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 SOPHIE BRARD

le jeudi 8 octobre 2015

Titre :

QUEL CADRE THEORIQUE ET PRATIQUE POUR L'UTILISATION DE LA SELECTION GENOMIQUE DANS L'AMELIORATION GENETIQUE DES CHEVAUX?

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GenPhySE)

Directeur(s) de Thèse :

MME ANNE RICARD

Rapporteurs :

Mme MATHILDE DUPONT NIVET, INRA JOUY EN JOSAS

Mme PASCALE LE ROY, INRA RENNES

Membre(s) du jury :

Mme SOPHIE DANVY, IFCE EXMES, Président

Mme ANNE RICARD, INRA TOULOUSE, Membre

M. STEVEN JANSSENS, UNIVERSITE CATHOLIQUE DE LOUVAIN, Membre

Remerciements

Mes remerciements vont en premier lieu à ma directrice de thèse. Merci Anne pour tout ce que vous avez pu me transmettre, et pour votre encadrement toujours positif de mon travail. Ça a été un vrai plaisir de travailler avec vous, et si c'était à refaire je re-signerai mon contrat de thèse sans hésiter !

Merci aux membres de mon comité de thèse : Didier Boichard, Jean-Michel Elsen, Andres Legarra, Laurence Moreau, pour l'apport de vos regards extérieurs sur l'orientation à donner à la thèse et sur les résultats obtenus. Grâce à vous nous avons pu valoriser dans une publication la méta-analyse !

Merci également aux rapporteurs, Pascale Le Roy et Mathilde Dupont-Nivet, et aux deux autres membres de mon jury de thèse, Sophie Danvy et Steven Janssens.

Merci aussi aux gestionnaires, Nancy et Valérie, pour leur aide dans la préparation de mes déplacements.

Merci également aux organismes qui ont rendu cette thèse possible en la finançant : l'Institut Français du Cheval et de l'Equitation, et l'INRA via le méta-programme SelGen.

Je remercie aussi les amis rencontrés à l'INRA au cours de ma thèse : Céline, Diane, Charlotte, Maxime, Yoannah, Mathilde, Morgane… Grâce à vous les pauses café et les déjeuners ont toujours été de bons moments (mention spéciale à Céline qui a parfois eu à gérer des formalités de déplacements pour moi!).

Un grand merci également à mon compagnon et à ma famille. Mon compagnon pour sa présence au quotidien. Ma famille pour m'avoir poussée et accompagnée dans mes études. Aujourd'hui je ne regrette pas d'avoir renoncé à devenir palefrenière, et je sais que j'ai de la chance en pleine recherche d'emploi de bénéficier du même soutien moral (et logistique !) que quand je passais mon bac ou les épreuves de l'agro.

Résumé

La sélection génomique substitue à la connaissance de la généalogie celle des séquences d'ADN et connait un succès spectaculaire dans la sélection des bovins laitiers. En équin, le gain de précision pour les valeurs génétiques en CSO a été estimé faible entre la généalogie et la génomique, éventuellement à cause des particularités des populations d'apprentissage et de validation. L'objectif est de définir pour les races équines les conditions d'efficacité et de fonctionnement de la sélection génomique. La partie théorique de la thèse a consisté en une méta-analyse afin de comprendre le lien entre précision théorique et observée en fonction des paramètres des populations. L'étude a montré l'importance du nombre efficace de marqueurs M_e. Ce paramètre spécifique de la population, de la structure génomique et de la parenté doit être évalué, au même titre que l'héritabilité en génétique classique. D'un point de vue pratique, la 1ère voie d'amélioration était de rechercher des gènes à effet majeur sur l'aptitude au concours de saut d'obstacles (CSO) ou au concours complet. Aucun gène majeur n'a été localisé malgré des détections significatives. Le 2nd levier pour améliorer l'estimation des valeurs génétiques en CSO était d'utiliser le Single-Step, méthode qui combine l'information génomique des étalons génotypés et la généalogie de l'ensemble des chevaux non génotypés utilisés pour l'indexation. L'évaluation pour le CSO a donc été revisitée. Malgré le re-calcul de l'héritabilité et l'application des points sur toute la période, le gain en précision reste faible. La sélection génomique a également été testée sur des chevaux d'endurance, mais comme pour le CSO les précisions obtenues pour le moment ne sont pas assez élevées pour justifier une utilisation de la sélection génomique. Récemment, un gène majeur agissant sur l'aptitude à trotter (*DMRT3*) a été identifié. Malgré l'effet très négatif d'un allèle sur la qualification et les performances précoces, le Trotteur français (TF) est polymorphe pour le gène à cause d'un effet positif de ce même allèle sur les performances tardives. La sélection classique et la sélection génomique ont été comparées en incluant ou non dans le modèle un marqueur lié à *DMRT3*, nous permettant d'identifier la meilleure combinaison de modèle et de méthode à utiliser pour estimer les valeurs génétiques du TF. Enfin, le paramètre M_e a été estimé dans les populations de chevaux utilisées au cours de la thèse, et les résultats des évaluations génomiques ont été comparés en fonction de M_e et des autres paramètres influant sur la précision de la sélection génomique. Deux nouveaux projets prévoyant de génotyper des chevaux de CSO d'une part et des TF d'autre part devraient permettre respectivement d'améliorer la précision de l'évaluation génomique en CSO et de confirmer l'intérêt de la prise en compte de *DMRT3* dans l'évaluation génomique des TF.

Abstract

Genomic selection uses genotypes information instead of pedigree information for the estimation of breeding values. In dairy cattle, the selection schemes were greatly improved with this method. In horses, a first attempt of genomic selection showed that the evaluation accuracy was not much improved when using genotypes information compared to classic evaluation, possibly because of the structure of the reference and validation populations. The objective of the thesis was to define the theoretical and practical conditions for the use of genomic selection in horses. The theoretical work of the thesis consisted in a meta-analysis to understand the relation between observed and theoretical accuracy depending on the parameters of the population. We proved the importance of the effective number of independent segments in the genome M_e. This parameter is specific of the population and of the genomic structure and relationship structure. We recommend to estimate this parameter before genomic evaluation, just like heritability that is estimated before genetic evaluation. Regarding practical tasks of the thesis, the first solution to improve the breeding values estimation for jumping performances was to look for genes having a major effect on performances in jumping competitions and three-day's events, but no major gene was evidence in spite of significant detections. The 2nd solution was to perform a single-step evaluation. This method combines information from genotyped stallions and from the pedigree of the whole population. Even if the heritability was re-estimated and points distributed to all horses to have a homogeneous criteria, the accuracy of genomic evaluation was not much improved. Genomic selection was also tested on horses running endurance races, but as for jumping the accuracy was not high enough. Recently, a major gene having a huge effect on the ability of horses to trot was evidenced (*DMRT3*). Even if one allele has a negative effect on qualification and early earnings, French Trotter (FT) is still heterozygote because of a positive effect of this allele on late performances. Genetic and genomic evaluations were compared with or without using in the model a SNP linked to *DMRT3* as a fixed effect. This study allowed identifying the best combination of model and method to use for estimation of FT breeding values. Finally, the parameter M_e was estimated in the populations of horses used in the thesis. The results of genomic evaluations were compared according to M_e and the other parameters having an influence on the accuracy of genomic evaluations. Two new projects will genotype more jumping horses and FT, they should allow to improve the accuracy of genomic evaluation for jumping horses and to acknowledge the interest of using *DMRT3* in the genomic evaluation of FT.

Table des matières

Introduction

Des disciplines équestres très variées sont pratiquées en France. D'une discipline à l'autre, les qualités requises pour les chevaux diffèrent : endurance pour des courses de plusieurs dizaines de kilomètres, adresse, puissance et rapidité pour les concours de saut d'obstacle, capacité à trotter à vive allure pour les courses au trot… Si les conditions de vie du cheval, son entrainement et les circonstances dans lesquelles se déroulent les épreuves auxquels il participe ont un effet sur ses performances, la part due à la génétique est loin d'être négligeable. Plusieurs races de chevaux sont donc élevées dans le but de produire les animaux ayant les bonnes caractéristiques pour réussir dans la discipline visée. A cette fin, les meilleurs reproducteurs sont choisis afin d'obtenir des descendants plus performants que les individus de la génération actuelle. Le progrès génétique d'une génération à l'autre dépendra de la précision avec laquelle on estime la capacité de l'individu à transmettre ses qualités à sa descendance (précision de l'estimation des valeurs génétiques), de la proportion de reproducteurs retenus parmi les candidats à la sélection (intensité de la sélection), et du temps nécessaire pour obtenir une nouvelle génération (intervalle de génération).

L'article de Meuwissen *et al.* (2001) a montré qu'il est possible d'estimer les valeurs génétiques à partir de marqueurs répartis sur le génome, les SNPs (Single Nucleotid Polymorphisms), suffisamment nombreux pour capturer les effets des gènes responsables de la variabilité génétique des performances. Deux modèles existent. L'un consiste à estimer dans une population de référence l'effet de chacun des marqueurs sur la performance et à en déduire la valeur génétique de l'individu connaissant les marqueurs qu'il porte. L'autre remplace l'apparentement « classique » connu grâce au pédigrée par l'apparentement « génomique » révélé par les marqueurs dans l'évaluation des valeurs génétiques des individus. On parle dans les deux cas d'évaluation génomique, et sous certaines hypothèses les deux modèles sont équivalents. Il y a quelques années, la sélection génomique a révolutionné l'amélioration génétique des bovins laitiers. Alors qu'avant il fallait attendre qu'un taureau ait plusieurs dizaines de filles en lactation pour estimer correctement sa valeur génétique, il est maintenant possible d'estimer suffisamment précisément la valeur génétique de l'individu dès sa naissance. Ceci a permis de mettre fin au testage sur descendance long et coûteux des taureaux.

Chez les chevaux, la sélection repose actuellement sur les performances propres des individus, c'est-à-dire sur la réussite en courses ou en compétitions, et sur les informations apportées par l'ascendance. Pour sélectionner un cheval pour la reproduction, il faut donc attendre qu'il soit en âge de concourir et qu'il ait suffisamment de performances. La sélection génomique pourrait permettre d'obtenir des valeurs génétiques aussi précises plut tôt dans la vie du cheval, réduisant ainsi l'intervalle entre les générations. L'objectif de cette thèse est de définir pour les races équines les conditions d'efficacité et de fonctionnement de la sélection génomique. La thèse est co-financée par l'Institut Français du Cheval et de l'Equitation et par le méta-programme INRA SelGen, et s'appuie sur les génotypages recueillis au cours des projets JUMPSNP, GENEQUIN et GENENDURANCE.

Les aspects théoriques de la mise en place de la sélection génomique sont abordés dans un chapitre (1) bibliographique présentant le principe de l'évaluation génomique ainsi que les différents leviers identifiés jusqu'à présent pour en améliorer la précision. Le chapitre 2 replace ces leviers en fonction des contraintes et des atouts des populations équines pour lesquelles la sélection génomique a été testée au cours de la thèse : cycle d'élevage, discipline de prédilection, et règlement des stud-books. Pour contribuer à la compréhension des mécanismes qui sous-tendent la précision de l'évaluation

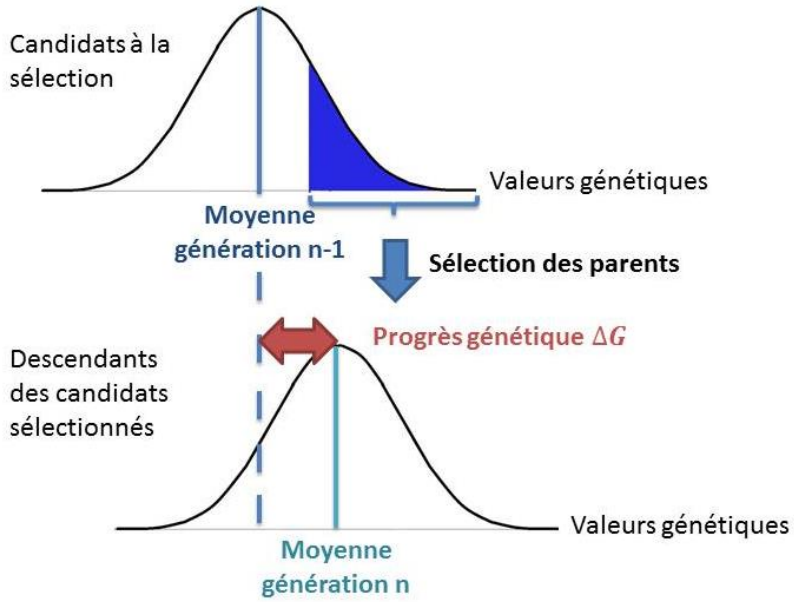
génomique, une méta-analyse reprenant les formules déterministes de calcul de la précision a été réalisée (3). Puis nous avons cherché une solution pour chacune des populations : recherche de marqueurs à effet important (4) et utilisation de l'ensemble de la population dans l'évaluation génomique (5) pour les chevaux de sport, étude d'une population moins sélectionnée (performeurs et non étalons) pour les chevaux d'endurance (6), utilisation d'un gène majeur en complément de l'évaluation génomique pour les trotteurs (7). Enfin, un dernier chapitre (8) reprend l'ensemble des résultats obtenus lors des tests de sélection génomique, et les compare à nos hypothèses concernant l'importance d'un nouveau paramètre génétique : le nombre de segments indépendants (M_e).

