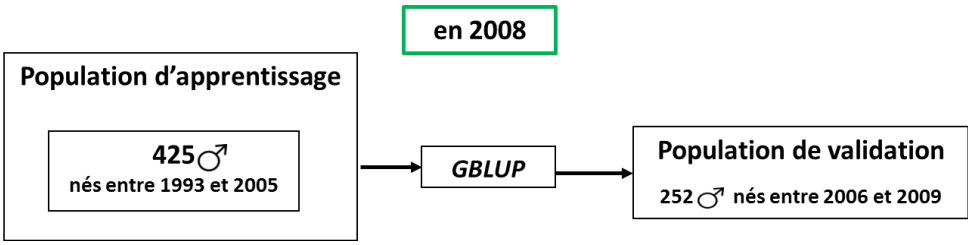


Figure 3.25 : Schéma de la validation croisée utilisée dans l'article I

La population des 67 mâles génotypés était trop petite pour envisager de réaliser sur elle seule une validation croisée. De plus, les femelles ne pouvaient pas être intégrées dans la population d'apprentissage, car étant nées entre 2008 et 2009, elles n'avaient pas de descendants connus nés en 2010, dernière année de naissance des descendants des mâles. La qualité des prédictions a été évaluée à l'aide des corrélations et des pentes de régression entre les GEBV (obtenues à partir des évaluations génomiques basées sur les DYD de 2008) et les DYD (disponibles en janvier 2013) pour les 252 mâles de validation afin de quantifier respectivement la précision et la surestimation ou la sous-estimation des GEBV prédites.

Article I : Premiers pas vers la sélection génomique pour la population multiraciale caprine française

Carillier C, Larroque H, Palhière I, Clément V, Rupp R, Robert-Granié C, 2013. A first step toward genomic selection in the multi-breed French dairy goat population. Journal of Dairy Science, 96: 7294-7305.

J. Dairy Sci. 96:7294–7305
http://dx.doi.org/10.3168/jds.2013-6789

A first step toward genomic selection in the multi-breed French dairy goat population

C. Carillier,*[1] H. Larroque,* I. Palhière,* V. Clément,† R. Rupp,*and C. Robert-Granié*
*Institut National de la Recherche Agronomique (INRA), UR631, Station d'Amélioration Génétique des Animaux (SAGA), 24 Chemin de Borde Rouge, Auzeville CS 52627, 31326 Castanet-Tolosan cedex, France
†Institut de l'élevage (Idele), 24 Chemin de Borde Rouge, Auzeville CS 52627, 31326 Castanet-Tolosan cedex, France

ABSTRACT

The objectives of this study were to describe, using the goat SNP50 BeadChip (Illumina Inc., San Diego, CA), molecular data for the French dairy goat population and compare the effect of using genomic information on breeding value accuracy in different reference populations. Several multi-breed (Alpine and Saanen) reference population sizes, including or excluding female genotypes (from 67 males to 677 males, and 1,985 females), were used. Genomic evaluations were performed using genomic best linear unbiased predictor for milk production traits, somatic cell score, and some udder type traits. At a marker distance of 50 kb, the average r^2 (squared correlation coefficient) value of linkage disequilibrium was 0.14, and persistence of linkage disequilibrium as correlation of r-values among Saanen and Alpine breeds was 0.56. Genomic evaluation accuracies obtained from cross validation ranged from 36 to 53%. Biases of these estimations assessed by regression coefficients (from 0.73 to 0.98) of phenotypes on genomic breeding values were higher for traits such as protein yield than for udder type traits. Using the reference population that included all males and females, accuracies of genomic breeding values derived from prediction error variances (model accuracy) obtained for young buck candidates without phenotypes ranged from 52 to 56%. This was lower than the average pedigree-derived breeding value accuracies obtained at birth for these males from the official genetic evaluation (62%). Adding females to the reference population of 677 males improved accuracy by 5 to 9% depending on the trait considered. Gains in model accuracies of genomic breeding values ranged from 1 to 7%, lower than reported in other studies. The gains in breeding value accuracy obtained using genomic information were not as good as expected because of the limited size (at most 677 males and 1,985 females) and the structure of the reference population.

Received March 11, 2013.
Accepted July 24, 2013.
[1]Corresponding author: celine.carillier@toulouse.inra.fr

Key words: dairy goat, genomic evaluation, linkage disequilibrium, female genotype

INTRODUCTION

Selection in the French Alpine and Saanen dairy goat breeds has been implemented by a single breeding organization. The objectives of this breeding scheme were to improve milk composition, milk yield, and udder morphology. Selection for these characteristics was based on a combined index calculated from EBV for milk yield, fat and protein yields, fat and protein contents, and various udder-type traits. This total merit index, which differs for Alpine and Saanen breeds (Clément et al., 2006), will change in 2013 to introduce selection on SCS (Rupp et al., 2011). Genetic evaluation of milk production traits has been carried out simultaneously in the 2 breeds using a BLUP animal model and considering all female performance records since 1980. Genetic evaluation of type traits (based on performance recorded since 2000) and SCS is performed separately for the 2 breeds.

The 2 breeds originated from the same single breed. The white coat variety of the Alpine goat, bred in the northern area of the Swiss Alps, was selected centuries ago to create the Saanen breed (Babo, 2000). When Alpine and Saanen were introduced in France in the 1910s, they were largely crossbred. In the 2000s, the percentage of Alpine genes in Saanen goats was 3.6 (Piacere et al., 2004), and genetic distance between the 2 breeds was <0.13 (Araujo et al., 2006).

The core selection population was composed of 1,000 dams of bucks selected each year for their reproductive ability, genetic level, and morphology. The AI rate was 20% in all goat herds. For health (e.g., no Q fever, tuberculosis), reproductive ability, growth, and genetic level, only 20% of all males (40 Alpine and 35 Saanen) born from assortative mating with the dams of bucks were progeny tested each year. Among those progeny-tested males, 25 Alpine and 15 Saanen were used as AI bucks. Progeny testing was performed on at least 60 daughters per buck over 18 mo. This led to a short generation interval of less than 4 yr in the sire-daughter

pathway, but a longer one (>5.5 yr) in the sire-son pathway (Danchin-Burge, 2011).

The availability of the Illumina goat SNP50 BeadChip (Illumina Inc., San Diego, CA; Tosser-Klopp et al., 2012) and recent genotyping methods means it is now feasible to assess genomic selection in this species. In dairy cattle, genomic selection has led to decreased generation intervals in the male pathway because of the early selection of males and improved breeding value accuracies for young animals at birth (Schaeffer, 2006; de Roos et al., 2010). Although the generation interval in the sire-son pathway in French dairy goats is shorter than that in dairy cattle, it is expected to be reduced with genomic selection, because of higher breeding value accuracies of young males at birth. In French dairy goats, AI bucks were largely used by breeders with more than 1,000 daughters per males, which led to accurate breeding values of males. Breeding values of young males at birth were 62% accurate on average, using average parent EBV accuracy. The aim of using genomic selection in this species would be to obtain breeding value accuracies for young males at least as accurate as the pedigree-derived breeding value accuracies to limit progeny testing.

The quality of genomic predictions depends on the number of phenotypes and genetic markers, heritability of traits, the reference population size (Goddard, 2009; Hayes and Goddard, 2010; Liu et al., 2011), relationships within the reference population, and relationships between reference and candidate populations (Habier et al., 2010).The number of bucks progeny tested each year in the French dairy goat breeding scheme limited the number of genotyped bucks for this study. The small reference population of this study consisted of all bucks progeny tested from 1993 to 2009. To maintain the link between the reference and candidate population, it was not possible to increase the number of male genotypes by genotyping more generations of young males. For this reason, we assessed the use of genotyped females on breeding value accuracy of candidates, using genotypes from commercial females available at the time of the study.

A first objective of this study was to examine the structure of the reference population by considering the level of linkage disequilibrium (**LD**) within the population and between the 2 breeds. A second objective was to study the size and structure of the reference population on the accuracy of genomic EBV (**GEBV**).

MATERIALS AND METHODS

Animals and Genotypes

The Saanen and Alpine purebred animals included in this study were obtained from 2 populations. The first population had previously been used for QTL detection for new traits such as fine milk composition (Maroteau et al., 2012). This population consisted mostly of females from commercial herds: 2,254 females (938 Saanen and 1,316 Alpine) born between 2008 and 2009 and their 20 sires (9 Saanen and 11 Alpine). The best goats of this population will be used in future goat breeding as dams of bucks. The second population genotyped was composed of 852 bucks (369 Saanen and 483 Alpine born between 1993 and 2011). All bucks were progeny tested except young buck candidates born between 2010 and 2011. From 2003, all French bucks progeny tested in the breeding scheme were genotyped, of which 60% were Alpine bucks.

Animals were genotyped using the Illumina goat SNP50 BeadChip (Tosser-Klopp et al., 2012). Of the 53,347 SNP on the chip, 46,959 were validated after quality control. Missing SNP genotypes were not imputed, and the marker effect of the missing SNP genotypes was set to zero for affected animals. Quality control consisted of obtaining a call rate threshold of 98%, a minor allele frequency >1%, and checking for Hardy-Weinberg equilibrium. Validation was carried out separately for the 2 breeds, and SNP that were not retained in both Saanen and Alpine breeds were discarded. Because of poor DNA quality and the animal call frequency threshold being set at 99%, 2,810 genotyped animals (1,164 Saanen and 1,646 Alpine) were available for this study.

Phenotypes

Five milk production traits were considered: milk yield, fat yield, and protein yield ($h^2 = 0.3$ for milk, fat, and protein yields) and fat content and protein content ($h^2 = 0.5$) (Bélichon et al., 1999). Somatic cell score ($h^2 = 0.20$; Rupp et al., 2011) and 5 udder type traits were also studied: udder floor position ($h^2 = 0.29$), udder shape ($h^2 = 0.32$), rear udder attachment ($h^2 = 0.27$), fore udder ($h^2 = 0.30$), and teat angle ($h^2 = 0.31$) (Clément et al., 2002). All heritabilities defined for SCS and udder type traits were averages from Alpine and Saanen breeds. In general, values for milk production traits of Saanen goats were different (Bélichon et al., 1999) from those of Alpine goats (783 vs. 733 kg for milk yield in Saanen and Alpine respectively; Institut de l'élevage, 2010).

The phenotypes used for the genomic evaluation were daughter yield deviations (**DYD**) for the 677 AI bucks and yield deviations (**YD**) for the 1,985 females. Yield deviations were calculated from the official genetic evaluation of January 2012 (Clément et al., 2002) using Genekit software (Ducrocq, 1998) as performance corrected for fixed effects. Daughter yield deviations were

calculated from DYD averages corrected for environmental effects and the merit of their dams (VanRaden et al., 2009). Each female of this study had 2 lactations; that is, 2 YD per female, weighted by 1 for first lactation and by 0.8 for second lactation, as in the official genetic evaluation. Each male's DYD was weighted by effective daughter contributions (**EDC**; Fikse and Banos, 2001), calculated from all daughters considered in the national genetic evaluation. The EDC were calculated separately in each breed for SCS and type traits, and simultaneously for the other traits using crEDC software (Sullivan, 2010). Average EDC ranged from 36.5 for the teat angle trait to 65.9 for milk yield in the whole phenotyped male population.

Cases Studied

The aim of this study was to examine the effect of population size and female genotypes on the accuracy of predictions, using several reference and candidate populations. The first population (A) consisted of 67 males in the reference population and 148 candidates. The candidates were young males born in 2010 and 2011 with no daughters at the time of the present study. The 67 reference males were all born between 1999 and 2009; 54% were Alpine and 46% were Saanen (Table 1; case A). Among the 148 candidates, 15 were half-sibs of genotyped females. This small reference population was used to investigate the usefulness of adding genotyped females in a small male population. The second population (B) consisted of a reference population with the same previous 67 males plus 1,985 females, and the same 148 candidates as in case A. Cases C and D consisted of the same animals as cases A and B, respectively, plus 610 males in the reference population and the same candidate population as in previous cases. All the additional males in cases C and D were related (ancestors or half sibs) to the males in cases A and B. Of the 610 additional males in the reference population, 26% were born before 2001 (Table 1). In this study, case B was not compared with case C, because it is difficult to compare the addition of both males and females because of the different accuracies of their phenotypes.

Description of the Population

Extent of LD and Persistence of LD Phases. Estimations of the extent of LD between markers in the whole reference population and in each breed, as well as estimations of the persistence of LD phases among Alpine and Saanen breeds were calculated. Because phases of chromosomes in this study were unknown, the extent of LD between markers was measured between genotypes for each pair of SNP within chromosome. Thus, the measure of LD used was the correlation across diploid genotypes as proposed by Rogers and Huff (2009):

$$r^2 = \frac{\left[\mathrm{cov}(g_i, g_j)\right]^2}{\mathrm{var}(g_i) \times \mathrm{var}(g_j)},$$

where g_i and g_j were the genotypes at SNP i and j, respectively. Average r^2 (squared correlation coefficient) values were calculated for 20-kb intervals. For the persistence of LD phases between breeds, average correlations of signed r-values between Alpine and Saanen were derived over intervals of the same distance.

Table 1. Number of genotyped males for each case (population) by year of birth

	Cases[1]			
	A and B		C and D	
Year	Alpine	Saanen	Alpine	Saanen
1993–1998	0	0	24	19
1999	0	2	23	21
2000	0	0	19	18
2001	0	0	23	25
2002	2	6	28	29
2003	7	3	36	28
2004	4	0	41	25
2005	7	5	42	29
2006	5	7	40	25
2007	3	3	40	29
2008	0	0	28	17
2009	8	5	40	28
2010	43	33	43	33
2011	43	29	43	29

[1]Cases A and B: 67 males in reference population, 148 male candidates; cases C and D: 677 males in reference population, 148 male candidates.

The interest in using r-values instead of r^2-values for persistence of phases between breeds was the use of a signed value. The r-value can be different in 2 breeds even if the absolute value is similar. The extent of LD and persistence of LD phases were evaluated for a population of 677 AI bucks and 148 young bucks (Table 1; cases C and D).

Relationships and Inbreeding Between and Within Populations. Inbreeding and relationship coefficients were calculated using Pedig software (Boichard, 2006) for both reference and candidate populations in all cases studied. The relationship coefficient between 2 animals is the probability that, at a given locus, the 2 individuals share alleles identical by descent from the same ancestor. The inbreeding coefficient of an individual is the probability that, at a given locus, an individual has received similar alleles from both parents. In this study, it was calculated using the Meuwissen and Luo (1992) method.

Statistical Model for Genomic Evaluation

To estimate GEBV for both females and males, genomic BLUP (**GBLUP**) using genomic BLUPf90 software (Misztal et al., 2002) was implemented. The mixed model considered was $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$, where $\mathbf{y}$ is a vector of phenotypes weighted by EDC for males (DYD phenotypes) and the official weights of lactation (1 for first lactation and 0.8 for second lactation) for females (YD phenotypes), $\mathbf{X}$ is the incidence matrix relating breed effect ($\boldsymbol{\beta}$) to the individuals, $\mathbf{Z}$ is a design matrix allocating observations to breeding values ($\mathbf{u}$), and $\mathbf{e}$ is a vector of random normal errors. Genomic values $\mathbf{u}$ were normally distributed with $\mathrm{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$, where $\mathbf{G}$ is the genomic relationship matrix as defined by Van Raden (2008):

$$\mathbf{G} = 0.95 \times \frac{\mathbf{WW}'}{2\sum_{j=1}^{p} q_j(1-q_j)} + 0.05 \times \mathbf{A},$$

where p is the number of loci considered, q_j is the frequency of an allele of marker j estimated across Alpine and Saanen, $\mathbf{W}$ is a centered incidence matrix of SNP genotypes, and $\mathbf{A}$ is the pedigree-based relationship matrix. The genomic relationship matrix was derived from genomic and pedigree relationships to make $\mathbf{G}$ and $\mathbf{A}$ compatible (Christensen et al., 2012). Combining pedigree and genomic information in the relationship matrix avoids bias in the hypothesis of no selection in the base generation, which is true considering the $\mathbf{A}$ matrix but not considering $\mathbf{G}$ (Legarra et al., 2009; Vitezica et al., 2011). The $\mathbf{G}$ matrix was computed using allele frequencies across breed for computation simplicity reasons. Although considering the difference of allele frequencies in breed reduced the relationship coefficient between distant individuals, it did not affect the results on accuracy of GEBV (Makgahlela et al., 2013). Single nucleotide polymorphism marker effects were assumed to have a prior normal distribution and mixed model equations were used with the genomic relationship matrix (VanRaden, 2008). To study the effect of including genotyped females, YD and DYD phenotypes were taken into account together in the model. By definition, variances of YD and DYD were not the same:

$$\mathrm{r} \qquad \mathrm{Va}\,(2DYD_i) = \sigma_u^2 + \frac{1}{d_i}\left(2\sigma_u^2 + 4\sigma_e^2\right)$$

and

$$\mathrm{Var}(YD) = \sigma_u^2 + \sigma_e^2,$$

where d_i is the EDC of animal i. To take into account this difference, each EDC is multiplied by a coefficient

$$k = \frac{\sigma_e^2}{2\sigma_u^2 + 4\sigma_e^2},$$

where σ_u^2 is the genetic variance and σ_e^2 is the residual variance.

Because of the small population size (i.e., <400 male genotypes available per breed), Alpine and Saanen populations were analyzed together, considering the 2 breeds as 1, as in Bélichon et al. (1999). The genetic parameters considered were the official parameters for milk production traits (Alpine and Saanen treated together) and the average parameters of Alpine and Saanen goats for SCS and type traits.

Accuracy and Bias of Genomic Evaluation. Accuracy and bias of genomic evaluation were estimated by splitting the 677 males from the total reference population into a training set and a validation set. The training set consisted of 425 males born between 1993 and 2005. The DYD of these males were obtained from official 2008 genetic evaluations. This set was used to predict the GEBV of the validation population of 252 males (i.e., 37% of total population) born between 2006 and 2009. Accuracies of genomic selection were derived from a correlation between GEBV and DYD of validation males, where DYD were estimated from official genetic evaluation of January 2012. Pedigree-derived accuracies of validation males were estimated from correlation between EBV and DYD. The EBV were obtained from training males using the same

model as for GEBV, except that $\text{Var}(u) = \mathbf{A}\sigma_u^2$, where **A** is the pedigree relationship matrix, and excluding genotype information. The prediction equation used to compute EBV of young males was derived from DYD (and YD) of reference animals and pedigree. The gain of using genomic information was derived from the difference between correlations between GEBV and DYD and correlations between EBV and DYD for validation males.

Taking into account that females were born in 2008 and 2009, and that no males with known phenotypes were born after them, the effect of adding females on genomic evaluation accuracy could not be estimated by cross validation. Model accuracies, derived from prediction error variance (**PEV**), were used to investigate the benefit of adding female genotypes.

Model Accuracy of Young Buck Breeding Values Estimated from PEV. Estimates of additive genetic values and PEV were obtained for all animals using the GBLUP model. The model accuracy considered (ρPEV) was

$$\rho_{PEV} = \sqrt{\frac{\sigma_u^2 - PEV}{\sigma_u^2}},$$

where PEV was the variance of prediction error and σ_u^2 is the genetic variance estimated on GEBV obtained (Bijma, 2012).

Average model accuracies were calculated for the 148 young buck candidates not yet progeny tested. Because the genomic relationship matrix was built using pedigree information, pedigree and genomic accuracy derived from PEV could be compared, calculating gain of model accuracy using genomic evaluation (Legarra et al., 2009; Christensen et al., 2012). The gain in model accuracy was estimated using genomic information for young bucks by comparing pedigree-derived model accuracy with genomic model accuracy. Pedigree-derived accuracy was obtained using the method described in the Accuracy and Bias of Genomic Evaluation section: using the same pedigree information and phenotypes of males and females from a reference population as for GBLUP evaluation but without molecular information. The objective of this study was to investigate how average genomic prediction accuracy varies with reference population size and addition of males or females.

RESULTS AND DISCUSSION

Description of the Reference Population

Extent of LD in the Population. Average r^2 calculated for each breed separately (Alpine, Saanen) and for the multi-breed population (Alpine + Saanen) as a function of marker distance are presented in Figure 1. For the 3 populations studied, average r^2 decreased with increasing marker distance. This decrease was less substantial for marker distances >150 kb. Average r^2 was constant for distances >1,200 kb, and was 0.07 and 0.04, respectively, for the 2 single breed populations and the multi-breed population. Extent of LD estimated in Saanen was close to that estimated in Alpine. Average r^2 values obtained in the multi-breed population were lower than in the single-breed populations. In this study, the extent of LD between 2 consecutive SNP (i.e., 50 kb: average distance between 2 SNP on the chip) was 0.17 for single-breed populations and 0.14 for the multi-breed population.

Average r^2 values in dairy goat at 50 kb were similar to the values reported in Lacaune sheep [0.12; G. Baloche, INRA-Station d'Amélioration Génétique des Animaux (SAGA), Toulouse, France, personal communication] but lower than those reported in Holstein dairy cattle (from 0.18 to 0.3, de Roos et al., 2008; Habier et al., 2010) and in Landrace, Duroc, Hampshire, and Yorkshire pigs (from 0.46 to 0.36, Badke et al., 2012). Similarities in LD extent estimated in Saanen and Alpine breeds could be explained by their common ancestor. The lower estimates of average r^2 in the multi-breed population (Alpine + Saanen) than in the single breed populations are in agreement with results in dairy cattle (Toosi et al., 2010; Hozé, 2012). In the European multi-breed dairy cattle population, extent of LD was 0.15 at 70 kb compared with 0.19 and 0.25 in Montbeliarde and Brown Swiss breeds, respectively (Hozé, 2012). As expected, the difference in LD extent between the multi-breed population and the single breed populations increased with marker distance. For small marker distances, it was due to the common origin of the 2 breeds. For higher marker distances, it could be associated with the management of Alpine and Saanen as purebred for more than 40 yr. Indeed, LD calculated for small marker distance, when fewer recombinations are possible, reflects the former history of breeds. For larger distances, extent of LD reflects more recent history (Hayes et al., 2003).

Using simulation, Habier et al. (2007) demonstrated that part of genomic accuracy was due to LD, using the decay of accuracy and LD per generation. In German Holstein dairy cattle (with LD extent of 0.3 for 60 kb), the part of accuracy due to LD ranged from 10% for protein yield with a reference population of 1,048 dairy bulls to 47% for fat yield with a reference population of 2,960 bulls (Habier et al., 2010). Based on these results, the relatively low extent of LD measured in the dairy goat population in the current study should not lead to high values of genomic evaluation accuracies. However,

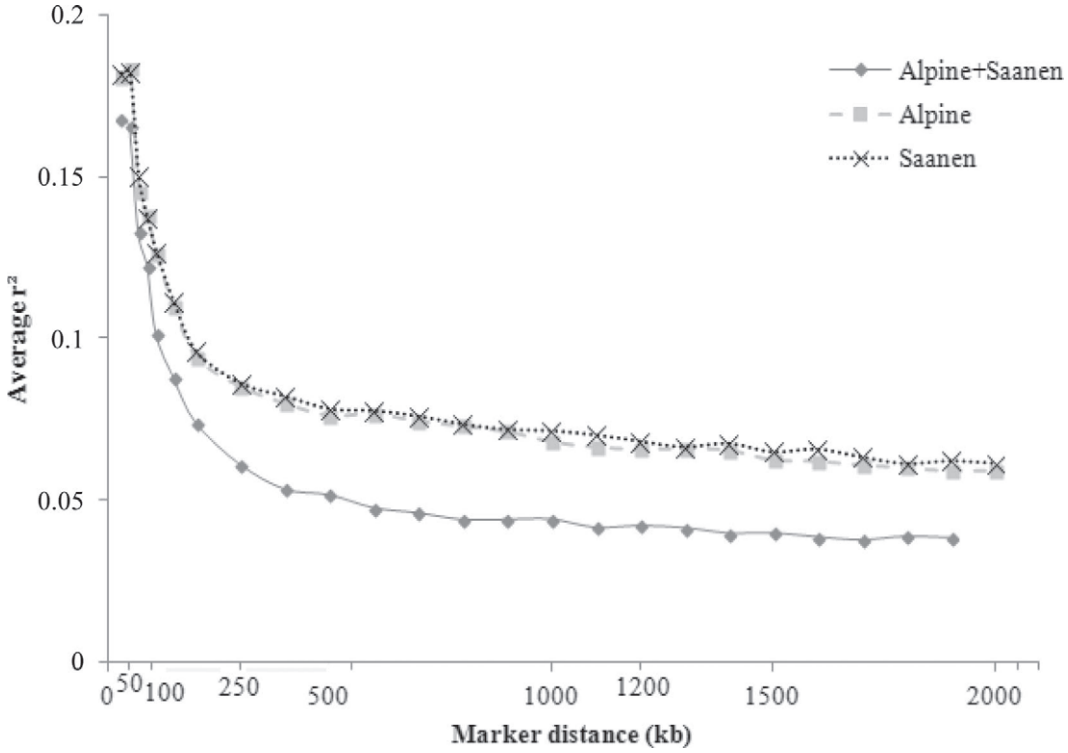


Figure 1. Linkage disequilibrium (average r^2) in Saanen and Alpine breeds and in the whole population (Alpine + Saanen).

accuracy of genomic evaluation was not the only parameter that influenced genomic evaluation accuracy.

Persistence of LD Phases Among Saanen and Alpine Breeds. Figure 2 shows the correlations between signed r-values of extent of LD among Alpine and Saanen breeds as a function of the distance between markers. Persistence of LD phases among Alpine and Saanen breeds decreased with genomic distance. At marker distance <50 kb, correlations of r among Alpine and Saanen breeds ranged from 0.88 to 0.56. This means that 2 SNP had the same level of LD in the Alpine breed and in the Saanen breed. The persistence of LD phases at 50 kb (i.e., average distance between 2 SNP) among Alpine and Saanen breeds was 0.56. Correlations of signed r-values estimated in Saanen and Alpine breeds decrease with increasing genomic distance between markers.

For short marker distances, persistence of LD phases among Alpine and Saanen was similar to the values reported between French Manech red-faced and black-faced sheep [0.5; G. Baloche, INRA-Station d'Amélioration Génétique des Animaux (SAGA), Toulouse, France, personal communication]. The 2 goat breeds (Alpine and Saanen) were genetically close centuries ago, as were the 2 Manech sheep breeds.

For greater marker distances, correlations of r-values in the reference population (0.08 for 600 kb) were close to those found between Lacaune and Manech black-faced sheep [0.09 for 600 kb; G. Baloche, INRA-Station d'Amélioration Génétique des Animaux (SAGA), Toulouse, France, personal communication]. But they were lower than that reported in dairy cattle between Jersey and Holstein for 600 kb (de Roos et al., 2008), in beef cattle between Charolais and Angus (Lu et al., 2012), and between Landrace and Yorkshire pigs (Badke et al., 2012).

Combining several breeds in a single reference population was considered when persistence of LD phases was high, as for Jersey and Holstein. However, the moderate level of LD phase persistence for 2 consecutive markers in Alpine and Saanen goats did not prohibit combining both breeds.

Relationships and Inbreeding Between and Within Populations. The pedigree file of 37,669 animals common to the 2 breeds took into account 26 generations. The kinship coefficient in the whole population was, on average, 1.6%. The kinship coefficient calculated within the 4 reference populations ranged from 1.6 to 1.8% (Table 2). The highest coefficients were obtained for case B, because of the addition of daughters of males from case A, and for case C, because of the addition of strongly related males from case A. Nevertheless, the addition of the daughters of 20 males from case C to case D did not increase the kinship

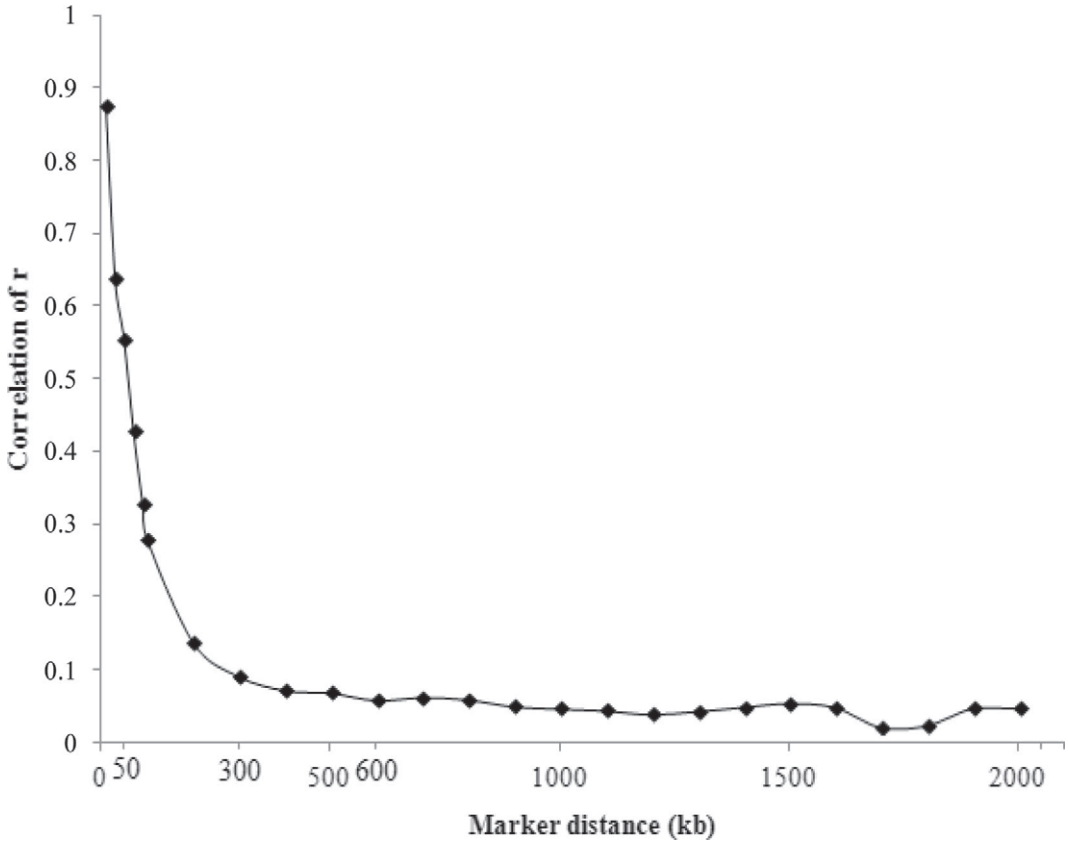


Figure 2. Persistence of linkage disequilibrium (LD) among Alpine and Saanen breeds derived as correlation of signed r LD values between the 2 breeds.

coefficient, because these females were not strongly related to all of the males. Kinship coefficients were 1.3% within the candidates and ranged from 1.3 to 1.4% between reference and candidate populations. Kinship coefficients reported in this study were low compared with that observed in cows (~5%; Habier et al., 2010; Pszczola et al., 2012).

Inbreeding coefficients (F_z in Table 2) were 2.1, 2.8, 2.2, and 2.6% within reference populations for cases A, B, C, and D, respectively, and 2.1% within candidate populations. They were lower than in Holstein dairy cattle (Miglior, 2000). The reported inbreeding in case B was caused by a higher proportion of females in the population, these being more inbred than the males. The relatively low levels of kinship and inbreeding coefficients within the populations can be attributed to the implementation of a new scheme of inbreeding management (optimizing contribution methods) in the French goat population in 2002 (Colleau et al., 2004). In this scheme, selection of AI bucks is managed within families to maintain genetic progress and minimize the average pairwise relationship coefficient in the population. However, among the 20 sires of the females, only a few had different ancestral origins. Inbreeding coefficients of those males and their daughters were higher (2.8 compared with 1.9%; results not shown) than those observed in the other males of the study.

Genomic Evaluation

Accuracy, Bias, and Gain in Accuracy of Genomic Evaluation Estimated by Cross Validation. Correlations between DYD and GEBV estimated in the validation population of 252 males ranged from 32.1% for SCS to 53.3% for fat content (Table3). The highest correlations were obtained for the most heritable traits (i.e., fat and protein contents).

These results were lower than those reported in the French Holstein dairy cattle population (Fritz et al., 2010) for similar traits: 39 versus 59% for milk yield, 36 versus 60% for protein yield, and 37 versus 63%, for udder floor position. These differences in accuracy could not be explained by DYD accuracy in French dairy goats (average EDC of 390), which is slightly higher than in dairy cattle, but are explained by the structure and the size of the reference population. Correlations

Table 2. Average (μ) and standard error (SE) kinship (F_{ij}) and inbreeding (F_z) coefficients within reference and candidate populations and between candidate and reference populations

	Case[1]							
	A		B		C		D	
Item	F_{ij}	F_z	F_{ij}	F_z	F_{ij}	F_z	F_{ij}	F_z
Within reference population								
μ (%)	1.6	2.1	1.8	2.8	1.8	2.2	1.7	2.6
SE (%)	2.1	1.0	3.2	1.4	2.6	1.1	2.9	1.3
Within candidate population								
μ (%)	1.3	2.1	1.3	2.1	1.3	2.1	1.3	2.1
SE (%)	2.3	0.7	2.3	0.7	2.3	0.7	2.3	0.7
Between reference and candidate populations								
μ (%)	1.4	—	1.3	—	1.4	—	1.3	—
SE (%)	2.1	—	2.0	—	2.2	—	2.0	—

[1]Case A: 67 males in the reference population, 148 male candidates; Case B: 67 males and 1,985 females in the reference population, 148 male candidates; Case C: 677 males in the reference population, 148 male candidates; Case D: 677 males and 1,985 females in the reference population, 148 male candidates.

between DYD and GEBV for our validation bucks were similar to those reported in Manech red-faced dairy rams (38 and 37% for milk and fat yields) with a training population of around 1,000 rams (Barillet et al., 2012) and in Normande dairy bulls (36 and 33% for milk and protein yields) with 930 training bulls (Fritz et al., 2010). Nevertheless, the results for SCS and type traits, between 32 and 43%, were slightly lower than in other species (48% for SCS of Manech, Barillet et al., 2012; 47% for udder floor position of Normande breed, Fritz et al., 2010).

Gains of accuracy using genomic information in our study (Table 3) ranged from 3.4% for protein content to 21.3% for fore udder. These gains for milk production traits were lower than those obtained for milk yield in other species (41% in Manech, Barillet et al., 2012; 41.7% in Normande, Fritz et al., 2010). This finding can be explained by the high pedigree-derived accuracies of young buck breeding values for those traits because of a high number of daughters per sire (388 in average). For udder floor position, gains were similar to those reported in the Normande dairy cattle breed (23.7%) from Fritz et al. (2012).

Regression coefficients presented in Table 3 ranged from 0.73 to 0.96. They were higher for fat and protein contents (96.2 and 94.9%), close to those reported in French Lacaune dairy sheep (85 to 86%; Duchemin et al., 2012) and French dairy cattle (71 to 113%; Karoui et al., 2012). A coefficient of 1 (indicating the absence of bias) was expected if the animals in the validation set were not selected. Biases of genomic breeding value estimations were low for fat and protein contents and for type traits, with regression coefficients up to 90%, except for fore udder.

Model Accuracy of Genomic Predictions for Candidates Estimated from PEV. Figure 3 shows the average model accuracy of genomic prediction (derived from PEV) calculated for the 148 candidates without progeny test results, in each case for milk yield (identical results were obtained for protein and fat yields), fat content (identical results were obtained for protein content), and teat angle (identical results were

Table 3. Correlations between daughter yield deviations (DYD) and genomic (G)EBV for males from validation population and regression coefficient of DYD onto GEBV

Trait	DYD × GEBV	DYD × EBV	Gain (%)	Regression coefficient
Milk yield	0.391	0.372	5.1	0.786
Fat yield	0.373	0.350	6.2	0.784
Protein yield	0.362	0.345	4.9	0.762
Fat content	0.533	0.495	7.7	0.962
Protein content	0.519	0.502	3.4	0.949
Somatic cell score	0.321	0.305	5.2	0.742
Udder floor position	0.367	0.304	20.7	0.918
Udder shape	0.339	0.280	21.1	0.899
Rear udder attachment	0.425	0.396	7.3	0.923
Fore udder	0.325	0.268	21.3	0.726
Teat angle	0.352	0.324	8.6	0.908

obtained for all type traits and SCS). Similar results were obtained when heritability and DYD accuracy were the same. For SCS and type traits, accuracy of DYD (EDC) for all of those traits was 35% lower than EDC of milk production traits. Accuracies ranged from 9% for type traits and SCS in case A to 56% for fat and protein contents in case D with highest heritabilities. The lower values observed for GEBV accuracies for the young males in case A could be explained by the small size of the reference population and the absence of all fathers of candidates in the reference population.

The highest accuracies obtained with a reference population of 677 males and 1,985 females were similar to those observed in Merinos sheep (ρPEV from 50 to 57%, for ultrasound-scanned traits) with an average relationship of 0.5 between animals (Clark et al., 2012) and in Jersey dairy cattle (from 52 to 57% for milk production traits; Hayes et al., 2009). These results were higher than those reported in hens (ρPEV: 42%, for presence test of *Salmonella* in spleen) because of the small number (1,342) of SNP used in Calenge et al. (2011).

The model accuracies obtained for the 257 males of the training population used in the previous section were lower than the accuracies derived from cross validation (from 9 to 56%, results not shown) as in other studies (Clark et al., 2012). This could be explained by not taking into account the genetic selection of candidates, which led to overestimations of model accuracy (Gorjanc et al., 2012).

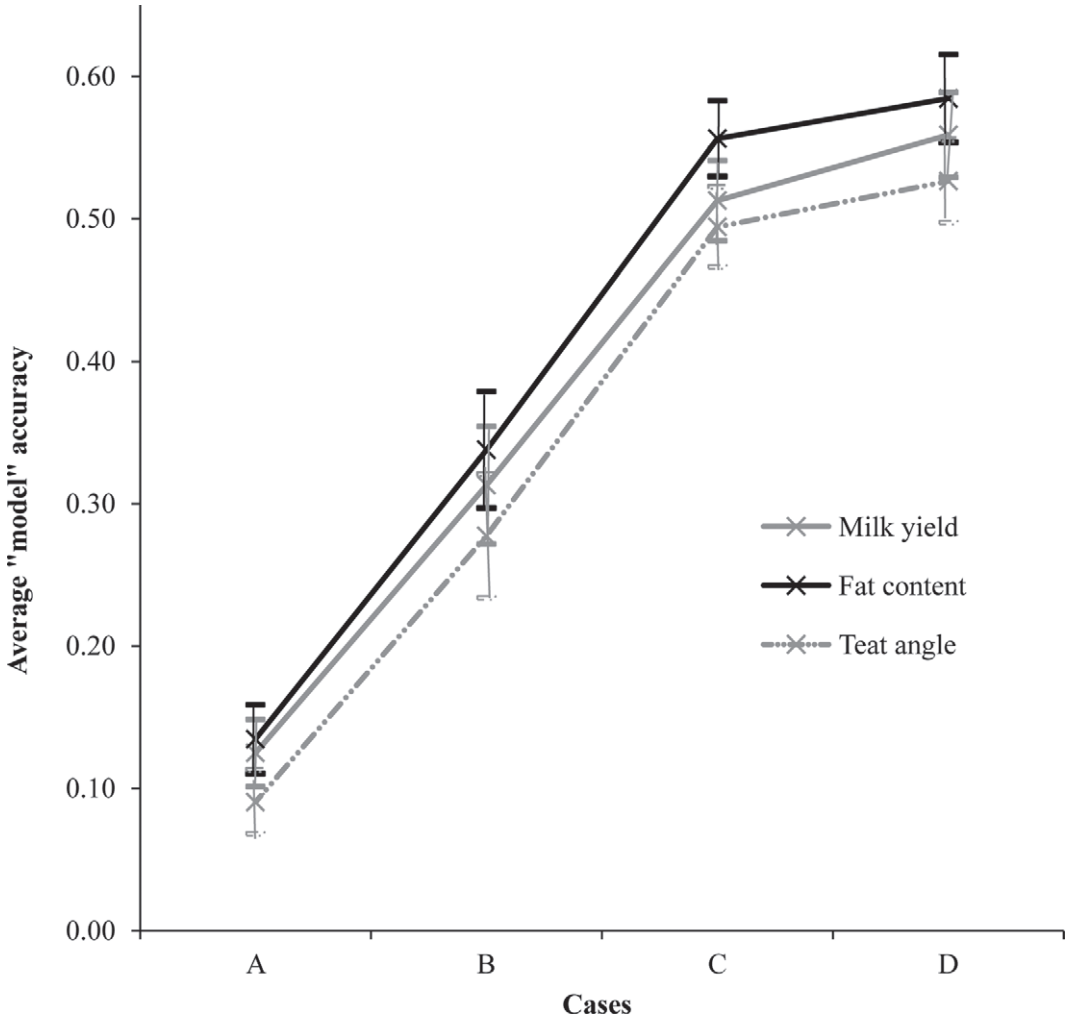


Figure 3. Average "model" accuracy derived from prediction error variance of genomic predictions for candidates in fat content, milk yield, SCS, and rear udder attachment. Case A: 67 males in the reference population, 148 male candidates; case B: 67 males and 1,985 females in the reference population, 148 male candidates; case C: 677 males in the reference population, 148 male candidates; case D: 677 males and 1,985 females in the reference population, 148 male candidates. The same results were obtained for protein yield and fat yield as are shown for milk yield; the same results were obtained for protein content as are shown for fat content; and the same results were obtained for udder floor position, udder shape, and fore udder as are shown for teat angle.

Table 4. Average (μ) and standard error (SE) of differences between breeding value "model" accuracies and pedigree accuracies derived from prediction error variance for the 148 male candidates

	Case[1]			
Trait	A	B	C	D
Milk yield[2]				
μ	0.03	0.05	0.03	0.05
SE	0.01	0.03	0.02	0.03
Fat content[3]				
μ	0.05	0.04	0.03	0.07
SE	0.02	0.01	0.02	0.04
Somatic cell score[4]				
μ	0.01	0.06	0.03	0.04
SE	0.02	0.02	0.03	0.03

[1]Case A: 67 males in the reference population, 148 male candidates; Case B: 67 males and 1,985 females in the reference population, 148 male candidates; Case C: 677 males in the reference population, 148 male candidates; Case D: 677 males and 1,985 females in the reference population, 148 male candidates.

[2]The same results were obtained for protein yield and fat yield.

[3]The same results were obtained for protein content.

[4]The same results were obtained for udder floor position, udder shape, fore udder, and teat angle.

Adding animals increased GEBV accuracy for all traits and in all cases (Figure 3). The addition of females increased the accuracy by 5% (case C vs. D). But accuracies for udder type traits and SCS increased greatly by 206% (case A vs. B, Figure 3) when females were added to the population of 67 males. However, in case D, the addition of females led to a greater improvement of GEBV accuracies for their 15 half-sibs than for the 148 other candidates; that is +30% for SCS (results not shown). In dairy cattle, the addition of dams of bulls slightly altered GEBV reliabilities derived from correlation between DYD and GEBV for candidate bulls, from −4.9% for fat yield in Holstein (Dassonneville et al., 2013) to 5.8% (Wiggans et al., 2011) and 8% (Pryce et al., 2012) for protein and fat contents. These lesser values could be due to the preferential treatment of some cows and lead to errors in the phenotypes of the bulls (Dassonneville et al., 2013); this is not the case in large herds such as in goats.