1. La sélection génomique : contexte bibliographique

1.1. Introduction : principe de la sélection classique

L'amélioration génétique des animaux repose sur un modèle qui décompose le phénotype P en une part expliquée par la génétique G qui se transmet d'une génération à l'autre et une part due à l'environnement E dans lequel l'animal réalise ses performances, de telle sorte que : $P = G + E$. La part génétique G se décompose elle-même en $G = A + D + I$, où A représente les effets additifs, D les effets de dominance et les I effets d'interactions. L'héritabilité d'un caractère, c'est-à-dire la part du phénotype qui est d'origine génétique et de nature additive est $h^2 = V(A)/V(P)$. Un caractère héritable peut potentiellement être amélioré par la sélection. La sélection utilise des valeurs génétiques, qui estiment la capacité d'un individu à transmettre ses qualités à sa descendance. La sélection peut être basée sur les performances individuelles (sélection massale), sur les performances des parents (sélection sur ascendance), sur les performances de descendants (sélection sur descendance), ou bien sur les performances des pleins-frères (sœurs) et/ou demi-frères (sœurs) (sélection sur collatéraux). Au cours d'une étape de sélection, les individus ayant les meilleures valeurs génétiques sont retenus parmi un groupe de candidats, et sont accouplés pour obtenir la génération suivante. Cette sélection améliore la valeur génétique moyenne de la population (Figure 1.1).

Figure 1.1 : Amélioration de la valeur génétique moyenne de la population lors de la sélection



Le progrès de la sélection d'une génération à l'autre ΔG se calcule de la façon suivante : $\Delta G = (i\, r \sigma_A)/T$. i est l'intensité de la sélection (la part des individus retenus parmi les candidats), r la précision des valeurs génétiques estimées, σ_A l'écart-type génétique additif du caractère, et T l'intervalle de temps entre deux générations. Les valeurs génétiques peuvent être estimées avec un modèle animal : $\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{Xb} + \boldsymbol{Wa} + \boldsymbol{e}$, où $\boldsymbol{Y}$ est un vecteur qui contient les performances des individus, μ est une moyenne (effet fixe), $\boldsymbol{b}$ est un vecteur contenant les effets fixes, $\boldsymbol{a}$ est un vecteur qui contient les valeurs génétiques des animaux (effets aléatoires), tel que $V(\boldsymbol{a}) = \boldsymbol{A}\sigma_a^2$, où $\boldsymbol{A}$ est la matrice d'apparentement entre les individus. $\boldsymbol{X}$ est une matrice d'incidence qui relie performances aux effets fixes, et $\boldsymbol{W}$ est une matrice d'incidence qui relie les performances aux animaux. $\boldsymbol{e}$ est un

terme résiduel. Pour estimer les valeurs génétiques, on minimise la variance résiduelle afin d'obtenir le BLUP (Best linear unbiased predictor), ce qui conduit à la résolution d'un système d'équations connu sous le nom de modèle mixte. Ces valeurs génétiques sont estimées et donc accompagnées d'un CD (Coefficient de détermination) qui varie entre 0 et 1 et indique la fiabilité de la valeur génétique. Plus le CD est proche de 1 et plus la valeur génétique est précise.

Il y a une quinzaine d'années, Meuwissen *et al.* (2001) ont proposé d'utiliser des marqueurs répartis sur l'ADN pour estimer les valeurs génétiques des animaux, marquant l'apparition de la sélection génomique. Cette partie présente dans un premier temps les marqueurs et leurs utilisations possibles, le principe de la sélection génomique et les opportunités et risques liés à cette méthode de sélection. Une seconde partie est consacrée à la précision de la sélection génomique, facteur clé du progrès génétique qui dépend de beaucoup de paramètres.

1.2. Principe de la sélection génomique, aperçu des méthodes disponibles et résultats attendus

1.2.1. La sélection génomique utilise des marqueurs répartis sur l'ADN

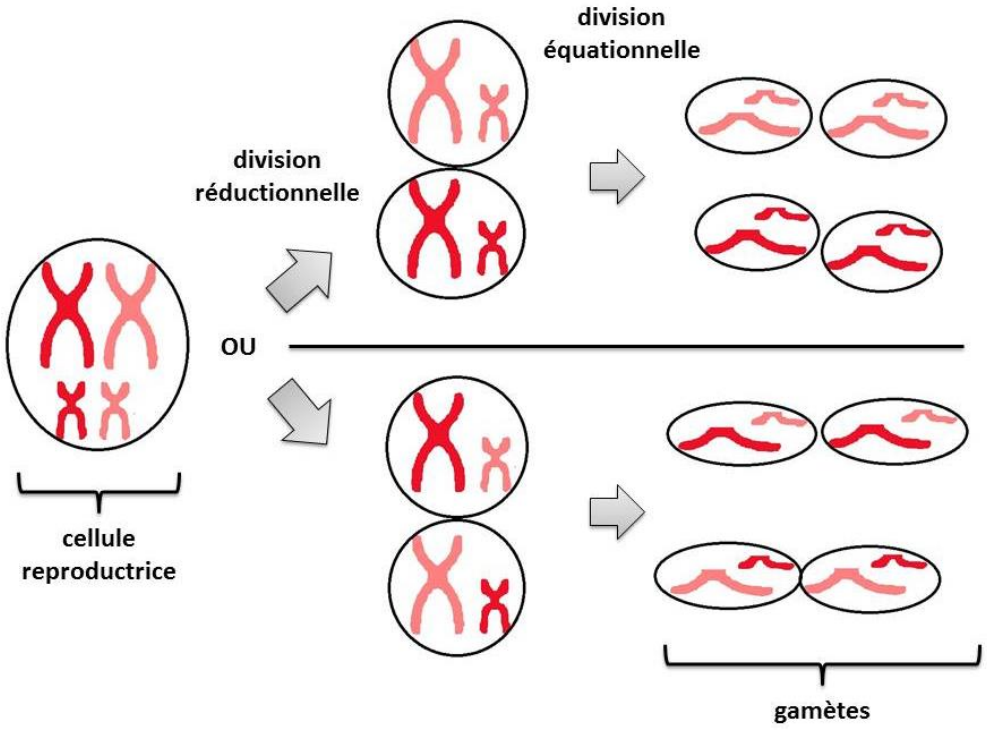
Qu'est-ce qu'un SNP (Single Nucleotid Polymorphism)?

L'ADN (Acide désoxyribonucléique) est la molécule qui, condensée sous la forme de chromosomes, est le principal support de l'hérédité. Un chromosome est constitué de deux chromatides identiques, portant les mêmes informations. Chez les eucaryotes les molécules d'ADN sont situées dans le noyau des cellules. Une molécule d'ADN est composée de deux brins complémentaires constitués de séquences de nucléotides portant des bases azotées : adénosine, cytosine, thymine, guanine. Certaines séquences peuvent être transcrites en ARN messagers qui quitteront le noyau et seront traduit par les ribosomes, conduisant à l'obtention de protéines constituées d'acides aminés. Ces séquences d'ADN sont dites codantes. Leur traduction est possible grâce au code génétique, redondant, non-ambigu et universel, qui à un codon de trois bases azotées associe un acide aminé. On peut définir un gène comme une séquence de l'ADN codante et située à un endroit précis de l'ADN, appelé locus. Les animaux étant diploïdes, au sein de chaque cellule chaque gène est présent en deux exemplaires : un sur chacun des chromosomes. Un gène peut exister en différentes versions, appelées allèles. Le terme d'allèles s'utilise pour désigner les différentes versions d'un gène, mais aussi plus généralement les différentes versions à un locus donné. Grâce aux processus de la méiose et de la fécondation, un animal possède pour chaque gène un allèle transmis par son père (gamète mâle) et un allèle transmis par sa mère (gamète femelle). Le génotype d'un individu, c'est-à-dire les versions des allèles qu'il porte, sera responsable de la part héréditaire de son phénotype. La diversité des génotypes d'un individu à l'autre est en partie le résultat du brassage inter-chromosomique qui a lieu lors de la méiose et de la fécondation : une cellule contient les chromosomes transmis par le père et par la mère (en couleurs différentes sur la Figure 1.2). Plusieurs combinaisons de chromosomes suivant leur origine paternelle ou maternelle sont possibles. La diversité des allèles existants pour un même gène est elle-même le résultat de mutations : modifications de la séquence non-détectées et non réparées au cours de la réplication de l'ADN. Si elles surviennent dans les cellules reproductrices elles sont transmises à la descendance. Du fait de la redondance du code génétique, une mutation peut-être silencieuse et conduire à la même protéine une fois l'ARN messager correspondant transcrit. En revanche si la protéine codée change et que sa fonction est modifiée, la variabilité apparue dans le génotype pourra avoir un effet sur le phénotype. Suivant

l'avantage ou le handicap éventuellement apporté par la mutation, l'allèle se répandra ou non dans la population.

Les mutations peuvent être le résultat d'une substitution, d'une insertion ou d'une délétion d'un nucléotide. Ces mutations sont des polymorphismes et sont très nombreuses sur le génome. Quand dans une séquence de nucléotides une variation est observée en un seul locus, il s'agit d'un SNP (Single Nucleotid Polymorphism). The 1000 Genome Project Consortium (2010) décrivent 15 millions de SNPs dans le génome humain. Les SNPs bi-alléliques sont utilisés comme marqueurs. Nous allons décrire dans la partie suivante le phénomène qui rend les SNPs utilisables en tant que tels.

Figure 1.2 : Exemple de brassage inter-chromosomique au cours des divisions de la méiose



En quoi les SNPs sont-ils informatifs ?

Les SNPs peuvent être informatifs de deux façons. Soit le SNP est confondu avec une mutation causale ayant un fort effet sur une performance, soit le SNP est en déséquilibre de liaison (DL) avec une mutation. Le déséquilibre de liaison est une association non-aléatoire entre deux loci qui s'observe par les fréquences des combinaisons d'allèles présents en deux loci. Soit un loci bi-allélique dont les allèles peuvent être A ou a (de fréquences respectives p_A et $p_a = 1 - p_A$), et un autre loci bi-allélique dont les allèles peuvent être B ou b (de fréquences respectives p_B et $p_b = 1 - p_B$). p_{AB} est la fréquence de la combinaison de l'allèle A sur le premier locus avec l'allèle B sur le second locus. Si $p_{AB} = p_A \times p_B$ alors les deux loci sont en équilibre de liaison. Sinon, les deux loci ne sont pas indépendants, et le déséquilibre de liaison peut se mesurer par $D = p_{AB} - p_A \times p_B$.

Le DL est dû à la structure de la molécule d'ADN et à son mode de transmission d'une génération à l'autre. La séquence de nucléotides sur une molécule d'ADN constitue un lien physique entre les loci. Au cours de la méiose, des échanges de segments chromosomiques peuvent avoir lieu lors de crossing-over : il s'agit d'un brassage intra-chromosomique. Au cours de ces recombinaisons génétiques, 2 loci très proches auront moins de chance d'être séparés que 2 loci très éloignés (Figure

1.3). On appelle r le taux de recombinaison, il s'agit de la fréquence de recombinaison entre deux loci. Plus deux loci sont liés, plus r est faible, et inversement pour des loci éloignés. Il est ainsi possible de calculer une distance génétique entre marqueurs en fonction du taux de recombinaison. Cette mesure est exprimée en centiMorgan (cM), 1cM correspondant à un taux de recombinaison de 1%. Ce taux de 1% signifie que 2 loci situés à 1cM l'un de l'autre seront séparés par un crossing-over une fois sur 100 méioses. Cependant il est aussi possible d'observer du DL entre des SNPs très éloignés ou ne se trouvant pas sur le même chromosome. Ce déséquilibre de liaison peut être induit par la sélection, si celle-ci porte simultanément sur deux caractères ou plus dont le déterminisme dépend de différentes régions du génome : les fréquences alléliques dans ces régions du génome sélectionnées en même temps évolueront conjointement, créant une relation statistique entre les fréquences alléliques dans ces zones sans qu'elles soient nécessairement proches les unes des autres.

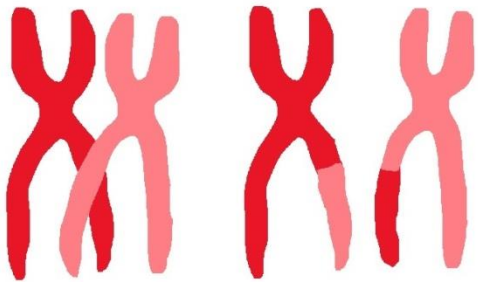
Une mesure courante du DL est r^2. Cette mesure est légèrement différente du D présenté précédemment. r^2 est le coefficient de corrélation entre les génotypes au marqueur et au QTL (ou à un 2ème marqueur), il représente la proportion de la variance expliquée par le QTL qu'on peut observer avec le marqueur (Hill et Robertson, 1968). Il se calcule de la façon suivante :

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

Le déséquilibre de liaison permet d'appréhender le passé d'une population, car il évolue en fonction de certains évènements. Il peut par exemple apparaitre quand deux populations avec des fréquences alléliques différentes en plusieurs loci fusionnent. Il peut aussi être créé quand la sélection porte conjointement sur plusieurs gènes. Dans ce cas ça ne sera pas la proximité physique des loci mais l'intérêt des allèles des différents gènes pour la population qui seront la cause du DL.

Le déséquilibre de liaison permet d'utiliser les SNPs comme des marqueurs répartis sur le génome. La sélection génomique repose sur l'hypothèse suivant laquelle les SNPs sont en déséquilibre de liaison avec les régions du génome dont le polymorphisme est responsable du phénotype étudié, et qu'ils sont suffisamment nombreux et bien répartis pour capturer toute la variance génétique (Meuwissen *et al*. 2001). Les parties suivantes décrivent deux utilisations possibles des SNPs, orientées vers la compréhension du déterminisme génétique des caractères et vers l'évaluation des individus pour la sélection.

Figure 1.3 : Représentation schématique d'un échange de segments chromosomiques lors d'un crossing-over



1.2.2. Quelles utilisations pour les SNPs en amélioration génétique ?

Localisation de régions du génome expliquant la variabilité

Les SNPs peuvent être utilisés pour détecter des QTL (Quantitative Trait Loci). Un QTL est une région du génome dont le polymorphisme cause une partie de la variabilité du caractère étudié. Une détection de QTL par analyse d'association exploite le déséquilibre de liaison en supposant que chaque QTL est en déséquilibre de liaison avec au moins un SNP. L'analyse d'association est une exploration sans *a priori* de l'ensemble du génome qui consiste à tester des possibilités d'association entre le polymorphisme des SNPs et la variabilité du phénotype. Le résultat est une cartographie des QTL ayant un effet sur les performances étudiées. Si un QTL très significatif est détecté, des investigations supplémentaires peuvent être menées dans la région du génome où il se trouve afin d'identifier un gène candidat dont la mutation expliquerait la variabilité des phénotypes observés pour le caractère. Cette méthode permet d'aboutir à une compréhension fine du déterminisme génétique de certains caractères. Un gène à effet majeur peut être identifié, comme *DGAT1* pour la production laitière chez les bovins par exemple (Grisart *et al*. 2002, Schennink *et al.* 2007). La connaissance des effets de gènes à effet majeur permettent d'enrichir le modèle d'évaluation animal classique, en ajoutant en effet fixe le ou les effets de substitution des allèles au(x) SNP(s) lié(s) au(x) QTL détecté(s). On parle de modèle assisté par gène quand le SNP est la mutation causale elle-même.

Si un marqueur est lié à un polymorphisme ayant un fort effet sur la performance, il peut être utilisé même si le gène à effet majeur n'est pas identifié. On parle dans ce cas de sélection assistée par marqueurs. Le marqueur est utilisé pour approximer le gène. Le principe est le même que celui de la sélection assistée par gène, mais le QTL n'est pas identifié et son génotype est remplacé par le génotype du marqueur en déséquilibre de liaison avec le QTL.

Si la variance génétique expliquée par le QTL vaut $\sigma^2{}_{QTL}$, alors le marqueur explique une variance génétique égale à $r^2\sigma^2{}_{QTL}$. Si le DL est très grand (proche de 1), le fait de ne pas connaitre le QTL n'est pas très pénalisant. Un inconvénient de cette méthode est que sur données réelles on ne peut pas connaître le déséquilibre de liaison réel entre le marqueur et le QTL.

Les marqueurs expliquent rarement plus de 10% de la variance génétique totale et ne sont donc pas suffisants pour sélectionner les animaux s'ils sont pris individuellement : les méthodes de génétique classique conservent leur intérêt. Aujourd'hui l'évaluation est réalisée en utilisant l'ensemble des SNPs, et plusieurs modèles ont été proposés et testés dans cette optique. Les plus courants sont présentés dans la partie suivante. La connaissance de l'architecture génétique des caractères reste cependant importante, et au cours de ma thèse j'ai réalisé une détection de QTL pour la performance en saut d'obstacle afin de vérifier l'existence d'éventuels gènes majeurs.

Estimation de valeurs génétiques

L'estimation précise des valeurs génétiques requiert une bonne estimation des effets des marqueurs. La sélection génomique suppose que tout QTL, quelle que soit son importance, peut être approché par les marqueurs situés à proximité. Pris tous ensemble, ces marqueurs devraient expliquer toute la variance génétique due aux QTL (Meuwissen *et al*. 2001). Les modèles utilisés pour l'estimation peuvent être classés en deux groupes suivant leurs hypothèses sur les effets des marqueurs à estimer. D'une part, les modèles linéaires supposent que chaque SNP explique une part identique de la variance génétique, égale à la variance génétique totale divisée par le nombre de marqueurs. D'autre part, les modèles non-linéaires supposent que certains marqueurs expliquent une part de la

variance génétique additive importante alors que d'autres expliquent une part faible voire nulle. Les modèles les plus courants sont présentés ici, la question de leur efficacité relative dans différentes situations sera abordée plus tard dans la section 3.2. de ce chapitre.

Le modèle linéaire suppose que tous les SNPs expliquent une part égale de la variance

Le modèle linéaire utilise tous les marqueurs, ce qui doit permettre de prendre en compte tous les QTL expliquant en général une grande part de la variance génétique. Meuwissen *et al.* (2001) présente ce modèle comme une extension du modèle de sélection assistée par marqueurs :

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{Xb} + \boldsymbol{Zg} + \boldsymbol{e},$$

Avec $\boldsymbol{Y}$ la matrice des performances, $\mathbf{1}$ un vecteur de 1, μ la moyenne (effet fixe), $\boldsymbol{b}$ les effets fixes, $\boldsymbol{g}$ l'effet de substitution des allèles aux SNPs (effets aléatoires en raison du très grand nombre de SNPs à effets faibles), et $\boldsymbol{e}$ la résiduelle. Comme le modèle contient à la fois des effets fixes et des effets aléatoires, les solutions sont obtenues en utilisant les équations du modèle mixte d'Henderson (1975). $\boldsymbol{X}$ est une matrice d'incidence. $\boldsymbol{Z}$ contient les génotypes aux SNPs. Il n'y a normalement pas d'effet polygénique car les SNPs sont sensés capturer toute la variance génétique. Dans ce modèle, les génotypes ne sont pas codés 0, 1, 2 mais sont standardisés de façon à ce que la moyenne des génotypes soit 0 et l'écart-type 1. La variance par SNP est supposée être la variance génétique totale divisée par le nombre de SNPs. On verra par la suite que cette hypothèse est régulièrement discutée dans les travaux portant sur la sélection génomique. Ce modèle est le SNP-BLUP, aussi appelé « modèle marqueurs ». La valeur génétique estimée d'un individu i vaut: $\hat{u}_i = \sum_j z_{ij}\hat{g}_j$.

Ce modèle est équivalent au « modèle animal ». Le modèle animal se déduit du modèle animal classique utilisé en sélection qui est le suivant :

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{Xb} + \boldsymbol{Wa} + \boldsymbol{e},$$

avec $\boldsymbol{b}$ les effets fixes, **a** les valeurs génétiques des individus telles que $V(a) = \boldsymbol{A}\sigma_a^2$, $\boldsymbol{A}$ étant la matrice d'apparentement. $\boldsymbol{W}$ et $\boldsymbol{X}$ sont des matrices d'incidence. La version génomique du modèle animal est :

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{Xb} + \boldsymbol{Wu} + \boldsymbol{e},$$

avec $\hat{u}_i = \sum_j z_{ij}\hat{g}_j$, d'où l'équivalence du modèle animal et du modèle marqueur. La variance de $\boldsymbol{u}$ vaut $\sigma_g^2\boldsymbol{ZZ}'/n$, $\boldsymbol{Z}$ étant une matrice dont les colonnes contiennent les génotypes à chaque SNP. $\boldsymbol{ZZ}'/n$ peut être interprétée comme une matrice d'apparentement génomique $\boldsymbol{G}$ qui remplace $\boldsymbol{A}$ dans le BLUP (Goddard 2009), donnant ainsi le GBLUP dans lequel les valeurs génétiques sont calculées en résolvant les équations du modèle mixte. Les étapes de calcul des valeurs génétiques diffèrent entre les deux modèles, mais les résultats sont les mêmes.

Les modèles bayésiens proposent d'utiliser des distributions a priori des effets des SNPs pour mieux approcher la réalité

L'objectif des modèles bayésiens en sélection génomique est d'estimer des valeurs génétiques en se basant sur une distribution *a priori* des effets des marqueurs plus proche de la réalité qu'avec les modèles infinitésimaux. Certains modèles sont présentés dans l'article de Meuwissen *et al.* (2001). Ces modèles supposent qu'en réalité il y a peu de mutations causales responsables de la variance

génétique d'un caractère, et que donc peu de SNPs auront réellement un effet important sur le caractère. Les effets des SNPs sont considérés comme des effets aléatoires.

Dans le modèle Bayes A, les SNPs peuvent avoir des effets supérieurs ou inférieurs aux effets autorisés par une distribution normale. La variance des effets des SNPs n'est plus identique pour tous, elle est estimée avec un échantillonnage de Gibbs à partir d'une distribution *a priori* et des informations apportées par les données. Cette distribution suit une loi de Student, la queue de la distribution est plus épaisse que celles d'une loi normale.

Le modèle Bayes B diffère du modèle Bayes A car il considère que beaucoup de loci ne ségrégent pas et qu'ils n'expliquent donc pas la variance génétique (Meuwissen *et al*. 2001). La distribution *a priori* est la même que pour le Bayes A, mais une part $1 - \pi$ des SNPs aura un effet nul.

Dans le modèle Bayes Cπ aussi on suppose qu'une fraction π des SNPs a un effet et que $1 - \pi$ des marqueurs n'ont pas d'effet, mais la proportion π est estimée à parti des données (Habier *et al*. 2011).

La méthode du Lasso (Tibshirani 1996) suppose que les effets suivent une loi exponentielle double, symétrique. Les effets des SNPs les plus faibles sont régressés à 0.

Les modèles avec réduction de dimension réduisent la taille du système à évaluer
Ces modèles ne font pas d'hypothèse sur la distribution des effets des SNPs.

La Principal component analysis (PCA) réduit la taille de la matrice des SNPs en identifiant quelques variables expliquant la plus grande part possible de la variance génétique additive (Solberg *et al*. 2008).

La partial least square regression (PLSR) (Solberg *et al*. 2008) fait de même mais avec un conditionnement sur les phénotypes.

Ces méthodes sont celles qui ont été les plus testées et comparées depuis les débuts de la sélection génomique chez les animaux. Nous verrons par la suite leurs avantages respectifs dans différentes situations en lien avec l'architecture génétique des caractères évalués.