Gain in "Model" Accuracy of Genomic Predictions for Candidates. Table 4 shows the gains in the accuracy of EBV of the 148 candidates observed when using genomic information or only pedigree and phenotypic information. These gains ranged from 1% for SCS and type traits in case A to 7% for fat and protein contents in case D. These values were similar to gains of theoretical accuracy obtained for the presence of *Salmonella* in hens (from 0 to 15%; Calenge et al., 2011). The gains observed in this study were lower than those obtained for Merinos sheep: from 76% for ultrasound-scanned eye muscle depth to 468% for ultrasound-scanned traits (Clark et al., 2012).

The addition of females to the reference population of 67 males was less advantageous for gains in GEBV accuracy (+33% for SCS and type traits, case B vs. case A; Table 4) than adding them in a larger reference population (+500% for SCS and type traits, case D vs. case C; Table 4). This improvement in accuracy gains could be explained by an increase in GEBV accuracy between case C and case D, whereas EBV accuracy (calculated using only phenotypes and pedigree information) was similar in both cases.

All females used in this study were sired by 20 sires. Genetic diversity of these females was not wide enough that the addition of genotyped females actually improved the prediction model. An interesting point for the future would be to examine the benefit of genotyping a set of buck dams chosen to represent genetic diversity of the whole set of dams.

The current reference population of 677 males used in this study comprised all bucks progeny tested in the breeding scheme. Adding new generations of genotyped males to the reference population will increase the reference population size but will not improve relationship between the candidate individuals and the reference individuals. The main way to increase size and improve structure of the French dairy goat reference population could be done essentially by genotyping females. The choice of these females should be further investigated.

CONCLUSIONS

The present study is the first report to be published on genomic evaluation in dairy goats. The results describe the characterization of the French reference population available currently with an extent of LD of 0.14 between 2 consecutive SNP. Accuracies of genomic evaluation were similar to values reported in other

species, but gains in using genomic information were slightly low because of the structure and size of the reference population. Accuracies and gains in accuracy could be improved by adding genotyped females. The use of the multiple-trait model, models using haplotypes instead of SNP, and single-step genomic BLUP models will be examined in the future.

ACKNOWLEDGMENTS

This work was funded by the French Genovicap and Phenofinlait programs (ANR, Apis-Gène, CASDAR, FranceAgriMer, France Génétique Élevage, and French Ministry of Agriculture, Paris, France) and the European 3SR project. The first author benefitted from financial support from the Midi-Pyrénées region (Toulouse, France) and the SELGEN program of the French National Institute of Research in Agronomy (INRA, Paris, France). We thank Helen Munduteguy (Pyratus, Tarbes, France) and Wendy Brand-Williams (INRA, Jouy-en-Josas, France) for checking the English of our article and the two reviewers for their very constructive comments. This study would not have been possible without the goat SNP50 BeadChip developed by the International Goat Genome Consortium (IGGC; www.goatgenome.org).

REFERENCES

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 score. J. Dairy Sci. 93:743–752.

Araujo, A. M. D., S. E. F. Guimaraess, T. M. M. Machado, P. S. V. Lopes, C. S. Pereira, F. L. R. D. Silva, M. T. Rodrigues, V. D. S. Columbiano, and C. G. A. D. Fonseca. 2006. Genetic diversity between herds of Alpine and Saanen dairy goats and the naturalized Brazilian Moxoto breed. Genet. Mol. Biol. 29:67–74.

Babo, D. 2000. Les Races Ovines et Caprines Françaises. 1st ed. France Agricole, Paris, France.

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, and J. P. Steibel. 2012. Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics 13:24.

Barillet, F., G. Baloche, G. Lagriffoul, H. Larroque, R. Robert-Granié, A. Legarra, and J. M. Astruc. 2012. Genomic selection in French Lacaune and Manech dairy sheep breeds: Comparison of BLUP and GBLUP accuracies. Page 3 in Proc. 38th ICAR session, Cork, Ireland.

Bélichon, S., E. Manfredi, and A. Piacère. 1999. Genetic parameters of dairy traits in the Alpine and Saanen goat breeds. Genet. Sel. Evol. 31:529–534.

Boichard, D. 2006. Pedig, logiciel d'analyse de généalogies adapté à de grandes populations; version 2007. INRA SGQA, Jouy-en-Josas, France.

Calenge, F., A. Legarra, and C. Beaumont. 2011. Genomic selection for carrier-state resistance in chicken commercial lines. BMC Proc. 5(Suppl. 4):S24.

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.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. Van Der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4. http://dx.doi.org/10.1186/1297-9686-44-4.

Clément, V., D. Boichard, A. Piacere, 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.

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. Rumin. 13:209–212.

Colleau, J. J., S. Moureaux, M. Briend, and J. Béchu. 2004. A method for the dynamic management of genetic variability in dairy cattle. Genet. Sel. Evol. 36:373–394.

Danchin-Burge, C. 2011. Bilan de variabilité génétique de 9 races de petits ruminants laitiers et à toison. Compte rendu no. 001172004, Institut de l'Élevage collection résultats, juin 2011. Institut de l'Élevage, Paris, France.

Dassonneville, R., A. Baur, S. Fritz, D. Boichard, and V. Ducrocq. 2013. Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment. Genet. Sel. Evol. 45:40. http://dx.doi.org/10.1186/1297-9686-44-40.

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.

de Roos, A. P. W., C. Schrooten, R. F. Veerkamp, and J. A. M. van Arendonk. 2010. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. J. Dairy Sci. 94:1559–1567.

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.

Ducrocq, V. 1998. Genekit, BLUP software; June 2011 version. INRA SGQA, Jouy-en-Josas, France.

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

Fritz, S., F. Guillaume, P. Croiseau, A. Baur, C. Hoze, R. Dassonneville, M. Y. Boscher, L. Journeaux, D. Boichard, and V. Ducrocq. 2010. Mise en place de la sélection génomique dans les trois principales races françaises de bovins laitier. Renc. Rech. Rumin. 17:455–458.

Goddard, M. E. 2009. Genomic selection: Prediction of accuracy and maximisation of long-term response. Genetica 136:245–257.

Gorjanc, G., J. M. Hickey, and P. Bijma. 2012. Reliability of breeding values in selected populations. Page 15 in Interbull Bulletin, 5, Cork, Ireland.

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

Habier, D., 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. Genet. Sel. Evol. 42:5. http://dx.doi.org/10.1186/1297-9686-42-5.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and E. Goddard. 2009. Accuracy of genomic breeding values in multibreed dairy cattle populations. Genet. Sel. Evol. 41:51. http://dx.doi.org/10.1186/1297-9686-41-51.

Hayes, B. J., and M. E. Goddard. 2010. Genome-wide association and genomic selection in animal breeding. Genome 53:876–883.

Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13:635–643.

Hozé, C. 2012. High density chip brings new opportunity for multibreed genomic evaluations in dairy cattle. Page 6 in 16th QTL MAS Workshop, Alghero, Sardinia, Italy.

Institut de l'élevage. 2010. Résultats de Contrôle Laitier des espèces bovine et caprine (campagne 2010). Accessed Mar. 27, 2013. http://idele.fr/recherche/publication/idelesolr/recommends/resultats-de-controle-laitier-des-especes-bovine-et-caprine-campagne-2010.html.

Karoui, S., M. J. Carabano, C. Diaz, and A. Legarra. 2012. Joint evaluation of French dairy cattle breeds using multiple-trait models. Genet. Sel. Evol. 44:39. http://dx.doi.org/10.1186/1297-9686-44-39.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663.

Legarra, A., and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic and pedigree data in a single-step best linear unbiased prediction. J. Dairy Sci. 95:4629–4645.

Liu, Z., F. R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genet. Sel. Evol. 43:19. http://dx.doi.org/10.1186/1297-9686-43-19.

Lu, D., M. Sargolzaei, M. Kelly, C. Li, G. Vander Voort, Z. Wang, G. Plastow, S. Moore, and S. P. Miller. 2012. Linkage disequilibrium in Angus, Charolais, and crossbred beef cattle. Front. Genet. 3:1–10.

Makgahlela, M. L., I. Strandén, U. S. Nielsen, M. J. Sillanpää, and E. A. Mäntysaari. 2013. The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. J. Dairy Sci. 96:5364–5375. http://dx.doi.org/10.3168/jds.2012-6523.

Maroteau, C., I. Palhière, H. Larroque, V. Clément, G. Tosser-Klopp, and R. Rupp. 2012. Genetic parameter estimation for major fatty acids in French dairy goats. Page 366 in 63rd Annu. Mtg. European Federation of Animal Science (EAAP), Bratislava, Slovakia.

Meuwissen, T. H. E., and Z. Luo. 1992. Computing inbreeding coefficients in large populations. Genet. Sel. Evol. 24:305–313.

Miglior, F. 2000. Impact of inbreeding-Managing a declining Holstein gene pool. Pages 108–113 in Proc. 10th World Holstein Friesian Federation Conf., Sydney, Australia.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUPF90 and related programs (BGF90). Pages 743–745 in Session 28, Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France.

Piacere, A., I. Palhière, H. Rochambeau, and D. Allain. 2004. Analysis of the genetic variability of the French Alpine and Saanen breeds using genealogical data. Page 30 in Proc. 8th Int. Goat Conf., Pretoria, South Africa.

Pryce, J., B. Hayes, and M. E. Goddard. 2012. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. Page 28 in Proc. 38th ICAR Conf., Cork, Ireland.

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. J. Dairy Sci. 95:389–400.

Rogers, A. R., and C. Huff. 2009. Linkage disequilibrium between loci with unknown phase. Genetics 182:839–844.

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. J. Dairy Sci. 94:3629–3634.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223.

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

Sullivan, P. 2010. Description of Usage for CrEDC_5e.c. Canadian Dairy Network, Guelph, Canada.

Toosi, A., R. L. Fernando, and J. C. M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88:32–46.

Tosser-Klopp, G., P. Bardou, C. Cabau, A. Eggen, T. Faraut, H. Heuven, S. Jamli, C. Klopp, C. T. Lawley, J. Mcewan, P. Martin, C. Moreno, P. Mulsant, I. Nabihoudine, E. Pailhoux, I. Palhière, R. Rupp, J. Sarry, B. Sayre, A. Tircazes, J. Wang, W. Wang, T. P. Yu, and W. Zhang. 2012. Goat genome assembly, Availability of an international 50K SNP chip and RH panel: An update of the International Goat Genome Consortium projects. Pages 1–14 in Plant and Animal Genome Conf., San Diego, CA.

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

VanRaden, P. M., C. P. Van Tassel, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. (Camb.) 93:357–366.

Wiggans, G. R., T. A. Cooper, P. M. Vanraden, and J. B. Cole. 2011. Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. J. Dairy Sci. 94:6188–6193.

Bilan

L'étude de la structure génétique de la population de référence caprine présentée dans l'article est un résumé de ce qui a été décrit plus en détail dans le chapitre 2. La structure de population a été décrite à partir de différents critères : DL et coefficients de consanguinité et de parenté. La population caprine génotypée présente des caractéristiques de plus grande diversité génétique que celles des bovins laitiers Holsteins en France. Elle est cependant similaire à celle observée chez les ovins laitiers Lacaune. La taille de la population de référence multiraciale (Alpine + Saanen), de 677 mâles et 1 985 femelles au maximum, est cependant faible comparée à celle des ovins laitiers Lacaune (1 886 mâles).

Les analyses de validation croisée ont montré une qualité de prédiction plutôt bonne avec des corrélations de validation génomiques supérieures aux précisions estimées sur pédigrée de 3% pour le taux protéique à 21% pour la forme de l'avant pis, le profil de la mamelle et la distance plancher-jarret. Les pentes de régression des GEBV prédites pour les mâles de validation sont également correctes, entre 0,73 pour la forme de l'avant pis et 0,96 pour le taux butyreux. La qualité des prédictions est légèrement inférieure à celle estimée en ovins Lacaune.

Les précisions génomiques théoriques des candidats augmentent avec la taille de population de référence considérée mais restent inférieures aux précisions estimées sur ascendance quel que soit le caractère considéré. En effet la valeur maximale de la précision génomique théorique obtenue pour la quantité de lait (0,55 avec une population de référence de 677 mâles et 1 945 femelles) est nettement inférieure à la précision sur ascendance des candidats obtenue en indexation classique (0,62). L'ajout de femelles génotypées dans la population de référence permet une nette amélioration de ces précisions théoriques des candidats dans le cas d'une population de référence de 67 mâles. Cependant l'intérêt relatif des génoytpages des femelles est plus faible, bien que non nul (+ 5% en moyenne), lorsque la population de référence de mâles considérée est plus grande (677 mâles).

En conclusion, la structure et la taille de la population de référence caprine française ne semble pas être un frein à une sélection génomique. La bonne qualité des prédictions qui en découle encourage l'utilisation des évaluations génomiques. Cependant le niveau de précision théorique des GEBV des candidats obtenu avec la méthode GBLUP en deux étapes est trop faible pour que cette méthode d'évaluation génomique soit envisagée en caprins laitiers français. D'autres approches méritent d'être testées.

3.I.1.B Evaluation génomique uni-raciale basée sur les DYD

L'article I propose une analyse des évaluations génomiques multiraciales dans l'espèce caprine. Cette stratégie a été la première adoptée en raison du faible nombre d'individus génotypés par race (355 Saanens et 470 Alpins). Cependant, la race Alpine et la race Saanen se distinguent du point de vue de leur structure génétique (chapitre 2). Cet éloignement génétique pose la question du gain réel apporté par le multiracial. Cette étude a donc consisté à réaliser une évaluation pour chaque race et à comparer les résultats à ceux obtenus dans le cas d'une évaluation multiraciale.

Une approche en deux étapes basée sur les DYD des animaux a été utilisée avec le modèle suivant : $y_s = \mu_s + Z_s u_s + e_s$ pour la race Saanen et $y_a = \mu_a + Z_a u_a + e_a$ pour la race Alpine avec **y** le vecteur des DYD, μ le vecteur de l'effet moyenne, **Z** la matrice d'incidence reliant les phénotypes (DYD) au vecteur des valeurs génomiques **u** et **e** le vecteur des résidus. Le vecteur des valeurs génomiques **u** est supposé normalement distribué de variance $\mathrm{Var}(u) = (0{,}95\ G + 0{,}05\ A_{22})\ \sigma^2_u$ combinant la matrice génomique **G** et la matrice de parenté **A**. La matrice génomique est estimée comme dans l'article I selon la méthode de VanRaden (VanRaden, 2008) excepté que les fréquences alléliques considérées sont celles obtenues dans chaque race séparément. Les caractères analysés sont les mêmes que ceux étudiés dans l'article I : les quantités de lait, de matière grasse (MG), de matière protéique (MP), les taux butyreux (TB) et protéique (TP), les comptages de cellules somatiques (LSCS), la distance plancher-jarret (PLA), le profil de la mamelle (PRM), la qualité de l'attache arrière (AAR), la forme de l'avant-pis (AVP) et l'orientation des trayons (ORT).

Les analyses conduites dans cette étude sont les mêmes que celles réalisées précédemment : validation croisée et estimation des moyennes des CD génomiques pour les candidats. La validation croisée a été réalisée séparément dans chaque race à l'aide de 232 mâles d'apprentissage et 152 de validation pour la race Alpine, et de 193 boucs d'apprentissage et 100 de validation pour la race Saanen (Figure 3 .26). A des fins de comparaison, les corrélations et les pentes des GEBV en fonction des DYD ont été ré-estimées par race à partir de l'évaluation multiraciale.

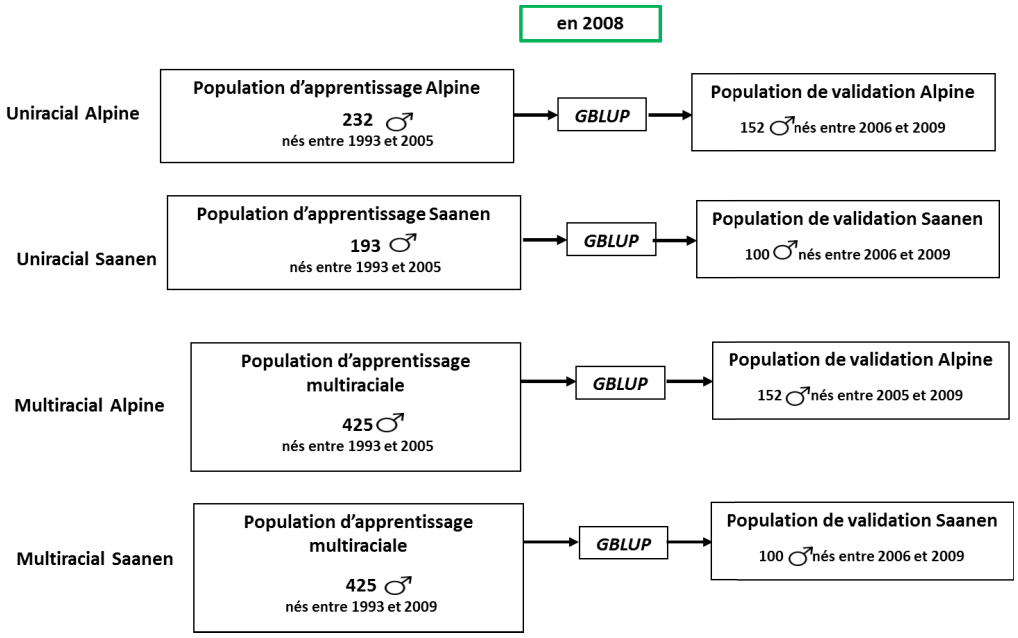


Figure 3.26 : Schéma des validations croisées réalisées pour l'étude du two steps en uniracial

Les coefficients de détermination des GEBV des jeunes mâles candidats ont également été estimés en utilisant une population de référence contenant soit la totalité (825) des génotypes des mâles et des femelles (cf. article I), soit uniquement les génotypes de chaque race. Les précisions théoriques des GEBV ($\sqrt{CD}$) ont été estimées séparément pour les 86 candidats Alpins et les 62 candidats Saanens dans chaque cas (Figure 3 .27).

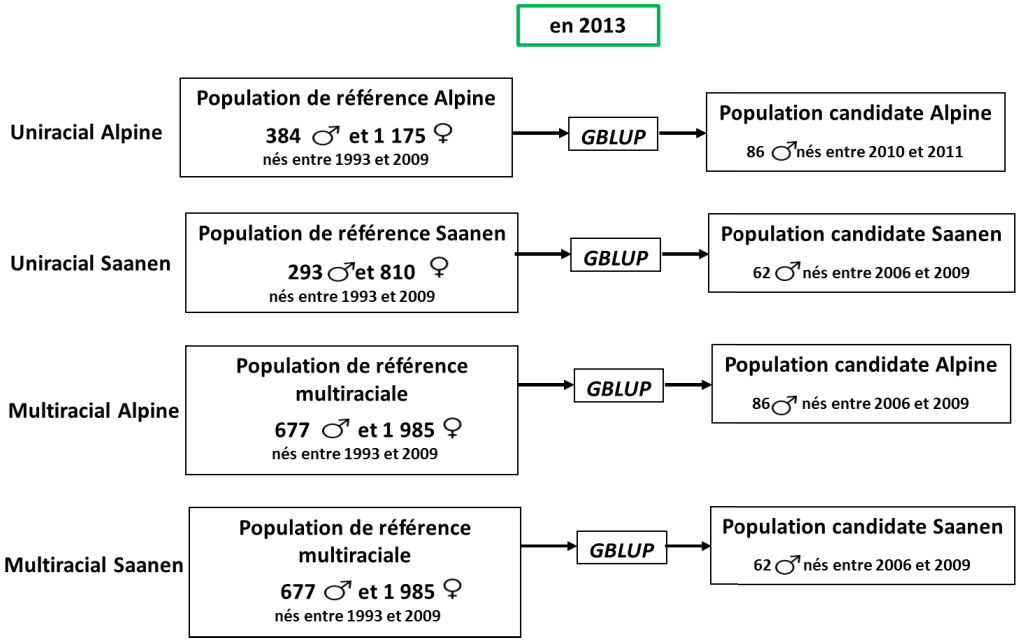


Figure 3.27 : Schéma des populations de référence et candidate utilisées pour l'estimation des précisions théoriques des GEBV des candidats dans le cas du two steps uniracial

La Figure 3 .28 présente les corrélations entre les DYD et les GEBV estimées sur les mâles de validation avec les différents modèles étudiés, y compris avec la population

multiraciale et pour l'ensemble des mâles de validation Alpine et Saanen (multiracial total, de l'article I). Les corrélations obtenues séparément dans chaque race ont été estimées entre 0,25 pour la matière protéique (MP) avec le modèle Saanen uni-racial et 0,51 pour le taux protéique (TP) avec le même modèle. Ces corrélations sont plus faibles que celles obtenues en ovins Lacaune, estimées entre 0,42 pour la quantité de lait et 0,56 pour le taux butyreux (TB) (Duchemin et al., 2012). Cependant la population ovine Lacaune française est de taille plus importante que la population considérée dans cette étude puisqu'elle comprend 1 886 mâles d'apprentissage et 681 mâles de validation.

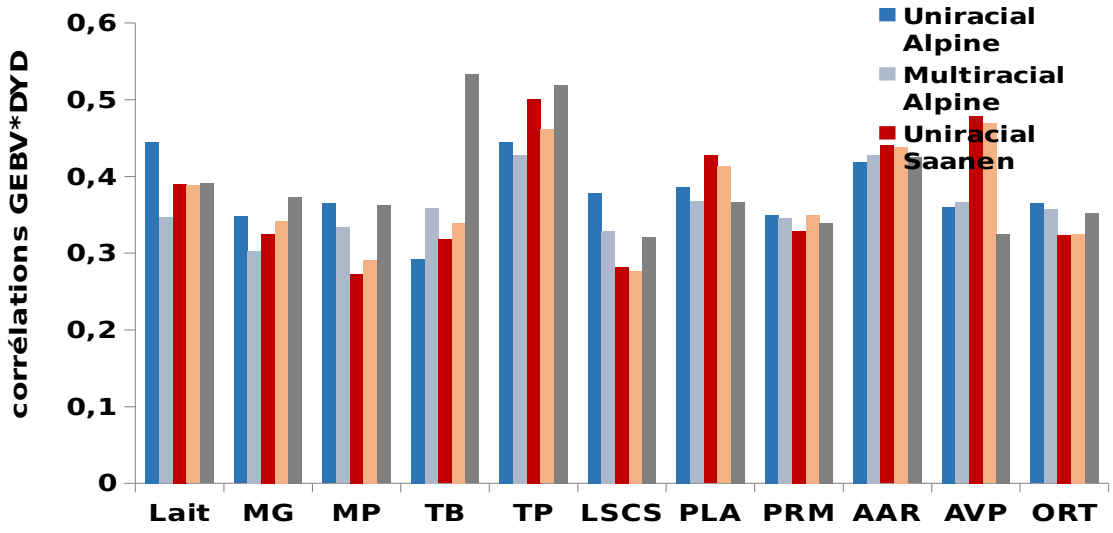


Figure 3.28 : Corrélations entre les GEBV et les DYD pour les mâles de validation de race Alpine ou Saanen, ou les deux (total) selon la population de référence utilisée (uniracial ou multiracial)

Les corrélations intra-race entre les GEBV et les DYD pour la population de validation obtenues dans le cas d'évaluations génomiques uniraciales (Uniracial Saanen et Uniracial Alpine, Figure 3 .28) sont similaires à celles obtenues dans le cas des évaluations génomiques multiraciales (Multiracial Saanen et Multiracial Alpine, Figure 3 .28) sauf pour la quantité de lait en Alpine. De plus, les corrélations obtenues intra-race pour l'évaluation multiraciale sont proches de celles trouvées pour l'ensemble des deux races (Multiracial total, Figure 3 .28 qui correspond aux résultats de l'article I) exceptées pour le TB et la forme de l'avant pis (AVP). Pour le TB les corrélations obtenues intra-race sont moins élevées que celles en multiracial alors que c'est le cas inverse pour l'AVP.

D'autre part, les différences de corrélations de validation entre la race Alpine et la race Saanen sont en général faibles, entre 1% sur PRM et 10% sur le TP, excepté sur le lait (14% en uni-racial), la MP (30% en uni-racial), les LSCS (35% en uni-racial) et l'AVP (37%). Ces différences de prédiction entre Alpine et Saanen ne sont pas au profit d'une race, elles existent dans les deux sens en fonction du caractère étudié. Ces résultats concordent avec ceux de la

littérature. En effet, l'utilisation d'une population multiraciale bovine Holstein-Jersey ne permet d'améliorer les corrélations de validation pour aucune des deux races et aucun caractère étudié (Hayes et al., 2009a). En revanche, certaines publications montrent que combiner plusieurs populations de bovins de race « Nordic Red » étroitement liées d'un point de vue génétique, permet d'augmenter les corrélations de validation de 0,1% à 60% selon le caractère considéré (Brøndum et al., 2011; Zhou et al., 2014). De même mais plus modestement, le regroupement de plusieurs populations européennes de bovins de race Holstein (Consortium EuroGenomics) permet d'augmenter les corrélations de validation, entre 8% et 11% (Lund et al., 2011).

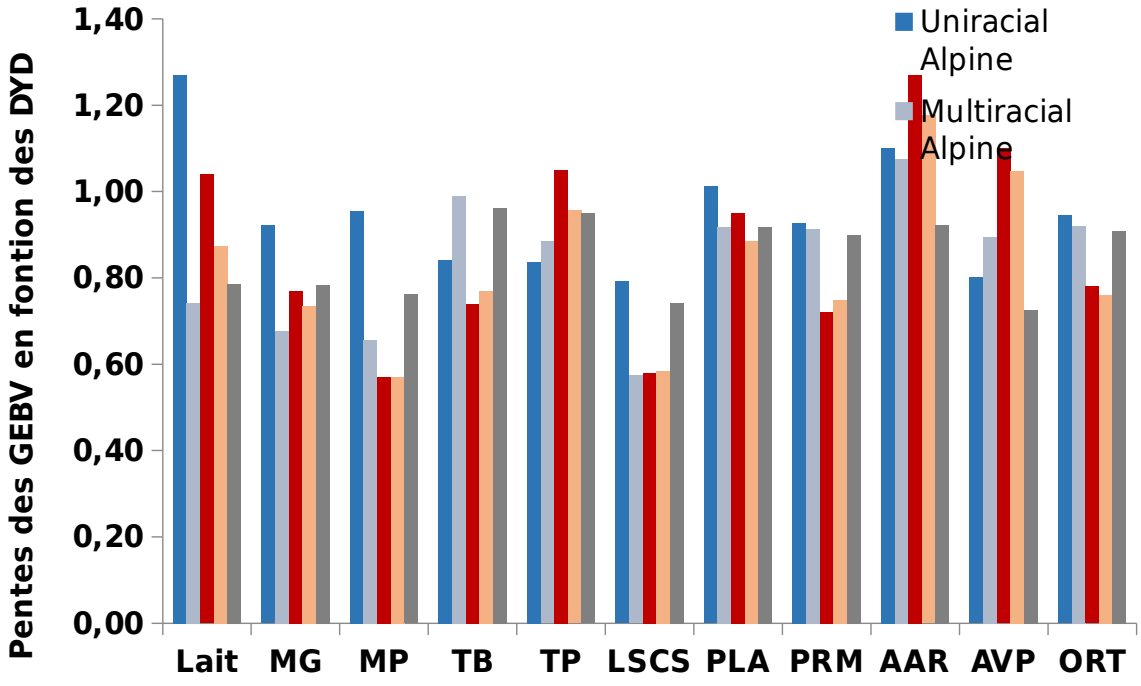


Figure 3.29 : Pentes de régression des GEBV en fonction des DYD pour les mâles de validation

Les valeurs des pentes de régression des DYD en fonction des GEBV pour les mâles de validation permettent d'analyser la sous-estimation ou la surestimation ainsi que la sur-dispersion ou la sous-dispersion des GEBV prédites. Dans cette étude les pentes de régression sont comprises entre 0,57 pour les LSCS avec un modèle multiracial analysé en race Alpine et 1,27 pour la quantité de lait avec le modèle uni-racial Alpin. On remarque que les pentes sont, pour la majorité des caractères, inférieures à 1 quelle que soit l'évaluation considérée (uniraciale ou multiraciale), ce qui signifie que les GEBV prédites sont sous-estimées. Le profil de ces pentes (Figure 3 .29) ne permet pas de conclure à une meilleure estimation des GEBV avec les évaluations multiraciales. En effet, les caractères pour lesquels les pentes sont améliorées (plus proches de 1) avec l'utilisation des évaluations multiraciales ne sont pas les

mêmes selon la race. Cependant, des pentes légèrement meilleures sont obtenues en évaluations uniraciales pour 5 caractères sur 11 en race Alpine (MG, MP, LSCS, PLA et ORT) et 4 caractères en race Saanen (lait, MG, PLA et ORT). Comme dans notre étude, considérant les pentes de régression, Brøndum et al (2011) n'ont pas trouvé d'avantage ou de désavantage à l'utilisation d'évaluations combinant plusieurs populations de bovins « Nordic red ». En revanche, Zhou et al (2014) ont montré que les évaluations combinant plusieurs populations de même race « Nordic red » et la race Ayrshire peuvent dégrader les coefficients de régression de l'ordre de 1% à 15% sur les caractères de production.

Les moyennes des précisions théoriques des GEBV des jeunes mâles candidats sont présentées à la Figure 3 .30. Elles sont estimées sur les candidats de chaque race dans le cas de l'évaluation uni-raciale correspondante et dans l'évaluation multiraciale. La précision sur ascendance est également présentée en Figure 3 .30, elle est estimée à partir des précisions des parents des 148 candidats. Les précisions théoriques des GEBV des candidats estimées avec une méthode two steps sont toujours inférieures aux précisions des EBV estimées sur ascendance quel que soit le caractère considéré. Les précisions estimées avec le modèle multiracial sont plus élevées que celles estimées avec les populations uni-raciales en raison d'une plus grande taille de population de référence. D'autre part, contrairement aux corrélations de validation, les précisions théoriques des GEBV des candidats sont plus fortes en race Alpine qu'en race Saanen. Ce résultat s'explique par la différence entre la taille de la population de référence dans les deux races (232 en Alpine vs 193 en Saanen). En effet, la taille de population semble être un des critères déterminants de la valeur des précisions lorsqu'elles sont estimées à partir des variances d'erreur de prédiction.

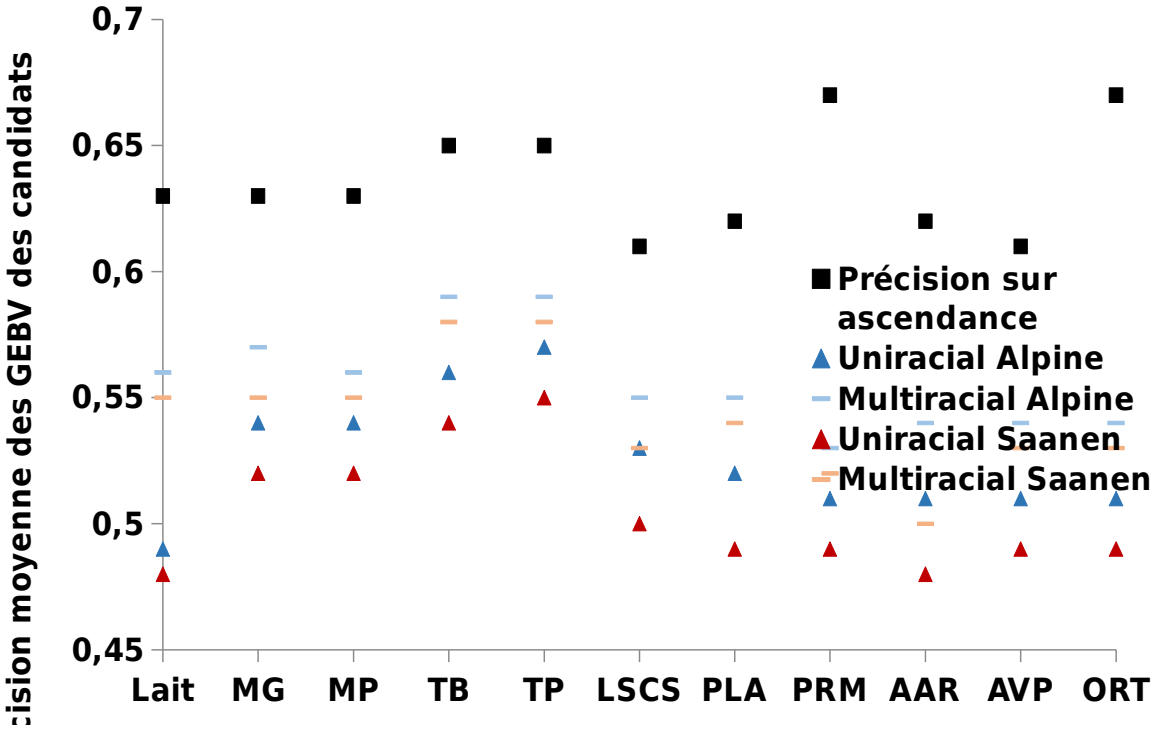


Figure 3.30 : Précision génomique théorique moyenne des valeurs génétiques des candidats de chaque race, calculées à partir des variances d'erreur de prédictions estimées avec la méthode GBLUP en 2 étapes à partir de populations de référence multiraciale ou uniraciale

En conclusion, malgré les petits effectifs de la population d'apprentissage considérés : 232 Alpins et 193 Saanens, la qualité de prédiction résultant des évaluations uni-raciales est similaire à celle obtenue avec une évaluation multiraciale. En revanche les précisions des GEBV des candidats estimés à partir des PEV sont plus élevées avec un modèle multiracial. Ces précisions étant inférieures aux précisions obtenues sur ascendance avec le modèle d'évaluation classique basée sur les performances brutes de femelles, la méthode en deux étapes considérant comme phénotypes les DYD, ne pourrait pas être envisagée pour l'évaluation génomique des caprins laitiers français. Cependant, d'autres phénotypes peuvent être utilisés avec cette méthode en deux étapes afin d'améliorer les qualités de prédictions ainsi que les précisions des GEBV.

3.I.2 Evaluation génomique multiraciale basée sur les EBV dé-régressés

Différents phénotypes peuvent être utilisés pour les indexations génomiques (cf. 1.II.2). Garrick et al. (2009) ont montré que l'utilisation des EBV dé-régressés associés à leur poids était plus appropriée. Dans l'étude présentée ici, la méthode et le modèle utilisés sont les mêmes que ceux de l'article I, les phénotypes sont :1) les EBV dé-régressés (DREBV)

corrigées des poids correspondants, ou 2) les EBV dé-régressés non corrigées (EBV/r^2, cf. 1.III.2), pour les mâles et les femelles.

Comme dans l'article I, une première analyse en validation croisée a été réalisée à partir des 677 mâles génotypés et phénotypés divisés en une population d'apprentissage de 425 mâles nés avant 2006 et une population de validation de 252 mâles nés entre 2006 et 2009 (Figure 3 .25). La qualité de prédiction a également été évaluée à l'aide des corrélations et des pentes de régression des GBEV sur les DYD.

D'autre part, les précisions théoriques des GEBV des candidats ont été estimés à l'aide de : 1) une population de référence de 677 mâles, 2) une population de 677 mâles et 1 985 femelles comme le montre le schéma en Figure 3 .31. Toutes les analyses ont été menées pour les 11 caractères indexés en caprins.

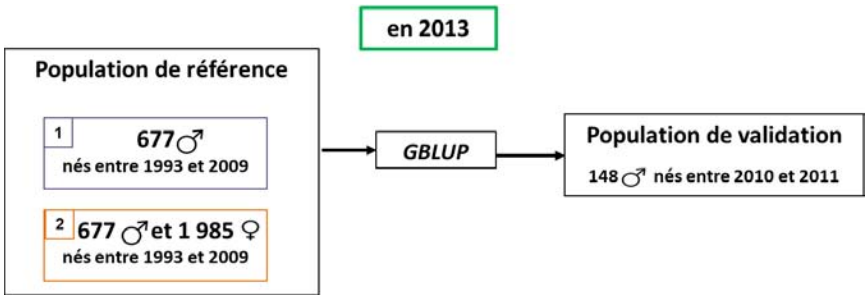


Figure 3.31 : Schéma des populations de référence et de candidats utilisés pour le calcul des précisions des GEBV des candidats dans le cas des évaluations génomiques two steps basées sur des DREBV

La précision des prédictions obtenue pour les mâles de validation (Figure 3 .32) avec les DREBV corrigées ou non n'est pas différente de celles obtenues avec les DYD pour les caractères laitiers excepté pour le TB. En effet, la corrélation entre GEBV et DYD pour les mâles de validation sur le taux butyreux est nettement meilleure (+ 51%) en utilisant les DYD plutôt que les DREBV corrigées ou pas. Pour les LSCS et les caractères de morphologie, les résultats obtenus avec les DREBV corrigées ou non sont en général légèrement meilleurs que ceux obtenus avec les DYD, excepté pour les LSCS et PRM. Les résultats obtenus à l'aide des DREBV corrigés sont similaires à ceux obtenus avec les DREBV non corrigées excepté pour l'AAR et l'AVP dont les résultats sont moins bons avec la correction. L'utilisation des DREBV corrigés ou non, n'améliore pas de manière conséquente les corrélations de validation. Ces résultats sont conformes à la faible amélioration observée de la précision (+4% en moyenne) avec l'utilisation des index dé-régressés comparée à celle des DYD, sur des données simulées à partir des schémas de sélection en bovins laitiers (Guo et al., 2010). En revanche, sur données réelles, (Gredler et al., 2010) ont montré une très légère

amélioration des prédictions avec l'utilisation des index dé-régressés par rapport aux DYD, pour des populations de bovins Fleckvieh.

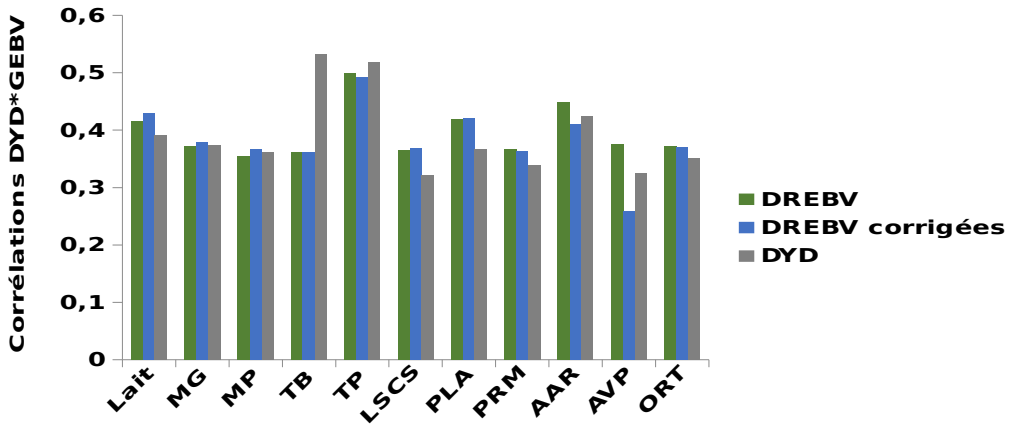


Figure 3.32 : Corrélations des DYD et des GEBV obtenus pour les 252 mâles de validation à partir des phénotypes DYD, DREBV ou DREBV corrigées des 425 mâles d'apprentissage

L'utilisation des DREBV améliore les pentes de régression des GEBV en fonction des DYD pour les caractères de production laitière et l'AVP, entre 3% sur le TP et 29% sur la MG (cf. Figure 3 .33). Les pentes de régression des GEBV sont peu affectées par le type de phénotypes utilisés pour les autres caractères étudiés (LSCS et caractères de morphologie) excepté pour l'AAR pour lequel les pentes sont plus proches de 1 avec les DREBV qu'avec les DREBV corrigées. Les différences de pentes de régression obtenues avec les DREBV corrigées ou non sont très faibles, entre moins de 1% pour le TB et 14% pour l'AAR. La correction des DREBV n'entraine pas toujours une amélioration de la pente puisque celle-ci est dégradée pour la MP, le profil de la mamelle (PRM). Il existe très peu de publications comparant les résultats d'évaluation génomiques avec différents types de phénotypes et aucune ne traite des pentes de régression.

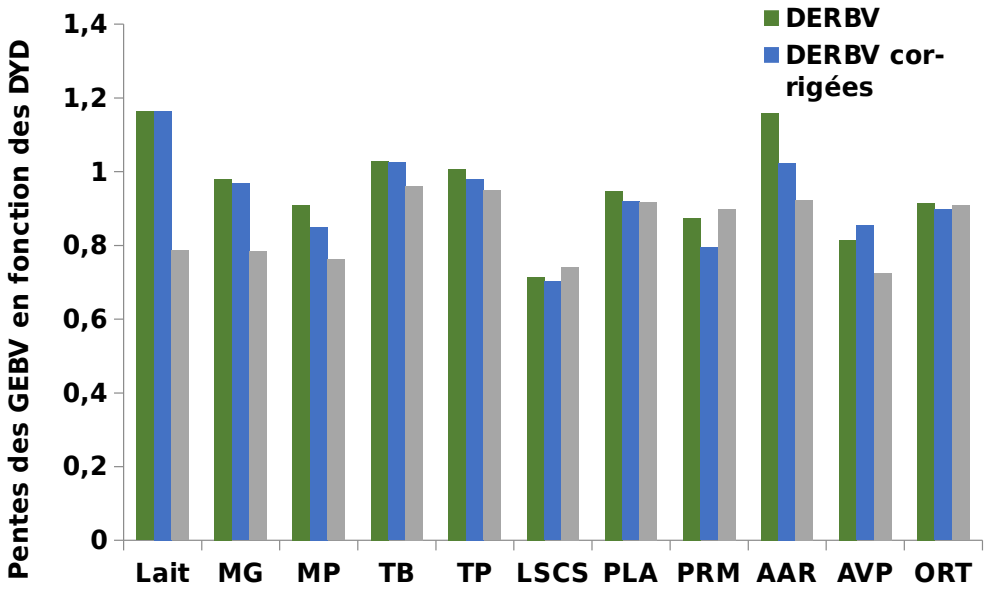


Figure 3.33 : Pentes de régression des GEBV en fonction des DYD obtenues pour les mâles de validation à partir des phénotypes DYD, DREBV ou DREBV corrigés des 677 mâles

Comme dans le cas de l'utilisation des DYD comme phénotype, l'utilisation des index dérégressés ne permet pas d'obtenir des précisions génomiques théoriques ($\sqrt{CD}$) des candidats supérieures au niveau de précision sur ascendance obtenu en évaluation classique. Dans le cas d'une population de référence ne contenant que des mâles (cf. Figure 3 .34 à gauche), l'utilisation des DREBV permet d'augmenter la moyenne des précisions théoriques des jeunes mâles candidats, entre 2% pour les taux butyreux et protéique, et 34% pour l'AAR. Le fait d'ajouter les femelles dans la population de référence en utilisant les DREBV non corrigées comme phénotypes améliore la précision théorique des GEBV des candidats pour tous les caractères sauf pour LSCS et PLA. L'amélioration des précisions génomiques théoriques des jeunes mâles avec l'ajout de femelles lorsque les DREBV corrigées sont utilisées comme phénotype n'est observée que pour la moitié des caractères : le lait, les matières, les LSCS et la distance plancher-jarret. Ces résultats peuvent s'expliquer par le poids choisi pour les DREBV corrigées. En effet dans cette étude le poids est celui développé par Garrick et al. (2009) adapté aux individus ayant un grand nombre de descendants. Il n'est donc pas approprié aux femelles qui en général n'ont pas un grand nombre de descendants. Ces mêmes auteurs ont également développé un poids spécifique pour des données répétées comme les lactations. Dans un souci d'homogénéité, nous avons choisi d'utiliser malgré tout, les mêmes poids pour les phénotypes mâles et femelles.