Le modèle en une étape utilise toute l'information disponible
Les modèles présentés précédemment font des hypothèses sur la distribution des effets des marqueurs, qui si elles sont fausses auront des répercussions sur l'estimation des valeurs génétiques. De plus, un biais peut apparaître dans les évaluations car tous les animaux ne peuvent être génotypés, et on se limite en général aux meilleurs, donc à une population sélectionnée. Enfin dans des populations à petits effectifs (bovins allaitants, chevaux) la quantité de données disponibles n'est pas toujours suffisante pour estimer correctement les valeurs génétiques à partir des marqueurs. Comme on le verra dans la partie 1.3., une quantité d'information importante est nécessaire en entrée du modèle pour estimer les valeurs génétiques précisément. Le modèle du single-step a l'avantage par rapport aux autres modèles de prendre en compte dans l'évaluation les génotypes d'individus non-phénotypés et les phénotypes d'individus non-génotypés (Legarra *et al*. 2009). J'ai utilisé ce modèle pour cette raison au cours de ma thèse pour une évaluation des chevaux de saut d'obstacle.

Mistzal *et al.* (2009) ont proposé de modifier la matrice d'apparentement $\boldsymbol{A}$ de façon à prendre en compte à la fois l'apparentement basé sur le pédigrée et la différence entre l'apparentement attendu basé sur le pédigrée (matrice $\boldsymbol{A}$) et l'apparentement dit « réalisé » observé à partir des marqueurs (matrice $\boldsymbol{A}_\Delta$). La matrice $\boldsymbol{H}$ obtenue en sommant $\boldsymbol{A}$ et $\boldsymbol{A}_\Delta$ ne fonctionnait pas car les termes non-diagonaux de $\boldsymbol{H}$ ne dépendaient pas de la matrice d'apparentement génomique $\boldsymbol{G}$. Legarra *et al*. (2009) ont ensuite proposé une amélioration bayésienne de cette matrice en dérivant conjointement la densité des valeurs génétiques d'individus génotypés (notés 2) et non-génotypés (notés 1) : $p(u_1, u_2) = p(u_1|u_2)p(u_2)$. $p(u_1|u_2)$ est basée sur le pédigrée grâce à l'index de sélection, et $p(u_2)$ ne dépend que du génotype. Avec ces développements $\boldsymbol{H}$ contient la covariance des distributions conjointes de u_1 et u_2.

Ces travaux ont été réalisés à partir du modèle animal. La même matrice $\boldsymbol{H}$ a été développée en parallèle à partir du modèle marqueurs équivalent par Christensen et Lund (2010). Leur objectif était d'imputer les génotypes considérés comme manquants (ceux des individus non-génotypés du pédigrée) à partir des données disponibles, en prenant en compte la distribution jointe des génotypes inférés et des génotypes connus. Pour cela ils ont considéré les génotypes comme des caractères quantitatifs. Ils obtiennent la matrice d'apparentement correspondante $\widehat{\boldsymbol{Z}}_\mathbf{1} = \boldsymbol{A}_{\mathbf{12}}\boldsymbol{A}_{\mathbf{12}}^{-\mathbf{1}}\boldsymbol{Z}_\mathbf{2}$ (là aussi 1 sont les individus non-génotypés et 2 les individus génotypés). La distribution conjointe des génotypes inférés permet de retrouver la matrice $\boldsymbol{H}$.

L'utilisation de SNPs permet donc grâce à leur capacité à capturer les effets des QTL via le déséquilibre de liaison d'analyser le déterminisme de caractères et d'estimer des valeurs génétiques. La partie suivante présente la mise en œuvre de la sélection génomique.

1.2.3. Application de la sélection génomique

La sélection génomique consiste à estimer les effets des marqueurs/remplacer la matrice d'apparentement génétique par la matrice d'apparentement génomique pour l'estimation des valeurs génétiques. Pour cela une population de référence est nécessaire. S'il n'est pas possible pour des questions économiques ou pratiques de génotyper tous les individus, la constitution de la population de référence doit être réfléchie de sorte à contenir un nombre suffisamment important d'individus représentatifs de la population afin d'entrainer correctement le modèle (Figure 1.4). Les équations de prédiction peuvent ensuite être utilisées pour estimer les valeurs génétiques des candidats à la sélection qui sont génotypés mais n'ont pas encore de performances pour le caractère étudié.

Une amélioration de la précision des valeurs génétiques estimées est attendue lors du passage de la sélection classique à la sélection génomique car la matrice génomique est supposée décrire plus précisément l'apparentement que le pédigrée. La précision de la sélection est la corrélation entre les valeurs génétiques vraies et les valeurs génétiques estimées. Il n'y a que sur des données simulées que l'on connait la valeur génétique vraie. Si toutes les hypothèses du modèle sont vérifiées, la précision peut être déduite de l'inverse de la matrice d'information. Empiriquement, on peut l'approcher en utilisant par exemple les valeurs génétiques d'individus très bien connus sur descendance. Il est aussi possible d'utiliser un autre critère, comme la corrélation entre la valeur génétique estimée et la performance, même si cette solution se rencontre peu fréquemment dans les travaux cités dans cette partie. La précision de la sélection génomique qu'on peut espérer peut se vérifier par validation croisée dans une population génotypée et phénotypée: les données sont

séparées en sous-échantillons et tour à tour les phénotypes des animaux de chaque sous-échantillon sont masqués et les valeurs génomiques estimées à partir des informations conservées (Figure 1.5). La précision de la sélection génomique est alors mesurée comme la corrélation entre les EBVs obtenus et les phénotypes réalisés, qui ont été masqués pour l'estimation des effets des SNPs. Quelle que soit la valeur obtenue pour la précision, la corrélation entre les valeurs génétiques et les performances ne peut excéder h (h^2 étant l'héritabilité du caractère étudié). Il faut donc diviser la corrélation entre les performances et les valeurs génétiques par h pour avoir une estimation non-biaisée de la précision de la sélection génomique. Une autre approche consiste à travailler avec des performances obtenues sur plusieurs années et à masquer celles des animaux les plus jeunes afin de tester la sélection génomique dans des conditions proches de son application réelle. Au cours de ma thèse j'ai utilisé la validation croisée chez les trotteurs et réalisé un essai de sélection génomique sur les plus jeunes chevaux performeurs en CSO.

Figure 1.4 : Principe de la sélection génomique

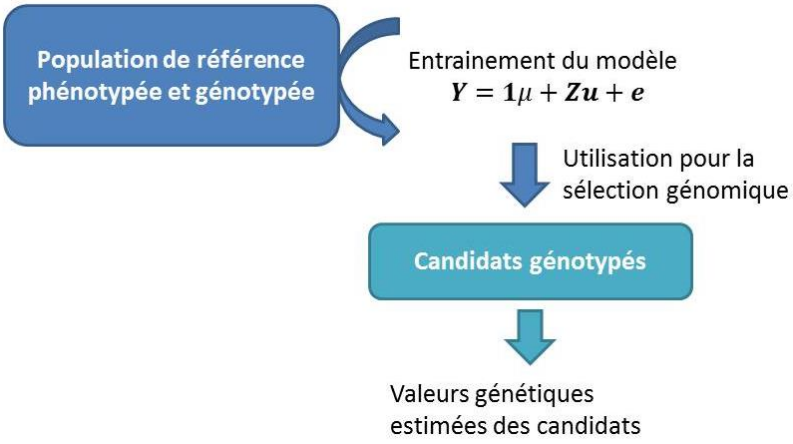
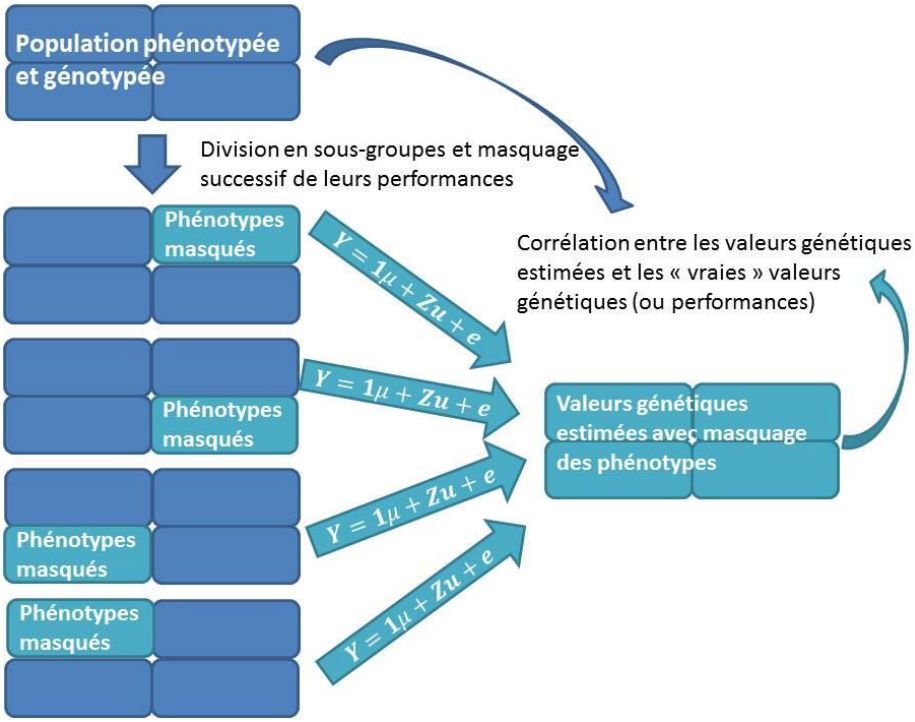


Figure 1.5 : Illustration du principe de la validation croisée



1.2.4. Opportunités et risques

Dès l'article de Meuwissen *et al*. (2001) les attentes d'amélioration de la sélection grâce au passage à la sélection génomique sont nombreuses. Nous allons voir dans cette partie quels sont ces avantages attendus (et observés), ainsi que les risques liés à l'utilisation de ces méthodes d'évaluation.

Améliorations attendues grâce au passage de la sélection classique à la sélection génomique

La sélection génomique doit permettre d'améliorer la précision des valeurs génétiques estimées. Le gain en précision apporté par la sélection génomique a été vérifié dans plusieurs espèces. Pour les bovins laitiers VanRaden *et al*. (2009) ont montré un gain en précision de 20% à 29% pour les caractères laitiers. Les gains possibles en race Lacaune pour les ovins laitiers ont été démontrés par Duchemin *et al.* (2012). Chez des poules pondeuses (Liu *et al*. 2014) la précision de la sélection est doublée pour des caractères de productions d'œufs. Les résultats en terme de progrès génétique sont aussi encourageants pour les ovins viande (Banks *et al*. 2009), ainsi que pour les bovins allaitants (Weber *et al*. 2012).

L'intérêt de la sélection génomique est aussi dans certaines espèces de réduire l'intervalle de génération. L'ADN pouvant être obtenu dès la naissance, voire avant, il est possible d'obtenir une valeur génétique pour un animal très jeune, sans avoir à mettre en place un testage systématique de sa descendance. Cet avantage attendu n'est vérifié que si l'intervalle de génération biologique est inférieur à l'intervalle de génération nécessaire pour évaluer précisément les reproducteurs sans sélection génomique. C'est le cas chez les bovins laitiers, où la sélection génomique a révolutionné la sélection en mettant fin à un testage sur descendance long (une dizaine d'années) et coûteux. Dans leur cas l'intervalle de génération est réduit au minimum. Pour les espèces qui sont dans ce cas, il peut devenir intéressant de tenter de réduire l'intervalle de génération biologique vu que celui-ci est devenu le facteur limitant. Mais en réalité il y a cependant un équilibre à trouver entre l'intervalle de génération possible grâce à la disponibilité précoce des données ADN et l'intervalle de génération pré-sélection génomique : en effet en réduisant cet intervalle au minimum on dispose de moins de générations d'animaux avec des performances pour évaluer les reproducteurs, ce qui se répercute sur la précision des valeurs génétiques.

La sélection génomique améliore la précision des valeurs génétiques estimées dans les populations animales citées précédemment, mais cette amélioration peut être faible dans des populations où la précision est déjà élevée. En revanche la sélection génomique est intéressante quand les animaux d'élite qui sont candidats à la sélection ne sont pas ceux qui réalisent les performances. Pour ces individus, la sélection génomique peut être une solution (Muir *et al.* 2007). La sélection génomique serait aussi avantageuse pour les caractères qui se mesurent une fois l'animal abattu, comme la qualité de la carcasse du poulet de chair par exemple (Liu *et al*. 2014), ou pour la résistance aux maladies vu que les animaux destinés à la reproduction sont dans des élevages où les aspects sanitaires sont très maitrisés.