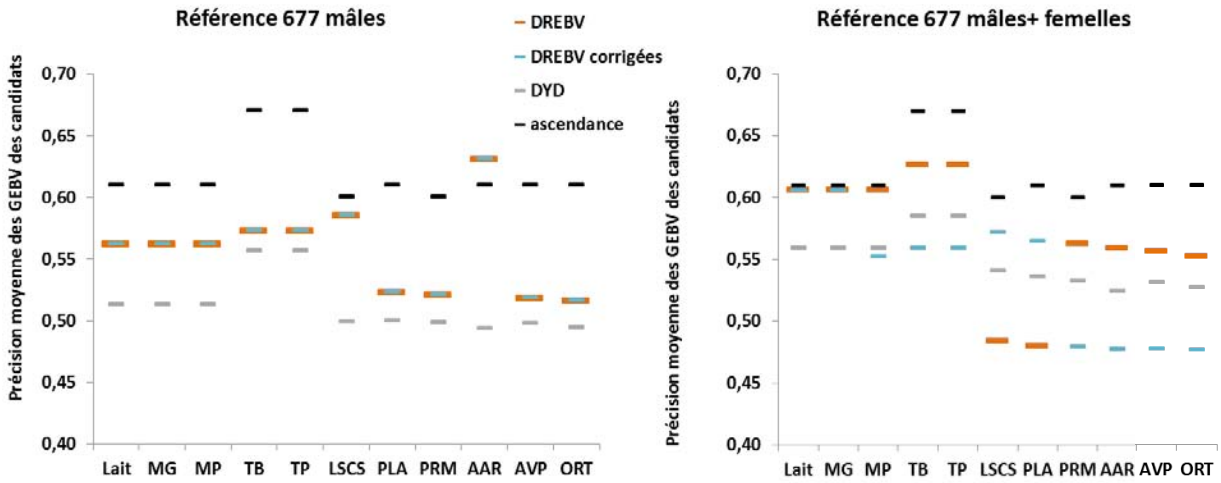


Figure 3.34 : Moyenne des précisions génomiques théoriques des 148 mâles candidats estimés dans le cas d'une population de référence contenant uniquement des mâles à gauche et des mâles et des femelles à droite pour les différents types de phénotypes utilisés

En conclusion, l'utilisation d'index dé-régressés ne permet pas d'améliorer la qualité des prédictions et n'améliore que très peu les précisions théoriques des GEBV des mâles candidats. La méthode en deux étapes incluant uniquement les phénotypes des individus génotypés permet d'obtenir une qualité de prédiction des EBV limitée ainsi que des précisions des GEBV des candidats inférieures à celles estimées sur ascendance.

3.I.3 Evaluation génomique multiraciale et uniraciale basée sur l'approche pseudo single-step

Les résultats obtenus en utilisant l'approche en deux étapes peuvent être améliorés par l'ajout dans le modèle de l'information phénotypique des autres mâles connus sur descendance (mâles de testage). En effet, les DYD des mâles génotypés ne prennent en compte que les performances de leurs filles, or l'information des femelles issues d'autres pères n'est pas considérée bien qu'elles puissent être apparentées aux mâles et donc que leur information puisse améliorer les prédictions des GEBV des candidats. Cette approche appelée pseudo single step (Baloche et al., 2013) a été appliquée à notre population en utilisant comme phénotypes les DYD de tous les mâles. Comme précédemment, une première analyse a consisté à réaliser une validation croisée afin d'étudier la qualité de prédiction. Pour cette validation croisée (schéma en Figure 3 .35), basée sur les phénotypes utilisés pour l'indexation de janvier 2008, 54 412 DYD de mâles nés entre 1980 et 2005 ayant plus de 10 filles ont été ajouté aux 425 DYD des mâles génotypés utilisés dans le paragraphe 3.I.1.A. Parmi les phénotypes additionnels, 20 522 sont des DYD d'individus de race Saanen, 31 024 de race Alpine et 2 866 d'animaux croisés. Pour les estimations des précisions des GEBV des

jeunes mâles candidats, basées sur les données de l'évaluation de janvier 2013, les DYD des mâles nés entre 2006 et 2011 sont pris en compte (soit 2 885 DYD, dont 55% sont de race Alpine, 42% de race Saanen et 3% d'animaux croisés). La méthode utilisée pour l'estimation des valeurs génétiques/génomiques des animaux est la même que celle utilisée dans les études précédentes, le GBLUP. Le modèle utilisé est le même que celui décrit dans l'article I mais la matrice de parenté entre individus est différente puisqu'elle relie tous les mâles phénotypés (génotypés ou non) entre eux. Le modèle utilisé ici est le suivant : $y = \mathbf{X\beta + Zu} + e$ où **y** est le vecteur des observations contenant les DYD de tous les boucs testés sur descendance, **β** le vecteur de l'effet fixe race, **X** la matrice d'incidence reliant les observations aux effets fixes, **u** le vecteur des valeurs génétiques/génomiques, **Z** la matrice d'incidence reliant les observations aux individus avec $\mathrm{Var}(u) = H\sigma_u^2$ où **H** est défini en 1.III.1.1.1 selon la méthode de Christensen et al. (2012), et **e** le vecteur des erreurs.

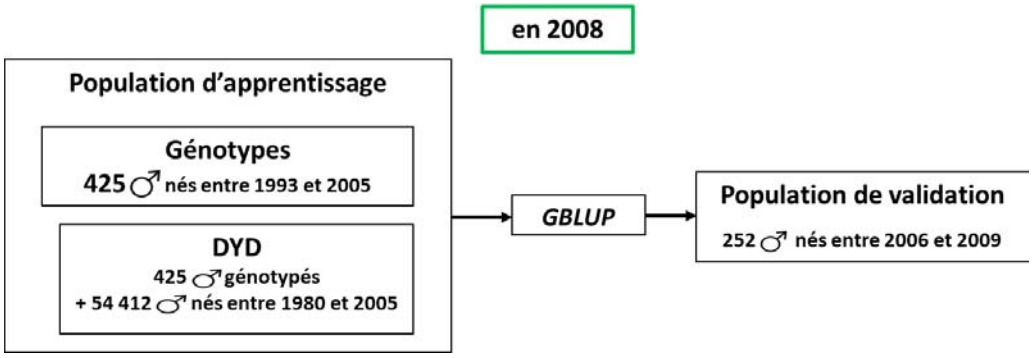


Figure 3.35 : Schéma de la validation croisée réalisée dans le cas des évaluations génomiques pseudo single-step multiraciales

La Figure 3 .36 présente les corrélations de validation entre les EBV ou GEBV et les DYD des 252 mâles de validation pour les trois modèles étudiés : 1) le pseudo single step pédigrée incluant les données phénotypiques (DYD) des animaux génotypés ou non et le pédigrée (en bleu), 2) le pseudo single step génomique incluant en plus par rapport au 1) les données de génotypages des 677 mâles génotypés et phénotypés (en orange), et 3) le two steps génomique qui ne considère que les DYD des mâles génotypés comme phénotypes (en gris). L'ajout de génotypes des 677 mâles (pseudo single step génomique vs pseudo single step pédigrée) ne permet pas d'améliorer de manière significative les corrélations de validation pour la majorité des caractères. Les corrélations de validation obtenues avec la méthode pseudo single step sont faiblement dégradées avec l'ajout de l'information génomique pour 5 caractères : le TB (-10%), la MP (-6%), la quantité de lait (-5%), l'ORT (-5%) et les LSCS (-3%). Elles sont en revanche légèrement améliorées (entre 2 et 6%) dans le cas des caractères de morphologie excepté l'ORT. Ces résultats sont différents de ceux

obtenus avec la méthode deux étapes (article I) où l'ajout des données de génotypes permet d'améliorer les corrélations entre 3% et 21% pour tous les caractères considérés. Les gains de corrélations, lorsqu'ils sont observés ici, sont en général plus faibles que ceux observés en deux étapes.

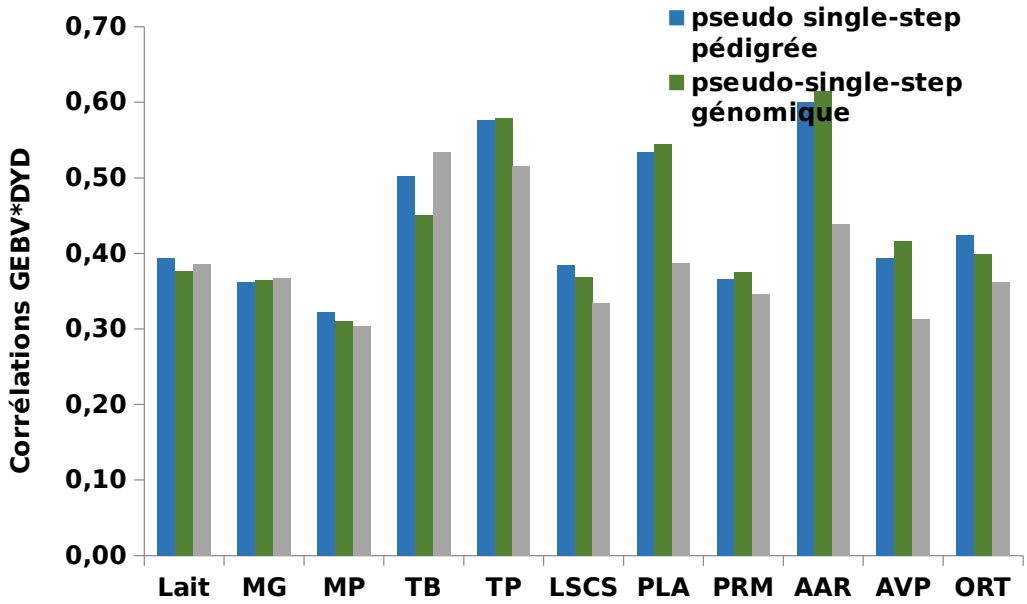


Figure 3.36 : Corrélations entre GEBV et DYD pour les 252 mâles de validation obtenues dans le cas du pseudo single-step sur données phénotypiques seulement (pédigrée), sur données phénotypiques et génomiques (génomique) et dans le cas du two-steps génomique (cf. article I)

Cependant, les corrélations de validation obtenues avec l'approche pseudo single step génomique sont plus élevées, de 3% pour la MP à 47% pour l'AAR, par rapport à celles de l'approche two-steps pour la majorité des caractères étudiés. L'ajout des phénotypes des mâles de testage ne permet pas d'améliorer les précisions de validation pour les caractères de production laitière sauf pour la MP (3%) et le TP (14%). Pour ces caractères, les phénotypes des femelles sont enregistrés depuis plus longtemps que pour les caractères de morphologie mammaire et les LSCS. De plus, les phénotypes sont répétés dans le cas des caractères de production ce qui n'est pas le cas des caractères de morphologie mammaire. L'information contenue dans les DYD des mâles d'apprentissage génotypés pour les caractères laitiers est donc plus précise ce qui peut expliquer le fait qu'ajouter les mâles de testage, très anciens par rapport aux mâles de validation, ne permet pas d'augmenter les précisions pour ces caractères, contrairement à ce qui est observé sur les caractères de morphologie mammaire et les LSCS.

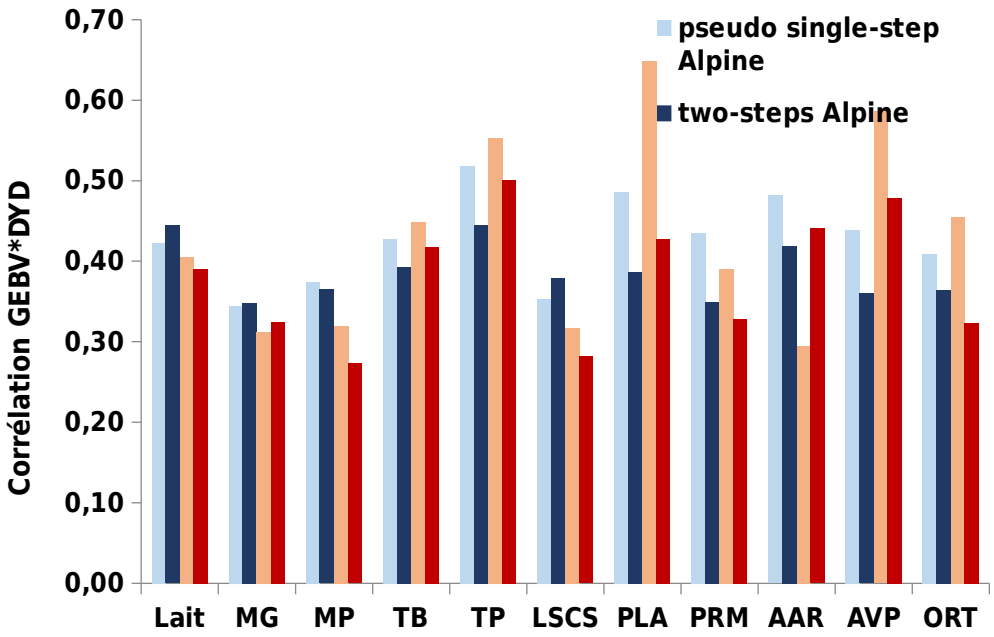


Figure 3.37 : Corrélations entre GEBV et DYD pour les 101 mâles Saanen et les 151 mâles Alpins de validation, obtenues dans le cas du pseudo single-step et le cas du two steps uniracial génomiques (évaluation Alpine ou Saanen)

Les corrélations entre GEBV et DYD obtenues pour les mâles de validation dans le cas des évaluations génomiques pseudo single step uniraciales (Saanen ou Alpine) sont présentées à la Figure 3 .37. Comme dans le cas des évaluations multiraciales, le pseudo single step uniracial permet d'obtenir des corrélations de validation plus élevées par rapport au two-steps uniracial pour la majorité des caractères : MP, TP, PLA, PRM, AVP et ORT. En revanche, contrairement à ce qui était obtenu dans le cas des évaluations multiraciales (Figure 3 .36), les corrélations de validation sur le TB obtenues avec les évaluations uniraciales pseudo single step sont supérieures à celles obtenues en two steps en race Alpine comme en Saanen. L'ajout des phénotypes des mâles de testage donne des résultats différents dans les deux races pour les caractères quantité de lait, comptage de cellules somatiques et qualité de l'attache arrière. Le pseudo single step est avantageux en race Saanen pour le lait et les LSCS, c'est le contraire pour le caractère AAR.

Les résultats trouvés dans cette étude sont différents de ceux observés en race ovine Lacaune (Baloche et al., 2013; Duchemin et al., 2012) pour laquelle ajouter des phénotypes d'animaux non génotypés permet d'augmenter les précisions quel que soit le caractère considéré de 3% pour la quantité de lait à 34% pour les LSCS. En ovins Lacaune, ajouter les phénotypes des mâles d'IA non génotypés dans le modèle apporte plus d'information que dans notre étude. En effet, contrairement aux caprins, les mâles nés avant 1998 n'étant pas génotypés, l'ajout de ces DYD permet de capter l'information de la voie femelle pour les

grands-pères maternels les plus anciens. De plus, l'ajout des DYD des collatéraux permet d'augmenter la précision des DYD des mâles d'apprentissage moins élevés en ovins Lacaune que dans notre étude. De même en bovins laitiers Holstein (Gao et al., 2012), ajouter 9 374 phénotypes de taureaux non génotypés permet d'augmenter les précisions de validation de 1% pour la MG à 30% pour la longévité. Ce n'est cependant pas le cas pour les caractères de conformation de la mamelle et d'aplomb avec une dégradation des corrélations de 2 à 5%.

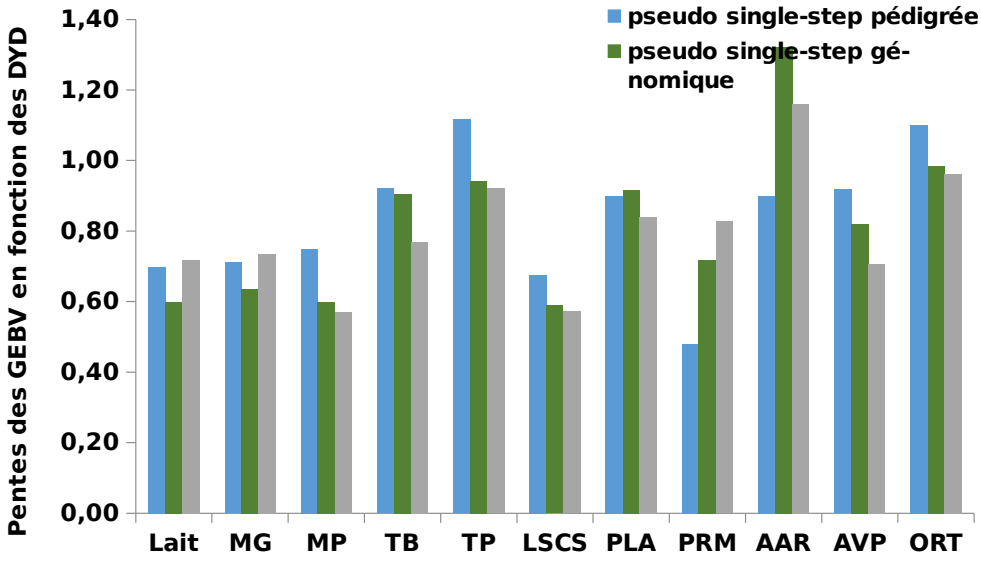


Figure 3.38 : Pentes de régression des GEBV en fonction des DYD pour les mâles de validation avec l'approche two steps incluant l'information génomique et avec l'approche pseudo single step incluant ou non l'information génomique

Les pentes des droites de régressions des GEBV en fonction des DYD estimées pour les 252 boucs de validation dans le cas du two steps génomique et de l'approche pseudo single step pédigrée et génomique sont présentées à la Figure 3 .38. Ces pentes sont estimées entre 0,48 pour le profil de la mamelle avec le pseudo single-step pédigrée et 1,32 pour la qualité de l'attache arrière avec le pseudo single-step génomique. Ces pentes sont améliorées par l'ajout de phénotypes mâles (pseudo single-step génomique vs two-steps génomique), c'est-à-dire que les GEBV prédites sont moins sous-estimées et sous-dispersées pour la moitié des caractères considérés (MP, TB, TP, LSCS, PLA, AVP et ORT). En revanche, l'ajout d'information génomique dans les évaluations pseudo single-step (génomique vs pédigrée) dégrade les pentes pour presque tous les caractères sauf pour l'orientation des trayons, le profil de la mamelle et la distance plancher-jarret. Ces résultats sont similaires à ceux observés en ovins Lacaune (Baloche et al., 2013; Duchemin et al., 2012) pour lesquels, les pentes estimées en pseudo single step sont moins bonnes que celles estimées en two steps pour deux caractères (lait et LSCS) sur trois. En revanche Gao et al. (2012) ont montré qu'en

bovins laitiers les pentes étaient améliorées par l'ajout des phénotypes de mâles non génotypés.

Dans cette étude, les pentes obtenues avec l'approche pseudo single step sont détériorées lorsque l'information génomique est prise en compte sauf pour le PLA et le TB. Ces résultats sont contraires à ceux obtenus en ovins Lacaune (Baloche et al., 2013) pour lesquels les pentes sont améliorées avec l'apport de l'information génomique sauf pour le caractère LSCS.

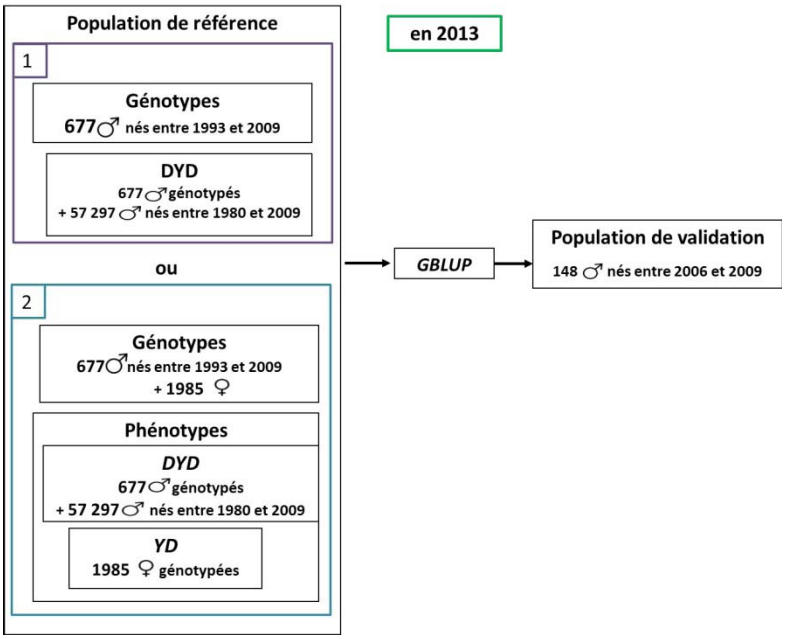


Figure 3.39 : Schéma des populations de référence et de candidats utilisées dans le cas des évaluations génomiques pseudo single step

Les précisions génomiques théoriques ($\sqrt{CD}$) des 148 jeunes mâles candidats ont été calculées à partir des variances d'erreur de prédiction estimées avec la méthode pseudo single step génomique dans deux cas différents. Dans le premier cas, la population de référence contient les 677 mâles génotypés et phénotypés ainsi que les mâles phénotypés mais non génotypés. Dans le deuxième cas, elle contient les mâles du premier cas ainsi que les 1 985 femelles génotypées utilisées dans l'article I (Figure 3 .39). Les précisions théoriques des GEBV estimées avec l'approche pseudo single step sont meilleures que celles obtenues avec l'approche two steps.

Pour l'ensemble des caractères étudiés, l'ajout de l'information génomique et phénotypique des femelles permet d'augmenter les précisions théoriques des GEBV des candidats de 0,1% pour la distance plancher-jarret (PLA) à 32% pour la qualité de l'attache arrière (AAR) (Figure 3 .40). Les gains observés avec l'ajout de femelles sont plus importants que ceux estimés avec le two steps génomique (Carillier et al., 2013), qui étaient

au maximum de 8% pour la quantité de lait. De plus, dans le cas des caractères de production laitière, l'ajout des génotypes des femelles permet aux précisions des GEBV des candidats de dépasser la précision sur ascendance. L'augmentation de cette précision, avec l'ajout des femelles, est donc plus forte que dans le cas du two steps.

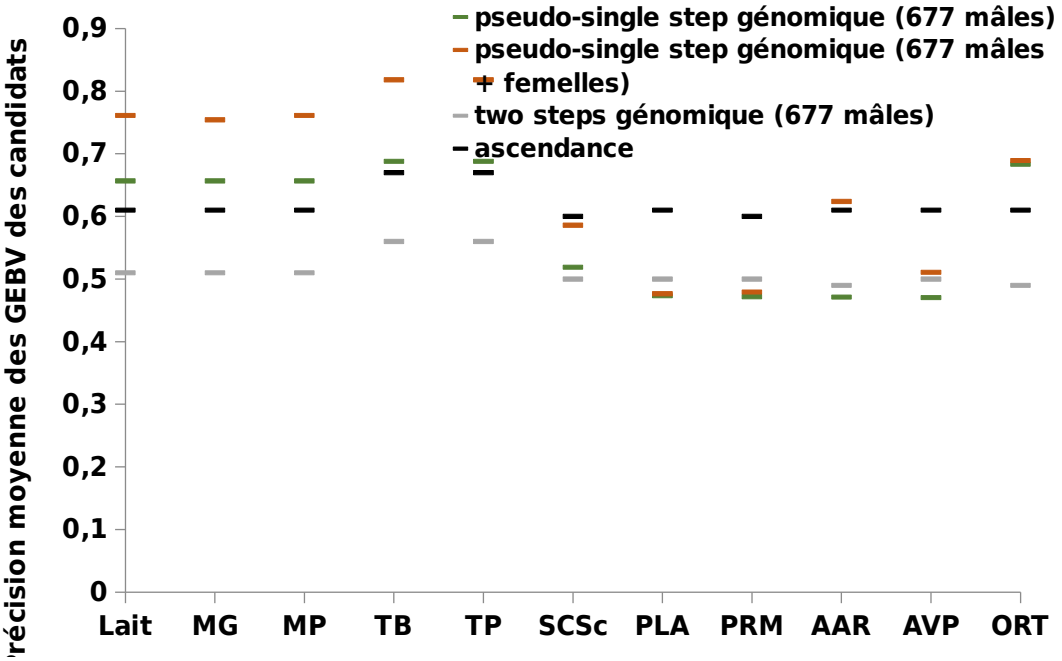


Figure 3.40 : Précisions théoriques moyennes des GEBV estimées pour les 148 jeunes candidats avec la méthode en 2 étapes ou la méthode pseudo single step*
*** pseudo single step à partir d'une population de référence de 677 mâles (« pseudo single step 677 mâles ») ou de 677 mâles et 1 985 femelles (« pseudo single step mâles+femelles »)**

En conclusion de cette étude, l'ajout de phénotypes de mâles de testage (pseudo single step vs two steps) ne permet pas d'augmenter les précisions des GEBV de façon significative pour l'ensemble des caractères en uniracial comme en multiracial.

3.I Evaluation génomique par des méthodes Bayésiennes

Dans cette étude, des méthodes de prédiction bayésiennes ont été testées afin d'analyser la qualité des prédictions par validation croisée lorsque seulement une proportion de SNP est considérée dans le modèle. Le schéma de validation croisée utilisé dans cet étude est le même que celui appliqué dans l'article I (Figure 3 .25). A l'aide du logiciel GS3 (Legarra, Ricard et Filangi, programme INRA), les méthodes GBLUP, Lasso Bayesien (BL) et Bayes Cπ (cf. 1.III.1) ont été utilisées sur les mêmes données de l'article I (two steps multiracial) en utilisant les DYD comme phénotypes. Dans l'approche Bayes Cπ, plusieurs valeurs de π ont été testées : 0,1%, 5%, 10%, 30% ainsi qu'une proportion estimée par le logiciel (BCPest). Le

modèle s'écrit sous la forme suivante : $y = \mathbf{Xb} + \mathbf{Za} + \mathbf{Tg} + e$ avec **y** le vecteur des 425 DYD des mâles d'apprentissage, **X** la matrice d'incidence reliant les phénotypes aux effets fixes (**b**), **Z** la matrice d'incidence reliant les phénotypes aux valeurs polygéniques aléatoires (**a**), telles que **a** ~ $N(0, A\sigma_a^2)$, **T** la matrice reliant les phénotypes aux effets des marqueurs (**g**) et **e** le vecteur des résidus. La distribution des effets des marqueurs dépend de la méthode utilisée. Dans le cas du GBLUP, la variance de l'effet d'un SNP σ_g^2 peut s'écrire en fonction de la variance expliquée par l'ensemble des SNP (σ_u^2) telle que : $\sigma_g^2 = \frac{\sigma_u^2}{2\sum_i p_i q_i}$. Pour la méthode BayesCπ, la loi des effets des marqueurs est :

$$\begin{cases} p(g|\pi,\sigma_g^2) = 0 & \text{avec une probabilité } 1-\pi \\ p(g|\pi,\sigma_g^2) & N(0,\sigma_g^2) \text{ avec une probabilité } \pi \end{cases}$$

. Dans le cas du Lasso Bayésien, la loi *a priori* de l'effet d'un marqueur est : $p(g_i,\lambda) = \frac{\lambda}{2}\exp(-\lambda|g_i|)$ où λ est un paramètre d'échelle permettant de définir l'intensité de sélection des SNP dont la valeur initiale est fixée à $\frac{2}{\sigma_g^2}$.

La Figure 3 .41 présente les corrélations entre GEBV et DYD pour les 252 mâles de validation obtenues à partir d'évaluations génomiques multiraciales selon les différents modèles : GBLUP, Bayesian Lasso (BL) et Bayes Cπ (BCP). La valeur de π utilisée pour les évaluations génomiques avec la méthode Bayes Cπ a un impact non négligeable sur les corrélations de validation obtenues. Les différences dépendent fortement du caractère considéré. Cette conclusion a également été observée sur des données bovines Holstein Danoises (Su et al., 2010) avec différentes valeurs de π testées (5, 10, 20 et 50%). Les différences observées entre corrélations de validation en fonction de la valeur de π utilisée se situent entre 16% pour les LSCS (entre le BCP 5% et le BCP 0,1%) et 80% pour le TP (entre le BCP 0,1% et le BCP 30%). Le modèle BCP 0,1% permet d'obtenir les corrélations de validation les plus élevées pour la majorité des caractères excepté la MG, le TB, le profil de la mamelle (PRM), les LSCS pour lesquels la valeur de π a peu d'effet et la distance plancher-jarret (PLA) et pour laquelle les corrélations de validation les plus élevées sont obtenues avec

π = 5 et 10%. Ces résultats peuvent être expliqués par un nombre différent de QTL pour chacun des caractères étudiés (Maroteau et al., 2013). Le peu de différences entre les corrélations de validation obtenues en BayesC π avec différentes valeurs de π testés (5, 10, 20 et 50%) a également été faite sur des données bovines Holstein Danoises (Su et al., 2010).

On remarque également que les corrélations de validation les plus élevées ne sont pas obtenues avec le modèle BCPest où la proportion de SNP ayant un effet est estimée par le programme. Ce type de résultat a également été constaté chez les ovins laitiers Lacaune pour lesquels peu de différences étaient observées entre les résultats obtenus avec un π estimé par le modèle ou avec un π fixé à 10% (Duchemin et al., 2012). Les valeurs de π estimées pour chaque caractère sont présentées au Tableau 3 .13. Elles se situent entre 11% pour le score de cellules somatiques et 30% pour la qualité de l'attache arrière. Ces résultats sont difficiles à relier aux nombres très différent de QTL trouvés (au seuil 5% chromosome) pour chacun des caractères, de 2 pour les LSCS (Maroteau et al., 2013) à plusieurs dizaines pour les caractères de morphologie (Palhiere et al., 2012).

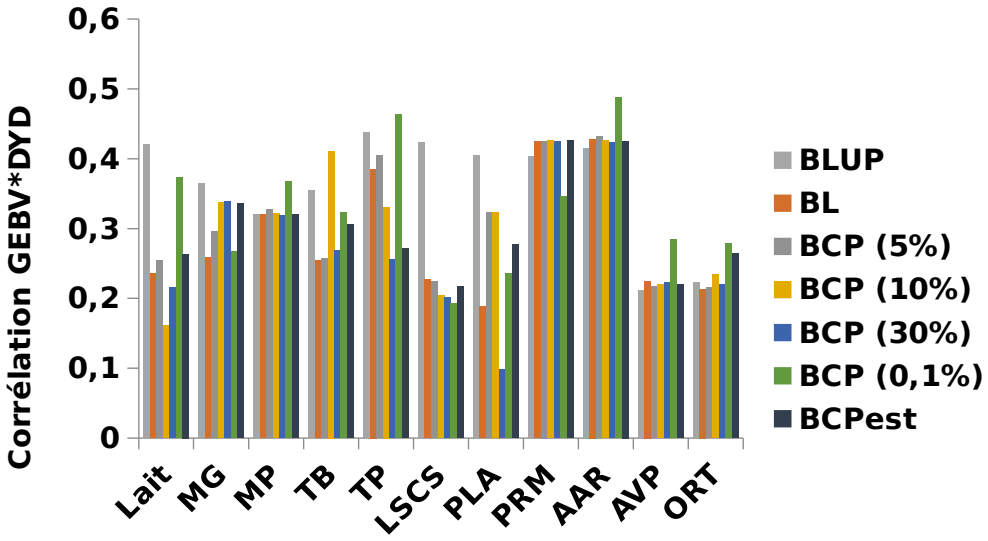


Figure 3.41 : Corrélations entre DYD et GEBV pour les 252 mâles de validation estimées avec différents modèles : GBLUP (estimé avec le logiciel GS3), Bayesian Lasso (BL), Bayes Cπ (BCP) pour différentes valeurs de π*

* différents niveaux de π sont utilisés : 0,1% ; 5% ; 10% ; 30% et estimé par le logiciel (BCPest)

Les corrélations de validation obtenues avec le modèle Bayesian Lasso (BL) sont similaires à celles obtenues avec le Bayes Cπ 5% sauf pour le caractère PLA pour lequel le BL donne des résultats intermédiaires entre le BCP 5% et le BCP 30%. De plus, les résultats obtenus avec le modèle BL sont toujours inférieurs à ceux obtenu avec le modèle BCP. Dans cette étude, les méthodes Bayésiennes donnent de moins bons résultats que la méthode

fréquentiste (GBLUP) pour la moitié des caractères considérés. Pour les autres caractères (MP, TP, AAR, AVP et ORT), la méthode Bayes Cπ avec un π fixé à 0,1% surpasse modestement la méthode GBLUP de 2% pour le TP à 20% pour l'AVP. En revanche dans le cas du caractère TP, c'est le Bayes Cπ avec π fixé à 10% qui permet d'obtenir les corrélations de validation les plus élevées. Ces résultats semblent différents de ceux trouvés dans la littérature. En effet, en ovins Lacaune les corrélations de validation obtenues avec la méthode GBLUP sont similaires à celles obtenues avec la méthode Bayes Cπ (Duchemin et al., 2012). De même en bovins laitiers Holstein (Jiménez-Montero et al., 2013) et Montbéliarde (Colombani et al., 2013), le Lasso Bayésien fournit les mêmes résultats que le Bayes Cπ et le GBLUP pour la plupart des caractères. En bovins laitiers Simmental (Hozé et al., 2014b), les précisions de validation moyennes obtenues sur l'ensemble des caractères avec le Bayes Cπ sont similaires à celles obtenues avec le GBLUP. Elles sont cependant légèrement meilleures que celles obtenues avec le GBLUP pour la quantité de lait et la quantité de matière grasse.

Tableau 3.13 : Proportion de SNP ayant un effet sur les caractères (π), estimée à l'aide du programme GS3 avec la méthode Bayes Cπ estimée

	Lait	MG	MP	TB	TP	LSCS	PLA	PRM	AAR	AVP	ORT
π (%)	21	24	17	13	14	11	14	15	30	15	19

La Figure 3 .42 présente les pentes de régression des GEBV en fonction des DYD obtenues pour les 252 mâles de validation avec les différents modèles testés (GBLUP, Bayes Cπ et Lasso Bayésien). Ces pentes sont estimées entre 0,10 pour le caractère PLA avec la méthode Bayes Cπ (π = 30%) et 1,35 pour les caractères PRM et AAR avec les méthodes Bayésiennes (BL et BCP). Pour la majorité des caractères les pentes estimées avec l'approche GBLUP sont plus proches de 1 que celles estimées avec les méthodes Bayésiennes. Au sein des méthodes Bayésiennes, les pentes les plus proches de 1 sont obtenues avec le Bayes Cπ, le paramètre π étant fixé à 0,1% pour la grande majorité des caractères excepté le TP et la distance plancher-jarret (PLA). Ces résultats vont dans le même sens que ceux obtenus pour les corrélations de validation. Dans la littérature, les pentes de régression obtenues avec les méthodes Bayésiennes sont soit meilleures que celles obtenues avec le GBLUP (de 8% pour le TP à 17% pour la quantité de lait avec le Bayes Cπ, π étant estimé ou fixé à 10% en ovins laitiers) (Duchemin et al., 2012), soit similaires (en bovins laitiers (Colombani et al., 2013; Jiménez-Montero et al., 2013)).

Dans notre étude, le Lasso Bayésien et le Bayes Cπ ne permettent d'améliorer pour tous les caractères ni les corrélations de validation, ni les pentes de régression des GEBV en fonction des DYD pour les 252 mâles de validation. Parmi les approches Bayésiennes, les meilleurs résultats sont cependant observés avec un Bayes Cπ où π est fixé à 0.1%. Ce qui signifie qu'une très faible proportion de SNP à fort effet peut permettre d'améliorer les modèles d'évaluation génomique. Il pourrait donc être envisagé par la suite de tester d'autres modèles incluant ces SNP avec des poids plus forts associés. Les faibles différences observées entre les méthodes Bayésiennes et la méthode GBLUP sont également observées dans la littérature. Cependant, il n'est pas fréquent d'observer une dégradation des corrélations avec l'utilisation de méthodes Bayésiennes. Les diminutions des corrélations observées dans cette étude avec l'utilisation des méthodes Bayésiennes peuvent être expliquées par l'aspect multiracial des évaluations. En effet, il a été montré (Palhiere et al., 2012; Maroteau et al., 2013) que les QTL détectés dans l'espèce caprine n'étaient pas toujours communs aux deux races considérées (Alpine et Saanen). Les méthodes Bayésiennes sélectionnant les SNP ayant un effet sur le caractère, les évaluations multiraciales sélectionneront les SNP ayant un effet sur au moins une des races. Ce phénomène peut alors engendrer des erreurs de prédiction pour la race pour laquelle le SNP n'a pas d'effet sur le caractère.

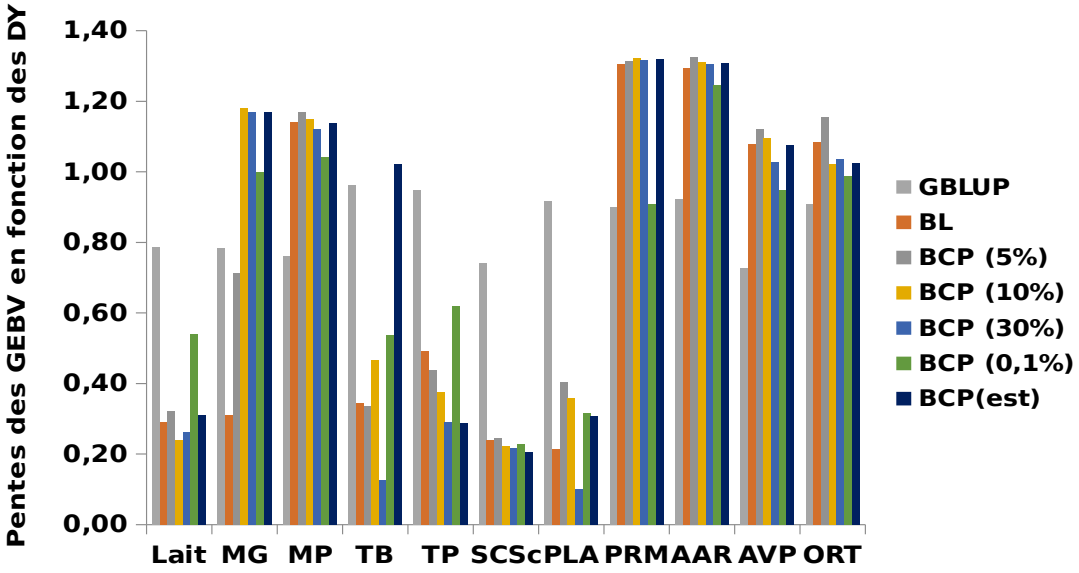


Figure 3.42 : Pentes de régression des GEBV en fonction des DYD pour les mâles de validation obtenues avec les différents modèles testés : GBLUP (estimé avec le logiciel GS3), Lasso Bayesien (BL), Bayes Cπ (BCP)

En conclusion de ce troisième chapitre, la méthode GBLUP semble être la plus appropriée dans le cas de la population caprine française. Le choix des phénotypes (DYD des

animaux génotypés, DYD de l'ensemble des animaux génotypés ou non, EBV dérégressés pondérés ou non) dans l'approche en deux étapes (two steps) n'a pas d'impact sur les précisions obtenues. Les corrélations de validation obtenues restent toutefois inférieures à celles observées dans la littérature en ovins ou bovins laitiers ce qui peut être expliqué en partie par la petite taille de la population de référence caprine. Les précisions théoriques des GEBV des jeunes mâles, nés en 2010 et 2011et n'ayant pas de performance, restent inférieures aux précisions sur ascendance. L'évaluation génomique dans ce cas ne permet donc pas d'augmenter les précisions théoriques. D'autres méthodes comme les méthodes dites single step basées sur les performances brutes des femelles doivent être envisagées pour tenter de pallier ce problème. Enfin les résultats obtenus avec les évaluations génomiques two steps uniraciales sont différents de ceux obtenus avec les évaluations multiraciales. Dans le cas des évaluations uniraciales, la taille de la population de référence est divisée par deux, malgré tout, les corrélations de validation ne sont pas si mauvaises et peuvent même être supérieures à celles obtenues avec les évaluations multiraciales. Il est cependant difficile de conclure à l'avantage de l'une ou l'autre des méthodes. Ces deux types d'évaluations seront donc également testés dans une approche single step dans le chapitre suivant.

Chapitre 4 : Étude d'évaluations génomiques basées sur les performances brutes

4.I Comparaison des évaluations génomiques uniraciales et multiraciales

Introduction

Les résultats du chapitre 3, basés sur une méthode en deux étapes, montrent que même si nous obtenons des corrélations entre DYD et GEBV proches de celles obtenues en ovins laitiers Lacaune ou en bovins laitiers (à taille de population de référence similaire à nos données), les précisions des valeurs génomiques estimées pour les 148 jeunes boucs sont inférieures à celles estimées sur ascendance. L'utilisation d'une méthode en une seule étape (single step cf. 1.III.1.1) basée sur les performances brutes des animaux n'était pas envisageable jusqu'en 2009 (Misztal et al., 2009) pour des raisons liées à la complexité des calculs. Les précisions obtenues avec cette méthode surpassent celles obtenues avec la méthode two steps dans de nombreux articles de la littérature (Aguilar et al., 2010; Legarra et al., 2014b; Vitezica et al., 2010; Li et al., 2014; Gao et al., 2012). Le but de cette étude est donc, dans un premier temps, de mesurer l'intérêt d'une évaluation génomique single step en population caprine laitière française notamment pour la précision théorique des GEBV des jeunes candidats.