La sélection génomique est aussi un outil permettant une meilleure gestion de la diversité génétique dans une population. En effet, dans la sélection classique, ne sachant pas quels allèles ont hérité 2 plein-frères de leurs parents, si la sélection a lieu avant l'obtention de performances ces 2 animaux seront sélectionnés conjointement car ils auront la même valeur génétique sur ascendance. Dans ce cas de figure, avec la sélection génomique, on connait le génotype aux marqueurs de chaque animal, et il est possible de différencier 2 plein-frères et donc d'exercer une sélection plus fine (Hayes *et al*.

2009a). Sonesson *et al.* (2012) montrent par ailleurs que l'utilisation de la sélection génomique n'augmente la consanguinité que très localement sur le génome au niveau des loci sélectionnés. La consanguinité qui peut être recherchée en certains points, par exemple quand c'est un génotype homozygote en un loci qui donne les meilleures performances, affecte peu le reste du génome. La sélection génomique serait donc un bon outil pour gérer la consanguinité, et la sélection basée sur les marqueurs n'entrainerait pas une augmentation globale de la consanguinité (Sonesson *et al.* 2012).

La sélection génomique a donc des avantages indéniables déjà observés dans plusieurs espèces. Cependant son utilisation comporte aussi quelques risques, décrits dans la partie suivante.

Risques liés à la sélection génomique à garder à l'esprit

Hayes *et al.* (2009a) soulignent que la sélection génomique utilise des marqueurs sensés capturer toute la variance génétique. Cependant si ce n'est pas le cas la sélection ne sera réalisée que sur les QTL dont les effets sont effectivement capturés par les SNPs. De plus, comme seule une partie de la population est génotypée, les QTL dont la fréquence serait très faible et qui ne seraient pas portés par les animaux génotypés ne pourront pas être sélectionnés. Il y a donc un risque en utilisant la sélection génomique d'ignorer les effets de QTL rares mais participant à la variabilité du caractère.

Une particularité de la sélection génomique, qui sera développée dans la suite de ce chapitre, est que pour qu'elle soit précise plusieurs conditions doivent être remplies concernant la quantité des données et leur structure. Si la sélection génomique fonctionne bien dans les plus grandes populations d'animaux d'élevage, certaines espèces ont peu d'individus en production, ou bien des structures de populations peu adaptées à une utilisation simple de la sélection génomique. Ces espèces ou races ne peuvent pas appliquer la sélection génomique aussi facilement que les autres ou obtiennent de moins bons résultats (Aguilar *et al.* 2009), quelle que soit la qualité de la sélection classique utilisée jusqu'à présent.

La sélection génomique permet de sélectionner plus précisément et plus rapidement en réduisant l'intervalle de génération, mais ce progrès accéléré comporte des risques. Meuwissen *et al.* (2013) mettent en garde contre l'augmentation plus rapide de la consanguinité : la réduction de l'intervalle de génération signifie que les animaux se reproduisent plus rapidement. Comme le taux de consanguinité augmente à chaque génération, la consanguinité augmente donc plus vite de façon mécanique quand la sélection génomique réduit l'intervalle de génération par rapport à la sélection classique. Une sélection plus rapide signifie aussi que les allèles d'intérêt sont fixés plus rapidement (Zhang et Hill 2004), particulièrement dans les populations de petite taille, ce qui conduit à une diminution de la variance génétique additive pour le caractère. Bijma *et al.* (2012) montrent d'ailleurs qu'il faut tenir compte de cette diminution de la variance génétique additive dans le calcul de la précision, au risque sinon de sous-estimer le gain en précision apportée par la sélection génomique. Enfin une sélection plus rapide risque aussi de modifier le déséquilibre de liaison dans les régions du génome sélectionnées (Calus 2010), ce qui va nécessiter une ré-estimation régulière des effets des SNPs afin de ne pas détériorer la précision. Or, en raccourcissant l'intervalle de génération le temps pour collecter des enregistrements de performance diminue (Meuwissen *et al*. 2013): il faut tenir compte de la nécessité de mettre à jour le modèle et trouver un compromis entre la réduction de l'intervalle de génération et le temps nécessaire à l'acquisition de nouveaux phénotypes.

La sélection génomique a donc des avantages reconnus sur les facteurs du progrès génétique (diminution de l'intervalle de génération, estimation plus précise des valeurs génétiques, base de sélection augmentée si beaucoup de candidats sont génotypés), permettant ainsi d'améliorer les schémas de sélection. Elle a cependant aussi des inconvénients liés à sa faisabilité dans des populations de petite taille ou de structure particulière, aux risques entrainés par une sélection accélérée et aux hypothèses sur lesquelles elle repose. Depuis ses débuts la sélection génomique a été testée dans beaucoup de populations. Une question récurrente (que pose mon sujet de thèse chez les équidés) est celle des conditions à réunir pour obtenir la meilleure précision possible. L'objectif de la partie suivante est de répertorier ces questions et de faire un état des lieux des réponses déjà apportées.

1.3. Comment obtenir la meilleure précision possible avec la sélection génomique ?

La sélection génomique constitue une opportunité pour l'amélioration génétique car elle permet suivant les schémas existant d'améliorer un ou plusieurs des facteurs du progrès génétique. Une étape d'optimisation des schémas est cependant nécessaire car les paramètres du progrès génétique interagissent, et l'amélioration d'un paramètre peut en dégrader un autre. Parmi les paramètres du progrès génétique, la précision de l'évaluation est peut-être le critère le plus étudié dans les publications comparant la sélection classique et la sélection génomique. En effet, la précision de la sélection génomique est sensible à de nombreux facteurs. Certains comme l'héritabilité ne peuvent être modifiés. D'autres comme les caractéristiques de la population de référence, le modèle choisi pour l'estimation ou encore le nombre de marqueurs utilisés peuvent être optimisés. Le but de cette partie est de présenter les questionnements sur la précision de la sélection génomique par le biais de ces trois facteurs.

1.3.1. Quelle population de référence utiliser ?

Combien d'individus sont nécessaires ?

Dès l'article fondateur de la sélection génomique, Meuwissen *et al*. (2001) précisent qu'un nombre important d'individus devra constituer la population de référence afin d'estimer correctement les effets des marqueurs, en particulier pour les caractères les moins héritables.

L'effet positif sur la précision de la sélection génomique d'une augmentation de la taille de la population de référence a depuis été largement vérifié : chez les bovins laitiers (Schaeffer *et al*. 2006, Luan *et al.* 2009, VanRaden *et al*. 2009, Pszczola *et al*. 2011), chez les bovins allaitants (Brito *et al*. 2011), chez les ovins (Daetwyler *et al.* 2010), chez les végétaux (Zhong *et al*. 2009, Jannink 2010, Asoro *et al*. 2011).

Dans plusieurs études, le faible nombre d'animaux disponibles pour constituer la population de référence est un frein pour la mise en place de la sélection génomique, car il s'agit d'un facteur limitant pour améliorer la précision des valeurs génétiques estimées. C'est notamment le cas des bovins laitiers en Irlande (Berry 2009) : malgré l'utilisation de taureaux bien phénotypés les précisions obtenues dans d'autres pays ne sont pas atteintes car leur population de référence ne compte que 600 animaux, contre plusieurs milliers dans d'autres pays. Le faible nombre de taureaux dans la population de référence est aussi une hypothèse avancée par VanRaden *et al.* (2009) pour expliquer la grande variabilité des CD obtenus lors d'un essai de sélection génomique sur des taureaux Nord-Américains. Brito *et al.* (2011) ont observé dans une simulation sur des bovins

allaitants qu'avec un trop faible nombre d'animaux dans la population de référence la précision de la sélection génomique augmente très peu quand l'héritabilité du caractère augmente. A l'inverse de ces résultats, Liu *et al*. (2014) obtiennent chez des poulets de chair des précisions supérieures aux valeurs qu'ils attendaient compte-tenu de la petite taille de leur population de référence. Ils supposent que ce résultat inattendu pourrait être dû à une faible variabilité génétique dans leur lignée : le nombre de segments chromosomiques indépendants et donc le nombre d'effets à estimer serait faible, réduisant ainsi la quantité d'information nécessaire en entrée du modèle.

L'importance croissante de la taille de la population de référence pour utiliser la sélection génomique sur des caractères peu héritables a été observée par Hayes *et al*. (2009a) en bovins laitiers. Le même constat a été fait par Luan *et al.* (2009). Cependant Brito *et al*. (2011) trouve chez les bovins allaitants que pour un caractère trop peu héritable la multiplication par 4 de la taille de la population de référence ne suffit pas pour améliorer la précision de la sélection génomique.

Liu *et al*. (2011) ont voulu quantifier chez des bovins laitiers l'effet d'une augmentation de la taille de la population de référence sur l'estimation des effets des marqueurs. Quand leur population de référence passe de 700 individus à 5 000, la variance des effets estimés des SNPs est multipliée par 5. Mais la relation entre la précision de la sélection génomique et le nombre d'animaux dans la population de référence n'est pas linéaire : Erbe *et al.* (2013) par exemple trouvent qu'au-delà de 5 000 individus la précision n'est plus améliorée.

Le nombre d'animaux dans la population de référence apparait donc comme un facteur important dans la précision de la sélection génomique. Cependant le nombre d'individus en tant que tel ne suffit pas pour caractériser une population de référence.

La quantité d'information disponible sur les individus est un facteur important : Hayes *et al.* (2009a) montrent par exemple que pour un caractère déterminé par beaucoup de QTL à effets faibles il faudra beaucoup d'enregistrements de phénotypes pour estimer correctement les effets des SNPs. Luan *et al*. (2009) trouvent chez des bovins allaitants que les valeurs génétiques des pères sont mieux estimées connaissant les phénotypes de leurs descendants. Or, en pratique le coût des génotypages peut limiter le nombre d'animaux qui seront inclus dans la population de référence.

Mais la quantité de données n'est pas le seul facteur à prendre en compte dans la constitution d'une population de référence. Dans le cadre du GBLUP, l'intérêt d'inclure beaucoup d'animaux dans la population de référence est nuancé par Habier *et al*. (2013). Ils montrent qu'en ajoutant beaucoup d'animaux non-apparentés aux candidats dans la population de référence le risque d'erreur dans l'estimation de leur apparentement basé sur les marqueurs augmente. Les écarts trop importants entre les matrices d'apparentement génétique et génomique causent des erreurs dans les estimations des valeurs génétiques, ce qui réduit le gain en précision attendu par rapport à l'augmentation de la taille de la population de référence. Comme beaucoup d'autres, Liu *et al*. (2011) trouvent que les candidats à la sélection ont des CD plus élevés quand leur père fait partie de la population de référence. Lund *et al*. (2011) testent l'utilisation d'une population de référence commune en bovins laitiers obtenue par l'agrégation d'animaux de même race élevés dans différents pays européens. Contrairement aux résultats attendus, la fertilité bénéficie peu de cette augmentation du nombre d'individus. Les auteurs supposent que ce résultat est dû à des corrélations génétiques faibles pour ce caractère entre les populations des différents pays. Ces résultats mettent en lumière un point clé qui doit être pris en compte dans la constitution d'une population de

référence : l'apparentement entre les individus candidats et la population de référence, mais aussi à l'intérieur de la population de référence elle-même. La question de l'apparentement entre la population de référence et la population de validation, et à l'intérieur de la population de référence font l'objet des points suivants.

Importance de l'apparentement sur la précision de l'évaluation génomique

Meuwissen *et al*. (2009) estiment que le nombre d'individus dans la population de référence devrait être de $2N_eL$, N_e étant la taille efficace de la population et L la longueur du génome en Morgan, soit au moins 6 000 individus dans une population avec une taille effective de 100 si l'on veut atteindre une précision de 0.9. Si chez les Holstein N_e est en général faible (autour de 50), dans plusieurs espèces comme les chevaux la taille effective de la population peut atteindre plusieurs centaines, nécessitant selon la formule de Meuwissen *et al.* (2009) des dizaines de milliers d'individus. Mais ce résultat a été obtenu en supposant les individus non apparentés. Clark *et al*. (2012) trouvent que plus la taille de la population de référence est grande et moins l'apparentement entre les candidats et la population de référence a un effet sur la précision de la sélection génomique. Contrairement au cadre de la simulation de Meuwissen *et al.* (2009), dans la réalité les individus de la population de référence et de la population de validation sont apparentés. Cette partie présente des résultats obtenus sur la prise en compte de l'apparentement dans la constitution des populations de référence avec pour objectif d'atteindre la meilleure précision possible.

Quel apparentement entre la population de référence et les candidats ?

L'importance de l'apparentement entre la population de référence et les candidats pour améliorer la précision de la sélection génomique a été constatée à plusieurs reprises (Habier *et al*. 2007, Legarra *et al*. 2008). Liu *et al.* (2014) remarquent que la précision de la sélection génomique vérifiée par validation croisée est plus élevée quand les animaux sont répartis aléatoirement dans les groupes comparée à la précision obtenue quand les groupes sont constitués de façon à minimiser l'apparentement entre groupes. Ils proposent comme explication que la répartition aléatoire des animaux dans les différents groupes leur permet d'avoir des plein-frères (sœurs) et/ou des demi-frères (sœurs) dans la population de référence, et cet apparentement entre la population de référence et de validation permet une meilleure estimation des valeurs génétiques. Cleveland *et al*. (2012) trouvent chez des bovins que quand l'apparentement entre la population de référence et la validation augmente, la précision de la sélection génomique est moins sensible à la variation d'autres paramètres comme l'héritabilité ou bien la méthode d'estimation utilisée.

D'autres travaux ont cherché à décortiquer les sources de la précision en lien avec l'apparentement entre les populations de référence et de validation. Les notions d'apparentement et de déséquilibre de liaison sont très liées, et les effets de ces deux composantes sur la précision de la sélection génomique peuvent être étudiés conjointement. En effet, l'étendue du déséquilibre de liaison dans la population dépend de la taille efficace de la population N_e. Plus N_e est faible et plus les individus sont apparentés, et donc plus l'étendue du déséquilibre de liaison dans la population est importante, et moins il y aura de segments indépendants à estimer. Wientjes *et al*. (2013) montrent que l'apparentement entre la population de référence et les candidats explique une part plus importante de la précision que le déséquilibre de liaison quand la population est de petite taille. En revanche quand la population de référence est de grande taille ils trouvent que le déséquilibre de liaison a un

effet sur la précision plus important que l'apparentement. Le même résultat est obtenu par Habier *et al*. (2013).

Habier *et al*. (2013) vont plus loin en identifiant la co-ségrégation comme une source de précision pour la sélection génomique, au même titre que le déséquilibre de liaison et l'apparentement. Il s'agit d'une ségrégation non-indépendante des allèles d'un même gamète due à des liaisons entre loci. Ils précisent qu'il ne peut pas y avoir de co-ségrégation sans déséquilibre de liaison, mais qu'il ne s'agit pas de la même chose car la co-ségrégation mesure le déséquilibre de liaison chez les fondateurs de la population uniquement. Pour identifier la part de la précision due à la co-ségrégation, une simulation est faite de façon à avoir à la fois du déséquilibre de liaison et de la co-ségrégation (les individus sont apparentés, et les SNPs et les QTL sont simulés sur un même chromosome pour garantir la liaison) ou seulement du DL (les individus sont non-apparentés). Leurs résultats montrent que la co-ségrégation a un effet plus important sur la précision quand la taille de la population de référence est faible. Ils soulignent aussi que la précision due à la co-ségrégation et aux relations génétiques additives dépend beaucoup de l'apparentement, et que ces deux sources d'information peuvent donner la limite basse de la précision quand le déséquilibre de liaison dans la population est faible.

Plusieurs travaux ont donc montré que l'apparentement entre les populations d'apprentissage et de validation permet d'estimer les valeurs génétiques plus précisément : les SNPs pourront capturer les relations de parenté. Un apparentement important (en fait une faible diversité génétique) dans la population réduira le nombre de segments indépendants à estimer. Cependant, ces résultats ne signifient pas pour autant que la variabilité génétique doit être réduite. L'apparentement à l'intérieur de la population de référence a aussi été étudié.

Quel apparentement à l'intérieur de la population de référence ?

Pszczola *et al*. (2012) montrent que les précisions obtenues sont semblables pour différents niveaux d'apparentement moyen à l'intérieur de la population de référence. En revanche, ils trouvent un effet de la taille des familles de demi-frères : pour un niveau d'apparentement donné à l'intérieur de la population de référence, plus les familles de demi-frères au sein de la population de référence sont de petite taille, plus la précision de la sélection génomique augmente. En apparence, ce résultat pourrait signifier qu'il faut limiter l'apparentement à l'intérieur de la population de référence. Cependant, la méthode de simulation utilisée dans cette étude est telle qu'elle revient en fait à répartir les animaux dans des groupes pour une validation croisée de façon aléatoire ou en limitant l'apparentement entre les groupes, comme dans l'étude de Liu *et al*. (2014) citée précédemment. Quand l'apparentement entre les groupes est limité, ils contiennent de grandes familles de plein-frères et demi-frères et sont donc peu apparentés entre eux. Quand la répartition est aléatoire, ces familles sont réparties dans plusieurs groupes, et donc les candidats auront des plein-frères ou des demi-frères dans la population de référence. Ce travail redémontre l'importance de l'apparentement des candidats et de la population de référence, mais pas d'un apparentement réduit à l'intérieur de la population de référence.

Rincent *et al*. (2012) ont exploré différentes méthodes d'obtention de la population de référence chez le maïs: soit une minimisation de l'apparentement dans la population de référence, soit un algorithme qui teste l'effet de l'ajout d'individus à la population de référence sur le CD moyen des candidats restants. Ils expliquent qu'utiliser le CD serait plus intéressant que d'utiliser les erreurs

d'estimations des valeurs génétiques (ce qui a été fait dans d'autres travaux), car le CD prend en compte l'erreur d'estimation et la valeur génétique additive capturée. La méthode basée sur le CD moyen obtenu par les candidats est celle qui donne les meilleurs résultats. Les individus retenus dans la population de référence ne sont pas les mêmes suivant la taille de la population: quand elle est petite ce sont plutôt des individus extrêmes qui sont choisis (et la méthode de l'apparentement minimal donne les mêmes résultats), alors que quand la population est de grande taille les individus sont pris dans toute la population. Ils observent que comme la méthode qui minimise l'apparentement, la méthode basée sur le CD moyen obtenu par les candidats choisit pour la population de référence les individus les moins apparentés. Ces résultats sont similaires à ceux de Pszczola *et al*. (2012), car la minimisation de l'apparentement dans la population de référence augmente l'apparentement entre la population de référence et les candidats, ce qui entraine une meilleure estimation de leurs valeurs génétiques et donc une augmentation de leur CD. Isidro *et al*. (2015) utilisent également la méthode consistant à maximiser le CD moyen des candidats, et ils constatent eux aussi que ce critère réduit l'apparentement dans la population de référence pour augmenter l'apparentement entre la population de référence et les candidats. Isidro *et al.* (2015) appliquent cette méthode au blé et au riz, et la comparent à une méthode stratifiée. Cette méthode consiste à identifier les sous-groupes présents dans la population en étudiant la matrice d'apparentement génomique. Ensuite des individus sont pris au hasard dans chaque sous-population avec un nombre par sous-population proportionnel à leurs tailles respectives, ce qui doit assurer une variabilité importante dans la population de référence. Isidro *et al*. (2015) combinent aussi les deux méthodes en appliquant la méthode dite du CD moyen obtenu par les candidats au choix des individus à l'intérieur de chacune des sous-populations. En général leur méthode stratifiée donne de bons résultats, quelle que soit la taille de la population. Finalement ils montrent que dans une population où les sous-groupes sont bien distincts la méthode avec stratification donne une bonne précision. Quand la population a une structure moins tranchée la méthode du CD moyen des candidats serait préférable. Ces résultats sont cependant à nuancer car ils dépendent aussi du caractère étudié. Isidro *et al*. (2015) supposent donc qu'il y a un lien entre l'architecture génétique du caractère et la méthode à utiliser pour choisir les individus de la population de référence. La méthode du CD moyen requiert de plus un temps de calcul plus long que les autres méthodes testées. Ils concluent en revanche sur l'utilisation des erreurs d'estimation des valeurs génétiques des candidats comme critère pour inclure les individus dans la population de référence : avec cette méthode les individus retenus dans la population de référence sont plus apparentés qu'avec les autres méthodes, et donc l'apparentement avec les candidats est plus faible et la précision risque de décroitre plus rapidement au cours des générations.

Une autre façon d'aborder la composition de la population de référence est de s'intéresser au nombre de générations d'individus qui devraient en faire partie. Muir *et al.* (2007) trouvent avec une simulation d'une population animale que la précision de la sélection génomique est plus élevée quand le nombre de générations utilisées pour estimer les effets des marqueurs augmente. Il vaudrait mieux utiliser plusieurs générations de petite taille plutôt qu'une seule génération de grande taille. En revanche Bastiaansen *et al*. (2012) trouvent le résultat inverse, avec une précision plus élevée quand la population de référence est composée d'une seule génération au lieu de plusieurs. Chez l'avoine, Asoro *et al*. (2011) montrent avec des données réelles qu'ajouter des générations plus anciennes à la population de référence augmente la précision de la sélection génomique, et quand elle n'augmente pas elle n'est pas dégradée non plus.

Il semblerait donc d'après les travaux cités dans cette partie que l'importance d'un apparentement réduit dans la population de référence soit la conséquence « naturelle » d'un apparentement important entre les individus de la population de référence et ceux de la validation. Dans des populations de petite taille, minimiser l'apparentement dans la population de référence reviendrait à y inclure les individus les plus différents les uns des autres, ce qui garantit indirectement d'avoir beaucoup de variabilité dans la population de référence, et que des candidats très différents auront des apparentés dans la population de référence. Cette population de référence doit être enrichie par de nouveaux individus au fil des générations, comme nous allons le voir dans la partie suivante.

Quel enrichissement de la population de référence au cours du temps?

Beaucoup de travaux ont montré que la précision de la sélection génomique décroit quand des générations successives d'individus sont évaluées à partir de la population de référence initiale. Muir *et al.* (2007) trouvent que 5 générations après la première génération de validation la sélection génomique cesse d'être efficace. Zhong *et al*. (2009) obtiennent une précision plus faible pour la 4ème génération après l'entrainement du modèle que pour la 1ère génération de validation. Sur des données simulées, Hayes *et al*. (2009c) quantifient la perte en précision par une diminution du CD de 2% par génération. Cette perte en précision est due au fait que l'estimation des effets des marqueurs dépend du pédigrée et plus généralement des individus utilisés, et cette estimation n'est pas applicable à un autre groupe d'animaux séparés par plusieurs générations de la population de référence.

Pszczola *et al*. (2012) expliquent cette perte en précision par une diminution de l'apparentement entre la population de référence et les individus évalués. Habier *et al*. (2007) démontrent qu'au bout de plusieurs générations seul le déséquilibre de liaison apporte encore de l'information, et que donc pour mesurer la part de la précision due au DL présentée plus tôt on peut calculer la précision atteinte plusieurs générations après la première génération de validation. Habier *et al.* (2013) confirment ces résultats en montrant que le DL persiste plutôt bien au cours des générations, mais qu'en revanche la précision apportée par la co-ségrégation diminue au fil des générations. Ce résultat indiquerait que la co-ségrégation de SNPs et de QTL était réalisée sur des segments chromosomiques de grande taille, et que leur taille a diminué d'une génération à l'autre à cause de recombinaisons. Ce phénomène s'explique par le fait que la population est sélectionnée (Calus 2010) : la sélection est un moyen connu pour défaire le déséquilibre de liaison entre des SNPs et des QTL, ce qui survient quand la fréquence d'un allèle en un loci est beaucoup modifiée. Legarra *et al*. (2008) montrent que des apparentés éloignés apportent peu d'information sur les candidats comparés à des apparentés proches : or au fil des générations il y a bien une diminution de l'apparentement des nouveaux individus avec ceux qui composaient la population de référence. Des solutions ont été proposées pour maintenir la précision de la sélection génomique.

Meuwissen *et al*. (2001) ont montré sur des données simulées que la plupart de la variance génétique pouvait être capturée grâce au déséquilibre de liaison dans la population. Habier *et al.* (2007) obtiennent le même résultat sur données réelles. On a vu précédemment que cette source d'information n'était vraiment importante que pour des populations de référence de très grande taille. Dans ce cas une solution pourrait être d'utiliser une méthode Bayésienne comme le Bayes B (Habier *et al*. 2007, Hayes *et al*. 2009a, Meuwissen *et al*. 2009, Habier *et al.* 2007), qui est particulièrement adaptée pour capturer les informations apportées par le DL.

Cependant il est rare d'avoir une population de référence de grande taille. Il est plus recommandé de phénotyper et de génotyper régulièrement de nouveaux individus (Habier *et al*. 2007, Hayes *et al*. 2009a) apparentés aux nouveaux candidats (Legarra *et al*. 2008). Calus (2010) précise que cet enrichissement de la population de référence doit être réfléchi en fonction du temps nécessaire pour obtenir de nouveaux phénotypes. Si l'intervalle imposé par la durée du phénotypage est long, il pourra être plus intéressant d'utiliser un modèle bayésien qui donnera une précision plus stable dans le temps grâce à sa capacité à mieux capturer le DL.