D'autre part, les analyses réalisées au chapitre 3 n'ont pas permis de conclure clairement sur le gain relatif d'une évaluation multiraciale comparée à une évaluation par race. Ces deux types d'évaluations (multiraciale et uniraciale) ainsi qu'une évaluation bicaractère Alpine-Saanen telle que proposée par (Karoui et al., 2012) ont été testées dans ce chapitre en utilisant la méthode single step. Les deux caractères considérés dans l'évaluation bicaractère étaient par exemple la quantité de lait en race Saanen et la quantité de lait en race Alpine. La corrélation génétique entre les mesures du même caractère dans chaque race a été estimée dans une sous population de 113 980 performances de première lactation d'environ 49 000 chèvres Saanen et 65 000 chèvres Alpine par l'algorithme Espérance-Maximisation du maximum de vraisemblance restreinte (EM-REML), en utilisant le logiciel remlf90 (Misztal et al., 2002) pour chaque caractère étudié dans le chapitre précédent. Trois valeurs différentes de corrélations entre les caractères Alpine et Saanen ont été considérées pour les évaluations génomiques bicaractères : (i) une corrélation estimée par EM-REML, (ii) une corrélation fixée à 0,99 pour être proche d'une évaluation multiraciale et (iii) une corrélation fixée à 0 pour être

proche d'une évaluation uniraciale. La méthode d'estimation des effets du modèle utilisée dans cette étude était le GBLUP implémentée dans le logiciel blup90iod (Misztal et al., 2002).

Comme dans l'article I (Chapitre 3), des corrélations de validation entre les GEBV estimées dans les différentes évaluations testées et les DYD disponibles en janvier 2013 ont été calculées pour les 252 mâles de validation nés entre 2006 et 2009. Les phénotypes en entrée du modèle pour calculer ces DYD sont les performances de l'indexation officielle de 2013, année pour laquelle aucun des mâles de validation n'a de filles. Les pentes de régression des GEBV en fonction des DYD pour les mâles de validation ont également été estimées afin d'évaluer la dispersion des valeurs génomiques estimées. D'autre part, les mêmes modèles ont été testés à partir des performances de l'indexation officielle de janvier 2013 afin de pouvoir estimer les précisions des GEBV des jeunes candidats nés en 2010 et 2011. Ces précisions ont été obtenues à l'aide du logiciel accf90 (Misztal et al., 2002) via une approximation de la variance d'erreur de prédiction. La moyenne de ces précisions pour l'ensemble des 148 jeunes mâles obtenues avec le modèle multiracial a été comparée à celles obtenues avec les autres modèles testés. Les estimations des corrélations de validation et des pentes de régression ont été réalisées pour l'ensemble des animaux (population multiraciale) puis séparément dans chacune des deux races (cf. Figure 4 .43). Il a été fait de même pour les précisions théoriques des GEBV (Figure 4 .44).

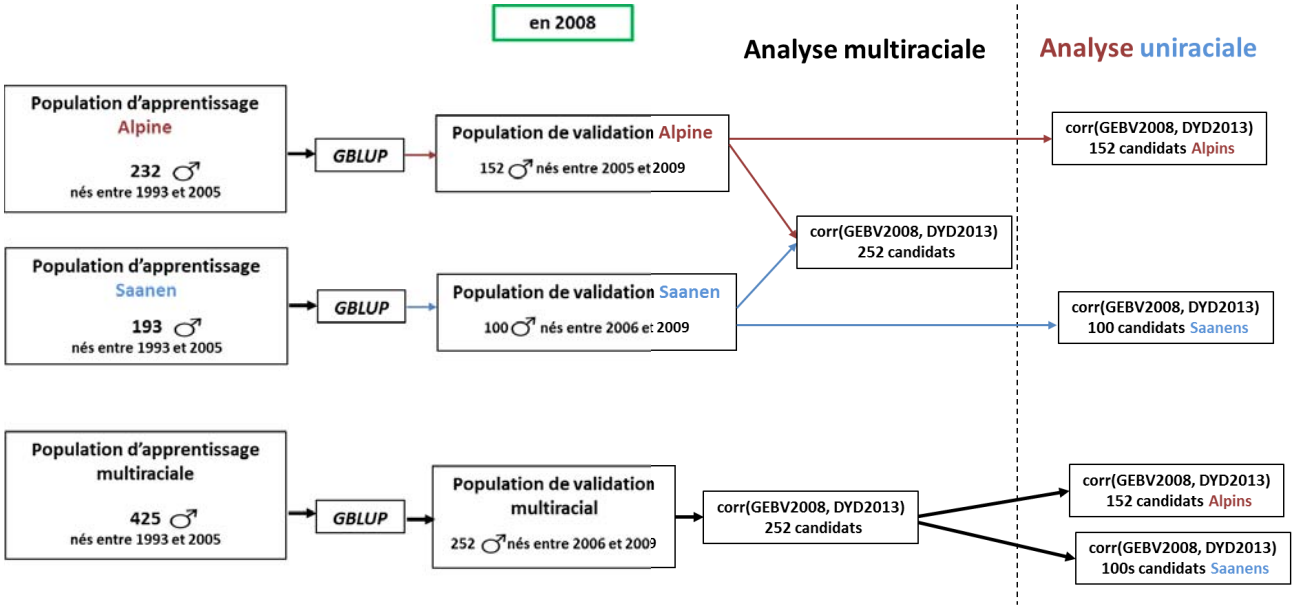


Figure 4.43 : Schéma des validations croisées réalisées dans l'article II

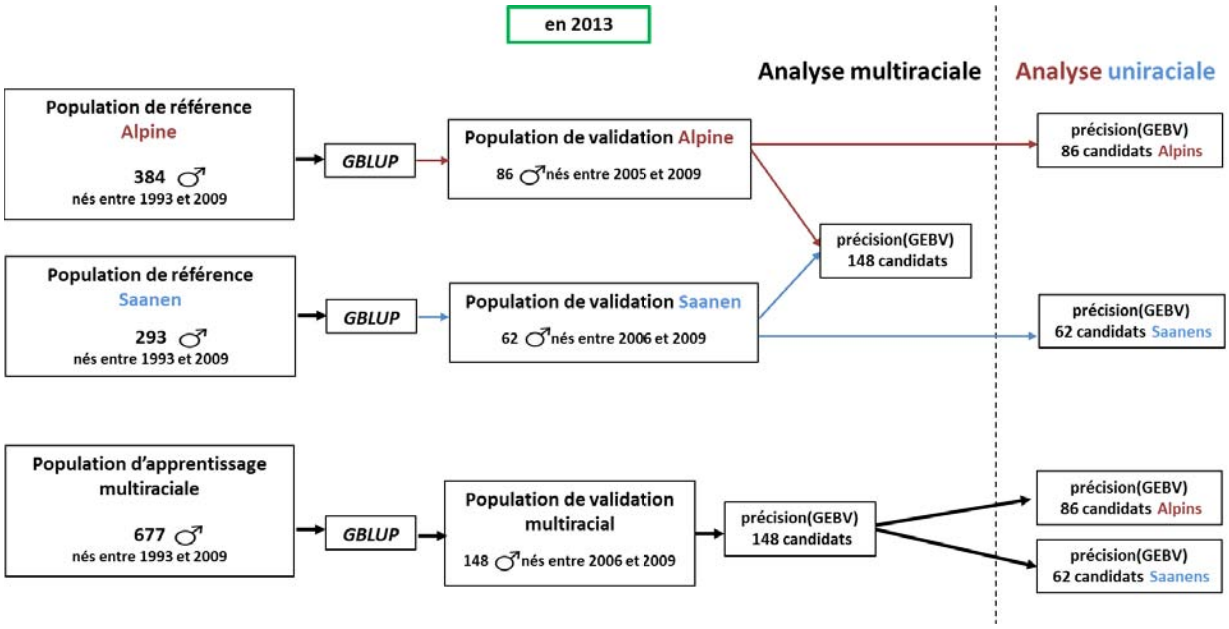


Figure 4.44 : :Schéma des populations de référence et de candidats utilisées dans l'article II

Article II : Comparaison des évaluations single-step multiraciales et uniraciales pour la population de caprins laitiers français

Carillier C, Larroque H, Robert-Granié C, 2014. Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. Genetics Selection Evolution, 46:67.

Carillier et al. Genetics Selection Evolution 2014, 46:67
http://www.gsejournal.org/content/46/1/67

GSE Genetics Selection Evolution

RESEARCH Open Access

Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population

Céline Carillier[1,2,3*], Hélène Larroque[1,2,3] and Christèle Robert-Granié[1,2,3]

Abstract

Background: All progeny-tested bucks from the two main French dairy goat breeds (Alpine and Saanen) were genotyped with the Illumina goat SNP50 BeadChip. The reference population consisted of 677 bucks and 148 selection candidates. With the two-step approach based on genomic best linear unbiased prediction (GBLUP), prediction accuracy of candidates did not outperform that of the parental average. We investigated a GBLUP method based on a single-step approach, with or without blending of the two breeds in the reference population.

Methods: Three models were used: (1) a multi-breed model, in which Alpine and Saanen breeds were considered as a single breed; (2) a within-breed model, with separate genomic evaluation per breed; and (3) a multiple-trait model, in which a trait in the Alpine was assumed to be correlated to the same trait in the Saanen breed, using three levels of between-breed genetic correlations (ρ): ρ = 0, ρ = 0.99, or estimated ρ. Quality of genomic predictions was assessed on progeny-tested bucks, by cross-validation of the Pearson correlation coefficients for validation accuracy and the regression coefficients of daughter yield deviations (DYD) on genomic breeding values (GEBV). Model-based estimates of average accuracy were calculated on the 148 candidates.

Results: The genetic correlations between Alpine and Saanen breeds were highest for udder type traits, ranging from 0.45 to 0.76. Pearson correlations with the single-step approach were higher than previously reported with a two-step approach. Correlations between GEBV and DYD were similar for the three models (within-breed, multi-breed and multiple traits). Regression coefficients of DYD on GEBV were greater with the within-breed model and multiple-trait model with ρ = 0.99 than with the other models. The single-step approach improved prediction accuracy of candidates from 22 to 37% for both breeds compared to the two-step method.

Conclusions: Using a single-step approach with GBLUP, prediction accuracy of candidates was greater than that based on parent average of official evaluations and accuracies obtained with a two-step approach. Except for regression coefficients of DYD on GEBV, there were no significant differences between the three models.

Background

With the recent availability of the Illumina goat SNP50 BeadChip [1], it has become possible to study the implementation of genomic selection in French dairy goat. In this species, as in dairy cattle, genomic selection has the ability to shorten the generation interval for the sire-son pathway (5.5 years with progeny-testing [2]) by selecting males shortly after birth. The utility of genomic selection is dictated by the accuracy of the genomic breeding values (GEBV) of the selection candidates, which has to be greater than the parent average accuracy for genomic ($\sqrt{\frac{1}{4}\text{ reliability sire} + \frac{1}{4}\text{ reliability dam}}$) selection to be effective.

However, although the French population of genotyped goats is the largest worldwide, it only counts around 900 males of the Alpine and Saanen breeds. Analysis of the genetic structure of this population [3], based on estimates of linkage disequilibrium and inbreeding and kinship coefficients, showed a high level of genetic diversity. The

* Correspondence: celine.carillier@toulouse.inra.fr
[1]INRA, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, 31326 Castanet-Tolosan, France
[2]Université de Toulouse INPT ENSAT, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, 31326 Castanet-Tolosan, France
Full list of author information is available at the end of the article

number of effective founders was estimated to be equal to 191 and 124 for the Alpine and Saanen breeds, respectively, based on females born between 2010 and 2011 [2]. Considering the relatively small sample size and the genetic structure of this population, the prediction accuracy of GEBV is not expected to be as high in these populations as in dairy cattle [4].

Within the French dairy goat breeding scheme, bucks used for breeding have more than 2000 daughters at the end of their productive life. Females used as buck dams have at least three lactations. This yields accurate parent estimated breeding values (EBV), and consequently fairly high parent average accuracies of young selection candidates (0.63 on average for milk yield) [3].

A first study on genomic prediction in dairy goats [3] used a two-step genomic approach. Step one consisted of deriving average daughter performance corrected for fixed and non-genetic random effects and for genetic effects of the dams (daughter yield deviations, DYD). Step two involved a genomic evaluation based on these DYD. These steps are especially essential for dairy traits on which males are selected and genotyped but for which they are not phenotyped [5]. This approach resulted in GEBV with accuracies as high as those of parent average EBV for young selection candidates that were not yet progeny-tested [3].

GEBV accuracies of candidates depend on the genetic characteristics of the reference population i.e. number of individuals, effective population size, linkage disequilibrium, inbreeding level, and relationship of candidates to the reference population and heritability of the phenotype [6-8]. The structure of the reference population could not be optimized by either genotyping other males because all progeny-tested males were genotyped or by choosing in the reference population the highly related males because of its size. In this context, GEBV accuracy could only be improved by using the most suitable model for genomic evaluation. Several studies [9-11] have shown that the single-step approach proposed by Legarra et al. [12] provides greater accuracy than the two-step approach used in [3]. The single-step approach allows all recorded phenotypes to be used, without pre-adjustment for fixed effects, and to evaluate all animals, regardless of whether they have been genotyped or phenotyped [5]. Another approach, the pseudo-single-step approach [11] can also deal with non-genotyped individuals by adding records of non-genotyped males in a two-step procedure. The pseudo-single-step approach is a two-step approach (i.e. based on DYD). However, it is an intermediate approach between the two-step and single-step approaches, because it includes information on all males with DYD, but ignores maternal information unlike the single-step approach [11]. Thus, it was expected that validation correlations using the pseudo-single-step approach would be intermediate to those obtained with the single-step and two-step approaches.

Our goat population was composed of two breeds that were evaluated together in the previous study because of their small population sizes [3]. These breeds are managed together in 10% of the French dairy goat herds. The white coat Saanen breed is a selected variety of the Alpine breed that originated several centuries ago [13]. Currently, selection in French dairy goats is done within-breed for the Alpine and Saanen breeds based on within-breed genetic evaluations, except for milk production traits, for which a multi-breed model is used. Alpine and Saanen breeds have similar heritability and repeatability parameters for milk production and type traits but differ in genetic and residual variances [14], persistence of linkage disequilibrium [3] and allele frequencies. These differences raise doubts on the benefits of multi-breed genomic evaluation for this population. Few studies dealing with multi-breed evaluation in other species have compared their results to within-breed models [15-17]. When several breeds are pooled in a single reference population, accuracies depend on the genetic characteristics and similarities between breeds. Recent research has not led to a consensus on the advantage of using multi-breed genomic evaluation. Very few studies have explored the use of relationships between breeds in multi-breed genomic evaluation [18,19] using, for instance, multiple-trait models.

Here, we tested a single-step approach using three models. The first model was a multi-breed model in which Alpine and Saanen populations were pooled together and considered as a single population (with only one set of genetic parameters). The second model was a multiple-trait model as described in Karoui et al. [18], where a trait recorded in the Alpine breed was considered to be different from, but correlated with, the same trait recorded in the Saanen breed. In this case, genetic parameters were specific to each breed. The third model was a within-breed model, where one model was used for the Saanen population and another for the Alpine population, with different genetic parameters for each breed.

Methods

Data

The SNP genotypes obtained using DNA samples extracted from blood were performed according to the French National Guidelines for the care and use of animals for research purposes. Animal genotypes used in this study were the same as in Carillier et al. [3]. After a quality check (MAF > 1%, call rate > 98%) that was done separately for the two breeds, 46 959 SNPs (out of 53 347 of the Illumina SNP50 BeadChip [1]) were validated. Missing SNP genotypes (0.1%) were not imputed but the GBLUP method used took missing data into account when estimating GEBV. From the 825 genotyped bucks

(355 Saanen and 470 Alpine individuals) born between 1993 and 2011, 148 (86 Alpine and 62 Saanen individuals born in 2010 and 2011) were not yet progeny-tested and could not be used for training.

Five milk production traits that were derived from a total lactation were analyzed, i.e. milk yield (kg), fat and protein yields (kg), fat and protein contents (g/kg), along with somatic cell score (SCS: log-transformed somatic cell counts), and five udder type traits that were scored on a linear scale of 1 to 9, i.e. udder floor position, udder shape, rear udder attachment, fore udder and teat angle. Udder type traits were recorded once for each female. Repeatability was not modeled for the udder type traits. Data originated from the official genetic evaluation of January 2013, using only records on purebred Alpine and Saanen goats. For milk production traits, 4 178 315 Alpine records (30.2% first lactations, 24.2% second lactations and 45.6% third or more lactations) and 3 173 516 Saanen records (31.1% first lactations, 24.5% second lactations and 44.4% third or more lactations) of females born between 1950 and 2012 were used. Recently, the number of records for SCS and type traits has been smaller than that for milk production traits. Weights for SCS and milk production records were as defined in the official genetic evaluation [20,21] according to lactation number (from 1 to 10) and length of lactation (up to 180 days or not).

The pedigree used in this study consisted of 2 981 809 animals (40% Saanen, 53% Alpine, 4% crossbred of Alpine and Saanen and 3% other breeds) born between 1950 and 2012 and considered up to 29 generations for males. It was completed by 43 unknown parent groups defined according to breed and birth year: one group every two years, with sires and dams pooled together.

Genetic models used for analysis

Animal GEBV were estimated using genomic best linear unbiased prediction (BLUP) with the blup90iod program [22] , using models as described in the following.

Multi-breed model

The first model used for multi-breed genomic prediction was:

$$\mathbf{y} \quad \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of all female records (ν) from the two breeds weighted by their lactation weights as defined above, and $\mathbf{X}$ is the incidence matrix relating fixed effects ($\boldsymbol{\beta}$) to individuals. The following fixed effects were considered for milk production traits and SCS: herd (within year and parity), age and month at delivery (within year and region), length of dry period (within year and region) and breed. For type traits, fixed effects were: herd (within year), age at scoring, lactation stage and breed. $\mathbf{W}$ is the incidence matrix relating permanent environmental effects ($\mathbf{p}$), which were normally distributed $N\ \mathbf{0}, \sigma_p^2\mathbf{I}_t$, to individuals, $\mathbf{Z}$ is a design matrix allocating observations to breeding values ($\mathbf{u}$), and $\mathbf{e}$ is a vector of random normal errors, normally distributed $N\ \mathbf{0}, \sigma_e^2\mathbf{I}_\mathrm{v}$. Genomic breeding values $\mathbf{u}$ were assumed normally distributed with $Var\ \mathbf{u} \quad \mathbf{H}\sigma_u^2$, where $\mathbf{H}$ is the multi-breed genetic relationship matrix combining SNP marker information and pedigree data, implemented as in Legarra et al. [12]:

$$\mathbf{H} \quad \begin{matrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\ \mathbf{G} - \mathbf{A}_{22}\ \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{matrix},$$

where $\mathbf{A_{11}}$ is a sub-matrix of the pedigree-based relationship matrix ($\mathbf{A}$) for ungenotyped animals, $\mathbf{A_{22}}$ is a sub-matrix for genotyped animals, and $\mathbf{A_{12}}$ (or $\mathbf{A_{21}}$) is a sub-matrix that describes the pedigree-based relationship between ungenotyped and genotyped animals. The genomic relationship matrix ($\mathbf{G}$) was scaled to be comparable with the $\mathbf{A}$ matrix using the correction defined by Gao et al. [10]. Matrix $\mathbf{G}$ was derived as in [23].

$$\mathbf{G} \quad \frac{\mathbf{MM}'}{2 \sum_{j\ 1}^{p} \mathrm{q_j}\ 1-\mathrm{q_j}},$$

where p is the number of SNPs, $\mathrm{q_j}$ the allele frequency of the whole population (Alpine and Saanen) for SNP j and $\mathbf{M}$ is a centered matrix of SNP genotypes.

Within-breed model

The within-breed model was similar to the above multi-breed model except that Alpine and Saanen were evaluated separately. The relationship matrices ($\mathbf{H_s}$ for Saanen breed and $\mathbf{H_a}$ for Alpine breed) were defined as in the previous model except that they were derived from the allele frequencies ($\mathrm{q_j}$) observed in each breed. The same pedigree as defined above in the paragraph on data was used to derive the pedigree relationship matrix.

Multiple-trait model

The third model used in this study was the same as that used in [18]:

$$\mathbf{y}_b \quad \mathbf{X}_b\boldsymbol{\beta}_b + \mathbf{Z}_b\mathbf{u}_b + \mathbf{W}_b\mathbf{p}_b + \boldsymbol{\varepsilon}_b,$$

where $\mathbf{y}_b \quad \begin{matrix}\mathbf{y}_a \\ \mathbf{y}_s\end{matrix}$ and $\mathbf{u}_b \quad \begin{matrix}\mathbf{u}_{b_a} \\ \mathbf{u}_{b_s}\end{matrix}$ is the vector of true breeding values normally distributed with $Var\ \mathbf{u}_b \quad \begin{matrix}\sigma_{u_a}^2 & \sigma_{u_{s,a}} \\ \sigma_{u_{s,a}} & \sigma_{u_s}^2\end{matrix} \otimes \mathbf{H}$. The fixed effects considered in this model $\boldsymbol{\beta}_b \quad \begin{matrix}\boldsymbol{\beta}_{b_a} \\ \boldsymbol{\beta}_{b_s}\end{matrix}$ were the same as in the within-breed model. The vector of permanent environmental

effects, $\mathbf{p}_b = \begin{matrix}\mathbf{p}_{b_a}\\ \mathbf{p}_{b_s}\end{matrix}$, was normally distributed with $Var\ \mathbf{p}b = \begin{matrix}\sigma^2_{p_a}\mathbf{I}_{t_1} & \mathbf{0}\\ \mathbf{0} & \sigma^2_{p_s}\mathbf{I}_{t_2}\end{matrix}$, where $\sigma^2_{p_a}$ and $\sigma^2_{p_s}$ are the values used in official evaluations (Table 1). The genomic relationship matrix was built as in the multi-breed model and estimated with allele frequencies derived across breed. In this model, $\boldsymbol{\varepsilon}_b = \begin{matrix}\boldsymbol{\varepsilon}_{b_a}\\ \boldsymbol{\varepsilon}_{b_s}\end{matrix}$ is the vector of random normal errors, defined as: $Var\ \boldsymbol{\varepsilon}_b = \begin{matrix}\sigma^2_{\varepsilon_a}\mathbf{I}_{n_1} & \mathbf{0}\\ \mathbf{0} & \sigma^2_{\varepsilon_s}\mathbf{I}_{n_2}\end{matrix}$.

Three levels of genetic covariance between the traits for the Alpine and Saanen breeds ($\sigma_{u_{s,a}}$) were used in this study: (1) the covariance estimated from the data (see the section on "Genetic parameter estimation" below), (2) the covariance such that the genetic correlation (ρ) was equal to 0, which leads to a model that is similar to the within-breed model, and (3) the covariance such that the genetic correlation was equal to 0.99, which results in a model that is close to the multi-breed model.

Estimation of genetic parameters

The genetic parameters used in the multi-breed model were those used in official genetic evaluations (Table 1). For the multiple-trait and within-breed models, genetic parameters were estimated from the data (Table 2), except for repeatabilities, which were those in Table 1 and considered similar for both breeds.

Genetic and residual variances $\sigma^2_{u_s}$, $\sigma^2_{u_a}$, $\sigma^2_{e_s}$ and $\sigma^2_{e_a}$ and the genetic covariance were estimated by restricted maximum likelihood (REML) with remlf90 software [22], using a multiple-trait model (see above) and the multi-breed **H** matrix described earlier. Standard deviations of heritabilities and genetic correlations were estimated using the approximation of Klei [24], using airemlf90 software [22]. Because of computational issues, we chose to estimate these parameters on a subset of first lactation records from the data, without considering repeatability. For milk production traits, these parameters were estimated using 113 980 first lactation records (49 201 for the Saanen and 64 779 records for the Alpine breed) from the population of all females born between 2008 and 2012. Since the data included fewer records for SCS and udder type traits, all first lactation records on all females born between 2002 and 2012 were used for these traits (i.e. 130 230 records for SCS and 202 102 for udder type traits). These analyses included 1985 females genotyped with the Illumina SNP50 BeadChip [1], using the same quality control as for males. Genotyped females were groups of 100 half-sibs from 20 different sires [3] involved in a design for quantitative trait locus (QTL) detection [25].

Table 1 Genetic parameters used in the multi-breed model

Trait	Genetic variance	Heritability	Repeatability
Milk yield	11665.48	0.30	0.50
Fat yield	14.26	0.30	0.50
Protein yield	9.04	0.30	0.50
Fat content	8.73	0.50	0.70
Protein content	2.32	0.50	0.70
Somatic cell score	0.30	0.20	0.47
Udder floor position	0.30	0.31	-
Udder shape	0.59	0.34	-
Rear udder attachment	0.42	0.27	-
Fore udder	0.34	0.29	-
Teat angle	0.20	0.29	-

Repeatability is not modeled for udder type traits.

Cross-validation analyses

Cross-validation consisted in splitting the population of 677 progeny-tested and genotyped bucks into two sets: the training set (425 males born between 1993 and 2005: 232 Alpine and 193 Saanen individuals) and the validation set (252 males born between 2006 and 2009: 100 Saanen and 152 Alpine individuals). Phenotypes used for the training set were records of females born before 2008, i.e. the first year of lactation for daughters of bucks born in 2005. Prediction quality was evaluated based on Pearson correlations between GEBV and DYD [26] for the validation males, and regression coefficients of DYD on GEBV. The DYD were obtained from official genetic evaluations (January 2013). These validation correlations serve as indicators of predictive ability and the regression coefficients (slopes) serve as indicators of the dispersion of GEBV; a slope above 1 indicates under-dispersion of GEBV and a slope below 1 indicates over-dispersion.

Prediction error variances (PEV) of GEBV were estimated as in Misztal et al. [27] by estimating the inverse of the coefficients matrix of the mixed-model equations [28] using FORTRAN program accf90. Average model accuracies were derived from PEV as in Carillier et al. [3] for the 148 young males that were not yet progeny-tested and born between 2010 and 2011. Prediction quality of validation males, and GEBV accuracy of young animals were analyzed both in the whole population (Alpine + Saanen animals) and also for each breed separately.

Results and discussion

Genetic parameters

Table 2 reports the estimates of heritability (h^2), genetic variances and genetic correlations between Alpine and Saanen breeds for all traits studied. Standard errors of heritability ranged from 0.008 for udder type traits for the

Table 2 Estimates of genetic parameters of type and production traits in Saanen and Alpine goat population using the multiple-trait model

Trait	Alpine goats		Saanen goats		Estimated genetic correlation between Alpine and Saanen
	Genetic variance	Heritability	Genetic variance	Heritability	
Milk yield	9, 962	0.31	7, 278	0.26	0.45
Fat yield	14.1	0.28	11.2	0.25	0.48
Protein yield	8.6	0.31	6.1	0.25	0.54
Fat content	7.3	0.48	8.0	0.51	0.46
Protein content	2.3	0.60	2.9	0.56	0.54
Somatic cell score	0.3	0.20	0.3	0.16	0.47
Udder floor position	0.6	0.51	0.6	0.57	0.76
Udder shape	0.7	0.40	0.9	0.47	0.59
Rear udder attachment	1.2	0.47	1.1	0.52	0.73
Fore udder	0.6	0.44	0.4	0.42	0.59
Teat angle	0.3	0.42	0.4	0.45	0.66

Alpine breed to 0.020 for milk production traits for the Saanen breed (results not shown). The highest heritabilities were found for protein content (around 0.6), udder depth (around 0.55), rear udder attachment (around 0.5) and fat content (0.5). SCS was the less heritable trait, with estimates of 0.2 and 0.16 in the Alpine and Saanen breeds, respectively.

Heritability estimates obtained for milk production traits were close to those reported by Belichon et al. [14] but they tended to be smaller, especially for fat yield in Alpine goats (0.28 *vs* 0.37), protein yield in Saanen goats (0.25 *vs* 0.34), and fat content in both breeds (0.48 *vs* 0.58 and 0.51 *vs* 0.60 in Alpine and Saanen goats, respectively). Our heritability estimates for udder type traits were much higher than those reported by Clément et al. [20]: 0.51 and 0.57 *vs* 0.34 and 0.37 for udder floor position in the Alpine and Saanen breeds, respectively. These differences could be explained by the more recent data that were used here, with females born between 2008 and 2012, compared to the females born between 1998 and 1997 [14] and 2000 and 2002 [20] in other studies. Genetic parameter estimates for SCS were fairly similar to those reported by Rupp et al. [29].

Heritability estimates were similar for Saanen and Alpine goats except for udder shape (0.47 *vs* 0.40), udder floor position (0.57 *vs* 0.51) and protein yield (0.25 *vs* 0.31) (Table 2). The largest between-breed differences in heritability were previously reported for udder shape (0.40 *vs* 0.28 in Alpine and Saanen goats, respectively) and protein content (0.58 *vs* 0.50) [14,20]. However, genetic variances in our study tended to differ between Alpine and Saanen populations (14.1 *vs* 11.2 for instance for fat yield). Thus, the similar heritability estimates in the two breeds were explained by a similar ratio between genetic and residual variances, rather than similar variances.

Estimates of the genetic correlation between traits in the Alpine and Saanen breeds ranged from 0.45 for milk yield to 0.76 for udder floor position (Table 2). Standard errors of estimates of the genetic correlation ranged from 0.1 for udder type traits and protein content to 0.3 for SCS (results not shown). Estimates were close to those estimated between Holstein and Normande and between Montbéliarde and Normande dairy cattle breeds [18], i.e. from 0.38 to 0.46 for milk yield and from 0.35 to 0.56 for fat content, but lower than the correlations between Holstein and Montbeliarde breeds (0.79 for milk yield and 0.66 for fat content). These results suggest that French Holstein and Montbeliarde dairy cattle breeds are genetically closer than the Alpine and Saanen goat breeds, perhaps due to the introgression of Red Holstein genes in the Montbéliarde breed in the 1980's. The highest correlations and the lowest standard errors of these correlations were obtained for udder floor position (0.76) and rear udder attachment (0.73). These high genetic correlations for udder type traits suggest that marker effects in the Alpine breed were closer to marker effects in the Saanen breed for these particular traits.

Analysis of the multi-breed population

Validation correlations

To make results from the models comparable, GEBV from the within-breed model were centered on the overall average across the two breeds. For the within-breed model, validation correlations and slopes (Tables 3 and 4) were estimated using the centered GEBV of all validation males (Alpine + Saanen).

Table 3 reports the Pearson correlations between GEBV and DYD for the 252 validation males for all models. The correlations ranged from 0.33 for protein yield using the multi-breed model to 0.70 for protein content using the

multi-breed or the multiple-trait model with ρ estimated or equal to 0. The highest correlations were obtained for traits with the highest heritabilities, i.e. for fat content (0.61), protein content (0.70), and rear udder attachment (0.64), as previously also reported by Carillier et al. [3].

Using the within-breed model, validation correlations were slightly greater than with the multi-breed model, except for protein content (–1%) and teat angle (–3%). The largest increase was for protein yield (+9%). Other studies using within-breed models [17,19,30] did not evaluate regression slopes and correlations of GEBV across breeds, which makes it difficult to compare results with our findings. The studies that compared multi-breed to within-breed genomic evaluations reported increases in validation correlations from 2% for milking ability in Finnish cattle to 50% for maternal calving in Swedish cattle [17] and from 52% for milk yield to 80% for protein yield in Chinese bulls [30]. However, grouping several breeds (Holstein, Jersey and Brown Swiss) in the same training set could reduce validation correlations by 2% to 3% [19].

Validation correlations obtained with the multiple-trait models were close to those obtained with the within-breed and multi-breed models, with differences ranging from 1% for udder shape to 6% for fat content. In dairy cattle, Olson et al. [19] found that the multiple-trait model outperformed (+9%) the multi-breed model in a two-step approach on de-regressed EBV. Using different values for the genetic correlation between Alpine and Saanen breeds (ρ = 0.99, ρ estimated and ρ = 0) did not have a major impact on correlations between DYD and GEBV (Table 3). Differences ranged from 0% for SCS, udder floor position and udder shape, to 3.4% for teat angle. These results are consistent with those of Karoui et al. [18] who observed no impact of the genetic correlation on validation correlations and regression coefficients.

Table 3 Pearson correlations between DYD and GEBV from three models for the 252 validation males regardless of breed

Trait	Multi-breed	Multiple-trait			Within-breed
		ρ = 0.99	ρ estimated	ρ = 0	
Milk yield	0.43	0.43	0.42	0.43	0.43
Fat yield	0.44	0.45	0.46	0.46	0.46
Protein yield	0.33	0.36	0.35	0.35	0.36
Fat content	0.61	0.59	0.60	0.60	0.63
Protein content	0.70	0.68	0.70	0.70	0.69
Somatic cell score	0.47	0.47	0.47	0.47	0.47
Udder floor position	0.59	0.59	0.59	0.59	0.59
Udder shape	0.55	0.56	0.56	0.56	0.56
Rear udder attachment	0.64	0.66	0.66	0.67	0.66
Fore udder	0.50	0.49	0.50	0.51	0.51
Teat angle	0.61	0.60	0.58	0.57	0.59

ρ is the genetic correlation between Alpine and Saanen goats used in the multiple-trait model.

Comparison of the results obtained with the multi-breed model and with two-step approach [3] shows that validation correlations using the single-step approach increased for all traits, by 10% for milk yield up to 74% for teat angle, expect for protein yield (–8%). Increases were greater for udder type traits (mean 59%) than for milk production traits (mean 14%). This is consistent with results found in a Lacaune dairy sheep population [11,31], for which validation correlations increased from 11% for milk yield to 53% for udder depth. Previous studies using the pseudo-single-step approach led to intermediate results with increases in validation correlations from 0.1% to 10% in Nordic Holstein cattle [10] and from 2% to 34% in Lacaune dairy sheep [11] compared to the two-step approach.

Regression coefficients of DYD on GEBV

Estimates of the regression coefficients of DYD on GEBV for validation males (Table 4) ranged from 0.43 for protein yield using the multi-breed model to 1.51 for rear udder attachment using the within-breed model. Standard errors of these estimates ranged from 0.06 for protein content with the multiple-trait model using the estimated ρ to 0.12 for milk yield with the multi-breed model. Based on these standard errors, no significant differences in slopes were observed between the models. The lowest regression coefficients of DYD on GEBV were found for protein yield (0.43 with the multi-breed model), fore udder (0.55 with the multiple-trait model using the estimated ρ) and fat yield (0.61 with the multiple-trait model using ρ = 0). Slopes that were the furthest from 1 were obtained for traits with the lowest validation correlations between DYD and GEBV. It is difficult to interpret these slopes when the estimated validation correlations are not sufficient. However, slopes of DYD on GEBV less than 1, which indicates over-dispersion of GEBV, were observed for almost all traits except for protein content with the within-breed model, and for rear udder attachment and teat angle with the multiple-trait model using ρ = 0.99, as well as with the within-breed model for these three traits.

Regression coefficients were slightly closer to 1 with the within-breed model than with the multi-breed model, i.e. by +6% for SCS to +57% for fore udder, except for fat (–8%) and protein contents and for rear udder attachment (Table 4). These results indicate less dispersion of GEBV with the within-breed model than with the multiple-trait model for almost all traits. Using the multiple-trait model, slopes were slightly greater when ρ was estimated or equal to 0 (by 1% for fat yield with ρ = 0 to 48% for fore udder with ρ estimated) compared to the model with ρ = 1, as in

Table 4 Regression coefficients of DYD on GEBV from three models for the 252 validation males regardless of breed

Trait	Multi-breed		Multiple-trait						Within-breed	
			ρ = 0.99		ρ estimated		ρ = 0			
	reg	SE	reg	SE	reg	SE	reg	SE	reg	SE
Milk yield	0.58	0.12	0.76	0.10	0.48	0.10	0.49	0.10	0.77	0.10
Fat yield	0.62	0.08	0.67	0.08	0.62	0.08	0.61	0.08	0.69	0.08
Protein yield	0.43	0.09	0.53	0.09	0.44	0.09	0.44	0.09	0.54	0.09
Fat content	0.86	0.07	0.97	0.08	0.76	0.08	0.76	0.08	0.79	0.08
Protein content	0.95	0.07	0.92	0.07	0.94	0.06	0.94	0.07	1.17	0.08
Somatic cell score	0.67	0.08	0.68	0.08	0.64	0.08	0.64	0.08	0.71	0.08
Udder floor position	0.76	0.08	0.86	0.07	0.80	0.08	0.80	0.08	0.90	0.08
Udder shape	0.66	0.09	0.96	0.09	0.65	0.09	0.65	0.09	0.96	0.09
Rear udder attachment	0.70	0.09	1.46	0.11	0.60	0.10	0.59	0.10	1.51	0.11
Fore udder	0.60	0.09	0.89	0.10	0.55	0.10	0.56	0.10	0.94	0.10
Teat angle	0.74	0.08	1.10	0.09	0.58	0.09	0.56	0.11	1.19	0.10

ρ is the genetic correlation between Alpine and Saanen goats used in the multiple-trait model.

the multi-breed model, except for some traits. Regression coefficients that were obtained with a genetic correlation estimated or equal to 0 were similar (equal or up to 2% different; Table 4). The slopes obtained with ρ = 0.99 were consistent with those obtained with the two-step approach [3], at 0.76 *vs* 0.79 for example for milk yield. Using a genetic correlation of 0.99 greatly reduced the dispersion of the GEBV compared to other correlation levels, i.e. by 6% for udder shape to up to 61% for teat angle, but not for protein content or rear udder attachment. In the study by Karoui et al. [18], slopes that were estimated using a between-breed genetic correlation of 0.95 were similar to those obtained with an estimated genetic correlation ranging from 0.38 to 0.79 for milk yield depending on the breeds considered.

Regression coefficients obtained with the multi-breed single-step model were lower than those estimated with the two-step approach [3], by 9% for udder floor position and by up to 43% for protein yield, except for protein content. The differences obtained in this study were consistent with those reported in the literature: 17% for final score in US Holstein bulls [5], and from 12% for milk yield to 14% for SCS in Lacaune dairy sheep [11,31]. These regression coefficients were not as good as expected, although allele frequency differences between the genotyped and base-population animals [5] were taken into account in the genomic relationship matrix using the approach of Christensen [32]. The correction of the genomic relationship matrix for differences between base-population and genotyped animals proposed by Vitezica et al. [33] gave similar results for slopes of DYD on GEBV (results not shown). Gao et al. [10,34] showed that in a Nordic Holstein dairy cattle population, the corrections done to the genomic relationship matrix as proposed in [33] did not significantly reduce the over-dispersion of GEBV (from 0% to 3%) and even increased it in some cases (from 1% to 2%).

Model-based accuracies

Figure 1 shows the average model accuracies estimated on the 148 candidates using predictions based on the 677 males of the reference population. These accuracies ranged from 0.54 for fore udder using the within-breed model to 0.74 for fat and protein contents using the multi-breed model. The highest accuracies (on average 0.5 and 0.56 for fat and protein content, respectively) were obtained for traits with the highest heritabilities.

With the within-breed model, model accuracies were similar to those obtained with the multiple-trait model at ρ equal to 0.99 for milk, fat and protein yields and for udder floor position (Figure 1). However, they were lower than with the multi-breed model, except for udder depth (+1.5%) and SCS (the same result was obtained with the three models). Reductions in accuracy when using the within-breed model compared to the multi-breed model ranged from −1% for protein content to −14% for fore udder. These results could be explained by the small size of the reference population of each breed. With the multiple-trait model, average model accuracy was higher when ρ was set equal to 0.99 than with the estimated ρ, which in turn was higher than with ρ set equal to 0. Accuracies with the multiple-trait model were lower than with the multi-breed model by −1% for protein yield to −5% for protein content, but slightly higher (by +1% to +3%) for teat angle, udder shape and rear udder attachment.

The best model-based accuracies were obtained using the multi-breed model for milk production traits and using the multiple-trait model with ρ equal to 0.99 for SCS and udder type traits (Figure 1). Compared to the

two-step approaches [3], the single-step approach increased model accuracy from 28% for udder type traits and SCS to 37% for milk, fat and protein yields. These model accuracies were greater than parent average accuracies for almost all traits, by 1% for rear udder attachment and teat angle with the multiple-trait model and the estimated ρ, to 14% for fat and protein contents using the multi-breed model. However, model-based accuracies did not exceed parent average accuracy for: (1) SCS and udder type traits with the multiple-trait model and ρ set to 0, (2) udder shape and fore udder with the within-breed model, and (3) SCS with the multiple-trait model and the estimated ρ.

Analysis by breed

In Tables 5 and 6, GEBV from the multi-breed model were deviated from the mean GEBV of each breed to compare results from the multi-breed and the within-breed models. Table 5 shows Pearson correlations between GEBV and DYD for the 152 Alpine and 100 Saanen validation males

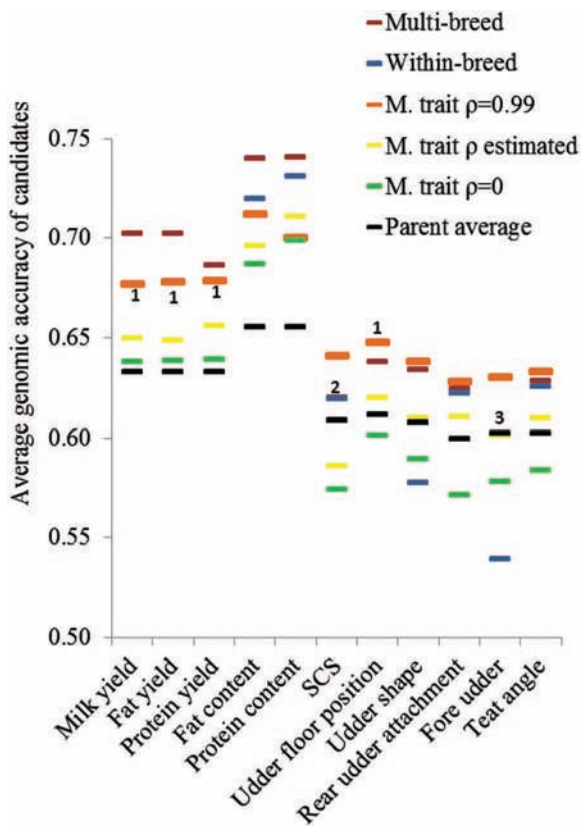


Figure 1 Average model accuracy for the 148 candidate bucks regardless of breed. 1 means similar results were obtained for the within-breed model and the multiple-trait model using ρ = 0.99; 2 means similar results were obtained for the within-breed model and the multi-breed model; 3 means similar results were obtained for the multi-breed model and the multiple-trait model using ρ = 0.99. M. trait means multiple-trait model.

using the within-breed and the multi-breed models. Correlations obtained with the multiple-trait models were similar to those with the multi-breed model (results not shown), which is consistent with Makgahlela et al. [34]. The validation correlations were similar for the multi-breed and the within-breed models, except for fat yield in Saanen, protein yield in Alpine, and fat content in both breeds. In dairy cattle, combining several breeds in the training set did not improve validation correlations [19,35] except for the Brown Swiss breed, which had the smallest population size [19]. The higher prediction ability obtained with the within-breed model compared to the multi-breed model (Table 5) for fat content (0.55 vs 0.47 in Alpine and 0.65 *vs* 0.53 in Saanen populations) could be explained by the presence of one of the mutations in the *DGAT1* gene in the Saanen breed but not in Alpine goats (C Maroteau, UNCEIA, Toulouse, personal communication).