Peut-on obtenir une bonne précision dans un contexte multiracial ?

Un nombre important d'individus doit constituer la population de référence pour que la sélection génomique soit suffisamment précise. Pour augmenter la taille d'une population de référence trop petite, il peut être tentant d'y inclure des individus d'une autre race. Il arrive aussi qu'une population soit multiraciale de par son histoire, ou encore que les pédigrées soient mal connus ou incomplet. Dans ces situations, les possibilités pour améliorer la population de référence en termes de taille ou d'apparentement sont limitées. Plusieurs solutions et développements ont déjà été proposés pour ces situations compliquées.

Hayes *et al*. (2009b) ont montré qu'il n'est pas possible d'estimer les valeurs génétiques d'animaux quand la population de référence est composées d'individus qui ne sont pas de la même race que les candidats. Ceci peut être dû au fait que des QTL peuvent ségréger dans une race mais pas dans les autres (Hayes *et al*. 2009b). Une solution peut être d'utiliser des populations mixtes rassemblant 2 races ou plus (Hayes *et al*. 2009a). De Roos *et al*. (2009) montrent que si des individus de 2 races sont évalués avec une même population de référence, il faut que les 2 races y soient représentées. Dans le cas contraire, les candidats de la race absente de la population de référence obtiendront des valeurs génétiques trop peu précises. La précision de la sélection génomique est d'autant plus faible que les races sont différentes (Daetwyler *et al*. 2008) ou que la divergence entre les 2 races est ancienne (Ibánez-Escriche *et al.* 2009).

Goddard et Hayes (2007) montrent qu'il faut un taux de recombinaison élevé dans les 2 races utilisées, mais aussi que les phases de liaison entre les QTL et les SNPs soient les mêmes dans les 2 races. Cette dernière condition, énoncée également par Hayes *et al*. (2009a) est remplie chez les Holstein et les Angus à condition que les SNPs soient séparés de moins de 10kb. Si ce n'est pas le cas, les SNPs ne captureront pas les mêmes effets aux QTL dans les différentes populations à cause de fréquences alléliques trop différentes dans ces 2 populations (Daetwyler *et al*. 2008). Ce résultat est cohérent avec celui de De Roos *et al*. (2009) qui montrent que l'évaluation de candidats à partir d'une population de référence contenant des animaux d'une race différente donnait de moins bons résultats quand la densité des marqueurs diminuait. Ibánẽz-Escriche *et al*. (2009) mettent en évidence le même effet positif d'une augmentation de la densité des marqueurs sur la sélection génomique multiraciale. Quand la densité de marquage est trop faible, comme chez le mouton par exemple, l'évaluation multiraciale ne peut pas être utilisée (Daetwyler *et al*. 2010).

Différentes solutions ont été proposées pour tenter de contourner ces inconvénients. Ibánẽz-Escriche *et al*. (2009) ont une population de référence et une population de validation pouvant contenir 2 races, et ils comparent un modèle où l'effet de l'allèle estimé est unique, et un modèle dans lequel l'effet de l'allèle est estimé suivant son origine paternelle ou maternelle. Le modèle estimant l'effet de l'allèle suivant son origine donne des valeurs génétiques qui ont la même

précision, voire une précision un peu meilleure qu'avec un modèle où l'effet estimé des allèles est unique. Son intérêt est plus important quand les races sont plus différentes, mais il diminue quand la densité de marqueurs augmente, car la densité plus importante des SNPs améliore l'estimation de leurs effets dans le modèle ou l'effet estimé de l'allèle est unique, sans distinction sur son origine paternelle ou maternelle. Thomasen *et al*. (2013) proposent pour des taureaux Danois de tenir compte de leur origine (réellement Danoise ou bien Nord-américaine) estimée à partir du pédigrée ou des marqueurs. Mais malgré l'utilisation en covariable de la proportion du pédigrée ou des marqueurs d'origine réellement Danoise, la précision reste très proche de celle obtenue avec une sélection génomique simple sans prise en compte de l'effet race.

Une modélisation comparable à celle de Thomsaen *et al.* (2013) consiste à utiliser des fondateurs ou groupes de parents inconnus (Miztal *et al*. 2013). Le principe des groupes de parents inconnus était déjà utilisé en sélection classique, avant l'arrivée de la sélection génomique. L'objectif initial était de prendre en compte le fait que des individus importés peuvent avoir des performances moyennes différentes de la moyenne de la population nationale. Dans la pratique, les individus dont on ne connait pas les parents sont répartis dans des groupes d'animaux nés de parents inconnus. Cela consiste à attribuer un même père et/ou une même mère fictifs aux individus que l'on considère comme issu du même groupe. La répartition peut se faire suivant la race de l'individu ou bien suivant sa période de naissance afin de prendre en compte le progrès génétique qui se traduira par des performances moyennes différentes pour des individus nés de parents inconnus à plusieurs générations de distance. Cette méthode est également utile quand le pédigrée est incomplet. En évaluation multiraciale, elle permet de tenir compte de l'origine des individus. Les groupes de parents inconnus sont inclus dans le modèle en tant que covariable. Chaque animal aura une pondération des groupes de parents inconnus en fonction de son pédigrée. Il est important de constituer les groupes de parents inconnus de façon à ce qu'ils soient suffisamment grands pour que leurs effets soient correctement estimés, et il faut prendre garde à ne pas construire des groupes qui se confondraient avec des effets fixes (Mistzal *et al*. 2013). Mais dans le cadre de la sélection génomique, Mistzal *et al.* (2013) rapportent plusieurs cas où l'utilisation de groupes de parents inconnus ne constitue plus une amélioration du modèle et au contraire produit des valeurs génétiques biaisées. Ils expliquent ces mauvais résultats par des différences trop importantes entre la matrice d'apparentement classique A et la matrice d'apparentement réalisé G. D'une part, la matrice G capture des relations d'apparentement qui ne sont pas visibles dans A et qui ne concordent pas avec les groupes de parents inconnus construits par les évaluateurs : ces groupes sont supposés non-apparentés, alors que les SNPs capturent en général des relations de parenté non-nulles. D'autre part, dans la plupart des cas on ne génotype pas la population entière, et la matrice G ne contient donc des informations que sur un échantillon récent de la population. Des améliorations ont été proposées pour remédier à ce problème. Christensen *et al*. (2012) ont proposé de modifier la matrice d'apparentement entre les groupes de parents inconnus ou fondateurs en introduisant un apparentement γ entre les fondateurs, et une consanguinité de $\gamma/2$ pour tous les fondateurs. La matrice d'apparentement entre les fondateurs n'est donc plus :

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

mais:

$$\boldsymbol{A^{\gamma}} = \begin{pmatrix} 1+\frac{\gamma}{2} & \gamma & \gamma \\ \gamma & 1+\frac{\gamma}{2} & \gamma \\ \gamma & \gamma & 1+\frac{\gamma}{2} \end{pmatrix}$$

Legarra *et al.* (2015) proposent de matérialiser cette modification de la matrice d'apparentement directement dans le pédigrée, en ajoutant un fondateur supplémentaire, le méta-fondateur, qui serait le père et la mère de tous les autres fondateurs. Ce fondateur n'est pas un individu mais le pool de gamètes dont sont issus les fondateurs. Pour prendre en compte des origines de races différentes, il est possible d'utiliser plusieurs méta-fondateurs, ce qui permet de connecter les groupes de parents inconnus dans le pédigrée. Ceci revient à adapter la matrice A à la matrice G, à l'inverse de ce qui est fait dans le single-step. Chacun des méta-fondateurs aura un coefficient de consanguinité propre, et des coefficients d'apparentement avec les autres méta-fondateurs du pédigrée.

Il est donc établi que la population de référence doit contenir un nombre important d'individus, avec un optimum au-delà duquel l'ajout d'individus supplémentaires ne permet pas de mieux estimer les valeurs génétiques. Il faut que les candidats soient apparentés à la population de référence pour avoir une bonne précision, et en conséquence la population de référence doit être enrichie de nouveaux individus au fil des générations de sélection. Il est possible d'utiliser plusieurs races dans la population de référence à condition que la race des candidats y soit représentée. Dans un contexte multiracial complexe (comme chez les chevaux de saut d'obstacle par exemple) les modèles peuvent être adaptés pour tenir compte de cette complexité. La partie suivante est consacrée aux principaux modèles qui ont été comparés depuis les débuts de la sélection génomique.

1.3.2. Comment choisir le modèle pour l'estimation des valeurs génétiques?

Une question qui se pose lors de la mise en place de la sélection génomique est celle du choix du modèle pour l'estimation des effets des SNPs. Meuwissen *et al.* (2001) en présentent plusieurs. Nous avons vu précédemment que ces modèles peuvent être classés en plusieurs groupes suivant leurs hypothèses sur les effets des marqueurs à estimer. D'une part certains modèles supposent que chaque SNP explique une part identique de la variance génétique, égale à la variance génétique totale divisée par le nombre de marqueurs. D'autre part, des modèles supposent que certains marqueurs expliquent une part de la variance génétique importante alors que d'autres expliquent une part faible voire nulle. Enfin d'autres modèles ne font aucune hypothèse sur la distribution des effets des SNPs. Les hypothèses sur les effets des marqueurs sont à mettre en relation avec l'architecture génétique du caractère pour lequel les valeurs génétiques sont estimées. Nous allons voir que depuis les débuts de la sélection génomique ces types de modèles ont été testés, comparés, voire mélangés.

Il existe des modèles sans *a priori* sur la distribution des effets des SNPs, comme la Partial Least Square Regression ou la Principal Component Analysis. Ces 2 méthodes font des régressions multi-variées et réduisent la dimension du jeu de variables à estimer (les SNPs) à un petit nombre de combinaisons linéaires. Solberg *et al*. (2009b) ont montré que ces méthodes donnent de moins bons résultats que le Bayes B. Elles sont en général peu utilisées, et la suite de cette partie sera plutôt axée sur le modèle linéaire et les méthodes bayésiennes.

Différence d'efficacité attendue entre les modèles suivant leurs hypothèses sur les effets des marqueurs

Les modèles présentés par Meuwissen *et al*. (2001) font des hypothèses sur la distribution des effets des QTL. D'une part le modèle infinitésimal suppose que chacun des SNPs explique une part de la variance génétique additive égale à la variance génétique additive totale divisée par le nombre de SNPs. D'autre part, les modèles bayésiens supposent qu'une faible part des SNPs a un effet sur la performance, et introduisent ce facteur *a priori* dans le modèle ou l'estiment à partir des données. Dans la pratique, ces différents modèles vont plus ou moins bien fonctionner, notamment en fonction de l'architecture génétique réelle du caractère.

Plusieurs travaux ont montré que le GBLUP donne des précisions aussi bonnes ou meilleures que les méthodes non-linéaires quand le déterminisme du caractère est polygénique, qu'il n'y a pas de gène à effet majeur pour le caractère. Zhong *et al*. (2009) obtiennent chez l'orge de bons résultats avec le GBLUP pour des caractères de ce type. Verbyla *et al.* (2009) font les mêmes observations chez des bovins laitiers : le GBLUP est le modèle le plus performant quand il n'y a pas de gène majeur expliquant le caractère. Hayes *et al.* (2010) observent les mêmes résultats pour des caractères de conformation chez des vaches laitières.

A l'inverse, les modèles utilisant des distributions *a priori* sur les effets des SNPs seraient plus performantes quand un nombre restreint de QTL a un effet sur le caractère. Zhong *et al*. (2009) trouvent chez l'orge que les valeurs génétiques de caractères connus pour être influencés par quelques QTL importants sont mieux estimées avec un modèle bayésien. Daetwyler *et al*. (2013) obtiennent des précisions un peu meilleures avec des modèles à sélection de variable quand peu de QTL influent sur le caractère. Goddard et Hayes (2007) expliquent que les méthodes bayésiennes utilisent des distributions *a priori* des effets des SNPs plus proches de la réalité, par exemple les caractères de production laitière gouvernés par environ 150 QTL. Daetwyler *et al*. (2010) ont étudié la relation entre architecture génétique et choix du modèle avec des simulations. Là encore avec la méthode Bayes B les résultats sont meilleurs quand le nombre de QTL est faible. Ils montrent que les résultats du Bayes B dépendent du nombre de segments indépendants dans le génome, qui dépend de la taille efficace de la population et du nombre de QTL (Daetwyler *et al*. 2010). Meuwissen *et al*. (2009) montrent que la supériorité du Bayes B sur le GBLUP est accrue quand la densité des marqueurs augmente, cette densité supérieure devant permettre de mieux capturer les effets des QTL les plus importants.