The validation correlations estimated in this study were higher for the Saanen breed than for the Alpine breed for almost all traits, from 18% for fat and protein content to 65% for fore udder (Table 5). Similar results for the two breeds were obtained only for SCS, udder shape and teat angle. The absence of differences in phenotypic variances and DYD accuracies between the two breeds did not help to explain the difference in accuracies. However, this could be due to a higher inbreeding level in the Saanen breed (2.3% in Saanen vs 1.8% in Alpine) and a higher kinship coefficient between the training and testing sets (2.4% in Saanen vs 1.1% in Alpine, using genomic data). Thus, the larger training set size available for the Alpine breed did not counterbalance the smaller relationships between training and testing sets it had compared to the Saanen breed.

Table 6 reports regression coefficients of DYD on GEBV for each breed with the multi-breed and the within-breed models. Slopes obtained with the multiple-trait model were similar to those obtained with the multi-breed model (results not shown). As mentioned previously, almost all slopes of DYD on GEBV were less than 1, which indicates over-dispersion of GEBV. Regression coefficients were closer to 1 in Saanen than in Alpine goats, except for protein content, SCS, rear udder attachment and teat angle. Since the validation correlations obtained for the Saanen breed were higher than for the Alpine breed, the best slopes were obtained for the Saanen breed. Differences between the two models were greater for regression coefficients than for accuracies and ranged from 1% for udder shape for the Alpine breed to 69% for fore udder for the Saanen breed. Nevertheless, these differences were small compared to the high standard errors of the slopes.

Average model accuracies of the 148 candidates analyzed separately for each breed (results not shown) were little affected by the model used and were close to the accuracies analyzed by pooling Alpine and Saanen breeds

Table 5 Breed-specific Pearson correlations between DYD and GEBV from three models for 152 Alpine and 100 Saanen goats

Trait	Alpine goats		Saanen goats	
	Multi-breed	Per breed	Multi-breed	Within-breed
Milk yield	0.35	0.36	0.50	0.49
Fat yield	0.35	0.39	0.47	0.53
Protein yield	0.34	0.26	0.34	0.37
Fat content	0.47	0.55	0.53	0.65
Protein content	0.62	0.62	0.72	0.73
Somatic cell score	0.45	0.45	0.43	0.43
Udder floor position	0.52	0.53	0.65	0.65
Udder shape	0.51	0.51	0.51	0.50
Rear udder attachment	0.51	0.51	0.69	0.69
Fore udder	0.33	0.34	0.55	0.56
Teat angle	0.46	0.45	0.42	0.44

together. Average model accuracies ranged from 0.62 for SCS, rear udder attachment and fore udder to 0.74 for protein content. Results on model accuracies were higher than expected given the small population size used in this study (86 Alpine and 62 Saanen individuals), and were slightly higher (by +1% for protein content to +8% for udder depth) for the Saanen than for the Alpine goats. The better results obtained for the Saanen breed compared to the Alpine breed could be explained by a higher inbreeding level [3] and kinship coefficient in the Saanen breed.

Genetic selection in the French breeding programs for Alpine and Saanen breeds is achieved through within-breed selection. Therefore, to compare the three models proposed in this study, we need to focus on the within-breed comparisons but because of the small size of the reference population (less than 400 males in each breed), multi-breed genomic evaluation has to be considered.

Conclusions

This study compared three models (multi-breed, within-breed and multiple-trait) using a single-step approach for genomic evaluation. Quality of the predictions was similar for the three models, except for the dispersion of the GEBV, which was better with the within-breed model. The single-step approach resulted in higher prediction accuracy and over-dispersion of GEBV compared to the two-step approach. Average model accuracy for the candidates using a single-step approach outperformed the accuracy derived from pedigree-based parent average information from official evaluations, except for udder shape and teat angle. The best accuracies were obtained with the multi-breed model. Considering the small size of the population used in the within-breed model, accuracies were not expected to be high. A major gene or causal mutation specific to each breed (*DGAT1* and casein variant) could explain the good results obtained for the within-breed model. Based on prediction quality, there was no difference between the three models compared in this study. The most convenient model for genomic evaluation in French dairy goats would be the multi-breed model using a single-step approach. This model is the easiest to implement since it requires just one evaluation instead of two (multi-breed *vs* within-breed) and less computing time than the multiple-trait model. However, the dispersion of the GEBV indicates that improvements are needed before this model can be viably implemented in official evaluations.

Table 6 Breed-specific regression coefficients of DYD on GEBV from three models for 152 Alpine and 100 Saanen goats

Trait	Alpine goats				Saanen goats			
	Multi-breed		Per breed		Multi-breed		Within-breed	
	reg	SE	reg	SE	reg	SE	reg	SE
Milk yield	0.54	0.12	0.65	0.14	0.84	0.15	0.89	0.16
Fat yield	0.49	0.10	0.72	0.12	0.67	0.13	0.93	0.14
Protein yield	0.52	0.12	0.45	0.12	0.54	0.15	0.56	0.15
Fat content	0.84	0.13	0.76	0.13	0.72	0.12	1.11	0.13
Protein content	0.98	0.10	1.02	0.10	1.19	0.12	1.27	0.12
Somatic cell score	0.68	0.11	0.72	0.12	0.66	0.14	0.69	0.14
Udder floor position	0.84	0.11	0.89	0.12	1.01	0.12	0.92	0.11
Udder shape	0.97	0.13	0.98	0.13	0.93	0.16	0.99	0.17
Rear udder attachment	1.15	0.16	1.21	0.13	1.70	0.18	1.71	0.18
Fore udder	0.54	0.17	0.62	0.17	1.05	0.13	1.11	0.14
Teat angle	0.99	0.18	0.99	0.19	0.85	0.16	0.93	0.16

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
CC analyzed the data and wrote the paper. CC, CRG and HL interpreted the results. CRG and HL revised and improved the manuscript. All authors have read and approved the final manuscript.

Acknowledgements
The authors thank the French Genovicap and Phenofinlait programs (ANR, Apis-Gène, CASDAR, FranceAgriMer, France Génétique levage, French Ministry for Agriculture) 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. We also thank the GenoToul bioinformatics facility in Toulouse for providing computing and storage resources. This study would not have been possible without the goat SNP50 BeadChip developed by the International Goat Genome Consortium (IGGC): http://www.goatgenome.org. The authors thank Ignacy Misztal for the availability of the blup90iod2 program.

Author details
[1]INRA, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, 31326 Castanet-Tolosan, France. [2]Université de Toulouse INPT ENSAT, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, 31326 Castanet-Tolosan, France. [3]Université de Toulouse INPT ENVT, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, 31076 Toulouse, France.

Received: 27 February 2014 Accepted: 18 September 2014
Published online: 29 October 2014

References

1. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, Donnadieu-Tonon C, Eggen A, Heuven HC, Jamli S: **Design and characterization of a 52 K SNP chip for goats.** *PLoS One* 2014, **9:**e86227.
2. Danchin-Burge C: *Bilan de variabilité génétique de 9 races de petits ruminants laitiers et à toison*, Compte rendu N 001172004 Institut de L©levage, collection résultats, juin 2011. 2011.
3. 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 7305.
4. Fritz S, Guillaume F, Croiseau P, Baur A, Hoze C, Dassonneville R, Boscher MY, Journeaux L, Boichard D, Ducrocq V: **Mise en place de la sélection génomique dans les trois principales races françaises de bovins laitiers.** *Renc Rech Ruminants* 2010, **17:**455 458 [http://www.journees3r.fr/IMG/pdf/2010_15_05_Fritz.pdf]
5. 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 score.** *J Dairy Sci* 2010, **93:**743 752.
6. Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, **136:**245 257.
7. Hayes BJ, Goddard M: **Genome-wide association and genomic selection in animal breeding.** *Genome* 2010, **53:**876 883.
8. Liu Z, Seefried FR, Reinhardt F, Rensing S, Thaller G, Reents R: **Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction.** *Genet Sel Evol* 2011, **43:**9.
9. Su G, Guldbrandtsen B, Gregersen VR, Lund MS: **Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population.** *J Dairy Sci* 2010, **93:**1175 1183.
10. Gao H, Christensen OF, Madsen P, Nielsen US, Zhang Y, Lund MS, Su G: **Comparison on genomic predictions using GBLUP models and two single-step blending methods with different relationship matrices in the Nordic Holstein population.** *Genet Sel Evol* 2012, **44:**1 8.
11. Baloche G, Legarra A, Sallé G, Larroque H, Astruc J-M, Robert-Granié C, Barillet F: **Assessment of accuracy of genomic prediction for French Lacaune dairy sheep.** *J Dairy Sci* 2013, **97:**1107 1116.
12. Legarra A, Aguilar I, Misztal I: **A relationship matrix including full pedigree and genomic information.** *J Dairy Sci* 2009, **92:**4656 4663.
13. Babo D: *Races ovines et caprines françaises*. France Agricole Editions, Paris, France; 2000. [http://books.google.fr/books?id=9OWWy8SJkwkC&printsec=frontcover&hl=fr#v=onepage&q&f=false]
14. Bélichon S, Manfredi E, Piacère A: **Genetic parameters of dairy traits in the Alpine and Saanen goat breeds.** *Genet Sel Evol* 1999, **31:**529 534.
15. Simeone R, Misztal I, Aguilar I, Vitezica ZG: **Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure.** *J Anim Breed Genet* 2012, **129:**3 10.
16. Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME: **Accuracy of genomic breeding values in multi-breed dairy cattle populations.** *Genet Sel Evol* 2009, **41:**1 9.
17. Br ndum RF, Rius-Vilarrasa E, Strandén I, Su G, Guldbrandtsen B, Fikse WF, Lund MS: **Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations.** *J Dairy Sci* 2011, **94:**4700 4707.
18. Karoui S, Carabano MJ, Diaz C, Legarra A: **Joint evaluation of French dairy cattle breeds using multiple-trait models.** *Genet Sel Evol* 2012, **44:**1 10.
19. Olson K, VanRaden P, Tooker M: **Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss.** *J Dairy Sci* 2012, **95:**5378 5383.
20. Clément V, Boichard D, Piacere A, Barbat A, Manfredi E: **Genetic evaluation of French goats for dairy and type traits.** In *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production: 19–23 August 2002; Montpellier.* 2002:4.
21. Clément V, Caillat H, Piacère A, Manfredi E, Robert-Granié C, Bouvier F, Rupp R: **Vers la mise en place d'une sélection pour la résistance aux mammites chez les caprins.** *Renc Rech Ruminants* 2008, **15:**405 408 [http://www.journees3r.fr/IMG/pdf/2008_13_genetique_02_Clement.pdf]
22. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH: **BLUPF90 and related programs (BGF90).** In *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production: 19–23 August 2002; Montpellier.* 2002:2.
23. VanRaden PM: **Efficient methods to compute genomic predictions.** *J Dairy Sci* 2008, **91:**4414 4423.
24. Klei B, Tsuruta S: **Approximate variance for heritability estimates.** [http://nce.ads.uga.edu/html/projects/AI_SE.pdf].
25. Maroteau C, Palhière I, Larroque H, Clément V, Tosser-Klopp G, Rupp R: **QTL detection for traits of interest for the dairy goat industry.** In *Proceedings of the 64th Annual Meeting of the European Federation of Animal Science (EAAP): 26–30 August 2013; Nantes.* 2013.
26. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Invited Review: Reliability of genomic predictions for North American Holstein bulls.** *J Dairy Sci* 2009, **92:**16 24.
27. Misztal I, Tsuruta S, Aguilar I, Legarra A, VanRaden PM, Lawlor TJ: **Methods to approximate reliabilities in single-step genomic evaluation.** *J Dairy Sci* 2013, **96:**647 654.
28. Meyer K: **Approximate accuracy of genetic evaluation under an animal model.** *Livest Prod Sci* 1989, **21:**87 100.
29. Rupp R, Clément V, Piacere A, Robert-Granié C, Manfredi E: **Genetic parameters for milk somatic cell score and relationship with production and udder type traits in dairy Alpine and Saanen primiparous goats.** *J Dairy Sci* 2011, **94:**3629 3634.
30. Zhou L, Ding X, Zhang Q, Wang Y, Lund MS, Su G: **Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population.** *Genet Sel Evol* 2013, **45:**7.
31. Duchemin SI, Colombani C, Legarra A, Baloche G, Larroque H, Astruc JM, Barillet F, Robert-Granié C, Manfredi E: **Genomic selection in the French Lacaune dairy sheep breed.** *J Dairy Sci* 2012, **95:**2723 2733.
32. Christensen OF: **Compatibility of pedigree-based and marker-based relationship matrics for single-step genetic evaluation.** *Genet Sel Evol* 2012, **44:**37.
33. Vitezica ZG, Aguilar I, Misztal I, Legarra A: **Bias in genomic predictions for populations under selection.** *Genet Res Camb* 2011, **93:**357 366.
34. Makgahlela ML, M ntysaari EA, Strandén I, Koivula M, Nielsen US, Sillanp MJ, Juga J: **Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle.** *J Anim Breed Genet* 2013, **130:**10 19.
35. Hayes BJ, Daetwyler HD, Bowman P, Moser G, Tier B, Crump R, Khatkar M, Raadsma HW, Goddard ME: **Accuracy of genomic selection: comparing theory and results.** *Anim Breed Genet* 2009, **18:**34 37.

doi:10.1186/s12711-014-0067-3
Cite this article as: Carillier *et al.*: Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. *Genetics Selection Evolution* 2014 **46**:67.

Bilan

Les corrélations génétiques estimées entre les caractères Saanen et Alpine, de 0,45 pour le lait à 0,76 pour le caractère distance plancher-jarret, sont du même ordre que celles obtenues entre les races bovines Holstein et Normande ou Normande et Montbéliarde pour les caractères de production laitière (Karoui et al., 2012). Les corrélations génétiques les plus fortes sont obtenues pour les caractères de morphologie mammaire, ce qui suggère un déterminisme génétique commun aux deux races pour ces caractères.

En ce qui concerne les corrélations de validation, peu de différences entre les modèles ont été observées que ce soit pour la population totale ou pour chaque race séparément. Les valeurs des corrélations varient entre 0,26 pour la matière protéique en race Alpine et 0,73 pour le taux protéique en race Saanen, et sont plus élevées que celles obtenues avec la méthode two steps (cf. Article I). Les corrélations de validation obtenues en race Saanen sont plus fortes que celles en race Alpine (excepté pour LSCS et ORT) ce qui n'était pas le cas avec l'approche two steps. Ce résultat est à mettre en lien avec une structure de population plus favorable (consanguinité et apparentement entre population de référence et candidate plus élevés) à la sélection génomique en race Saanen (cf. Chapitre 2). De plus, les corrélations de validation observées séparément dans chaque race, obtenues soit avec les évaluations uniraciales soit avec les évaluations multiraciales, sont similaires. Les pentes de régression des GEBV en fonction des DYD en single step sont quant à elles similaires à celles obtenues dans l'approche two steps (article I). Elles sont également proches de celles de l'étude de Karoui et al. (2012) chez des bovins laitiers. D'autre part, il semble que ces pentes soient plus proches de 1 avec un modèle uniracial qu'avec les autres modèles testés.

Les moyennes des précisions théoriques des GEBV estimées pour les 148 candidats sont comprises entre 0,54 pour la qualité de l'attache arrière avec le modèle par race et 0,74 pour le TP avec le modèle multiracial. Les précisions théoriques les plus fortes sont obtenues avec le modèle multiracial pour les caractères laitiers, et avec le modèle bicaractère prenant en compte une corrélation génétique de 0,99 pour les LSCS et les caractères de morphologie mammaire. Ces fortes précisions sont supérieures à celles obtenues sur ascendance pour l'ensemble des caractères testés. On peut noter que les modèles « par race » et bicaractère (corrélation Alpine-Saanen de 0 ou estimée) ne permettent pas d'obtenir le niveau de précision obtenu sur ascendance pour les LSCS et les caractères de morphologie. Comme pour les corrélations de validation, les précisions sont plus élevées en race Saanen qu'en race Alpine.

Si l'on compare ces résultats avec ceux de l'approche « pseudo single step » du chapitre 3, on remarque que les corrélations de validation obtenues ici sont toujours nettement supérieures. Cet écart entre les résultats du pseudo single step et ceux du single step peut être expliqué par une forte proportion (44%) de femelles issues de pères inconnus, qui ne sont prises en compte que dans l'approche single step. En revanche l'approche pseudo single-step permet de considérer une part importante de la monte naturelle, puisque 69% des femelles issues de monte naturelle sont en fait issues de grands-pères d'IA.

En conclusion, les précisions génomiques obtenues avec l'approche single step surpassent celles obtenues avec les approches « two steps » et « pseudo single step ». Les différents modèles testés en single step ne peuvent pas être discriminés aux vues des résultats. Nous pouvons noter que malgré la petite taille des populations considérées dans le cas des évaluations par race, les précisions génomiques obtenues sont similaires à celles de l'évaluation multiraciale. L'évaluation la plus simple et la plus rapide en pratique est l'évaluation multiraciale puisqu'une seule évaluation est nécessaire contre deux dans le cas de l'uniracial. Les résultats des précisions étant également peu différents entre évaluations, il semblerait plus judicieux de privilégier une évaluation multiraciale. L'évaluation par race pourrait être cependant privilégiée si l'on souhaite intégrer des QTL dont les effets diffèrent selon la race dans les évaluations génétiques ou génomiques.

Dans cette étude, les corrélations de validation entre DYD et EBV n'ont pas été présentées. Les gains de précisions entre les corrélations DYD-GEBV et DYD-EBV dans le cadre des évaluations réalisées dans l'article II (single step multiracial, uniracial et bicaractère) sont présentés dans la Figure 4 .45. En single step, ils se situent entre -13% pour la forme de l'avant pis et +68% pour le TB obtenus avec des évaluations uniraciales. Ils sont en moyenne plus faibles que ceux obtenus en bovins laitiers Holsteins américains (+38% pour l'index de synthèse) (Aguilar et al., 2010) et en race bovine Brune (entre +8% et +19% (Gray et al., 2012)) pour les caractères de production laitière. Les corrélations de validation prenant en compte l'information génomique (GEBV*DYD) ne sont meilleures par rapport aux corrélations estimées à partir du pédigrée (EBV*DYD) que pour certains caractères. En effet, les corrélations de validation génomiques sont plus faibles que les corrélations de validation estimées à partir du pédigrée (gains de précision négatifs) pour le taux protéique de -0,2% à -11,6%, excepté pour le single step bicaractère avec une corrélation fixée à 0. Les gains de corrélations de validation obtenus avec l'information génomique sont également négatifs pour

les caractères MP avec le single step uniracial, LSCS et ORT pour toutes les évaluations excepté le single step multiracial ainsi que pour l'AVP dans le cas de l'évaluation bicaractère avec une corrélation fixée à 0,99 et de l'évaluation uniraciale. De telles diminutions de corrélation de validation n'ont pas été observées dans la littérature, elles pourraient être expliquées par la petite taille de la population d'apprentissage disponible en validation croisée (425 mâles génotypés). La taille de cette population a pu être augmentée à 677 mâles début 2015. En effet, les valeurs génétiques estimées (sur descendance) des 148 jeunes mâles nés entre 2011 et 2011 sont désormais connues. Ces mâles peuvent donc être utilisés comme population de validation, permettant d'utiliser l'ensemble de 677 mâles nés avant 2010 comme population d'apprentissage. Les gains de précisions obtenus avec un modèle single step multiracial sont alors dans ce cas toujours positifs (de +4% pour les SCS à +22% pour la matière protéique, résultats non montrés).

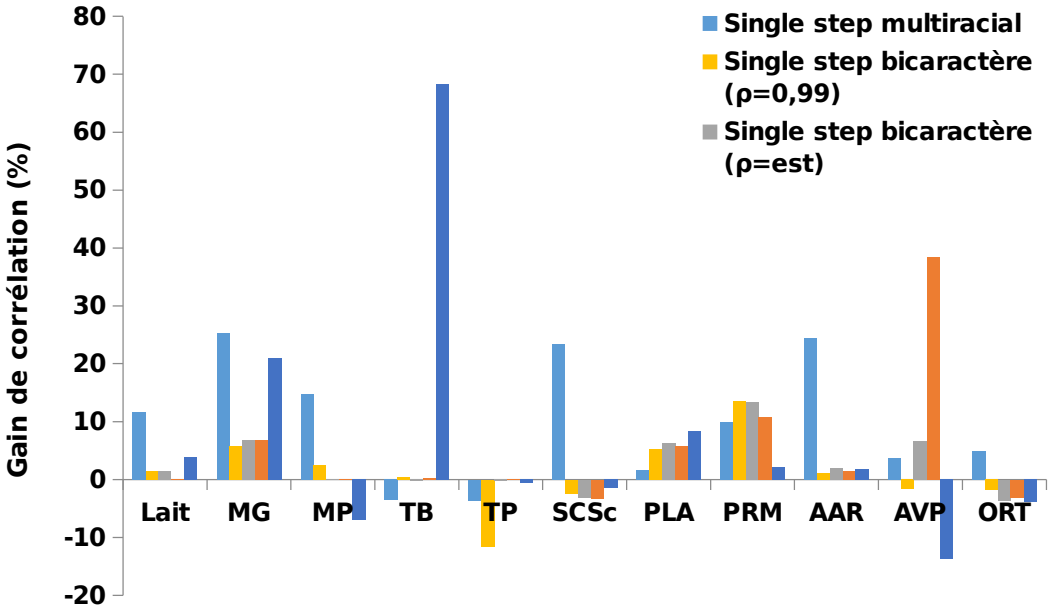


Figure 4.45 : Gain de précision (%) entre corrélations GEBV*DYD et EBV*DYD pour chacun des modèles de l'approche single step

4.II Intérêt de l'apport des génotypes femelles dans le modèle d'évaluation génomique

Dans le chapitre 3 (cf. Article I), l'intérêt de l'apport des génotypes des femelles du dispositif QTL sur les précisions génomiques semblait limité. Le single step permettant d'augmenter les précisions en s'affranchissant du calcul de performances corrigées, nous avons étudié l'intérêt d'introduire ces génotypes femelles dans une évaluation single step sur les données caprines.

Comme dans l'article précédent, la méthode utilisée est le GBLUP avec l'approche single step. Les valeurs génomiques ont été estimées à l'aide du logiciel blup90iod (Misztal et al., 2002) en utilisant les performances de l'article II pour estimer les précisions théoriques des GEBV des jeunes mâles candidats (soit les performances de l'indexation officielle de janvier 2013). Pour tester l'intérêt de l'apport des génotypes femelles, les génotypes utilisés sont soit ceux des 825 mâles, soit ceux des 825 mâles complétés de ceux des 1 985 femelles.

Le schéma de validation croisée utilisée dans cette étude ainsi que la description des populations candidate et de référence utilisées pour le calcul des précisions des candidats sont présentés en Figure 4 .46.

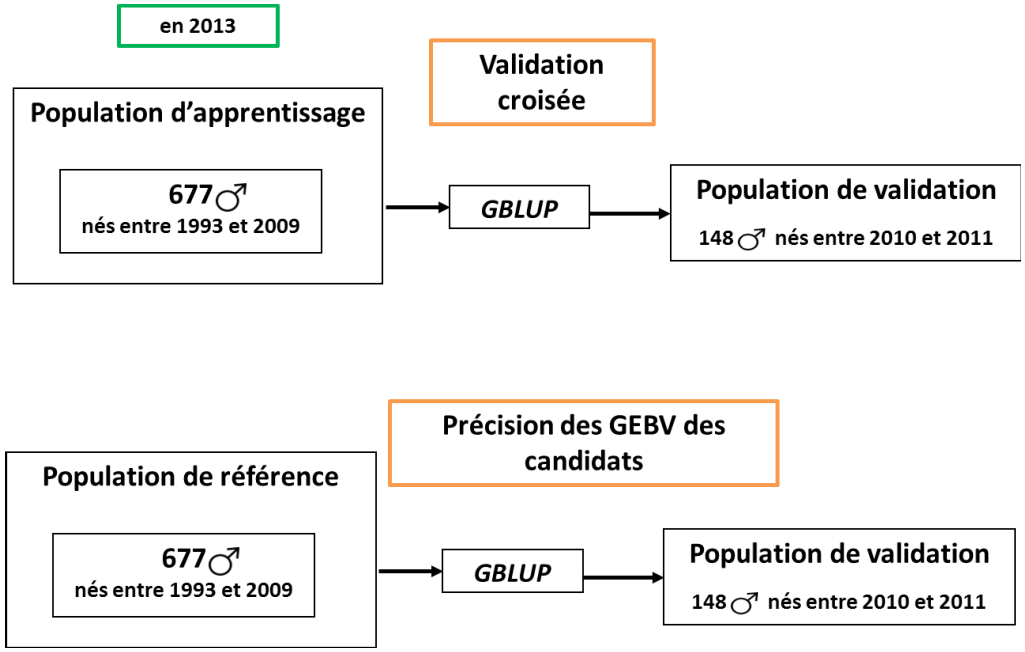


Figure 4.46 : Schéma de validation croisée et des populations de références et candidates utilisées pour l'estimation de l'apport du génotypage des femelles sur les précisions obtenues en single step

Les corrélations de validation obtenues pour les 148 jeunes mâles candidats (nés entre 2010 et 2011) entre leurs GEBV de 2013 et leurs EBV estimées avec l'indexation officielle de janvier 2015 sont présentées au Tableau 4 .14. On remarque que les précisions de validation sont améliorées avec l'apport des génotypes femelles pour la majorité des caractères, de 2% pour les LSCS, le PRM et l'AAR à 14% pour la matière grasse. Elles sont légèrement diminuées pour la quantité de lait de 3% et elles restent identiques pour la matière et le taux protéique. Ces augmentations de précisions avec l'ajout des génotypes femelles sont légèrement plus importantes que celles obtenues avec la méthode en deux étapes (cf. Article I) qui étaient en moyenne de 5%.

Tableau 4.14: Corrélations de validation entre les GEBV de 2013 et les EBV de 2015 pour les 148 jeunes mâles candidats nés entre 2010 et 2011 dans le cas d'une population de référence mâle ou mâles&femelles

	mâles &femelles	mâle
Lait	0,47	0,48
MG	0,41	0,36
MP	0,37	0,37
TB	0,72	0,70
TP	0,75	0,75
LSCS	0,45	0,44
PLA	0,50	0,48
PRM	0,51	0,50
AAR	0,55	0,54
AVP	0,59	0,54
ORT	0,37	0,36

Le gain moyen (en %) de précision théoriques des GEBV des candidats avec l'apport des génotypes des femelles pour les 148 jeunes mâles candidats est présenté dans la Figure 4 .47. Contrairement à ce qui est observé pour les corrélations de validation, l'ajout des génotypes des femelles permet d'augmenter les précisions calculées à partir des PEV pour tous les caractères. Les gains de précision constatés sont également plus élevés que ceux obtenus pour les corrélations de validation : entre 6,5% pour la matière protéique et 13,7% pour les taux butyreux et protéique. Les résultats présentés ici montrent le gain moyen estimé pour les 148 mâles candidats. Cependant le gain de précision est très variable selon les individus, il varie de 0,4% à 19,6% sur le TP par exemple. Cette différence de gain entre les individus ne semble pas être liée directement à leur degré de parenté avec les femelles génotypées (résultats non montrés).

En conclusion, l'ajout des génotypes des femelles du dispositif QTL permet d'augmenter légèrement les précisions génomiques obtenues en caprins. Cependant ces augmentations sont plus fortes pour les taux que pour les autres caractères. En bovins laitiers Holstein espagnols, utiliser les génotypes de femelles choisies au hasard donne les mêmes résultats sur les précisions génomiques que d'utiliser les meilleures femelles (Jiménez-Montero et al., 2010). Comme expliqué dans l'article I, les femelles du dispositif QTL sont les filles de 20 mâles bien connus (avec plus de 100 filles), et ne sont donc pas originales d'un point de vue génétique. Cependant dans le pédigrée caprin, il existe de nombreuses femelles

dont le père est inconnu, il pourrait donc être intéressant de génotyper certaines de ces femelles afin d'améliorer la parenté entre ces femelles et les individus du pédigrée.

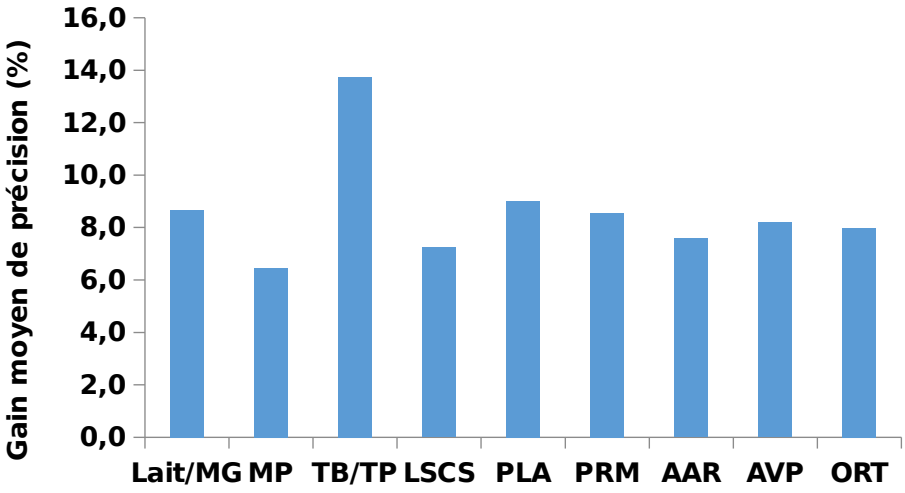


Figure 4.47 : Gain moyen de précision théorique des GEBV (%) avec l'apport des génotypes des femelles du dispositif QTL pour les 148 jeunes mâles candidats nés entre 2010 et 2011

4.III Quel poids pour la matrice génomique ?

La matrice de variance-covariance des valeurs génétiques considérée dans le chapitre 3 ainsi que dans l'article II a été construite avec un poids de 5% pour la matrice de parenté et un poids de 95% pour la matrice génomique. Certaines études ont montré des différences entre les corrélations de validation obtenues en fonction du poids donné à la matrice de parenté. En bovins laitiers Holstein, (Gao et al., 2012) soulignent qu'il existe un poids optimal pour la matrice de parenté qui se situe entre 0,15 et 0,20. Cependant d'autres études montrent que le poids utilisé a peu d'impact sur les précisions génomiques (Pribyl et al., 2012).

Nous avons donc testé différentes pondérations (25%, 50%, 75% et 95%) pour la matrice génomique et avons comparé les précisions génomiques obtenues. Les évaluations génomiques ont été réalisées comme dans la partie précédente (4.II) en utilisant un GBLUP single step multiracial sur les performances de l'indexation officielle de 2013 avec les génotypes des 825 mâles et 1 985 femelles. De même que précédemment, les corrélations de validation sont réalisées sur les 148 jeunes mâles entre les GEBV prédites en 2013et les EBV estimées en 2015. Les moyennes des précisions génomiques pour ces mâles (calculées à partir des PEV) ont également été étudiées.

Tableau 4.15: Corrélations des GEBV en 2013 et des EBV en 2015 pour les 148 jeunes mâles candidats pour tous les caractères évalués en caprins

	95%	75%	50%	25%
Lait	0,47	0,47	0,48	0,48
MG	0,41	0,40	0,41	0,39
MP	0,37	0,37	0,37	0,35
TB	0,72	0,72	0,72	0,71
TP	0,75	0,76	0,76	0,75
LSCS	0,45	0,42	0,44	0,44
PLA	0,50	0,50	0,50	0,49
PRM	0,51	0,52	0,52	0,51
AAR	0,55	0,55	0,55	0,54
AVP	0,59	0,59	0,59	0,56
ORT	0,37	0,37	0,37	0,37

Le Tableau 4 .15 présente les corrélations obtenues pour les 148 jeunes mâles pour l'ensemble des caractères étudiés. On remarque que les différences observées sur les corrélations de validation pour différents poids de la matrice génomique testés sont faibles entre 4% pour la quantité de lait et 7% pour la forme de l'avant pis. Ces résultats sont proches de ceux trouvés en bovins laitiers (Pribyl et al., 2012; Rodríguez-Ramilo et al., 2014) avec peu de différences observées. Cependant les corrélations de validation les plus fortes sont obtenues pour le poids de la matrice génomique testé le plus élevé (95%) excepté pour la quantité de lait et le profil de la mamelle (PRM). De même en bovins Holstein, le poids optimal pour la matrice génomique semble être un poids supérieur à 95% (Liu et al., 2011). Egalement en bovins Holstein, Gao et al. (2012) avaient eux conclu que le poids permettant d'obtenir les meilleures précisions se situait entre 0,75 et 0,80.

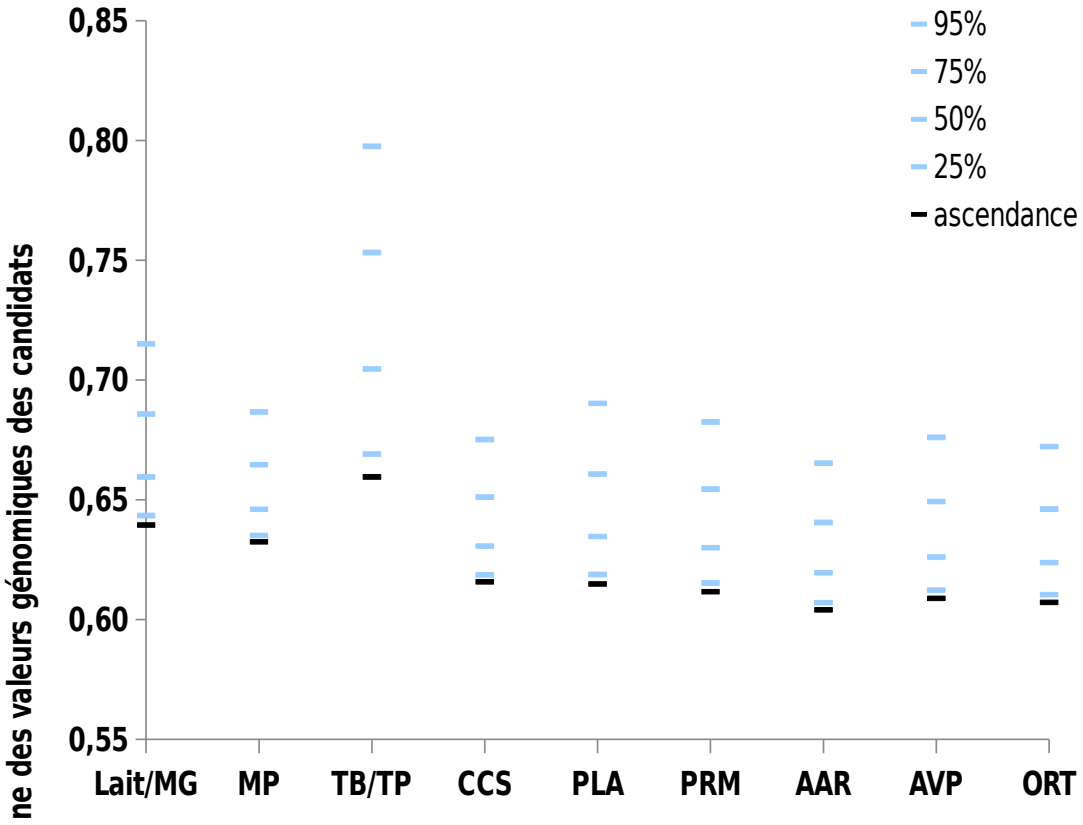


Figure 4.48 : Précision théorique moyenne des valeurs génomiques estimées des candidats selon le poids de la matrice génomique pour les différents caractères évalués en caprins laitiers

La Figure 4 .48 présente la moyenne des précisions théoriques ($\sqrt{CD}$) calculées à partir des PEV pour les 148 mâles candidats nés entre 2010 et 2011. Le classement de ces précisions en fonction du poids utilisé pour la matrice génomique est le même quel que soit le caractère considéré. Comme pour les corrélations de validation, les meilleures précisions sont obtenues pour un poids de la matrice génomique de 95%. Les différences observées sur les précisions en fonction des poids sont plus importantes (entre 10% pour la MP et 20% pour les taux) que celles observées sur les corrélations de validation. Cette divergence entre les précisions de validation et les précisions estimées à partir des variances d'erreur de prédictions avait également été constatée dans l'article II (cf. 4.1). Elle peut être expliquée par une surestimation des précisions lorsqu'elles sont calculées à partir des PEV (Clark et al., 2012). En effet, nous ne présentons que la moyenne des précisions estimées à partir des PEV, et non la variance de ces précisions qui semble augmenter avec le poids donné à la matrice G.

En conclusion, le poids optimal de la matrice génomique pour la précision est de 95% (soit 5 % pour la matrice de parenté). Cependant l'importance relative de ce poids semble faible notamment sur les précisions de validation. Certaines études (Gao et al., 2012) ont montré que le poids donné à la matrice G pouvait avoir un impact sur la dispersion des GEBV

obtenues, ce que nous n'avons pas étudié ici. Gengler et al. (2012) ont montré que les poids attribués à la matrice de parenté (estimée à l'aide du pedigrée) et à la matrice génomique équivalent à un modèle avec un effet polygénique et un effet SNP en considérant une distribution particulière entre variance polygénique et variance due aux SNP. Ce type de modèle ne nécessitant pas d'inversion de la matrice de parenté génomique est plus facile à implémenter, en particulier dans les cas plus complexes (multi-caractères, modèles de régression aléatoire) et lorsque les jeux de données sont importants

4.IV Prédiction d'une race à partir des génotypes d'une autre race

Etant donné le peu de différence observée entre les précisions génomiques obtenues avec des évaluations uniraciales et celles obtenues avec des évaluations multiraciales, la possibilité de prédire une race à partir des génotypes de l'autre race a été étudiée dans cette partie. Cette prédiction entre races pourrait permettre de réduire les coûts de génotypage qui sont pour l'instant encore élevés en caprins.

Des approches de type single step avec une méthode GBLUP ont été étudiées. Les performances utilisées sont celles de l'indexation officielle de 2008 multiraciale présentées dans l'article II pour la réalisation des analyses par validation croisée. Le schéma de validation croisée utilisée dans cette étude est présenté en Figure 4 .49. Dans un premier cas, les génotypes de 232 mâles Alpins, nés avant 2006 sont utilisés pour estimer la matrice **G** avec un poids de 95% dans la construction de la matrice de variance-covariance des valeurs génétiques **H** comprenant le pédigrée multiracial. Ces données sont utilisées pour prédire les GEBV de 100 mâles Saanen nés entre 2006 et 2009. Des corrélations de validation sont ensuite calculées entre les valeurs génétiques prédites pour les mâles Saanen et leurs DYD connus en 2013. Dans le deuxième cas, ce sont les génotypes de 193 mâles Saanen nés avant 2006 qui sont utilisés pour prédire les GEBV de 152 mâles Alpins nés entre 2006 et 2009 (cf. Figure 4 .49). Les corrélations de validation obtenues sont présentées dans la Figure 4 .50, le premier cas étant identifié par « predict Saanen » et le deuxième cas par « predict Alpine ». Enfin, des évaluations à partir des mêmes performances de l'indexation de 2008 ont été réalisées uniquement sur la base de la matrice de parenté afin de prédire les EBV des mâles de validation Saanen (« pédigrée Saanen », Figure 4 .50) ou des mâles de validation Alpins (« pédigrée Alpine », Figure 4 .50).

Il est difficile de conclure quant à la possibilité d'améliorer les prédictions génétiques des mâles d'une race à partir des génotypes des mâles de l'autre race. En effet, les corrélations de validation obtenues dans ce cas sont pour certains caractères plus élevées que celles

obtenues uniquement à partir du pédigrée (entre 0,1% pour l'ORT en race Alpine et 22% pour le TB en race Saanen) et pour d'autres plus faibles (entre -0.2% pour l'AVP et -43% pour la MP en Saanen). En bovins allaitants, Chen et al. (2013) n'ont pas montré d'amélioration des précisions des valeurs génétiques estimées des individus en prenant en compte les génotypes d'une autre race. Les précisions sont même dégradées d'environ 60% pour l'efficacité alimentaire en races Charolaise et Angus.

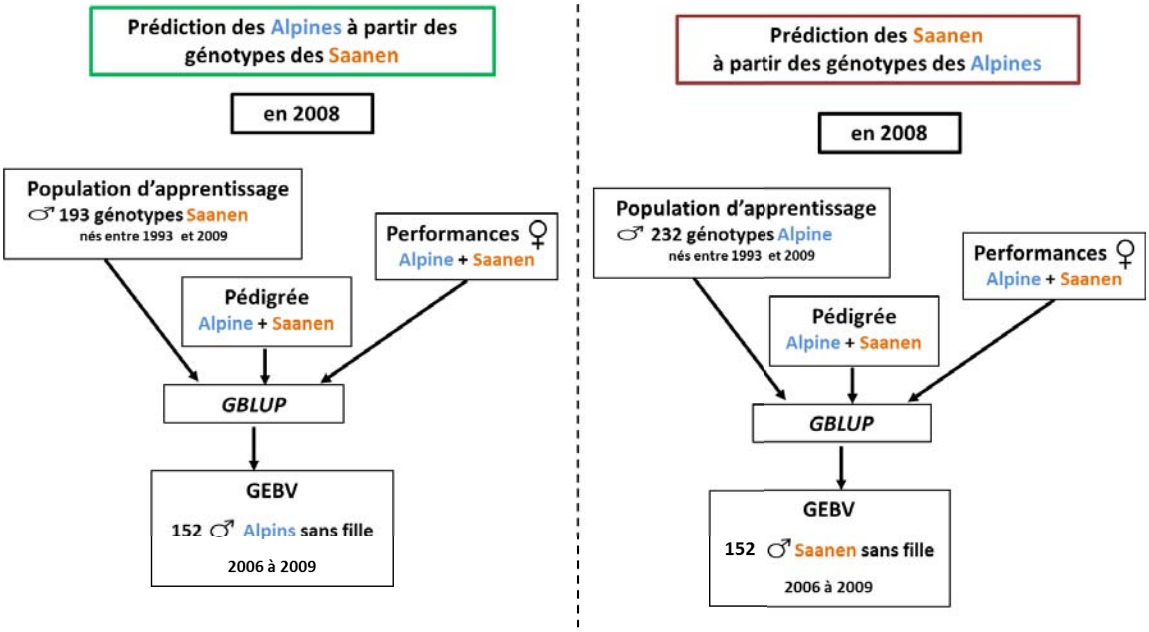


Figure 4.49 : Schéma de validation croisée utilisée pour prédire les Saanens à partir des génotypes des Alpines et pour prédire les Alpines à partir des génotypes des Saanens

Les corrélations obtenues pour prédire les Saanens à partir des génotypes des Alpines et celles obtenues pour prédire les Alpines à partir des génotypes des Saanens sont en général moins élevées que celles obtenues avec les évaluations multiraciales génomiques (Article II, table 5). Ces résultats étaient attendus dans la mesure où les mâles Alpins sont peu apparentés aux mâles Saanen (coefficient de parenté de 0,5% en moyenne). Des résultats similaires ont été constatés avec une population multiraciale de bovins laitiers : Fleckvieh, Hostein et Jersey. Les corrélations de validation obtenues pour les Fleckvieh ou les Jersey lorsque les génotypes utilisés sont ceux d'une autre race sont moins élevées (de plus de 90%) que celles obtenues dans le cas d'une population multiraciale (Pryce et al., 2011). La précision des valeurs génomiques des Jersey estimées à partir des Holstein (ou inversement) est nettement moins bonne (de 63% sur le TP à 99% sur la MP en Jersey) que celle obtenue avec une population multiraciale (Hayes et al., 2009a). En bovins laitiers , la qualité de prédiction des GEBV en race Simmentale à l'aide d'effets SNP estimés en race Montbéliarde (Hozé et al., 2014a) sont inférieures aux précisions obtenues en multiraciale de 33% en moyenne sur tous les

caractères. Ces précisions sont également inférieures à celles obtenues en évaluation uniraciale Simmental de 15% en moyenne sur les caractères de production laitière. Olson et al. (2012) ont également montré que prédire les valeurs génomiques d'une race (Jersey, Hostein ou Brune des Alpes) à partir d'une population de référence d'une autre race dégrade les précisions des GEBV de 10 à 62% par rapport à l'utilisation d'une population de référence constituée de la même race. Dans la littérature, les différences de précisions entre un modèle multiracial et celles obtenues dans le cas « prédiction d'une race à partir des génotypes de l'autre race » sont globalement plus élevées que celles constatées en caprins.

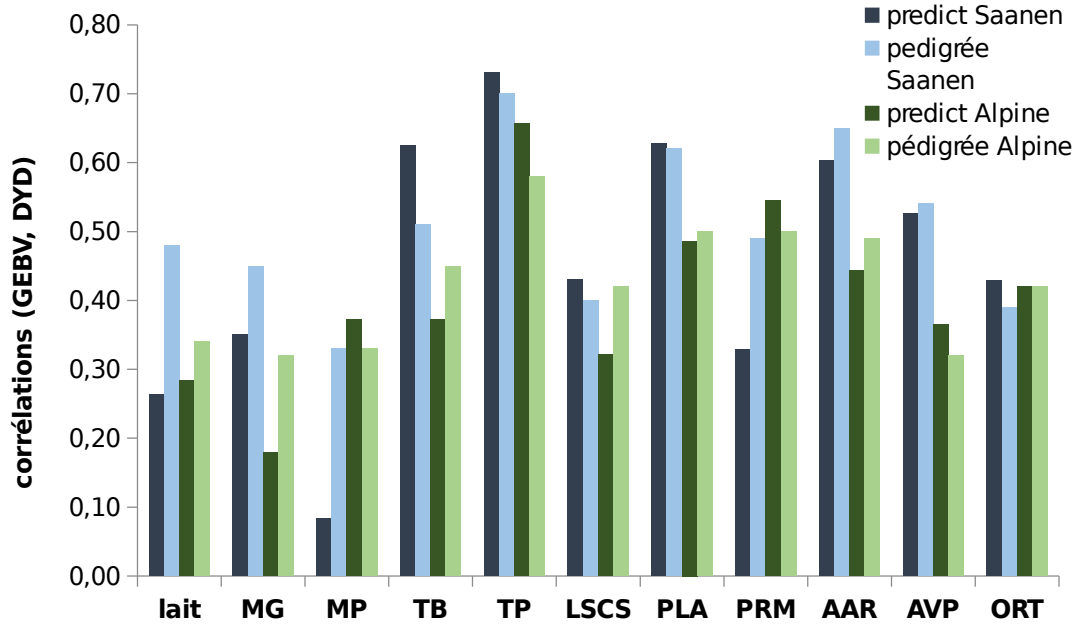


Figure 4.50 : corrélations de validation entre les GEBV et les DYD des 100 mâles Saanen ou des 152 mâles Alpins de validation dans le cas d'une population de référence multiraciale*
*corrélation estimée en race Saanen sur pédigrée uniquement (pédigrée Saanen, EBV*DYD), corrélation estimée en race Alpine sur pédigrée uniquement (pédigrée Alpine, EBV*DYD), ou à partir d'une population de référence Alpine (génotypes) pour prédire des mâles Saanen (pred Saanen, GEBV*DYD) ou d'une population de référence Saanen (génotypes) pour prédire des mâles Alpines (pred Alpine, GEBV*DYD)

L'étude des pentes de régression entre les GEBV prédites et les DYD connus en 2013 pour les mâles Alpins prédits à partir des Saanen et inversement (non présentées ici) montre des pentes plus éloignées de 1 (entre 14% pour les LSCS en Alpine et 56% pour le TP en Saanen) que dans le cas du multiracial.

En conclusion et comme cela a été montré notamment en bovins laitiers, il n'est pas envisageable, dans le cas des caprins laitiers français, de ne génotyper qu'une seule des deux races pour prédire les valeurs génomiques des deux races. Cela s'explique par la sélection réalisée en race pure (Alpine et Saanen) depuis de nombreuses années et par la parenté moyen

très faible (0,5%) qui en découle. Enfin, il a également été montré que les QTL identifiés dans l'espèce caprine ne sont pas nécessairement communs aux deux races (Maroteau et al., 2013).

En conclusion de ce chapitre, les évaluations génomiques de type single step permettent d'obtenir une meilleure qualité de prédiction des valeurs génétiques des candidats ainsi que de meilleures précisions. En revanche, il n'est pas possible de conclure à l'avantage d'un des modèles testés (multiracial, multicaractère ou uniracial) du point de vue des précisions de l'évaluation génomique. Le modèle le plus simple à implémenter et le plus rapide en temps de calcul étant le modèle multiracial, c'est ce modèle qui pourra être préconisé pour les futures évaluations génomiques caprines. Comme pour les évaluations génomiques « two steps », mais de manière plus modérée, l'ajout des génotypes des femelles du dispositif QTL permet d'améliorer les précisions des évaluations génomiques. Dans l'objectif d'augmenter davantage ces précisions, le génotypage d'autres femelles pourrait être envisagé comme nous le verrons plus en détail dans le chapitre « conclusions et perspectives ». Enfin le poids de la matrice de parenté génomique dans la matrice de variance-covariance des valeurs génétiques a peu d'impact sur les précisions des évaluations génomiques obtenues dans cette étude.

L'approche de type single step donnant des précisions satisfaisantes pour la mise en place d'évaluations génomiques en caprins laitiers français, le chapitre suivant est focalisé sur l'amélioration de cette méthode par une approche originale visant à intégrer l'effet d'un gène majeur dans les évaluations génomiques.

Chapitre 5 : Intégration du gène de la caséine αs1 dans les évaluations génétiques et génomiques

Introduction

Dans de nombreuses espèces, le récent développement des puces SNP permet désormais d'intégrer l'information moléculaire dans les évaluations génétiques via les méthodes d'évaluations génomiques. Cependant, il existe d'autres informations moléculaires disponibles comme celles issues du génotypage de gènes majeurs (caséine α_{s1} en caprins laitiers (cf.1.I.1), gène PRP en ovins laitiers) utilisé actuellement pour réaliser une présélection des animaux candidats au testage. Les SNP bi-alléliques ne permettent pas toujours de capter l'information de ces gènes majeurs parfois plus complexes (multi-alléliques par exemple). C'est le cas du gène majeur de la caséine α_{s1} en caprins laitiers, pour lequel les variants sont pour la plupart des additions ou délétions de base nucléotide qui ne sont pas captés par les SNP. De plus, les seuls variants pouvant être capté par des SNP sont éliminés lors du contrôle qualité réalisé sur les marqueurs en raison de leur caractère monomorphe dans notre population. En effet, étant donné l'importance de l'effet d'un gène majeur sur certains caractères, il serait intéressant de pouvoir intégrer cet effet dans les modèles d'évaluations génétiques afin d'améliorer la prédiction des valeurs génétiques des animaux. Les évaluations génomiques officielles réalisées en bovins laitiers en France (Boichard et al., 2012) intègrent l'information des QTL par le biais d'effets aléatoires des haplotypes aux QTL. Il a été montré que cette approche permet une légère amélioration des précisions de prédiction des GEBV comparé au GBLUP two steps (Ducrocq et al., 2014).

Le but de cette étude est donc d'évaluer l'intérêt d'inclure un effet du génotypage au locus de la caséine α_{s1} caprine dans les modèles d'évaluation génétiques et génomiques. Dans cette étude, 6 310 animaux (boucs candidats au testage sur descendance et mères à boucs) de race Alpine et Saanen ont été génotypés pour le locus de la caséine α_{s1}. La majorité des animaux dont les performances sont utilisées dans les évaluations génétiques officielles n'est donc pas génotypée. Aussi dans un premier temps, l'effet du génotype de la caséine α_{s1} a été intégré dans les évaluations génétiques et génomiques basées sur les pseudo-performances (DYD) des seuls mâles génotypés pour la caséine α_{s1} (two steps). Les caractères analysés sont ceux étudiés dans les autres chapitres de la thèse (les cinq caractères de production, les LSCS et les cinq caractères de morphologie). Plusieurs modèles ont été testés intégrant le génotype de la caséine α_{s1} en tant qu'effet fixe ou aléatoire dans les évaluations génétiques ou

génomiques, à partir d'une population multiraciale (Alpine + Saanen), ou de populations uniraciales (Alpine ou Saanen).

Afin de pouvoir intégrer un effet du génotype de la caséine α_{s1} dans les évaluations génétiques basées sur les performances propres de l'ensemble des femelles indexées (single step), deux solutions ont été envisagées. La première consiste à prédire le génotype pour toutes les femelles non génotypées. Pour cela, une méthode basée sur des algorithmes itératifs (Vitezica et al., 2005; Fernando et al., 1993) a été utilisée dans le but d'obtenir pour chaque femelle non génotypée ses probabilités pour tous les génotypes possibles. Deux modèles utilisant ces probabilités ont été testés. Le premier prend en compte la combinaison des probabilités pour tous les génotypes possibles comme un effet aléatoire dans le modèle. Le deuxième classe les génotypes possibles en 3 groupes selon leur degré d'effet sur le taux protéique : fort, moyen ou faible. Ces trois groupes constituent les trois niveaux d'un effet fixe. La deuxième solution envisagée est basée sur la méthode du « gene content » (Gengler et al., 2007) utilisant des groupes de parents inconnus dans le pédigrée. Cette méthode permet d'évaluer simultanément, par modèle multicaractère, la valeur génétique pour le taux protéique et pour le gene content c'est-à-dire le nombre de chaque allèle dans le génotype des individus. L'intérêt d'intégrer l'effet de ce gène dans les modèles est évalué par validation croisée à l'aide des corrélations entre les (G)EBV prédites par les modèles sur les données de 2013 et les EBV connues en 2015 (sur descendance) de 148 jeunes mâles nés entre 2010 et 2011.

Une première partie de l'étude a consisté à étudier les fréquences alléliques du génotype au locus de la caséine α_{s1}. Les fréquences des génotypes prédits par l'algorithme itératif ont également été comparées aux fréquences réelles. De plus, afin de comparer les méthodes, les génotypes ont été reconstruits à partir de l'estimation des gene content de chaque allèle de la caséine α_{s1} pour toutes les femelles non génotypées et comparés aux génotypes prédits par la méthode itérative. Des fréquences génotypiques ont ensuite été calculées à partir de ces génotypes reconstruits et comparées aux fréquences réelles et aux fréquences estimées pour les génotypes prédits avec la méthode iterative peeling. Une analyse de la significativité de l'effet du génotype dans les modèles ainsi qu'une estimation de la part de variance phénotypique expliquée par ce génotype a été réalisée.

Article III : Introduction de l'effet du gene majeur de la caséine αs1 dans les évaluations génétiques et génomiques des caprins laitiers français

Carillier-Jacquin C, Larroque H, Robert-Granié C, 2015. Including casein αs1 major gene effect on genetic and genomic evaluations of French dairy goats. Soumis à Genetics Selection Evolution en juin 2015.

Including α_{s1} casein major gene effect on genetic and genomic evaluations of French dairy goats

Céline Carillier-Jacquin[1,2,3§], Hélène Larroque[1,2,3], Christèle Robert-Granié[1,2,3]

[1] INRA, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, 31326 Castanet-Tolosan, France

[2] Université de Toulouse INPT ENSAT, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, 31326 Castanet-Tolosan, France

[3] Université de Toulouse INPT ENVT, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, 31076 Toulouse, France

[§]Corresponding author

Email addresses:

CC: celine.carillier@toulouse.inra.fr

HL: helene.larroque@toulouse.inra.fr

CRG: christele.robert-granie@toulouse.inra.fr

content approach performed better than other methods tested. It was only slightly better (from 1% to 9%) than a model without α_{s1} casein effect.

Conclusions

Including α_{s1} casein effect on genetic evaluation models in French is possible and has an interest to improve predictive ability. Difficulties to predict genotypes for ungenotyped animals limited the improvement of accuracy of estimated breeding values obtained.

Background

With the recent development of molecular technologies, genomic selection is planned to be introduced in several species, and major genes could be known and sequenced. Several major genes are already being selected like prion protein gene (PrP) for scrapie resistance in dairy sheep and goats [1, 2]. Hypothesis of a lot of quantitative trait loci (QTL) with small effect on genetic variation was considered in genomic selection process [3, 4]. Information on QTL with large effect (e.g. major gene) is ignored in genomic selection [5]. This information could be included in genetic or genomic evaluations to improve accuracies of estimated breeding values (EBV). In French dairy cattle, QTL were selected using linkage disequilibrium and linkage analysis for the largest one and using Elastic-Net approach for the others [5]. QTL effects are then added in genomic evaluations considering random effects of SNP haplotypes [6, 7]. Such an approach in French dairy cattle performed slightly better than classical genomic selection for accuracy of genomic selection [6]. It is easy to use QTL haplotypes or major genes information if this information is available for all animals. In cases of some animals are non-genotyped, it's possible to perform genetic evaluations based only on corrected performances of animals genotyped as in two steps genomic evaluation [8]. Nevertheless evaluation based on performances of all individuals led to better genomic accuracies [9]. Alleles at major gene have to be predicted for animal which are not genotyped. A solution to this problem is to deal with probabilities of carrying each possible allele. These

probabilities could be estimated using animals genotyped and pedigree based on iterative peeling method [10, 11]. Another possibility to deal with ungenotyped animal is to consider the gene content, i.e. the number of copies of a particular allele in a genotype of an animal, as a continuous variable [12]. However, the concept of gene content was developed for biallelic gene and has to be extent for multiallelic ones.

In French dairy goats, accuracy of genomic selection is not as high as in dairy cattle [13] due to the size and structure of the reference population [14]. Genomic selection is planned to be applied in this population but accuracy is expected to be improved by approaches considering well-known major genes like α_{s1} casein gene. In dairy goats, an effect of casein α_{s1} polymorphism has been found on protein content, protein yield, milk yield [15] and on fat content [7, 16]. Caprine α_{s1} casein polymorphism is one of the key factors which determines important technological properties of milk, such as cheese yield and cheese curd formation [17]. In French dairy goats breeding scheme, all candidate bucks for progeny testing born after 1986 were genotyped for casein α_{s1} gene using allele specific PCR (polymerase chain reaction) and RFLP (restriction fragment length polymorphism) technologies [18]. These genotypes are used to shortlist young candidates, eliminating males which carry alleles with poor effect on protein content. Until now, no effect of this genotype is included in genetic evaluation of French Alpine and Saanen goats.

The goat α_{s1} casein gene is a complex gene with at least seventeen possible alleles: A, B_1, B_2, B_3, B_4, C, H, L and M with strong effect on quantity of casein produced, E and I with intermediates effects, D, F and G with weak effects and three null alleles (O_1, O_2, N) [19, 20]. Alleles listed in French dairy goat populations were alleles A, B, C, E and F with, before 2000, a main proportion of alleles E and F for Saanen and Alpine breeds [17]. This complex gene could not be captured by SNP, in deed some alleles like allele E consist in insertion of several nucleotides [21].

The aim of this study is to take into account a α_{s1} gene effect, using available typing of α_{s1} casein gene information, in genetic and genomic evaluations. In French dairy goats, this typing is available for potential AI bucks and dam of bucks. First, a description of allele frequencies in the French population and their effects on production traits, somatic cell counts (SCS) and udder type traits was conducted. Then we compared several models to take into account typing of α_{s1} casein gene as a fixed or random effect in genetic and genomic evaluations based on males daughter yield deviation (two-step model). At last, α_{s1} casein gene effect was added to genetic and genomic evaluation models based on all females with true performances (single-step model) using probabilities of α_{s1} casein genotype for females or a gene content [12] approach.

Methods

Data

Genotypes of α_{s1} casein were obtained using RFLP and PCR technologies on blood DNA samples. The genotype consisted in 6 alleles (A, B, C, E, F or O) at each homologous chromosome and any incomplete genotype (20% in this study), with one missing allele for instance, was ignored. This genotype was available for 6 310 animals (3 441 Alpine and 2835 Saanen) born between 1982 and 2011: 823 males progeny tested including 677 males born before 2010 and 146 born between 2010 and 2011, 2 538 other males candidates to progeny testing which were not selected due to poor semen quality or growth defect and 2 949 dams of bucks (1 420 Saanen and 1 529 Alpine).

The SNP genotypes also obtained with the Illumina SNP50 BeadChip [22] were the same as described in [14]. After quality check (MAF >1%, call rate > 98% and call freq >99%) which was done separately in Alpine and Saanen breeds, 46 959 SNPs were validated. These SNP genotypes were available for 825 males progeny tested (470 Alpine and 355 Saanen) including the 823 males with genotypes on α_{s1} casein gene. The SNP dedicated to

capture α_{s1} casein polymorphism were removed from the data due to poor call rate. All SNP genotypes were performed according to the French National Guidelines for the care and use of animals for research.

Traits studied here were the five production traits (milk yield (kg), fat and protein yield (kg) and fat and protein contents (g/kg)), somatic cell score (SCS: log-transformed somatic cell counts), and five udder type traits (udder floor position, udder shape, rear udder attachment, fore udder, teat angle) used in [13, 14]. Genetic parameters used in this study were those described in [14]. Phenotypes recorded on females, their weights and pedigree used in the last part of this study (single step model) were obtained from official genetic evaluation of January 2013 using 4 178 315 Alpine records and 3 173 516 Saanen records and 2 981 809 individuals for genealogy [13]. Based on the same official genetic evaluation, daughter yield deviation (DYD), averages performances of daughters corrected for environmental effect and the merit of the dam, were computed and used as male phenotypes for the two-step models in this study. The DYD were obtained from female performances described previously and explained in Carillier et al. 2013. Model used to derived these DYD was the following: $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e}$, where $\mathbf{y}$ is the vector of all female performances from the two breeds (Alpine and Saanen) and $\mathbf{X}$ is the incidence matrix for the fixed effects ($\boldsymbol{\beta}$). The fixed effects considered for production traits and SCS were the herd (within year and parity), the age and the month at delivery (within year and area, the length of dry period (within year and region) and the breed. For udder type traits, fixed effects considered were the herd (within year), the age at scoring, the lactation stage and the breed. $\mathbf{W}$ is the incidence matrix for the permanent environmental effects ($\mathbf{p}$) normally distributed ($\mathbf{p} \sim N(\mathbf{0}, \sigma_p^2 \mathbf{I}_t)$) only used for production traits and SCS, $\mathbf{Z}$ is a design matrix allocating observations to breeding values ($\mathbf{u}$) normally distributed ($u \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$), where $\mathbf{A}$ is the additive relationship matrix

$\mathbf{y}_4 = \mathbf{X}_g\boldsymbol{\beta}_g + \mathbf{z}_1\boldsymbol{\alpha}_1 + \mathbf{z}_2\boldsymbol{\alpha}_2 + \mathbf{z}_3\boldsymbol{\alpha}_3 + \mathbf{z}_4\boldsymbol{\alpha}_4 + \mathbf{z}_5\boldsymbol{\alpha}_5 + \mathbf{z}_6\boldsymbol{\alpha}_6 + \boldsymbol{\varepsilon} + \mathbf{W}_g\mathbf{p}_g + \mathbf{e}_g$ where $\boldsymbol{\varepsilon}$ is the polygenic effect with $\text{var}(\boldsymbol{\varepsilon}) = \mathbf{A}\sigma^2_g$, and $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, $\boldsymbol{\alpha}_4$, $\boldsymbol{\alpha}_5$ and $\boldsymbol{\alpha}_6$ were the effects of alleles A, B, C, E, F and O respectively. Equivalence between multiple trait model and univariate model in which the effects of the gene are fit as a covariable was described in detail by [28, 29]. For the multiallelic loci, variance of genetic values ($\mathbf{u}_g$) were derived as: $\text{Var}(\mathbf{u}_g) = \mathbf{A}\sigma^2_{u_g} = \mathbf{A}\left[\sigma^2_e + 2\sum_{i=1}^{6} p_i q_i \alpha_i^2 - 2\sum_{i=1}^{6}\sum_{j\neq i} p_i p_j \alpha_i \alpha_j\right]$ with p_1, p_2, p_3, p_4, p_5, p_6 the allelic frequency for allele A, B, C, E, F and O of α_{s1} casein gene respectively and $q_i = 1 - p_i$. Covariances between genetic values ($\mathbf{u}_g$) and genetic effects of gene contents ($\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$, $\mathbf{u}_4$, $\mathbf{u}_5$, $\mathbf{u}_6$) could be estimated as: $\text{Cov}\begin{pmatrix}\mathbf{u}_g\\ \mathbf{u}_1\\ \dots\\ \mathbf{u}_6\end{pmatrix} = \begin{pmatrix}\mathbf{A}\sigma^2_{u_g} & \mathbf{A}\sigma_{u_{g,1}} & \dots & \mathbf{A}\sigma_{u_{g,6}}\\ \mathbf{A}\sigma_{u_{g,1}} & \mathbf{A}\sigma^2_{u_1} & \dots & \mathbf{A}\sigma_{u_{1,6}}\\ \dots & \dots & \dots & \dots\\ \mathbf{A}\sigma_{u_{g,6}} & \mathbf{A}\sigma_{u_{6,1}} & \dots & \mathbf{A}\sigma^2_{u_6}\end{pmatrix}$ where $\text{cov}(\mathbf{u}_g, \mathbf{u}_i) = 2\mathbf{A}p_i q_i \alpha_i - 2\mathbf{A}\sum_{i=1}^{6}\sum_{i\neq j}^{6} p_i p_j \alpha_j$ and covariances between effects of gene content effects were: $\text{cov}\begin{pmatrix}\mathbf{u}_1\\ \dots\\ \mathbf{u}_6\end{pmatrix} = 2\mathbf{A}\begin{bmatrix}p_1q_1 & -p_1p_2 & \dots & -p_1q_6\\ -p_1p_2 & p_2q_2 & \dots & -p_2q_6\\ \dots & \dots & \dots & \dots\\ -p_1p_6 & -p_2p_6 & \dots & -p_6q_6\end{bmatrix}$. Variances and covariances for this model were estimated by restricted maximum likelihood (REML) with remlf90 software [25].

Cross validation analyses

In this study, cross validation analyses consisted in splitting population of the 825 males genotyped with the 50K BeadChip in a training set: the 677 males born before 2010 (with performances of daughters recorded until January 2013), and a testing set: the 148 young males born between 2010 and 2011 with no daughter in January 2013. The breeding values predicted for these 148 young males were compared to their daughter yield deviations

(DYD) estimated with the official genetic evaluation of January 2015 with a mean of 53 daughters per sire. The validation correlations consisted in Pearson correlations between the EBV obtained in 2013 and the DYD obtained in 2015 for these 148 males.

Results and Discussion

Allele frequencies of α_{s1} casein genotype in the French dairy goat population

The frequencies of α_{s1} casein genotype were estimated in three populations of Alpine and Saanen animals: 1) the population of 823 progeny tested males and genotyped on 50K BeadChip, 2) the population of 2 538 other males 3) the population of 2 949 dams of bucks (Table 1). The most frequent α_{s1} genotypes in the French dairy goat population are genotypes AA and AE carried by more than 50% of the animals in the three subpopulations. In the two male populations, more than 25% of the males carried the AA genotype whereas the most frequent α_{s1} genotype in the female population is AE with 39.1% of dams of bucks. The third most represented genotype is also different between males and females populations, it's EE in the males population (13.7% for progeny tested males and 14.9% for the other ones) and AF for the dams of bucks (9.3%). These results were different than those obtained in American Saanen and Alpine dairy goats in which genotypes EF and FF were the most frequent [30]. But they were close to the one found for French dairy goats born before 1989 with supremacy of alleles A (41% in Alpine and 18% in Saanen breeds) and allele E (54% in Saanen and 26% in Alpine breeds) [21]. However proportion of allele A found here is higher than the one previously found in French dairy goats in 90's due to genetic selection promoting allele A [17]. In this study, allele C seemed to be a rare allele with less than 5% of animals carrying an allele C in the three subpopulations. These proportions of C allele in the populations studied here were close to the ones found in the French dairy goat populations, between 1% and 2% depending on the breed considered [21]. This also is consistent with studies on Mexican, Brazilian and American dairy goats breeds with no allele C found in these populations [30–

32]. Proportion of alleles F and O for highly selected animals (progeny tested males and dam of bucks) is lower than the one found in females born before 1989: 7.9% for progeny tested males and 18.9% for dams of bucks vs 28% in Alpine and 24% [21] in Saanen breeds. This difference is also observed for not highly selected animals, i.e. population of other males, which could be explained by the fact that their sires were used for artificial insemination (selected on α_{s1} casein genotype) and their dams were females selected for high protein content [17, 18, 21].

To compare frequencies of predicted genotypes of females to genotype frequencies found in the population of dams of buck truly genotyped, we considered only the 9 303 females with a probability to carry a given genotype of at least 75%. These frequencies were weighted by the probabilities of carrying genotype which were for the given females between 77 and 90%. For instance for the genotype AA, we count all females which probability to carry genotype AA was up to 75%, weighting them by their probability to carry genotype AA. These frequencies (predicted Alpine and predicted Saanen, Figure 1) were compared separately in each breed to the real frequencies observed in the population of truly genotyped females (True Alpine and True Saanen, Figure 1). First, based on true genotypes, large differences are observed between genotype frequencies in Saanen and Alpine breeds especially for genotypes AA, AE, CO and EE with differences in frequencies higher than 15%. These differences between allele frequencies were also observed in American dairy goats with 35.7% of allele E in Alpine breed and 70.5% in Saanen breed [30]. In this study, A allele was the most frequent allele in Alpine breed whereas it was the E allele in Saanen breed. Differences between frequencies of A and E alleles between Alpine and Saanen breeds were already observed by Martin and Leroux (2000) in the French dairy goat population: 14% of allele A in Alpine breed vs 7% in Saanen breed. This difference could be explained by a lower involvement of Saanen breeders in selection on protein yield and contents. Saanen

[16, 18]. In Norwegian dairy goats, α_{s1} casein haplotypes had a significant effect on protein content, fat yield and fat content [7]. The lack of effect on protein yield in this study could be explained by a high negative genetic correlation between milk yield and protein content: -0.28 in Alpine and Saanen breeds [7]. This negative correlation seemed to be emphasized by taking into account α_{s1} casein genotypes in the model from -0.42 to -0.48 [18].

Genetic parameters estimated with α_{s1} casein genotype effect considered as random in the model for milk yield, fat content and protein content were shown in Table 2. Largest part of phenotypic variance for α_{s1} casein effect was found for Alpine breed (for instance, on milk yield 6.1% in Alpine breed vs 3.3% in Saanen breed). The polymorphism of α_{s1} casein explained between 24.4% (Saanen breed) and 38.2% (Alpine breed) of protein content variance whereas it explained between 8.7% (Saanen) and 18.2% (Alpine) of fat content variance. These results for protein content were consistent with results obtained by Barbieri et al. (1995) where polygenic heritability moved from 0.66 to 0.38 when including α_{s1} casein effect in the model. Part of polygenic variance estimated on protein content was quite high (between 48.3% for multi-breed population and 51.7% for Alpine one). This demonstrates that variability in protein content is not only explained by α_{s1} casein polymorphism. In deed a lot of other proteins were involved in milk composition like α_{s2}, β and κ caseins and lactoglobulin controlled by other genes than α_{s1} casein one [20].

Integrate α_{s1} casein genotype effect in a two steps model

Figure 2 shows the correlations between DYD in 2015 and (G)EBV predicted in 2013 for the 148 males born between 2010 and 2011 in four cases using multibreed populations: 1) using only pedigree information in relationship matrix without considering α_{s1} casein effect in the model, 2) using only pedigree information in relationship matrix and considering α_{s1} casein as a fixed effect, 3) using genomic (SNP genotypes) and pedigree information in relationship matrix without considering α_{s1} casein effect and 4) using genomic (SNP

genotypes) and pedigree information in relationship matrix and considering α_{s1} casein as a fixed effect. These results were presented for the group of traits among which the gene has an effect, i.e. production traits. Validation correlations obtained considering α_{s1} casein as random effect were similar to the one obtained when it was considered as fixed effect (results not shown). Integrate α_{s1} casein as a fixed effect in the genetic or genomic models improved significantly (Hotelling-Williams test, [34]) validation correlations for all traits except fat and protein yields. Validation correlations were improved from 6% for fat content using genomic information to 27% for protein content using only pedigree information. These results were consistent with results obtained in Lacaune meat sheep when FeCL gene was including in genetic selection model for prolificacy. In French meat sheep, including this gene effect allowed better predictions of EBVs especially for heterozygous females [35]. For fat and protein yields, traits for which no significant effect of α_{s1} casein was found, decrease of validation correlations (not significant) when adding α_{s1} casein effect in the genetic or genomic model was from 12% for fat yield to 21% for protein yield.

Improvement of validation correlation when adding α_{s1} casein effect in genetic evaluation was similar to the one found in genomic evaluation for fat content and milk yield. For protein content, no significant differences between validation correlation using SNP genotypes and α_{s1} casein information and validation correlation using only casein information was found.

Validation correlations estimated separately in each breed for multi-breed and per breed training populations using genomic and pedigree information with α_{s1} casein considered as a fixed effect in the model were in Table 3. Results were slightly better using Alpine and Saanen animals (multi-breed population) than using only Saanen animals in the training population to predict DYD of Saanen validation males. To predict DYD of Alpine males, the use of the multi-breed training population performed better for fat content and slightly for fat yield. But milk yield, protein yield and protein content in Alpine breed were better predicted

when using the Alpine training population. Differences between using per-breed or multi-breed training populations were higher for Alpine breed except for protein yield: from 1% for fat yield to 49% for fat content vs from 0% for fat yield to 8% for fat content in Saanen breed. Better validation correlations seem to be obtained with multi-breed population for all the traits in Saanen breed and only for fat content in Alpine breed. Multi-breed training population performed better for Saanen population due to lower frequency of some genotype at α_{s1} casein locus in the Saanen training population, like genotypes AA, AB or AC, (results not shown) compare to the Alpine one. These three genotypes considered as rare in Saanen training population were more frequent in Saanen validation population: 0.3% vs 3.9%, (results not shown). Their effects were then not well predicted in Saanen per-breed training population compared to in the multi-breed population.

Integrate probabilities of α_{s1} casein genotype in one step models

Table 4 shows validation correlations between DYD in 2015 and EBV predicted for protein content in 2013 for the males born between 2010 and 2011 in three cases: 1) multi-breed training and validation populations, 2) Saanen population and 3) Alpine population, for three models tested. The two first models tested were based on probabilities of α_{s1} casein genotypes for females with performances. In the first model ("random probabilities"), the combination of the probabilities for the 19 α_{s1} casein possible genotypes were considered as a random effect (with 3553 levels). The second model ("3 groups of probabilities") considered three groups of probabilities as described in Methods section. The third model ("gene content") was based on a gene content approach [12]. Validation correlations obtained with the "random probabilities" model were similar to the ones obtained using 3 groups of α_{s1} casein fixed effect. Results of prediction ability were higher (from +1% to +16%) using gene content approach than using other models. This result was consistent with Gengler et al.' study in which gene content approach based on a biallelic marker performed better than

iterative peeling one [12]. These Pearson correlations between predicted and known DYD for validation males were higher in Saanen breed than in Alpine one for the three models used. It could be explained by the highest level of inbreeding and relationship in this breed [14]. Except for Saanen breed, validation correlations using the two first models including α_{s1} casein effect ("random probabilities" and "3 groups of probabilities" models) did not exceed the ones obtained without α_{s1} casein effect in the model. Using gene content approach the correlations validation were slightly higher (from +1% in multibreed population to +9% in Saanen population) than the one obtained excluding this effect especially for per breed evaluations (Alpine or Saanen). This results is consistent with the one found in Canadian Hostein dairy cattle with an increase from 0.3% for somatic cell counts to 0.5% for milk yield [36]. Better correlations in this study could be due to the important effect of α_{s1} casein genotype on protein content (between 24% and 38% of part of phenotypic variance). However results found in Saanen breed using α_{s1} casein effect in the model were largely higher (from 7% to 9%) than the ones found without including this effect. These better results obtained in Saanen breed could be explained by the dispersion of α_{s1} casein allele frequencies in the population. Indeed, in Saanen population, genotype AE and EE were carried by almost 65% of the population which make the situation closer to a biallelic case, easier to predict.

This improvement in validation correlations when including α_{s1} casein effect in the model was lower than the one found in the genetic evaluation based on animals genotyped (two steps model, in this study). It could be explained by the difficulty to predict α_{s1} casein genotypes for females non genotyped especially when a high proportion (40% in French dairy goats) of females came from unknown parents. Even in a subpopulation of females with known parents, from which one third have at least one parent genotyped (results not shown), improvements of validation correlations were lower than the one expected according to results of the two steps model. Predict α_{s1} casein genotypes for non-genotyped animals even using

gene content approach is particularly difficult in this case due to the high number of allele considered. Indeed even with one parent genotyped, number of possible genotyped was too high to make a good prediction of the possible genotype for an individual. Moreover, several α_{s1} casein allele (A, B and C) have the same effect on phenotypes (protein content, fat content or milk yield) which make tough the prediction using gene content approach.

Conclusions

This study investigated how to include α_{s1} casein major gene effect in genetic evaluation of French dairy goats. First, a description of allele frequencies in the Alpine and Saanen populations showed differences between males and females, and between Alpine and Saanen animals. The α_{s1} casein genotype had an effect on several production traits in French dairy goats: milk yield, fat content and protein content with a high part of phenotypic variance explained by the α_{s1} casein genotype for protein content. Including an effect of α_{s1} casein in a genetic evaluation based only of males genotypes improved predictive ability even when genomic (SNP genotypes) information was taken into account. Two approaches were tested to include the effect of α_{s1} casein genotype in genetic evaluations based on female genotypes. The first approach was based on iterative peeling equations given probabilities at each possible genotype for each female. Predictive abilities found for the several models studied for this approach were similar. The second approach was based on gene content and performed better than the previous approach. Improvement of validation correlations due to inclusion of α_{s1} casein effect in the models were better for genetic evaluations based on genotyped animals due to the difficulty to predict multi-allelic genotype for ungenotyped animals. Differences of predictive ability between Alpine and Saanen breeds were found due to difference in genetic structure and allele frequencies.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

CC analyzed the data and wrote the paper. CC, CRG and HL interpreted the results. CRG and HL revised and improved the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

The authors thank the French Genovicap and Phenofinlait programs (ANR, Apis-Gène, CASDAR, FranceAgriMer, France Génétique Élevage, French Ministry for Agriculture) 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 of Agronomics Research (INRA) SELGEN program (XGen). We also thank the GenoToul bioinformatics facility in Toulouse for providing computing and storage resources. This study would not have been possible without the goat SNP50 BeadChip developed by the International Goat Genome Consortium (IGGC): www.goatgenome.org.

References

1. Palhiere I, BRochard M, MOAZAMI-Goudarzi K, Laloeë D, Amigues Y, Bed'hom B, Neuts É, Leymarie C, Pantano T, Cribiu EP, Bibé B, Verrier É: **Impact of strong selection for the PrP major gene on genetic variability of four French sheep breeds**. *Genet Sel Evol* 2008, **40**:663–680.

2. Nagy B, Fésüs L, Safar L: **Breeding for scrapie resistance and controll strategie in Hungary**. In Proceedings of the 56th European Federation for Animal Science (EAAP) meeting, 5-8 June 2005, Uppsala, Sweden; 2005.

3. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps**. *Genetics* 2001, **157**:1819–1829.

4. Nicholas FW: **Discovery, validation and delivery of DNA markers**. *Aust J Exp Agric* 2006, **46**:155–158.

5. Ducrocq V, Croiseau P, Baur A, Saintilan R, Fritz S, Boichard D: **Genomic evaluations using QTL information**. In proceedings of the 10th World Congress on Genetics Applied to Livestock Production (WCGALP),17 -22 August 2014, Vancouver, Canada; 2014.

6. Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, Druet T, Genestout L, Colleau JJ, Journaux L: **Genomic selection in French dairy cattle**. *Anim Prod Sci* 2012, **52**:115–120.

7. Hayes B, Hagesæther N, Ådnøy T, Pellerud G, Berg PR, Lien S: **Effects on Production Traits of Haplotypes Among Casein Genes in Norwegian Goats and Evidence for a Site of Preferential Recombination**. *Genetics* 2006, **174**:455–464.

8. VanRaden PM: **Efficient methods to compute genomic predictions**. *J Dairy Sci* 2008, **91**:4414–4423.

9. Christensen OF, Lund MS: **Genomic prediction when some animals are not genotyped**. *Genet Sel Evol* 2010, **42**:1–8.

10. Vitezica ZG, Elsen JM, Rupp R, Diaz C: **Using genotype probabilities in survival analysis: a scrapie case**. *Genet Sel Evol* 2005, **37**, 403–415.

11. Fernando RL, Stricker C, Elston RC: **An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops**. *Theor Appl Genet* 1993, **87**:89–93.

12. 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**. *Anim Int J Anim Biosci* 2007, **1**:21–28.

13. 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.

14. 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**:7294–7305.

15. 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.

16. Mahé MF, Manfredi E, Ricordeau G, Piacère A, Grosclaude F: **Effets du polymorphisme de la caséine αs1 caprine sur les performances laitières: analyse intradescendance de boucs de race Alpine**. *Genet Sel Evol* 1994, **26**:151.

17. Martin P, Leroux C: **Le gène caprin spécifiant la caséine αs1: un suspect tout désigné aux effets aussi multiple qu'inattendus**. *INRA Production Animale* 2000, *special issue : "Génétique moléculaire : principes et application aux populations animales"*:125-132.

18. Barbieri ME, Manfredi E, Elsen JM, Ricordeau G, Bouillon J, Grosclaude F, Mahé MF, Bibé B: **Effects of the αs1-casein locus on dairy performances and genetic parameters of Alpine goats.** *Genet Sel Evol* 1995, **27**:437–450.

19. Leroux C, Le Provost F, Petit E, Bernard L, Chilliard Y, Martin P: **Real-time RT-PCR and cDNA macroarray to study the impact of the genetic polymorphism at the alphas1-**

casein locus on the expression of genes in the goat mammary gland during lactation. *Reprod Nutr Dev* 2003, **43**:459–469.

20. Selvaggi M, Laudadio V, Dario C, Tufarelli V: **Major proteins in goat milk: an updated overview on genetic variability**. *Mol Biol Rep* 2014, **41**:1035–1048.

21. Grosclaude F, Ricordeau G, Martin P, Remeuf F, Vassal L, Bouillon J: **From gene to cheese: The caprine αs1-casein polymorphism, its effects and its evolution**. *INRA Production Animale* 1994, **7**:3-19.

22. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, Donnadieu-Tonon C, Eggen A, Heuven HC, Jamli S: **Design and Characterization of a 52K SNP Chip for Goats**. *PloS One* 2014, **9**:e86227.

23. Ducrocq V: *GENEKIT, BLUP Software*: June 2011 version. INRA SGQA; Jouy en Josas, France, 1998.

24. Fikse WF, Banos G: **Weighting factors of sire daughter information in international genetic evaluations**. *J Dairy Sci* 2001, **84**:1759–1767.

25. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH: **BLUPF90 and related programs (BGF90)**. In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production: 19-22 August 2002, Montpellier, France, 2002.

26. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness**. *Nat Genet* 2006, **38**:203–208.

27. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies**. *Nat Genet* 2010, **42**:348–354.

28. Legarra A, Vitezica ZG: **Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using Gene Content multiple Trait BLUP**. *Submitted to Genetics Selection Evolution* 2015.

29. Kennedy BW, Quinton M, Arendonk V, A J: **Estimation of effects of single genes on quantitative traits.** *J Anim Sci* 1992, **70**:2000–2012.

30. Maga EA, Daftari P, Kültz D, Penedo MCT: **Prevalence of alphas1-casein genotypes in American dairy goats**. *J Anim Sci* 2009, **87**:3464–3469.

31. Torres-Vazquez JA, Vazquez Flores F, Montaldo HH, Ulloa-Arvizu R, Valencia Posadas M, Gayosso Vazquez A, Alonso Morales RA: **Genetic polymorphism of the αs1-casein locus in five populations of goats from Mexico**. *Electronic Journal of Biotechonoly* 2008, ***11**:3, 1-11* 2008.

32. Soares MAM, Rodrigues MT, Mognol GP, Ribeiro L de FC, Silva JL da C, Brancalhão RMC: **Polymorphism of alpha s1-casein gene in a dairy goat herd in the southeastern region of Brazil**. *Rev Bras Zootec* 2009, **38**:1026–1032.

33. Gipson T: **Preliminary Observations: Inbreeding in Dairy Goats and Its Effects on Milk Production.** Proc. 17th Ann. Goat Field Day, Langston University, Langston, Etats-Unis, 2002.

34. Williams EJ: **The Comparison of Regression Variables**. *Journal of the Royal Statistical Society. Series B Methodol* 1959, **21**:396–399.

35. Martin P, Raoul J, Bodin L: **Effects of the FecL major gene in the Lacaune meat sheep population**. *Genet Sel Evol* 2014, **46**:48.

36. Gengler N, Abras S, Verkenne C, Vanderick S, Szydlowski M, Renaville R: **Accuracy of prediction of gene content in large animal populations and its use for candidate gene detection and genetic evaluation**. *J Dairy Sci* 2008, **91**:1652–1659.

Figures

Figure 1 - α_{s1} genotype frequencies for dam of bucks.
True Alpine and True Saanen correspond to frequencies of truly genotyped females, and Predicted Alpine and Predicted Saanen correspond to frequencies of α_{s1} genotype predicted using peeling equations

Figure 2 – Validation correlations[1] for the 148 validation males using or not α_{s1} casein genotype as fixed effect
[1]correlations between DYD in 2015 and GEBV in 2013. Pedigree without casein and Pedigree with casein correspond to a model respectively without or with α_{s1} casein effect using only pedigree to construct relationship matrix. Genomic without casein and Genomic with casein correspond to a model respectively without or with fixed effect of α_{s1} casein genotype using pedigree and SNP genotype information to construct relationship matrix.

Tables

Table 1 - α_{s1} genotype frequencies for the males[1] and the females[2]
[1] Two populations of males were considered: 1) the 823 progeny tested and genotyped on 50K BeadChip, 2) the 2 538 other male candidates to progeny testing
[2]Females considered were the 2 949 dams of bucks

Table 2 - Part of phenotypic variance explained by polygenic and α_{s1} casein effects in multi-breed, Alpine and Saanen populations

Table 3 - Validation correlations[1] for validation males using α_{s1} casein effect[2] in multi-breed, Saanen and Alpine populations
[1]correlations between DYD in 2015 and GEBV in 2013. [2] α_{s1} casein was considered as fixed effect. Per-breed Alpine and Per-breed Saanen correspond respectively to Alpine training population to predict Alpine validation population and Saanen training population to predict

Saanen validation males. Multi-breed Alpine and Multi-breed Saanen correspond to multi-breed training population to predict Alpine and Saanen animals respectively.

Table 4 - Validation correlations[1] for the 148 validation males for models based on females performances (one step) for protein content.

[1]correlations between DYD in 2015 and EBV in 2013. The random probabilities model corresponds to the model using the combination of the 19 α_{s1} casein possible genotypes as a random effect. The 3 groups of probabilities model corresponds to a model where the effect of the 3 groups of possible genotypes (strong, mean and weak effect on protein content) was considered as fixed effect. The gene content model correspond the model using gene content approach without using predicted probabilities of α_{s1} casein genotypes for females. Multi-breed results consisted in training and validation multi-breed (Alpine+Saanen) populations. Alpine and Saanen results consisted in training and validation populations from Alpine and Saanen breeds respectively.

Table 1

α_{s1} genotype	Progeny tested males	Other males	Females
AA	31.2	27.2	17.6
AE	27.0	25.8	39.1
EE	13.7	14.9	8.3
AB	11.4	5.5	8.4
BE	4.0	0.2	2.4
AF	3.9	8.2	9.3
EF	3.0	8.0	7.3
AC	2.9	2.5	3.9
BB	0.7	0.3	0.3
BC	0.5	0.2	0.3
CE	0.5	0.8	0.8
BF	0.4	3.7	1.4
CF	0.4	0.2	0.3
AO	0.1	0.3	0.2
CC	0.1	0	0.1
FF	0.1	1.1	0.4
BO	0	0.1	0
EO	0	0.3	0
FO	0	0.1	0

Table 2

	Multi-breed		Alpine		Saanen	
	casein α_{s1}	polygenic	casein α_{s1}	polygenic	casein α_{s1}	polygenic
Milk yield	4.6	46.0	6.1	43.1	3.3	47.0
Fat content	13.7	54.0	18.2	56.5	8.7	43.7
Protein content	33.8	48.3	38.2	51.7	24.4	40.7

Table 3

	Per-breed Alpine	Per-breed Saanen	Multi-breed Alpine	Multibreed Saanen
Milk yield	0.338	0.324	0.328	0.333
Fat yield	0.269	0.204	0.271	0.205
Protein yield	0.363	0.178	0.264	0.269
Fat content	0.232	0.360	0.346	0.390
Protein content	0.470	0.690	0.452	0.703

Table 4

	random probabilities	3 groups of probabilities	gene content	without α_{s1} casein
Multi-breed	0.64	0.63	0.73	0.72
Alpine	0.64	0.62	0.67	0.63
Saanen	0.81	0.80	0.82	0.75

figure1-carillier.pdf

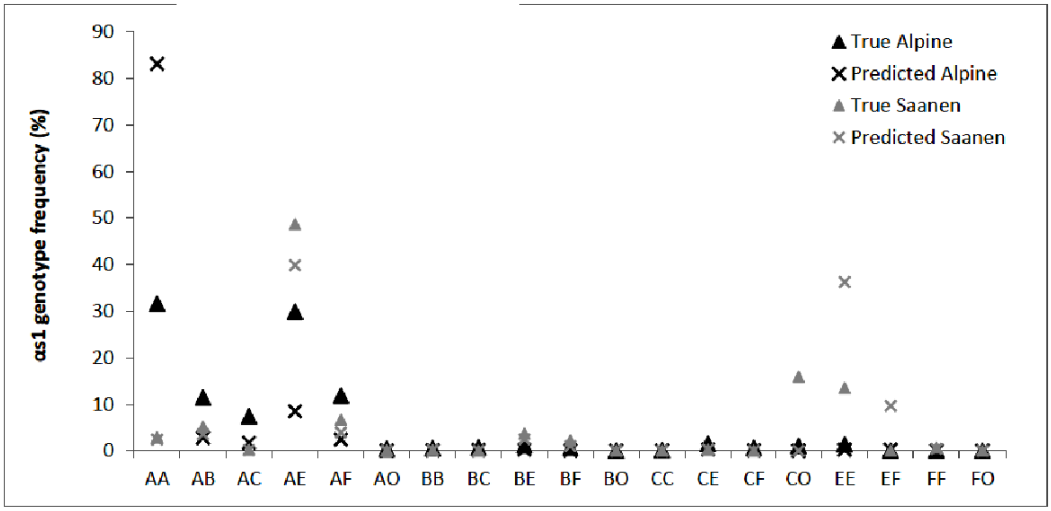
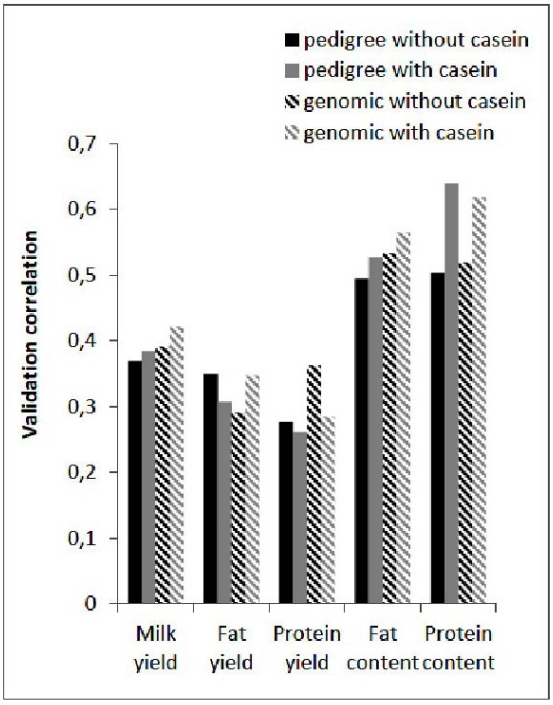


Figure
Click here to download Figure: figure2-carillier.pdf



Bilan

Dans la population caprine française étudiée, le génotype au locus de la caséine α_{s1} a un effet significatif sur les caractères quantité de lait, taux butyreux et taux protéique. Ce génotype n'a pas d'effet sur les autres caractères indexés (matières grasse et protéique, comptage de cellules somatiques et caractères de morphologie mammaire). Les parts de variance phénotypique expliquée par ce génotype sont estimées entre 3,3% pour le lait en race Saanen et 38,2% sur le TP en race Alpine.

Six allèles (et 19 génotypes) sont présents dans la population caprine française (A, B, C, E, F et O) sur les 17 existants dans l'espèce caprine. Les génotypes les plus représentés sont les génotypes AA et AE, avec plus que 50% des animaux génotypés. Il existe des différences entre les fréquences alléliques trouvées dans les populations femelles et mâles, ainsi qu'entre la population sélectionnée (mâles testés sur descendance et mères à bouc) et la population des mâles candidats au testage mais non retenus. Les fréquences alléliques obtenues en race Saanen sont différentes de celles obtenues en race Alpine notamment pour les génotypes AA, AE, EE et CO.

Pour les évaluations basées sur les pseudo-performances (two steps), les corrélations de validation obtenues en considérant le génotype de la caséine α_{s1} soit en effet fixe, soit en effet aléatoire sont similaires. L'inclusion de l'effet de ce génotype dans les évaluations two steps génétiques permet d'améliorer les corrélations de validation de 6% pour le taux butyreux à 27% pour le taux protéique. L'augmentation des corrélations est légèrement moins élevée lorsque l'effet polygénique est estimé à partir de l'information des typages SNP de la puce 50K. La prise en compte de l'effet de ce gène majeur dans les évaluations génétiques des caractères de production pour lesquels il n'a pas d'effet (matière grasse et protéique), ne perturbe pas les prédictions des valeurs génétiques des mâles de validation. Les plus fortes corrélations pour la population de validation Saanen sont obtenues en utilisant la population d'apprentissage multiraciale et non la population Saanen. Ce n'est pas le cas de la population Alpine, pour laquelle il semble plus judicieux d'utiliser la population d'apprentissage constituée d'animaux de race Alpine seulement. Ces différences de prédiction entre races Alpine et Saanen peuvent être expliquées par les différences entre fréquences alléliques observées entre les deux races.

La prédiction des génotypes au locus de la caséine α_{s1} pour les femelles non génotypées ne semble pas optimale. En effet, les fréquences alléliques des génotypes prédits avec la méthode « iterative peeling » ne correspondent pas aux fréquences alléliques calculées

dans la population génotypée notamment pour les allèles rares (C et O). Des différences entre fréquences génotypiques réelles et reconstruites sont également observées avec la méthode gene content mais sont différentes de celles observées avec la méthode « iterative peeling » (résultats non montrés). La différence moyenne sur l'ensemble des génotypes, entre fréquences des génotypes réels et prédits/reconstruits est similaire pour la race Saanen quelle que soit la méthode de prédiction utilisée (gene content ou iterative peeling). En revanche, cette différence moyenne est plus élevée pour la race Alpine avec la méthode « iterative peeling » qu'avec la méthode gene content (5,5% vs 2,3%). De plus, les fréquences obtenues en regroupant les génotypes (somme des fréquences de chaque génotype) reconstruits avec la méthode gene content par classe d'effet (forts : AA, AB, AC, BB, BC, CC, moyens : AE, AF, AO, BE, BF, BO, CE, CF, CO et faibles : EE, EF, FF, FO) sont plus proches des fréquences des génotypes réels, de 0,2% en race Alpine à 30,3% en race Saanen pour les génotypes à effets faibles . Cette méthode utilisant le TP comme variable corrélée prédit sans doute mieux les génotypes en fonction de leur effet.

Pour le taux protéique, les corrélations de validation obtenues dans le cas des évaluations génétiques (single step) utilisant les probabilités au génotype de la caséine α_{s1} pour les femelles non génotypées sont similaires quel que soit le modèle utilisé. Cependant, les résultats obtenus avec l'approche gene content sont meilleurs dans le cas de la population multiraciale. L'inclusion de l'effet du génotype caséine dans le cas des évaluations génétiques single step ne permet d'augmenter significativement la précision des prédictions que dans le cas de la population Saanen. Les difficultés rencontrées pour prédire un gène multi-allélique pour une population avec une forte proportion de parents inconnus peuvent expliquer les résultats décevants obtenus dans le cas de ces évaluations basées sur les performances brutes de toutes les femelles indexées.

Cette étude a permis de montrer qu'il existe un intérêt à ajouter les effets des gènes majeurs dans les évaluations génétiques et génomiques surtout lorsqu'il est difficile de les capter avec les marqueurs SNP. Cependant, dès lors que l'information sur le gène est manquante pour de nombreux animaux (et le pédigrée mal connu), les méthodes de prédictions des génotypes multi-alléliques ne sont pas assez précises pour que l'inclusion d'un effet du génotype puisse permettre d'améliorer la qualité des prédictions.

Conclusions et perspectives

I. Bilan des principaux résultats

Dans un contexte très marqué par la mise en place de la sélection génomique chez les bovins laitiers, l'arrivée de la puce 50k caprine a poussé la filière à s'interroger sur l'intérêt de la sélection génomique dans cette espèce. Etant donné la petite taille des populations caprines (30 et 40 mâles de race Saanen et Alpine testés sur descendance chaque année), un des premiers objectif de cette thèse était de caractériser la structure génétique de la population de référence. Cette étude a permis d'évaluer les atouts et les faiblesses de la population de référence pour la sélection génomique. Le second objectif de ma thèse était de définir le modèle et la méthode d'évaluation génomique les plus adaptés à la population caprine française c'est-à-dire ceux permettant d'obtenir les meilleures précisions génomiques.

Une première étude a montré que même si elle est représentative de la population en sélection, la population de référence n'a pas une structure optimale pour maximiser les précisions génomiques. En effet, le niveau de déséquilibre de liaison entre deux SNP est plutôt faible, la taille efficace élevée et la parenté entre la population candidate et la population de référence sont relativement moyens comparés aux grandes races bovines laitières. Cependant, bien que la taille de la population de référence caprine soit plus petite, sa structure génétique est comparable à celle de la population ovine Lacaune qui n'a pas été un d'obstacle à la mise en place de la sélection génomique. Enfin, bien qu'ayant une origine commune, les races Alpine et Saanen se distinguent clairement aujourd'hui d'un point de vue génomique (génotypes et fréquences alléliques), ce qui est un argument en défaveur d'évaluations multiraciales.

Dans un second temps, des évaluations génomiques two steps, basées sur des performances pré-corrigées ont été réalisées à partir de la population de référence caprine. Nous avons pu montrer que les précisions des évaluations génomiques sont similaires quels que soit la méthode (Lasso bayésien, BayesCπ ou GBLUP) ou les phénotypes utilisés (DYD, EBV dérégressés corrigés ou non). Les CD génomiques avec la méthode two steps n'atteignent pas le niveau de précision obtenu sur ascendance avec les évaluations génétiques classiques. Ces précisions sont cependant améliorées par l'ajout des femelles du dispositif de détection de QTL dans la population de référence. L'ajout des phénotypes (DYD) des mâles non génotypés (pseudo single step), permet une nette amélioration des précisions des évaluations génomiques pour les caractères de production, qui surpassent alors les précisions

sur ascendance obtenues en évaluation génétique classique. Les évaluations génomiques basées sur les performances brutes des femelles (single step) ont permis d'obtenir les meilleures précisions, supérieures de 1 à 14% selon le caractère considéré par rapport aux évaluations two steps. Les évaluations multiraciales ont été comparées aux évaluations uniraciales pour les trois méthodes (two steps, pseudo single step et single step) ainsi qu'aux évaluations bicaractères pour le single step. Les précisions des évaluations génomiques obtenues pour les évaluations multiraciales, uniraciales et bicaractère sont similaires malgré la petite taille de la population de référence considérée dans les évaluations uniraciales. Cette étude a également pu montrer qu'il n'est pas envisageable d'utiliser une population de référence uniraciale pour prédire les valeurs génomiques des animaux de l'autre race (Alpine pour prédire les Saanen par exemple).

Enfin, plusieurs modèles ont été testés afin d'intégrer dans les évaluations génétiques et génomiques le gène majeur de la caséine α_{s1} qui a un effet sur le TP (entre 24 et 38% de la variance phénotypique en Saanen et Alpine respectivement), le TB (entre 8 et 18% de la variance phénotypique) et la quantité de lait (entre 3 et 6% de la variance phénotypique). Prendre en compte l'effet de ce gène en effet fixe dans des évaluations pseudo single step permet d'améliorer les précisions des évaluations génétiques et génomiques. Cependant, intégrer un effet de ce gène dans les évaluations single step suppose de prédire les génotypes des femelles non génotypées. Nous avons utilisé deux méthodes : i) l'une consistant en une étape préalable de prédiction des génotypes manquants (« peeling » itérative), ii) et l'autre basée sur un modèle « gene content » proposé par Gengler et al. (2008) et adapté par Legarra et Vitezica (2015) qui permet d'analyser simultanément le caractère d'intérêt et d'estimer une valeur génétique des animaux pour chaque allèle du génotype. Prédire les génotypes des femelles (qui n'ont parfois aucun parent génotypés) est difficile et ne permet pas d'obtenir d'importants gains de précision avec l'ajout de l'effet du gène de la caséine α_{s1} dans les évaluations single step multiraciales. Cependant, un gain de 8% a été constaté en race Saanen sur le caractère taux protéique avec un modèle d'évaluation génomique single step uniracial et la méthode du gene content.

L'ensemble des résultats de mon travail de thèse ont permis de montrer que malgré une structure génétique non idéale de la population de référence caprine française, les précisions obtenues avec la méthode single step multiraciale ou uniraciale permettent d'envisager la sélection génomique dans cette espèce. Il est de plus possible d'améliorer ces précisions en incluant l'effet de gènes majeurs tels que celui de la caséine α_{s1}.

En perspective de ce travail, j'évoquerais quelques pistes à poursuivre pour mettre en place une sélection génomique efficace chez les caprins laitiers français. Nous discuterons de stratégies de génotypages, d'exploration de nouveaux modèles et méthodes d'évaluation ainsi que des changements à envisager dans le schéma de sélection caprin.

II. Augmenter la précision des évaluations génomiques

II.1. Explorer différentes stratégies de génotypages

II.1.1. Génotypage de femelles

Nous avons vu que la taille de la population de référence ainsi que le nombre de SNP considérés ont une influence sur les précisions des évaluations génomiques. La stratégie de génotypage mise en œuvre dans les races Saanen et Alpine en France est de génotyper tous les mâles des séries de testage (incluant les mâles testés et les mâles agréés) depuis les années 2000, ainsi que tous les anciens mâles agréés nés entre 1993 et 2000. Le nombre de mâles considérés dans la population de référence caprine est donc maximal et ne pourra être augmenté que par le génotypage des nouveaux mâles candidats à la sélection. Les génotypages sont actuellement réalisés à l'aide de la seule puce caprine disponible en 2011, la puce Illumina goat SNP50k BeadChip comptant environ 54 000 marqueurs. La taille de la population de référence caprine semble être un frein à l'augmentation conséquente des précisions des évaluations génomiques. Pour de telles petites populations, il existe peu de stratégies permettant d'augmenter la taille de population de référence. La première qui a été testée est de considérer une population multiraciale (Alpine + Saanen) mais elle ne permet pas vraiment d'augmenter les précisions des évaluations. La deuxième est l'ajout de génotypes femelles dans la population de référence. L'ajout des génotypes de femelles du dispositif de détection de QTL, étudié dans cette thèse, montre (chapitres 3 et 4) que les précisions des GEBV des candidats estimées à partir des PEV, peuvent être augmentées de 7 à 15% avec l'approche single step.

Une stratégie pour augmenter la taille de la population de référence caprine serait donc de génotyper des femelles. Le choix des femelles à génotyper n'est cependant pas aisé. En effet, les femelles du dispositif de détection de QTL sont fortement apparentées aux mâles de la population de référence (100 filles par père pour 20 pères de la population). L'ajout des génotypes de ces femelles apporte donc *a priori* peu d'information supplémentaire par rapport aux génotypes des mâles d'IA, sauf pour ce qui est de la part génétique de leurs mères. Les mères de ces femelles étant en grande partie des femelles issues d'IA, la variabilité génétique non encore captée par les génotypes des mâles d'IA remonte à plusieurs générations

antérieures (femelles dans l'ascendance qui seraient non issues d'IA). Le génotypage de l'ensemble des mères à bouc, toutes issues d'IA, ne serait donc pas non plus l'idéal de ce point de vue. Jiménez-Montero et al. (2010) ont montré que l'ajout des meilleures femelles dans la population de référence ne permet pas d'obtenir de meilleures précisions que l'ajout de femelles choisies au hasard. Ce phénomène peut principalement s'expliquer par le fait que choisir des femelles au hasard permet de maximiser la diversité génétique, ce qui n'est pas le cas quand on choisit les meilleures femelles. En effet, elles sont souvent issues d'un faible nombre de reproducteurs différents. Chez les caprins, les mères à boucs étant les mères des jeunes candidats à prédire, elles semblent essentielles à la précision des GEBV de ceux-ci. Ainsi, une solution envisageable serait d'identifier les mères à boucs (parmi les 1 060 femelles sélectionnées par an pour réaliser les accouplements programmés (Renard, 2008)) les moins apparentées avec les mâles d'IA afin de ne génotyper que ces femelles.

Le génotypage des mères à boucs les moins apparentées aux mâles d'IA pourrait être complété par le génotypage de quelques femelles clés (non mères à boucs) permettant de maximiser la diversité génétique dans la population de référence. Ces femelles clés pourraient être des demi-sœurs de candidats, donc filles de mâles d'IA dont la mère est issue de monte naturelle ou de père inconnu. Les femelles de père inconnu sont en fait issues d'un mâle de monte naturelle non identifié. Les mâles de monte naturelle étant très souvent issus d'IA, la variabilité génétique apportée par l'inclusion du génotypage de ces femelles dans la population de référence risque d'être faible. Afin d'identifier facilement les femelles les plus intéressantes d'un point de vue génétique, il serait plus stratégique de développer un outil d'assignation de parenté (Barbotte et al., 2012). Cet outil parait d'autant plus intéressant que le nombre de femelles issues de pères inconnus est en constante augmentation (cf. 1.I.2). La conception d'un tel outil est actuellement à l'étude dans le cadre d'un projet de recherche et de développement de l'INRA (OPA-Action Innovante, Isabelle Palhière, INRA, communication personnelle).

II.1.2. Constitution d'une population internationale

Une autre stratégie pour augmenter la taille de la population de référence serait d'envisager une collaboration internationale pour la mutualisation des populations de référence comme cela a été mis en place en bovins laitiers Holstein (Consortium Eurogenomics). La constitution de ce consortium en bovins a permis de constituer une population de taille 3 à 4 fois supérieure aux populations nationales. Les précisions des évaluations génomiques ont ainsi pu être augmentées de 8 à 12% (Lund et al., 2010 et 2011).

Chez les caprins, les races Saanen et Alpine ne font pas l'objet de programmes de sélection dans beaucoup de pays, ils sont essentiellement concentrés en Europe et en Amérique du Nord. Des collaborations pourraient être envisagées avec le Canada, dont 403 femelles Alpines et 318 Saanens sont actuellement génotypées avec la puce 50k d'Illumina (Brito et al., 2014), les Etats-Unis (Luo et al., 1997; Kennedy et al., 1982), l'Italie (Frattini et al., 2014), le Brésil dont la majeurs partie des animaux sont à l'origine importés de France (Araujo et al., 2006) ou la Grèce, mais dont seulement 0,1% des animaux sont de race Saanen ou Alpine (Katanos et al., 2005). L'Iran est également en train de constituer une population de chèvres laitières Alpine et Saanen conduites et sélectionnées en race pure (Nejhad et al., 2013), et envisage de génotyper leurs animaux (Abdollah Rezagholivand, université de Téhéran, communication personnelle). L'intérêt d'une population de référence caprine internationale repose sur la proximité génétique entre les populations. La constitution d'une telle population nécessiterait dans un premier temps d'étudier la distance génétique entre ces populations afin de savoir si les animaux de race Alpine et Saanen élevés dans les autres pays sont génétiquement proches de ceux élevés en France. Au vu de ces résultats, nous pourrions ensuite envisager une évaluation génomique commune, il faudra alors s'interroger sur le type de données phénotypiques qui pourront être mises en commun (DYD, EBV, performances brutes, …). Cependant, un organisme tel qu'Interbull en bovins laitiers n'étant pas encore créé en caprins, l'utilisation d'une telle population internationale pour l'évaluation génomique ne sera pas aussi facile.

D'autres collaborations pourraient être envisagées avec l'Ecosse dont la majorité des cheptels sont constitués d'animaux croisés entre les races Alpine et Saanen (Mucha et al., 2014), et avec le Mexique où les animaux sont issus de croisement entre les races Alpine ou Saanen avec des races locales (Torres-Vazquez et al., 2008). La distance génétique entre les races locales mexicaines et les races françaises étant probablement élevée, l'intérêt d'une collaboration avec le Mexique reste limité. En revanche, une collaboration avec l'Ecosse pourrait être plus intéressante dans la mesure où les animaux de races pures utilisés pour le croisement sont en partie d'origine française (Joanne Conington, Scotland's Rural College, communications personnelles). D'après Toosi et al. (2010), l'utilisation d'une population de référence constituée d'animaux croisés (que l'on cherche à améliorer) pour prédire des candidats de race pure (qui sont habituellement sélectionnés) permet d'obtenir des précisions similaires à celles obtenues dans le cas d'une population de référence constituée d'animaux de race pure. Nous avons montré que l'utilisation d'une population de référence multiraciale (Alpine + Saanen) ne dégradait pas les précisions comparées à celles obtenues avec des

évaluations génomiques uniraciales. Cependant dans le cas d'une éventuelle évaluation génomique basée sur une population composée d'animaux de 2 races pures et d'animaux croisés, une étude devrait être conduite afin d'identifier la parenté qui peut être capté à l'aide des animaux croisés écossais.

II.1.3. Utilisation de puces de différentes densités

Au démarrage du projet, et en raison des coûts élevés de génotypage avec la puce 50k, il était envisagé de génotyper des femelles sur une puce basse densité. Cette stratégie a été adoptée en bovins laitiers (Dassonneville et al., 2012a). Cependant, étant donné le prix récemment proposé pour la puce Illumina 50k caprine (44€), cette stratégie n'a pas été retenue (IGGC, Gwenola Tosser-Klopp, INRA, communication personnelle). Il est désormais envisagé de développer des puces SNP haute densité et d'exploiter les séquences complètes de quelques animaux. En effet, en utilisant un nombre élevé de marqueurs répartis sur tout le génome, il est plus probable de détecter des SNP fortement associés à un caractère d'intérêt. L'utilisation de ce type de puce permettrait également d'augmenter le déséquilibre de liaison entre les SNP et les QTL, donc d'accroitre les précisions des évaluations génomiques (cf.1.II.1), ce qui serait particulièrement intéressant dans le cas de populations de référence multiraciales. Erbe et al. (2012) ont montré une augmentation des précisions des évaluations génomiques de 3% pour la quantité de lait à 10% pour la MP avec l'utilisation d'une puce 624k au lieu d'une puce 39k dans le cas d'une population de référence multiraciale (Jersey + Holstein) de bovins laitiers. Cependant, Hozé et al. (2014b) ont montré qu'en bovins laitiers Normand, Holstein et Montbéliarde, les précisions ne sont que faiblement augmentées (entre 4 et 5%) avec l'utilisation d'une puce haute densité (777k) comparée à celle d'une puce 50k dans le cas des évaluations multiraciales.

L'idéal serait de disposer de la séquence complète des animaux permettant ainsi d'accéder à l'exhaustivité de leur information génomique. Il serait alors plus aisé d'identifier les gènes ou mutations causales impliqués dans la régulation des caractères d'intérêt et d'utiliser cette information pour prédire précisément les valeurs génétiques. En caprins, les 20 boucs du dispositif de détection de QTL sont actuellement séquencés. Cependant les techniques de séquençage étant relativement récentes et coûteuses, peu d'auteurs ont pu étudier l'intérêt d'utiliser de telles données pour l'évaluation génomique. Hayes et al. (2014) montrent un gain de précisions de 2 à 3% pour les caractères de production en bovins laitiers en passant d'une puce 800k aux données de séquence. L'utilisation de ces données de séquence comparée à celle d'une puce basse densité (7,5k non imputée) permet d'augmenter

les précisions des évaluations de 9% pour les caractères de production en bovins laitiers Holsteins (Pérez-Enciso et al., 2015).

Afin d'exploiter l'information génomique disponible issue de puces de densité différentes et de séquences, l'imputation (i.e. estimer les génotypes manquants) des données les moins denses (puces SNP) est nécessaire. Le principe de l'imputation est basé sur un modèle mathématique (chaines de Markov) permettant de prédire des haplotypes manquants à partir des haplotypes connus aux SNP voisins (Dassonneville, 2012). Les précisions d'imputation constatées chez les bovins laitiers américains se situent entre 94,2% et 98% pour l'imputation d'une puce 50k en une puce 800k (Erbe et al., 2012; VanRaden et al., 2011), elle est beaucoup plus forte chez les bovins laitiers (Hozé et al., 2013) et allaitants français (Dassonneville et al., 2012b) (entre 97,5% pour la race Simmental et 99,7% pour la race Normande). Le taux d'erreur d'imputation pour passer d'une puce basse densité à une puce haute densité est un peu plus élevé, environ 4% pour les bovins laitiers Holstein (Dassonneville et al., 2011). L'imputation des puces haute densité à partir des puces basse densité est cependant possible, avec des précisions allant de 90,5% pour une puce 3k à 96,1% pour une puce 6k dépendant de la méthode d'imputation utilisée (VanRaden et al., 2013). L'étude de Bouwman et al. (2014) a montré dans le cas de données simulées, qu'il est également possible d'imputer des données de séquences à partir de données de puce haute densité avec une précision de 89%.

La disponibilité simultanée des puces (50k, haute densité) et des données de séquence nécessite d'identifier les animaux à génotyper pour chaque support. Aux vues du coût de ces différents outils, la stratégie de génotypage à envisager serait de génotyper les candidats à la sélection ainsi que certaines femelles avec une puce 50k puis une fois leur descendance connue, de regénotyper les meilleurs mâles (futurs pères de candidats) sur une puce haute densité. La stratégie de séquençage adoptée en général est de séquencer les meilleurs mâles (mâles d'IA) connus sur descendance. Cependant le coup du séquençage pourrait conduire à ne séquencer qu'une partie des mâles d'IA, choisis de manière à obtenir la même diversité génétique que celle observée dans la population de référence actuelle.

Enfin, une fois l'ensemble des données imputées afin d'obtenir des données de séquences pour l'ensemble des animaux génotypés, celles-ci pourraient être exploitées pour détecter plus précisément les régions d'intérêt. Ces régions d'intérêt pourraient alors être intégrées dans un modèle d'évaluation génomique à déterminer.

II.1 Explorer de nouvelles modélisations

II.2.1. Modélisation basée sur la prise en compte d'un gène majeur

Nous avons vu dans ce manuscrit qu'il est possible d'intégrer les effets du gène majeur de la caséine α_{s1} afin d'augmenter la précision des évaluations génomiques pour les caractères sur lesquels ce gène a un effet (taux protéique en particulier). Un précédent travail de thèse réalisé par Cyrielle Maroteau (2014) a mis en évidence des QTLs ayant un effet sur les caractères de production et de morphologie dans l'espèce caprine.

Outre le gène majeur de la caséine α_{s1} détecté sur le chromosome 6 en race Alpine, la région du gène DGAT1 (pic à environ 11,5Mb sur le chromosome 14) a été détectée dans les deux races comme ayant un effet très fort (gène majeur) sur le TB. L'effet de substitution allélique de ce QTL est de 0,40 écart-type phénotypique en race Alpine contre 0,52 en race Saanen (Maroteau, 2014). Ce gène est connu pour son rôle dans la synthèse des triglycérides et donc des matières grasses du lait (Ozasa et al., 1989). Il a été identifié en bovins laitiers Holsteins (Grisart et al., 2002) et en ovins laitiers Manech tête rousse et tête noire et Lacaune (Barillet et al., 2005). L'effet du gène DGAT1 est désormais pris en compte dans les évaluations génomiques bovine laitières en France pour les races Holstein, Montbéliarde et Normande (Boichard et al., 2012). Chez les caprins laitiers français, deux mutations ont été identifiées pour ce gène, la mutation R396W présente en races Alpine et Saanen, ainsi que la mutation R251L qui n'a été observée que pour une seule famille de pères Saanen (Maroteau, 2014). La première mutation a un effet significatif et positif sur les caractères quantité de lait, MG et MP. L'effet sur le TB est significatif et négatif pour les deux mutations.

L'intégration de l'effet de ce gène majeur dans les évaluations génétiques et génomiques pour le TB pourra être envisagée selon l'approche « gene content » comme pour l'effet du gène de la caséine α_{s1}. Les génotypages pour le gène DGAT1 de l'ensemble des animaux génotypés sur puce 50k sont désormais disponibles et pourront être utilisés pour estimer la part de variance phénotypique expliquée par ce gène pour le TB notamment. Cependant, l'intégration de l'effet du gène DGAT1, gène plus simple que celui de la caséine αs1 puisque seulement triallélique, pourra être plus complexe dans la mesure où certains auteurs montrent qu'il existe un effet de dominance d'un des allèles du gène pour certains caractères (Molee et al., 2012) comme pour le TB (Koopaei et al., 2012; Kuehn et al., 2007) chez les bovins laitiers Holsteins. Un tel phénomène n'a pas été pris en compte dans les modélisations proposées au chapitre 5 de cette thèse. Cependant des modèles d'évaluations génétiques permettant d'intégrer un effet de dominance existent (Vitezica et al., 2013; Zhang et al., 2008) et pourraient être mis en œuvre. Ces auteurs montrent néanmoins que cette prise en compte

apporte une faible augmentation des précisions des évaluations (moins de 2% sur données simulées).

Streit et al. (2011) ont montré l'existence d'une interaction entre la part polygénique et le gène majeur DGAT1 en bovins laitiers Holstein pour le TB et le TP. Cette interaction s'explique d'un point de vue biologique puisque le gène DGAT1 agit sur l'activité enzymatique de la synthèse des triglycérides elle-même gouvernée par d'autres gènes. Cette interaction a pu être testée en comparant un modèle où le gène majeur DGAT1 était considéré comme un effet fixe, à un modèle où la part polygénique corrigée pour l'effet de DGAT1 était divisée en deux parties : une liée au premier allèle de DGAT1 ($\mathbf{u}_1$) et l'autre liée au deuxième ($\mathbf{u}_2$). Cette approche n'a pas été testée dans le cadre de ma thèse. Il pourrait donc être intéressant pour la suite d'envisager une telle modélisation pour la prise en compte du gène majeur de la caséine α_{s1} ou du gène DGAT1.

II.2.2. Modélisation prenant en compte plusieurs régions d'intérêt

Dans l'idéal, la connaissance de l'ensemble des gènes (mutations causales) affectant un caractère permettrait de faciliter la sélection génétique des animaux. En pratique, les gènes responsables sont rarement identifiés et l'on dispose seulement de quelques régions (QTL) ayant un effet sur les caractères. Des régions de ce type ont été détectées dans la cadre de la thèse de Cyrielle Marteau (2014) chez les caprins laitiers :

- Deux QTL ayant un effet sur le TP : un QTL en race Saanen correspondant à la zone des caséines β et α_{s2} sur le chromosome 6 et un QTL dans les deux races sur le chromosome 1 avec un effet de 0,30 écart-type phénotypique.
- Deux QTL ayant un effet sur la quantité de matière grasse: un QTL en races Saanen et Alpine sur le chromosome 21 avec un effet de 0,30 écart-type phénotypique et un QTL sur le chromosome 14 en race Alpine (0,29 écart-type phénotypique). Ce QTL a aussi un effet sur la quantité de lait, le TB et le TP.
- Un QTL ayant un effet sur le comptage de cellules somatiques a été mis en évidence sur le chromosome 19 en race Saanen. Cette zone, la première découverte comme ayant un effet sur la résistance aux mammites en caprins, est orthologue à une zone détectée sur le chromosome 11 en ovins (Rupp et al., 2014).
- Les caractères de morphologie mammaire sont également influencés par de nombreux QTL, notamment sur le chromosome 19. Sur ce chromosome, ces zones ont également un effet sur les caractères de production et de résistance aux mammites.

- Des QTL sur les chromosomes 3, 9 et 2 ont été détectés pour les caractères position du plancher et profil de la mamelle en race Alpine. Une région ayant un effet sur la forme de l'avant pis est située sur le chromosome 6, tandis qu'une autre ayant un effet sur l'orientation des trayons est localisée sur le chromosome 26 en race Saanen (Maroteau et al., 2014).

L'ensemble de ces QTL ou régions d'intérêt, pourrait être pris en compte dans les modèles d'évaluations génomiques. Une première solution serait de les intégrer dans un modèle two steps basé sur les phénotypes des seuls animaux génotypés. Dans ce cas, les effets des SNP détectés comme significatifs pourraient être intégrés en effet fixe ou aléatoire dans les modèles comme cela a été testé pour le gène de la caséine α_{s1} dans ma thèse. Une autre méthode appelée SAMG, pour sélection assistée par marqueurs génomiques, pourraient également être envisagée. Elle est mise en place depuis 2008 chez les bovins laitiers français de race Holstein, Normande et Montbéliarde et permet de capter 60 à 70% de la variabilité génétique. Ces évaluations génétiques combinent l'information de 300 à 700 QTL par caractère dont 20 à 40 gros QTL détectés par une approche LDLA (linkage disequilibrium and linkage analysis), les autres étant détectés par la méthode Elastic-Net (Boichard et al., 2012). Cette méthode d'évaluation génomique permet d'obtenir des précisions d'évaluations légèrement meilleures que celles obtenues avec un GBLUP two steps (entre 5% pour la MP et 13% pour le TB) sauf pour le TP (Boichard et al., 2012). Cependant Ducrocq et al. (2014) ont montré qu'en moyenne les précisions des évaluations sont similaires à celles obtenues avec le GBLUP. Ces méthodes pourraient être améliorées par la prise en compte des interactions entre QTL et entre QTL et part polygénique comme cela a été proposé pour l'effet du gène DGAT1 (cf. section II.1.1 de la conclusion).

Ces méthodes sont basées sur une approche two steps or nous avons montré dans cette thèse que l'approche d'évaluation génomique la plus adaptée à la population caprine est le single step. L'ensemble des chèvres caprines n'étant pas génotypées, il faudrait envisager une méthode pour prédire les allèles portés au QTL par ces femelles. Comme nous l'avons vu dans le chapitre 5, une approche de type « gene content » permet de prédire (bien que de façon imprécise) l'effet d'un QTL ou d'un gène majeur pour l'ensemble des femelles. Cette méthode semble cependant difficilement applicable pour des centaines de QTL car la prédiction n'est possible que QTL par QTL. Mulder et al. (2010) ont malgré tout adapté la méthode de Gengler et al. (2007, 2008) afin de permettre la prédiction simultanée de plusieurs haplotypes pour les animaux non génotypés. Cette méthode est cependant très gourmande en

temps de calcul car elle prend en compte un grand nombre de caractères (plusieurs dizaines de « gene content »). Des études supplémentaires devront être réalisées pour une application en caprins laitiers où des centaines de milliers de femelles devront être prédites à partir de seulement 825 mâles et 1 945 femelles génotypés.

Enfin, la prise en compte de régions d'intérêt dans les évaluations pourraient être envisagée à l'aide d'haplotypes plutôt que de SNP seuls, comme c'est le cas pour la méthode SAMG appliquée en bovins laitiers (Boichard et al., 2012). Les principaux avantages des méthodes basées sur les haplotypes sont de 1) mieux capturer le déséquilibre de liaison entre les SNP et les QTL (De Roos et al., 2011) et 2) de prendre en compte les possibles recombinaisons entre un QTL et un SNP (Calus et al., 2008). Quelle que soit la méthode utilisée pour définir les haplotypes (IBD (Calus et al., 2008; Croiseau et al., 2014; Villumsen and Janss, 2009), groupes de SNP (Moser et al., 2010) ou seuil de DL (Cuyabano et al., 2014)), leur utilisation dans des approches Bayésiennes basées sur les pseudo-phénotypes (DYD) ou les performances brutes permet d'augmenter les précisions d'évaluations génomiques de 2% à 9% pour la quantité de lait et la matière protéique en bovins laitiers (Makgahlela et al., 2013a). Ces méthodes présentent cependant deux difficultés, celle de trouver la taille idéale de l'haplotype à prendre en compte (Villumsen and Janss, 2009; Jonas et al., 2015), et celle de connaitre les génotypes pour l'ensemble des animaux phénotypés.

D'autres modélisations que celles étudiées dans cette thèse, sont donc envisageables afin de permettre une augmentation des précisions des évaluations génomiques en incluant l'information des zones d'intérêt. Un projet a actuellement été déposé dans le cadre du métaprogramme SELGEN (INCoMINGS) afin d'approfondir la question de l'inclusion des mutations causales dans les évaluations génomiques.

III. Evolutions des schémas de sélection caprins à envisager avec l'arrivée de la sélection génomique

Les intervalles de génération de la voie père sont plutôt longs en caprins (environ 4 ans pour l'intervalle père-fille et 5,5 pour l'intervalle père-fils). On pourrait donc imaginer que la sélection génomique permette de diminuer ces intervalles de génération, comme c'est le cas en bovins laitiers, en diffusant les mâles dès leur maturité sexuelle sur la base de leur index génomique. Malgré une maturité sexuelle précoce des boucs (5 mois (Leboeuf et al., 1998)), la période d'évaluation de leurs capacités sexuelles et de la qualité de leur semence est relativement longue et incompressible (entre 7 et 8 mois cf. Figure 6.1). Comme on peut le voir d'après le schéma des étapes clés de la vie d'un bouc testé sur descendance (Figure 6.1),

les boucs pourraient être agréés à 1 an et 7 mois (début du testage, Figure 6.1), ce qui réduirait considérablement les intervalles de générations de la voie père. Cependant, l'organisme de sélection caprin (Capgènes) ne souhaite pas diminuer les intervalles de génération de la voie père au détriment des précisions des valeurs génétiques estimées des mâles. Il souhaite donc conserver des coefficients de détermination très élevés (supérieur à 0,95) avant d'utiliser les mâles comme père à bouc c'est-à-dire attendre que les mâles soient testés sur descendance avec plus de 200 filles chacun. Ce type de scénario ne permet cependant pas d'augmenter plus le progrès génétique qu'un scénario où les mâles seraient utilisés dès la disponibilité de leur index génomique. De plus, ce scénario provoque un accroissement plus rapide du taux de consanguinité (Colleau et al., 2009). Une utilisation des mâles en ferme pour l'amélioration génétique des femelles est toutefois envisagée dès l'obtention de leur index génomique. Ainsi les mâles, utilisés de façon plus intensive que ce qu'ils l'étaient en période de testage sur descendance, obtiendront le nombre de filles nécessaires à l'agrément plus rapidement qu'en schéma de sélection classique. Les intervalles de générations de la voie père se verront donc légèrement réduits par le biais de la sélection génomique. Capgènes envisage également de revoir le schéma de sélection caprin afin de réduire la période de quarantaine. Elle peut actuellement s'étendre jusqu'à 8 mois pour certains mâles (Figure 6.1) car l'organisme de sélection attend d'avoir ramassé l'ensemble des mâles dans les élevages et une période d'au moins trois mois de quarantaine avant de les entrainer pour la collecte de semence. Il pourrait être envisagé de réaliser deux groupes de mâles pour l'entrainement à la collecte afin de réduire la période de quarantaine.

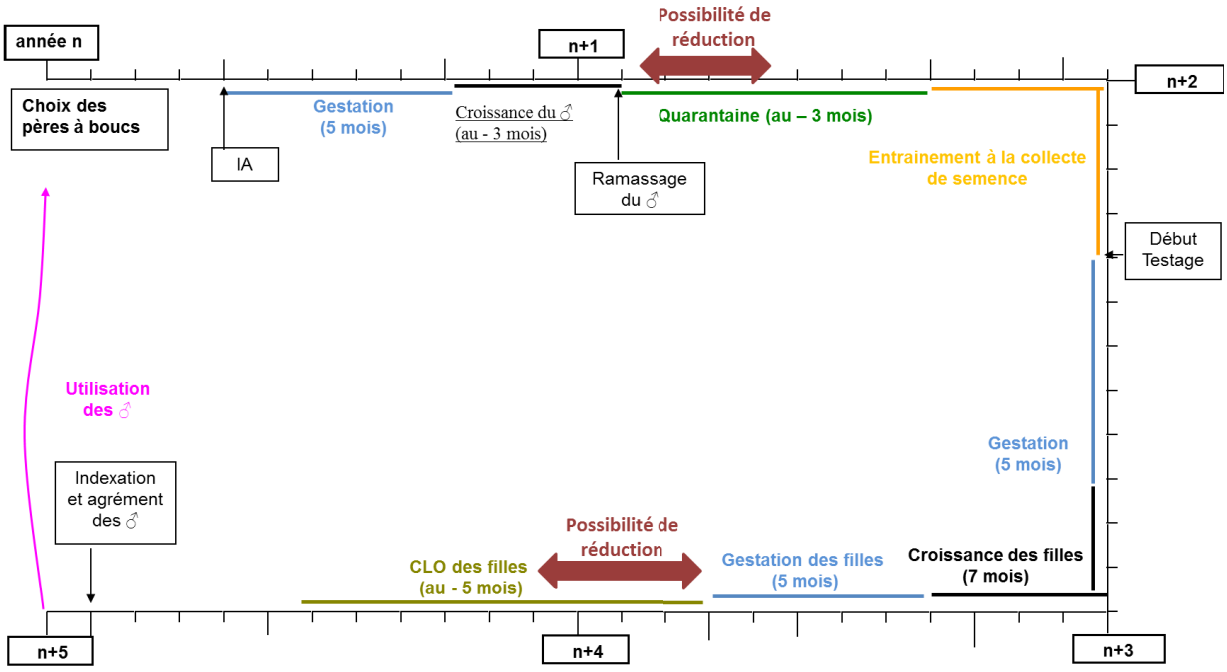


Figure 6.1 : Schéma des étapes clés de la vie d'un bouc testé sur descendance

Le génotypage des mâles candidats à la sélection permettra d'économiser les coûts de testage sur descendance des mâles notamment le coût d'entretien des mâles en attente d'index alors qu'ils ne sont pas utilisés. Contrairement au cas des ovins Lacaune où l'insémination animale est réalisée en semence fraiche, une sélection des mâles à la naissance permettrait de ne garder les mâles en station que le temps de produire suffisamment de doses d'IA. Le nombre de doses d'IA à réaliser par mâle est un des facteurs influençant la réduction du coût du testage avec la sélection génomique puisqu'elle détermine en partie le temps passé en station. Ce nombre est imposé par les besoins des éleveurs ainsi que par le nombre de mâles d'IA choisis pour le schéma. Ce nombre de mâles sélectionnés ainsi que le nombre de mâles candidats à génotyper (intensité de sélection) est un des seuls paramètres sur lesquels Capgène aura le plus de latitude pour améliorer le progrès génétique avec la sélection génomique. Le choix de l'intensité de sélection devra cependant tenir compte du coût de génotypage relativement élevé (70€), même s'il l'est moins que celui pratiqué en ovins laitiers lors de la mise en place de la sélection génomique (115€ (Buisson, 2012)). La réduction due à la suppression du testage qui pourra être observée en caprins pourrait permettre de financer l'augmentation de l'intensité de sélection. Des considérations économiques (coût du génotypage selon les différentes stratégies de génotypage citées plus haut) ainsi que le besoin des éleveurs en doses d'IA devraient permettre de faire un choix sur l'intensité de sélection à adopter pour un schéma génomique caprin.

L'arrivée de la sélection génomique en caprins laitiers ne permettra pas seulement d'augmenter le gain génétique annuel. En effet, elle peut également permettre d'intégrer de nouveaux caractères qui ne sont actuellement pas sélectionnés en raison de leur corrélation faiblement positive voire négative avec les caractères déjà en sélection ce qui engendrerait une diminution du progrès génétique pour les caractères principaux. Cette diminution du progrès serait contrebalancée par l'augmentation du progrès génétique engendré par la sélection génomique. En caprin, les nouveaux caractères envisagés pourraient être les caractères de fonction sexuelle et de morphologie des mâles. Le projet Maxi'mâle actuellement en cours (2015-2018) vise à collecter des phénotypes tels que le comportement des mâles (données de capteurs de mouvement et dosage hormonaux), les caractéristiques de production de semence (volume, motilité, capacité à congélation,…) et la morphologie (croissance, standard de race, malformations échographie testiculaire) afin de diminuer les pertes de candidats à la sélection due à des problèmes liés à la fonction sexuelle. Une des

tâches de ce projet consiste à utiliser les données actuellement collectées en centre de testage pour réaliser une évaluation génétique des futurs reproducteurs pour les caractères de production de semence.

La filière caprine est également intéressée par les caractères de robustesse qui pourraient être intégrés aux objectifs de sélection. Le projet RUSTIC (2016-2019) vise à étudier les caractères de longévité fonctionnelle et de persistance laitière dans le but de mettre en place une évaluation génétique de ces caractères. La variable de longévité considérée sera établie sur la base de données actuellement collectées (comme la date de naissance, les dates des évènements de production (mise bas, lactation), la date et les causes de réforme). Une évaluation de la longévité fonctionnelle est réalisée depuis les années 90 en bovins laitiers à l'aide du logiciel Survival kit (Ducrocq and Solkner, 1998). L'étude de la persistance et de la longévité devra également prendre en compte les phénomènes d'interaction génotype*milieu dont les effets sont particulièrement importants pour ce type de caractères.

Enfin, l'évaluation génétique d'autres caractères pourrait être envisages en caprins laitiers comme les caractères d'efficacité alimentaire (Puillet et al., 2012), ou de débit de traite (Ilahi et al., 1999), ou bien encore de fromageabilité ainsi que des caractères concernant le profil protéique du lait.

En conclusion, la sélection génomique chez les caprins laitiers est envisageable. Malgré la petite taille et la structure génétique non idéale de la population de référence, des modèles d'évaluations génomiques single step permettent d'obtenir des précisions acceptables y compris pour des évaluations uniraciales. Des études sont cependant nécessaires afin d'envisager d'autres stratégies de génotypage, de perfectionner le modèle d'évaluation et d'établir des scénarios d'évolution des schémas de sélection avec la sélection génomique.

Liste des tableaux

Liste des figures

Références

Abraham, K.J., L.R. Totir, and R.L. Fernando. 2007. Improved techniques for sampling complex pedigrees with the Gibbs sampler. *Genet. Sel. Evol.* 39:27–38.

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 score. *J Dairy Sci*. 93:743–752.

Araujo, A.M. de, S.E.F. Guimaraess, T.Mã.M. Machado, P.Sã¡. Lopes, C.S. Pereira, F.L.R. da Silva, M.T. Rodrigues, V. de S. Columbiano, and C.G. da Fonseca. 2006. Genetic diversity between herds of Alpine and Saanen dairy goats and the naturalized Brazilian Moxoto breed. *Genet. Mol. Biol.* 29:67–74.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, and J.L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitaive traits in elite north American oats. *Plant Genome*. 4:2:132–144.

Babo, D. 2000. Races ovines et caprines françaises. France Agricole Editions. 310 pp.

Badke, Y.M., R.O. Bates, C.W. Ernst, C. Schwab, and J.P. Steibel. 2012. Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics*. 13:24.

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

Barbieri, M.E., E. Manfredi, J.M. Elsen, G. Ricordeau, J. Bouillon, F. Grosclaude, M.F. Mahé, and B. Bibé. 1995. Effects of the αs1-casein locus on dairy performances and genetic parameters of Alpine goats. *Genet. Sel. Evol.* 27:437–450.

Barbotte, L., L. Genestout, S. Fritz, C. Chantry Darmond, and D. Boichard. 2012. Assignation de parenté par marqueurs SNP chez les bovins français. 19ème Rencontres des Recherches autour des Ruminants, 5-6 décembre 2012, Paris, France.

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. *Genet. Sel. Evol. GSE*. 37 Suppl 1:S109–123.

Baumung, R., and J. Sölkner. 2003. Pedigree and marker information requiremens to monitor genetic variability. *Genet Sel Evol*. 35:369–383.

Bélichon, S., E. Manfredi, and A. Piacère. 1999. Genetic parameters of dairy traits in the Alpine and Saanen goat breeds. *Genet Sel Evol*. 31:529–534.

Berry, D.P., F. Buckley, P. Dillon, R.D. Evans, and R.F. Veerkamp. 2004. Genetic relationships among linear type traits, milk yield, body weight, fertility and somatic cell count in primiparous dairy cows. *Ir. J. Agric. Food Sci. 43 161–176*.

Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *J Anim Breed Genet*. 129:345–358.

Boichard, D. 2006. Pedig, logiciel d'analyse de généalogies adapté à de grandes populations. INRA SGQA.

Boichard, D., N. Bouloc, G. Ricordeau, A. Piacere, and F. Barillet. 1989. Genetic parameters for first lactation dairy traits in the Alpine and Saanen goat breeds. *Genet. Sel. Evol.* 21:205–215.

Boichard, D., S. Fritz, M.N. Rossignol, M.Y. Boscher, A. Malafosse, and J.J. Colleau. 2002. Implementation of marker-assisted selection in French dairy cattle. 7ème World Congress of Genetics Applied to Livestock Production, 19-23 août 2002, Montpellier, France.

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.

Bouwman, A.C., J.M. Hickey, M.P. Calus, and R.F. Veerkamp. 2014. Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. *Genet. Sel. Evol.* 46:6.

Brard, S., and A. Ricard. 2014. Is the use of formulas a reliable way to predict the accuracy of genomic selection? *J Anim Breed Genet*. 132(3):207-17.

Brito, L.F., M. Jafarikia, D.A. Grossi, L. Maignel, M. Sargolzaei, and F.S. Schenkel. 2014. Characterization of Linkage Disequilibrium and Consistency of Gametic Phase in Canadian Goats. Genetics, 12(80):1-10.

Brøndum, R.F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen, W.F. Fikse, and M.S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. *J Dairy Sci*. 94:4700–4707.

Browning, S.R., and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data: inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 81:1084–1097.

Buch, L.H., Sørensen, M. K., Lassen, J., Berg, P., Sørensen, A. C. 2010. Dairy cattle breeding schemes with or without genomic selection and progeny testing. 9ème World Congress on Genetics Applied to Livestock Production, Leipzig, Allemagne, 2-6 août, 2010.

Buisson, D. 2012. Sélection génomique des races ovines laitières françaises : analyse des schémas actuels, première modélisation de scenario génomiques et bila technico-économique. Rapport de stage de fin d'étude d'ingénieur agronome, Agrocampus Ouest.

Calus, M.P.L., Y. De Haas, M. Pszczola, and R.F. Veerkamp. 2011. Predicted response of genomic selection for new traits using combined cow and bull reference population. Interbull Meeting, Stavanger, Norvège, 27-28 août, 2011.

Calus, M.P.L., A.P.W. de Roos, and R.F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*. 178:553–561.

Calus, M.P.L., and R.F. Veerkamp. 2011. Accuracy of multi-trait genomic selection using different methods. *Genet Sel Evol*. 43:1–14.

Campos, G. de los, H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics*. 182:375–385.

Carillier, C. 2012. Evaluation of a reference population in dairy goats for genomic selection. Rapport de stage de fin d'étude de master, AgroParis Tech.

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.

Christensen, O.F. 2012. Compatibility of pedigree-based and marker-based relationship matrics for single-step genetic evaluation. *Genet Sel Evol*. 44:1–10.

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H.J. Van Der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet Sel Evol*. 44:1–24.

Clément, V. 2012. Indexation caprine : Des nouveautés dans les index de synthèse caprins. *Note d'information aux organismes de l'élevage caprin*, 11.

Clément, V., D. Boichard, A. Piacere, A. Barbat, and E. Manfredi. 2002. Genetic evaluation of French goats for dairy and type traits. 7ème World Congress of Genetics Applied to Livestock Production, 19-23 août 2002, Montpellier, France.

Clément, V., H. Caillat, A. Piacère, E. Manfredi, C. Robert-Granié, F. Bouvier, and R. Rupp. 2008. Vers la mise en place d'une sélection pour la résistance aux mammites chez les

caprins. 15ème Rencontres des Recherches autour des Ruminants, 3-4 décembre 2008, Paris, France.

Clément, V., I. Palhiere, and H. Larroque. 2014. Évaluation génétique dans l'espèce caprine : Caractères de production laitière, de comptage de cellules somatiques et de morphologie. Compte-rendu n°00 14 202 041, Institut de l'élevage collection. résultats.

Colleau, J.J., S. Fritz, F. Guillaume, A. Baur, D. Dupassieux, M.Y. Boscher, L. Joumaux, A. Eggen, and D. Boichard. 2009. Simulation des potentialités de la sélection génomique chez les bovins laitiers. 16ème Rencontres des Recherches autour des Ruminants, 2-3 décembre 2009, Paris, France.

Colleau, J.J., S. Mattalia, and M. Brochard. 2004. A method for the dynamic management of genetic variability in dairy cattle. *Genet Sel Evol*. 36:373–394.

Colombani, C., A. Legarra, S. Fritz, F. Guillaume, P. Croiseau, V. Ducrocq, and C. Robert-Granié. 2013. Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. *J. Dairy Sci.* 96:575–591.

Croiseau, P., M.-N. Fouilloux, D. Jonas, S. Fritz, A. Baur, V. Ducrocq, F. Phocas, and D. Boichard. 2014. Extension to haplotypes of genomic evaluation algorithms. 10ème World Congress of Genetics Applied to Livestock Production, 17-22 août 2014, Vancouver, Canada.

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.

Daetwyler, H.D., K.E. Kemper, J.H.J. van der Werf, and B.J. Hayes. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. *J. Anim. Sci.* 90:3375–3384.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*. 185:1021–1031.

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*. 3:1–10.

Danchin-Burge, C. 2009. Estimation de la variabilité génétique de 19 races bovines à partir de leur généalogie. Compte rendu n°000972125, Institut de l'élevage collection résultats, Septembre 2009.

Danchin-Burge, C. 2011. Bilan de variabilité génétique de 9 races de petits ruminants laitiers et à toison. Compte rendu n°001172004, Institut de l'élevage collection résultats, Juin 2011.

Dassonneville, R. 2012. Genomic selection of dairy cows. Thèse de doctorat, Institut des Sciences et Industries du Vivant et de l'Environnement.

Dassonneville, R., A. Baur, S. Fritz, D. Boichard, and V. Ducrocq. 2012a. Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment. *Genet. Sel. Evol.* 44:40:1–8.

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. *J. Dairy Sci.* 94:3679–3686.

Dassonneville, R., S. Fritz, V. Ducrocq, and D. Boichard. 2012b. Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. *J. Dairy Sci.* 95:4136–4140.

Douguet, M., J.M. Astruc, and G. Thomas. 2013. Résultats de contrôle laitier, France 2013. Compte-rendu 0014 201 008, Institut de l'élevage collection résultats.

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.

Ducrocq, V., P. Croiseau, A. Baur, R. Saintilan, S. Fritz, and D. Boichard. 2014. 10ème World Congress of Genetics Applied to Livestock Production, 17-22 août 2014, Vancouver, Canada.

Ducrocq, V., and J. Solkner. 1998. Implementation of a routing breeding value evaluation of dairy cows using survival analysis techniques. 6ème World Congress on Genetics Applied to Livestock Production, 11-16 janvier 2006, Armidale, Australie.

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. *J. Dairy Sci.* 95:4114–4129.

Etancelin, C.M., J.M. Astruc, D. Porte, H. Larroque, and C. Robert-Granié. 2005. Multiple-trait genetic parameters and genetic evaluation of udder-type traits in Lacaune dairy ewes. *Livest. Prod. Sci.* 97:211–218.

Feingold, J. 1991. Le déséquilibre de liaison, *Médecine sciences* 7:161-168.

Fernando, R.L., C. Stricker, and R.C. Elston. 1993. An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. *Theor. Appl. Genet.* 87:89–93.

Fikse, W.F., and G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. *J. Dairy Sci.* 84:1759–1767.

France Agrimer. 2014. Les filières de l'élevage français : http://www.franceagrimer.fr/content/download/33636/305064/file/Les_fili%C3%A8res_de%20l_elevage_francais-sept-2014;pdf.pdf.

FranceAgriMer, et France Génétique Élevage. 2015. La France, un des pays leaders mondiaux de la production de lait de chèvre et de la génétique caprine : http://fr.france-genetique-elevage.org/IMG/pdf/fge_selection_des_races_caprines_laitieres_2.pdf

France Génétique Élevage. 2014. Des programmes de sélection efficace : Sélection des races caprines laitières : http://fr.france-genetique-elevage.org/Selection-des-races-caprines.html

Frattini, S., L. Nicoloso, B. Coizet, S. Chessa, L. Rapetti, G. Pagnacco, and P. Crepaldi. 2014. Short communication: The unusual genetic trend of αS1-casein in Alpine and Saanen breeds. *J. Dairy Sci.* 97:7975–7979.

Fritz, S., F. Guillaume, P. Croiseau, A. Baur, C. Hoze, R. Dassonneville, M.Y. Boscher, L. Journeaux, D. Boichard, and V. Ducrocq. 2010. Mise en place de la sélection génomique dans les trois principales races françaises de bovins laitiers. 17ème Rencontres des Recherches autour des Ruminants, 8-9 décembre 2010, Paris, France.

Fritz, S., F. Guillaume, J. Tarres, A. Baur, M. Boussaha, M.Y. Boscher, L. Journaux, A. Malafosse, M. Gautier, J.J. Colleau, A. Eggen, and D. Boichard. 2008. Utilisation des résultats de cartographie fine de QTL en sélection chez les bovins laitiers. 15ème Rencontres des Recherches autour des Ruminants, 3-4 décembre 2008, Paris, France.

Gao, H., O.F. Christensen, P. Madsen, U.S. Nielsen, Y. Zhang, M.S. Lund, and G. Su. 2012. Comparison on genomic predictions using GBLUP models and two single-step blending methods with different relationship matrices in the Nordic Holstein population. *Genet. Sel. Evol.* 44:1–8.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breedings values and weighting information for genomic regression analyses. *Genet Sel Evol*. 41:1–8.

Gengler, N, G., Nieuwhof, K. Konstantinov, M. Goddard. 2012. Alternative single -step type genomic prediction equation. 63ème Annual Meeting of the European Federation of

Animal Science, Bratislava, Slovaquie, 27-31 août 2012.
Gengler, N., S. Abras, C. Verkenne, S. Vanderick, M. Szydlowski, and R. Renaville. 2008. Accuracy of prediction of gene content in large animal populations and its use for candidate gene detection and genetic evaluation. *J. Dairy Sci.* 91:1652–1659.
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. *Anim. Int. J. Anim. Biosci.* 1:21–28.
Gianola, D., and J.B.C.H.M. van Kaam. 2008. Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. *Genetics*. 178:2289–2303.
Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*. 136:245–257.
Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. *J. Anim. Breed. Genet.* 124:323–330.
Goddard, M. e., B. j. Hayes, and T. h. e. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.* 128:409–421.
Gray, K.A., J.P. Cassady, Y. Huang, and C. Maltecca. 2012. Effectiveness of genomic prediction on milk flow traits in dairy cattle. *Genet. Sel. Evol.* 44:24.
Gredler, B., H. Schwarzenbacher, C. Egger-Danner, C. Fuerst, E. Emmerling, and J. Sölkner. 2010. Accuracy of genomic selection in dual purpose Fleckvieh cattle using three types of methods and phenotypes. 9ème World Congress on Genetics Applied to Livestock Production, Leipzig, Allemagne, 2-6 août, 2010.
Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 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 Res.* 12:222–231.
Guillaume, F., D. Boichard, V. Ducrocq, and S. Fritz. 2012. Utilisation de la sélection génomique chez les bovins laitiers. *INRA Prod. Anim.* 24:4:363–368.
Guo, G., M.S. Lund, Y. Zhang, and G. Su. 2010. Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. *J. Anim. Breed. Genet.* 127:423–432.
Habier, D., R.L. Fernando, and J.C.M. Deckkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genet. Soc. Am.* 177:2389–2397.
Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*. 12:186. doi:10.1186/1471-2105-12-186.
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. *Genet. Sel. Evol.* 42:1–12.
Harris, B.L., F.E. Creagh, A.M. Winkelman, and D.L. Johnson. 2011. Experiences with the Illumina High Density Bovine BeadChip. 44ème Interbull Bulletin, 26-29 août 2011, Stavanger, Norvège.
Harris, B.L., and D.L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J Dairy Sci*. 93:1243–1252.
Harris, B.L., A.M. Winkelman, and D.L. Johnson. 2014. Across breed genomic prediction in dairy cattle. 10ème World Congress of Genetics Applied to Livestock Production, 17-22 août 2014, Vancouver, Canada.

Hayes, B.J. 2011. Ben Hayes Course notes., INRA SAGA, 12-16 septembre 2011, Toulouse, France.

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. *Genet. Sel. Evol.* 41:1–9.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92:433–443.

Hayes, B.J., H.D. Daetwyler, P. Bowman, G. Moser, B. Tier, R. Crump, M. Khatkar, H.W. Raadsma, and M.E. Goddard. 2009c. Accuracy of genomic selection: comparing theory and results. *Anim. Breed. Genet.* 18:34–37.

Hayes, B.J., I. Macleod, H.D. Daetwyler, P.J. Bowman, A.J. Chamberlian, C.J. Vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, C. Schrooten, E. Mullaart, R. Fries, B. Guldbrandtsen, M.S. Lund, D. Boichard, R.F. Veerkamp, C.P. VanTassel, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D.J. deKoning, E. Santus, and M.E. Goddard. 2014. Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project. 10ème World Congress of Genetics Applied to Livestock Production, 17-22 août 2014, Vancouver, Canada.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009d. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91:47–60.

Hayes, B., P.M. Visscher, H. McPartlan, and M. Goddard. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13:635–643.

Henderson, C., 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.

Hill, W.G., and A. Roberson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226–231.

Hoeschele, I., and P.M. VanRaden. 1993. Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. *Theor. Appl. Genet.* 85:946–952.

Hozé, C., M.-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. *Genet Sel Evol*. 45:33.

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014a. Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds. 10ème World Congress of Genetics Applied to Livestock Production, 17-22 août 2014, Vancouver, Canada.

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014b. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. *J. Dairy Sci.* 97:3918–3929.

Ilahi, H., P. Chastin, F. Bouvier, J. Arhainx, E. Ricard, and E. Manfredi. 1999. Milking characteristics of dairy goats. *Small Rumin. Res.* 34:97–102.

Institut de l'élevage. 1997. Fertilité plus ou comment améliorer la fertilité des chèvres après insémination artificielle sur chaleur induite.

Institut de l'élevage. 2013. Chiffres clé 2013 : productions lait & viande.

Israel, C., and J.I. Weller. 1998. Estimation of candidate gene effects in dairy cattle populations. *J. Dairy Sci.* 81:1653–1662.

Israel, C., and J.I. Weller. 2002. Estimation of Quantitative Trait Loci Effects in Dairy Cattle Populations. *J. Dairy Sci.* 85:1285–1297.

Jiménez-Montero, J.A., O. Gonzalez-Recio, and R. Alenda. 2010. Genotyping strategies for genomic selection in dairy cattle. 15ème Reunion Nacional de Mejora Genetica Animal, 16-18 juin 2010, Vigo, Espagne.

Jiménez-Montero, J.A., O. González-Recio, and R. Alenda. 2013. Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle. *J. Dairy Sci.* 96:625–634..

Jonas, D., V. Ducrocq, M.-N. Fouilloux, and P. Croiseau. 2015. Alternative haplotype construction methods for genomic evaluation. Soumis à *Genet. Sel. Evol.*

Karoui, S., M.J. Carabano, C. Diaz, and A. Legarra. 2012. Joint evaluation of French dairy cattle breeds using multiple-trait models. *Genet Sel Evol*. 44:1–10.

Katanos, J., B. Skapetas, and V. Laga. 2005. Machine milking ability and milk composition of some imported dairy goat breeds and some crosses in Greece. *Czech J Anim Sci*. 50:9:394–401.

Kennedy, B.W., C.M. Finley, and G.E. Bradford. 1982. Phenotypic and genetic relationships between reproduction and milk production in dairy goats. *J. Dairy Sci.* 65:2373–2383. doi:10.3168/jds.S0022-0302(82)82511-3.

Kizilkaya, K., R.L. Fernando, and D.J. Garrick. 2009. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J Anim Sci*. 88:544–551.

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). *J Dairy Sci*. 95:4065–4073.

Koopaei, H.K., M.R.M. Abadi, S.A. Mahyari, A.E. Koshkoiyeh, A.R. Tarang, and P. Potki. 2012. Effect of DGAT1 variants on milk composition traits in Iranian Holstein cattle population. *Anim. Sci. Pap. Rep.* 30:231–239.

Kuehn, C., C. Edel, R. Weikard, and G. Thaller. 2007. Dominance and parent-of-origin effects of coding and non-coding alleles at the acylCoA-diacylglycerol-acyltransferase (DGAT1) gene on milk production traits in German Holstein cows. *BMC Genet.* 8:62.

Leboeuf, B., E. Manfredi, P. Boué, A. Piacère, G. Brice, G. Baril, C. Roqua, P. Humblot, and M. Terqui. 1998. L'insémination artificielle et l'amélioration génétique chezla chèvre laitière en France. *INRA Prod. Anim.* 113: 171-181.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J Dairy Sci*. 92:4656–4663.

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. *J. Dairy Sci.* 97:3200–3212.

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014b. Single Step, a general approach for genomic selection. *Livest. Sci.* 166:54–65.

Legarra, A., and I. Misztal. 2008. Technical Note: Computing Strategies in Genome-Wide Selection. *J. Dairy Sci.* 91:360–366.

Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, and S. Fritz. 2011. Improved Lasso for genomic selection. *Genet. Res.* 93:77–87.

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. Soumis à. *Genet. Sel. Evol.*

Lewontin, R.C. 1964. The interaction of selection and likage. General considerations; heterotic models. *Genetics*. 49:49–67.

Lewontin, R.C., and K.I. Kojima. 1960. The evolutionary dynamics of complex polymorphisms. *Evolution*. 14:458–472.

Li, M.H., I. Strandén, T. Tiirikka, M. Sevon-Aimonen, and J. Kantanen. 2011. A comparison of approaches to estimate the inbreeding coefficient and pairwise relatedness using genomic and pedigree data in a sheep population. *PloS One*. 6:e26256.

Liu, Z., F.R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet Sel Evol*. 43:9.

Li, X., S. Wang, J. Huang, L. Li, Q. Zhang, and X. Ding. 2014. Improving the accuracy of genomic prediction in Chinese Holstein cattle by using one-step blending. *Genet. Sel. Evol.* 46:1–5.

Lorenz, A.J. 2013. Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. *G3 GenesGenomesGenetics*. 3:481–491.

Luan, T., J.A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T.H.E. Meuwissen. 2009. The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics*. 183:1119–1126.

Lund, M., A.P.W. De Roos, A.G. De Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, and R. Reents. 2010. Improving genomic prediction by EuroGenomics collaboration. 9ème World Congress on Genetics Applied to Livestock Production, Leipzig, Allemagne, 2-6 août, 2010.

Lund, M.S., S.P.W. De Ross, A.G. De Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, and R. Reents. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet Sel Evol*. 43:1–28.

Lund, M.S., G. Su, U.S. Nielsen, and G.P. Aamand. 2009. Relation between accuracies of genomic predictions and ancestral links to the training data. Interbull Meeting, 21-24 août 2009, Barcelone, Espagne.

Luo, M.F., G.R. Wiggans, and S.M. Hubbard. 1997. Variance component estimation and multitrait genetic evaluation for type traits of dairy goats. *J. Dairy Sci.* 80:594–600.

Makgahlela, M.L., T. Knürr, G.P. Aamand, I. Strandén, and E.A. Mäntysaari. 2013a. Single step evaluations using haplotypes segments. 47ème Interbull bulletin, 23-25 août 2013, Nantes, France.

Makgahlela, M.L., E.A. Mäntysaari, I. Strandén, M. Koivula, U.S. Nielsen, M.J. Sillanpää, and J. Juga. 2012. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. *J Anim Breed Genet*. 130:10–19.

Makgahlela, M.L., I. Strandén, U.S. Nielsen, M.J. Sillanpää, and E.A. Mäntysaari. 2013b. The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. *J. Dairy Sci.* 96:5364–5375.

Malécot, G. 1948. Les mathématiques de l'hérédité, Masson éditions, Paris.

Manfredi, E., and T. Adnøy. 2012. Génétique des caprins laitiers. *INRA Prod. Anim.* 253 233-244.

Manfredi, E., A. Piacere, P. Lahaye, and V. Ducrocq. 2001. Genetic parameters of type appraisal in Saanen and Alpine goats. *Livest. Prod. Sci.* 70:183–189.

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.

Maroteau, C. 2014. Cartographie fine de QTL pour des caractères d'intérêt pour la filière caprine. Thèse de doctorat, Université de Toulouse.

Maroteau, C., I. Palhière, H. Larroque, V. Clément, G. Tosser-Klopp, and R. Rupp. 2013. QTL detection for traits of interest for the dairy goat industry. 64ème Annual Meeting of the European Federation of Animal Science, 26-30 août 2013, Nantes, France.

Mc Hugh, N., T.H.E. Meuwissen, A.R. Cromie, and A.K. Sonesson. 2011. Use of female information in dairy cattle genomic breeding programs. *J Dairy Sci*. 94:4109–4118.

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating Improvement of Livestock with Genomic Selection. *Annu. Rev. Anim. Biosci.* 1:221–237.

Meuwissen, T.H.E. 2009. Accuracy of breeding values of'unrelated'individuals predicted by dense SNP genotyping. *Genet Sel Evol*. 41:35.

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.

Minvielle, F. 1990. Principes d'amélioration génétique des animaux domestiques. Presses Université Laval. 240 pp.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92:4648–4655.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D.H. Lee. 2002. BLUPF90 and related programs (BGF90). 7ème World Congress in Genetics Applied to Livestock Production, 19-23 août 2002, Montpellier, France.

Mizstal, I., S. Tsuruta, I. Aguilar, A. Legarra, P.M. VanRaden, and T.J. Lawlor. 2013. Methods to approximate reliabilities in single-step genomic evaluation. *J Dairy Sci*. 96:647–654.

Molee, A., N. Duanghaklang, and P. Na-Lampang. 2012. Effects of acyl-CoA:diacylglycerol acyl transferase 1 (DGAT1) gene on milk production traits in crossbred Holstein dairy cattle. *Trop. Anim. Health Prod.* 44:751–755.

Morand-Fehr, P., D. Sauvant, and R. Baumont. 2012. Avant-propos : Un dossier sur l'élevage caprin : pourquoi? *INRA Prod. Anim. 253*.

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. *Genet Sel Evol*. 42:1–15.

Mucha, S., R. Mrode, M. Coffey, and J. Conington. 2014. Estimation of Genomic Breeding Values for Milk Yield in UK Dairy Goats.

Mulder, H.A., M.P.L. Calus, and R.F. Veerkamp. 2010. Prediction of haplotypes with missing genotypes and its effect on accuracy of marker-assisted breeding value estimation. *Genet Sel Evol*. 42:1–15.

Nejhad, R.K., S. Yousefi, and M.A. Azari. 2013. Survey of single strand conformation polymorphism of kappa-casein gene in Alpine and Saanen goats in Iran. *Anim. Sci. Pap. Rep.* 31:2:173–177.

Olson, K., P. VanRaden, and M. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci.* 95:5378–5383.

Ozasa, S., E.S. Kempner, and S.K. Erickson. 1989. Functional size of acyl coenzyme A:diacylglycerol acyltransferase by radiation inactivation. *J. Lipid Res.* 30:1759–1762.

Palhière, I. 2001. Bilan de la variabilité génétique des races Saanen et Alpine. Rapport de fin d'étude d'ingénieur agronome.

Palhiere, I., C. Maroteau, H. Larroque, V. Clément, G. Tosser-Klopp, and R. Rupp. 2012. QTL detection for milk production traits, fatty acids, udder morphology and milking speed in Alpine and Saanen goats. XI International Conference on Goats, 23-27 septembre 2012, Gran Canaria, Espagne.

Patier, C. 2012. Rapport relatif à l'état des lieux de la production de chevreau et de la commercialisation de viande caprine. Conseil général de l'alimentation, de l'agriculture et des espaces ruraux n°11110-2-2012.

Patry, C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci.* 94:1011–1020.

Pérez-Enciso, M. 1995. Use of the uncertain relationship matrix to compute effective population size. *J. Anim. Breed. Genet.* 112:327–332.

Pérez-Enciso, M., J.C. Rincón, and A. Legarra. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet. Sel. Evol.* 47:43.

Piccoli, M.L., J. Braccini, F.F. Cardoso, M. Sargolzaei, S.G. Larmer, and F.S. Schenkel. 2014. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. *BMC Genet.* 15:157.

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. *Genet. Sel. Evol.* 46:22.

Přibyl, J., J. Haman, T. Kott, J. Přibylová, M. Simeckova, L. Vostry, L. Zavadilova, V. Cermak, Z. Ruzicka, J. Splichal, M. Verner, J. Motycka, and L. Vondrasek. 2012. Single-step prediction of genomic breeding value in a small dairy cattle population with strong import of foreign genes. *Czech J. Anim. Sci.:* http://agris.fao.org/agris-search/search.do?recordID=CZ2012000435

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–909.

Pryce, J.E., B. Gredler, S. Bolormaa, P.J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Solkner, M.E. Goddard, and B.J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. *J Dairy Sci*. 94:2625–2630.

Pszczola, M., T. Strabel, J.A.M. Van Arendonk, and M.P.L. Calus. 2012a. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. *J Dairy Sci*. 95:5412–5421.

Pszczola, M., T. Strabel, H.A. Mulder, and M.P.L. Calus. 2012b. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J Dairy Sci*. 95:389–400.

Puillet, L., M. Tichit, O. Martin, and D. Sauvant. 2012. Apport d'un modèle du fonctionnement du troupeau pour évaluer la production laitière et l'efficacité alimentaire en élevage caprin. *INRA Prod. Anim.* 25(3): 291-304.

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.

Qanbari, S., M. Hansen, S. Weigend, R. Preisinger, and H. Simianer. 2010. Linkage disequilibrium reveals different demographic history in egg laying chickens. *BMC Genet.* 11:103:1–10.

Renard, J. 2008. Bilan des schémas de sélection Alpin et Saanen.

Rincent, r., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V.M. Rodriguez, J. Moreno-Gonzalez, A. Melchinger, C.C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, and L. Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). *Genetics*. 192:715–728.

Robert-Granié, C., A. Legarra, and V. Ducrocq. 2011. Principes de base de la sélection génomique. *INRA Prod. Anim.* 24:331–340.

De Rochambeau, H., E. Verrier, J.P. Bidanel, L. Maignel, F. Labroue, T. Tribout, I. Palhiere, J.M. Astruc, F. Barillet, and H. Chapuis. 2003. Mise en place de procédure de suivi de la variabilité génétique des populations animales domestiques sélectionnées et établissement de guides de gestion: application aux ovins laitiers et aux porcs. *Actes BRG*. 4:17–34.

Rodríguez-Ramilo, S.T., L.A. García-Cortés, and Ó. González-Recio. 2014. Combining Genomic and Genealogical Information in a Reproducing Kernel Hilbert Spaces Regression Model for Genome-Enabled Predictions in Dairy Cattle. *PLoS ONE*. 9:e93424.

Rogers, A.R., and C. Huff. 2009. Linkage disequilibrium between loci with unknown phase. *Genetics*. 182:839–844.

Roldán, D.L., A.E. Rabasa, S. Saldaño, F. Holgado, M.A. Poli, and R.J.C. Cantet. 2008. QTL detection for milk production traits in goats using a longitudinal model. *J. Anim. Breed. Genet.* 125:187–193.

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.

De Roos, A.P.W., C. Schrooten, and T. Druet. 2011. Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix. *J. Dairy Sci.* 94:4708–4714.

Rupp, R., G. Lagriffoul, J.M. Astruc, and F. Barillet. 2003. Genetic parameters for milk somatic cell scores and relationships with production traits in French Lacaune dairy sheep. *J. Dairy Sci.* 86:1476–1481.

Rupp, R., I. Palhiere, C. Maroteau, G. Baloche, G. Sallé, A. Tircazes, C. Moreno, G. Foucras, and G. Tosser-Klopp. 2014. Mapping QTL controlling milk somatic cell counts in sheep and goat support the polygenic architecture of mastitis resistance. 10ème World Congress of Genetics Applied to Livestock Production, 17-22 août 2014, Vancouver, Canada.

Saatchi, M., M.C. McClure, S.D. McKay, M.M. Rolf, J. Kim, J.E. Decker, T.M. Taxis, R.H. Chapple, H.R. Ramey, and S.L. Northcutt. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genet Sel Evol*. 43:40.

Schaeffer, L.R. 2006. Strategy for applying genome wide selection in dairy cattle. *J Anim Breed Genet*. 123:218–223.

Schrooten, C., H. Bovenhuis, J.A.M. Van Arendonk, and P. Bijma. 2005. Genetic progress in multistage dairy cattle breeding schemes using genetic markers. *J Dairy Sci*. 88:1569–1581.

Short, T.H., and T.J. Lawlor. 1992. Genetic Parameters of Conformation Traits, Milk Yield, and Herd Life in Holsteins. *J. Dairy Sci.* 75:1987–1998.

Shumbusho, F., J. Raoul, J.M. Astruc, I. Palhiere, and J.M. Elsen. 2013. Potential benefits of genomic selection on genetic gain of small ruminant breeding programs. *J. Anim. Sci.* 91:3644–3657.

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

Sonesson, A.K., J.A. Woolliams, and T.H.E. Meuwissen. 2012. Genomic selection requires genomic control of inbreeding. *Genet Sel Evol*. 44:1–10.

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genet. Res.* 35:131–155.

Streit, M., N. Neugebauer, T.H.E. Meuwissen, and J. Bennewitz. 2011. Short communication: evidence for a major gene by polygene interaction for milk production traits in German Holstein dairy cattle. *J. Dairy Sci.* 94:1597–1600.

Su, G., B. Guldbrandtsen, V.R. Gregersen, and M.S. Lund. 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. *J Dairy Sci*. 93:1175–1183.

Sved, J.A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol*. 2: 125-141.

Sved, J.A., A.F. McRae, and P.M. Visscher. 2008. Divergence between human populations estimated from linkage disequilibrium. *Am. J. Hum. Genet.* 83(6):737–743.

Szyda, J., E. Ptak, J. Komisarek, and A. Żarnecki. 2008. Practical application of daughter yield deviations in dairy cattle breeding. *J. Appl. Genet.* 49:183–191.

Tenesa, A., P. Navarro, B.J. Hayes, D.L. Duffy, G.M. Clarke, M.E. Goddard, and P.M. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17: 520-526.

Thomsen, H., N. Reinsch, N. Xu, C. Looft, S. Grupe, C. Kühn, G.A. Brockmann, M. Schwerin, B. Leyhe-Horn, S. Hiendleder, G. Erhardt, I. Medjugorac, I. Russ, M. Förster, B. Brenig, F. Reinhardt, R. Reents, J. Blümel, G. Averdunk, and E. Kalm. 2001. Comparison of estimated breeding values, daughter yield deviations and deâ€regressed proofs within a whole genome scan for QTL. *J Anim Breed Genet*. 118:357–370.

Toosi, A., R.L. Fernando, and J.C.M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. *J Anim Sci*. 88:32–46.

Torres-Vazquez, J.A., F. Vazquez Flores, H.H. Montaldo, R. Ulloa-Arvizu, M. Valencia Posadas, A. Gayosso Vazquez, and R.A. Alonso Morales. 2008. Genetic polymorphism of the αs1-casein locus in five populations of goats from Mexico. *Electron. J. Od Biotechonoly* 113: 1-11.

Tosser-Klopp, G., P. Bardou, O. Bouchez, C. Cabau, R. Crooijmans, Y. Dong, C. Donnadieu-Tonon, A. Eggen, H.C. Heuven, and S. Jamli. 2014. Design and Characterization of a 52K SNP Chip for Goats. *PloS One*. 9:e86227.

Vaiman, D., L. Schibler, F. Bourgeois, A. Oustry, Y. Amigues, and E.P. Cribiu. 1996. A Genetic Linkage Map of the Male Goat Genome. *Genetics*. 144:279–305.

Vandiest, P. 2006. Le schéma de sélection caprin français : exigeant, coûteux, ... mais performant! *Filière Ovine Caprine*. 15.

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

VanRaden, P.M., D.J. Null, M. Sargolzaei, G.R. Wiggans, M.E. Tooker, J.B. Cole, T.S. Sonstegard, E.E. Connor, M. Winters, J.B.C.H.M. van Kaam, A. Valentini, B.J. Van Doormaal, M.A. Faust, and G.A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* 96:668–678.

VanRaden, P.M., J.R. O'Connell, G.R. Wiggans, and K.A. Weigel. 2011. Genomic evaluations with many more genotypes. *Genet Sel Evol*. 43.

VanRaden, P.M., and G.R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. *J. Dairy Sci.* 74:2737–2746.

Verbyla, K.L., M.P.L. Calus, H.A. Mulder, Y. De Haas, and R.F. Veerkamp. 2010. Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information. *J Dairy Sci*. 93:2757–2764.

Villumsen, T.M., and L. Janss. 2009. Bayesian genomic selection: the effect of haplotype length and priors. *BMC Proc.* 3:S11.

Vitezica, Z.G., I. Aguilar, and A. Legarra. 2010. One-step vs. multi-step methods for genomic prediction in presence of selection. 9ème World Congress on Genetics Applied to Livestock Production, Leipzig, Allemagne, 2-6 août, 2010.

Vitezica, Z.G., J.M. Elsen, R. Rupp, and C. Diaz. 2005. Using genotype probabilities in survival analysis: a scrapie case. *Genet Sel Evol 37 403–415*.

Vitezica, Z.G., L. Varona, and A. Legarra. 2013. On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. *Genetics*. 195:1223–1230.

Weiss, K.M., and A.G. Clark. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18:19–24.

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus. 2013. The effect of linkage disequilibrium and family relationships on reliability of genomic prediction. *Genetics*. 193:621–631.

Wiggans, G.R., I. Misztal, and L.D. Van Vleck. 1988. Animal Model Evaluation of Ayrshire Milk Yield with All Lactations, Herd-Sire Interaction, and Groups Based on Unknown Parents1. *J. Dairy Sci.* 71:115–125.

Wright, S. 1969. Evolution and the genetics of populations: the theory of gene frequencies, Vol. II.

Zhang, L., H. Li, Z. Li, and J. Wang. 2008. Interactions Between Markers Can Be Caused by the Dominance Effect of Quantitative Trait Loci. *Genetics*. 180:1177–1190.

Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T.H.E. Meuwissen, M. Svendsen, H. Grove, U.S. Nielsen, and M.S. Lund. 2014. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. *J. Dairy Sci.* 97:4485–4496.