Il pourrait sembler au vu de ces résultats que le choix du type de modèle à utiliser pour la sélection génomique est simple et dicté par l'architecture génétique du caractère, en supposant qu'on la connaisse. Cependant, cette information n'est pas toujours disponible. Dans un cas comme dans l'autre, le choix du modèle décrivant une « réalité » fausse devrait avoir des conséquences sur la précision. Meuwissen *et al*. (2001) expliquent par exemple que le Bayes A et le Bayes B estiment bien les effets des gros QTL, mais que ces modèles risquent de ne pas bien estimer les effets des QTL moins importants qui contribuent pourtant à la variance génétique totale. Au contraire, le GBLUP n'étant pas sensible à l'architecture génétique du caractère (Daetwyler *et al*. 2010), il ne peut pas mettre à 0 les effets des SNPs n'influant pas sur le caractère, ce qui introduirait du bruit dans l'estimation des valeurs génétiques (Goddard et Hayes 2007).

Des propositions ont été faites pour améliorer ces modèles. Resende *et al*. (2012) combinent le GBLUP à une méthode bayésienne. Les effets des marqueurs sont estimés initialement avec un

GBLUP. Ensuite, les SNPs sont classés suivant l'importance de leurs effets et répartis dans des groupes. Les valeurs génétiques sont estimées en utilisant un nombre croissant de groupes en commençant par ceux contenant les SNPs ayant les effets les plus importants, jusqu'à ce que l'ajout d'un nouveau groupe de SNPs n'améliore plus la précision de la sélection génomique. Chez le pin, Resende *et al.* (2012) trouvent que cette méthode donne de meilleurs résultats que le Bayes B et le GBLUP.

Les méthodes comme le Bayes Cπ décrit précédemment ou encore le Lasso bayésien peuvent être plus performantes que le Bayes B à condition que le nombre de QTL expliquant le caractère soit très faible (Daetwyler *et al*. 2013). La Bayes Cπ a aussi l'avantage d'avoir un temps de calcul réduit comparé au Bayes B, et serait plus polyvalent dans les cas où l'architecture génétique n'est pas connue (Habier *et al*. 2011). Dans la méthode du Lasso bayésien, les effets des SNPs les plus faibles sont mis à 0. La précision de la sélection génomique n'est pas forcément améliorée, mais le temps de calcul est réduit vu que le nombre d'effets à estimer est plus faible (Verbyla *et al.* 2009). Cependant ces méthodes nécessitent que les SNPs soient effectivement en déséquilibre de liaison avec les QTL. Si la quantité de marqueurs n'est pas suffisante ou leur répartition pas adaptée, des effets des SNPs risquent d'être mis à 0 à tort par le modèle (Goddard et Hayes 2007, Su *et al*. 2010).

Wang *et al*. (2012) ont développé une méthode permettant de prendre en compte l'architecture génétique du caractère avec le modèle du single-step. Ils réalisent une première évaluation en une étape. En fonction des effets estimés des SNPs, ils créent une nouvelle matrice génomique qui est pondérée. Cette matrice est ensuite ré-utilisée dans un single-step donnant les valeurs génétiques estimées définitives. L'objectif était de proposer une méthode combinant les avantages du single-step et du Bayes C.

Une importance du choix du modèle à nuancer

Au vu de ces résultats, le choix d'un modèle inadéquat du point de vue de l'architecture génétique du caractère et de la quantité de données disponibles risque d'être préjudiciable pour la précision de la sélection génomique. Cependant, les différences de résultats attendues entre les modèles ne sont pas toujours observées. VanRaden (2008) compare des méthodes linéaires et non linéaires, et les animaux obtiennent quasiment les mêmes CD avec les 2 types de méthode. VanRaden *et al*. (2009) trouvent une corrélation entre les valeurs génétiques estimées avec des modèles linéaires ou non-linéaires proche de 1. Verbyla *et al*. (2010) comparent différentes distributions *a priori* pour des méthodes bayésiennes. Ils obtiennent dans tous les cas des valeurs génétiques estimées corrélées à plus de 85% avec les vraies valeurs génétiques, indiquant que les modèles bayésiens seraient peu sensibles aux changements de distributions *a priori*. Liu *et al.* (2014) comparent le GBLUP et le Lasso bayésien pour l'évaluation de caractères de croissance et de carcasse chez des poulets de chair et obtiennent des précisions similaires avec les deux méthodes. Lourenco *et al*. (2014) ont comparé le single-step et le single-step pondéré et trouvent une amélioration faible voire nulle quand les effets des SNPs sont pris en compte pour pondérer la matrice G. Les écarts de résultats entre des modèles faisant différentes hypothèses sur les effets des SNPs ne seraient donc pas toujours vérifiés. Daetwyler *et al*. (2010) ont montré avec des simulations que le GBLUP ne serait en fait pas sensible à l'architecture génétique du caractère, à moins que le nombre de QTL responsable de la variabilité génétique du caractère soit inférieur à 10. Meuwissen *et al*. (2013) expliquent ce phénomène par le fait que dans la pratique le nombre de QTL à effet faible est très élevé : l'hypothèse selon laquelle chaque SNP est en DL avec au moins un QTL et a donc un effet non-nul est vérifiée. De plus, très peu

de QTL seraient en déséquilibre de liaison parfait avec un seul SNP, d'où l'intérêt d'utiliser beaucoup de SNPs.

Même si le modèle infinitésimal n'est *a priori* pas le modèle qui décrit le mieux la réalité, les précisions obtenues avec ce modèle sont bonnes (Zhang *et al.* 2004), et nous avons vu que dans beaucoup d'études il permet d'atteindre des précisions proches de celles obtenues avec des modèles bayésiens supposés mieux décrire la réalité biologique. Les modèles bayésiens supposent une distribution des effets des marqueurs permettant de prendre en compte les effets très importants d'éventuels gènes majeurs. Ces modèles restent en général recommandés quand la présence de gènes majeurs est suspectée.

L'ajout d'un effet polygénique améliore-t-il la précision?

Les modèles précédents ont été développés et sont en général appliqués en supposant que les marqueurs sont suffisamment nombreux et bien répartis sur le génome pour être en déséquilibre de liaison avec les QTL et capturer leurs effets. Cependant dans la pratique ce n'est pas toujours le cas. Goddard et Hayes (2007) indiquent qu'un terme polygénique résiduel peut être ajouté dans le modèle. A partir du pédigrée, ce terme capture la variance génétique qui n'est pas capturée par les marqueurs. Ainsi, si un QTL a un effet sur le caractère mais est mal pris en compte dans l'évaluation car sa fréquence dans la population est trop faible, l'effet polygénique devrait pouvoir en tenir compte (Hayes *et al*. 2009a). Liu *et al.* (2011) ont étudié les conséquences de l'ajout d'un effet polygénique plus ou moins important dans le modèle. Ils observent qu'augmenter la variance polygénique résiduelle diminue la variance des effets des SNPs avec les effets les plus importants, et la valeur de leurs effets. D'après leurs résultats, la part de variance polygénique résiduelle optimale dépend du caractère : moins le caractère est héritable et plus la part de variance génétique résiduelle dans le modèle devrait être élevée. Gao *et al*. (2012) obtiennent des CD pour des bovins laitiers plus élevés de 0.3% pour les valeurs génétiques estimés avec un GBLUP incluant un effet polygénique comparés aux CD obtenus sans cet effet. Ils obtiennent aussi des valeurs génétiques estimées moins biaisées. Solberg *et al*. (2009b) avaient également observé une réduction du biais pour les valeurs génétiques estimées d'une population simulée. Dans leur cas, plusieurs générations de plus en plus distantes de la population de référence avaient été évaluées, et la réduction du biais persistait au cours des générations. Ils n'observaient cependant pas d'augmentation de la précision de l'évaluation génomique. En effet, l'intérêt de l'effet polygénique est conditionné par la qualité du déséquilibre de liaison entre SNPs et QTL. Liu *et al*. (2014) ont testé l'ajout d'un effet polygénique dans leur modèle pour des caractères d'héritabilité faible à intermédiaire (entre 0.19 et 0.44) chez le poulet de chair et n'obtiennent pas d'amélioration, ce qui les amène à conclure que leur puce 60K capture bien les effets des QTL.

La question du choix du modèle a donc été largement abordée depuis les débuts de la sélection génomique. Si les modèles bayésiens utilisent des distributions *a priori* sensées mieux décrire la réalité biologique, dans la pratique le modèle du GBLUP donne souvent des précisions très proches voire aussi bonnes.

1.3.3. Quel effet du choix des marqueurs sur la précision de l'évaluation génomique ?

Quelle quantité de marqueurs utiliser, comment les répartir sur le génome ?

La sélection génomique repose sur la capacité des marqueurs à capturer les effets des SNPs grâce au déséquilibre de liaison (Meuwissen *et al*. 2001). Il faut donc que les marqueurs soient suffisamment

nombreux et bien répartis sur le génome. En pratique les puces commercialisées sont uniques et ne permettent pas de choisir une densité ou une répartition des marqueurs. Cependant les effets des caractéristiques des marqueurs sur la précision ont été étudiés dans plusieurs travaux.

L'effet positif d'une augmentation du nombre de marqueurs sur la précision de la sélection génomique a été observé à maintes reprises : chez les bovins laitiers par VanRaden *et al.* (2009), chez les bovins allaitants par Brito *et al*. (2011), chez l'orge par Zhong *et al*. (2009), chez l'avoine (Asoro *et al*. 2011). En utilisant plus de marqueurs on augmente leur densité sur le génome, et donc les chances que les SNPs soient en déséquilibre de liaison élevé avec les QTL. La densité requise varie d'une espèce à l'autre et d'une race à l'autre en fonction de la longueur du génome et de l'étendue du déséquilibre de liaison (Goddard et Hayes 2007). Chez les bovins laitiers le déséquilibre de liaison entre des loci séparés de 50kb est de 0.35. Pour avoir ce déséquilibre de liaison entre les SNPs il faudrait une puce comptant au moins 60 000 marqueurs. Quand la variabilité génétique dans une population est plus grande, sa taille efficace N_e est plus grande également, et en conséquence le nombre de segments indépendants dans le génome est plus important aussi (Habier *et al*. 2009). Il faudra pour ces populations un nombre de marqueurs plus important afin d'estimer correctement les effets de chacun des segments.

Le nombre de SNPs à utiliser pour estimer correctement les valeurs génétiques dépend de la variabilité dans la population, mais aussi du modèle utilisé. Luan *et al.* (2009) et Solberg *et al.* (2009b) montrent que le modèle Bayes B en particulier est très sensible au nombre de marqueurs. Comme ce modèle suppose *a priori* une distribution des effets des SNPs variable, il faut que les marqueurs soient suffisamment nombreux pour identifier les QTL ayant les plus forts effets, mais également les QTL ayant des effets plus modérés mais participant tout de même à la variance génétique additive.

Cependant, passé un certain seuil, augmenter le nombre de SNPs ne va plus améliorer la précision de la sélection génomique. Moser *et al*. (2010) obtiennent ce résultat sur des données simulées. Sur des données réelles de bovins laitiers, Jensen *et al*. (2012) montrent que 44 000 marqueurs capturent 96% de la variance génétique additive, et qu'il y aurait donc peu d'intérêt à utiliser plus de SNPs. Erbe *et al*. (2013) vérifient cette hypothèse en comparant la précision de la sélection génomique obtenue avec une puce 50K et une puce imputée de 700K : la part de la variance génétique additive expliquée par les marqueurs augmente très peu alors que leur densité est multipliée par plus de 10. Dans la même étude ils retrouvent par ailleurs le fait que le nombre de marqueurs nécessaires varie en fonction de la race, puisque chez la Brune Suisse la part de la variance génétique additive capturée n'augmente plus au-delà de 20 000 marqueurs. Habier *et al.* (2013) ont étudié les sources de la précision de la sélection génomique. Ils montrent que le nombre de SNPs nécessaires pour atteindre une précision donnée varie en fonction du déséquilibre de liaison. Ce nombre de SNPs « idéal » est très inférieur au nombre de SNPs présents sur les puces à haute densité, ce qui expliquerait pourquoi la précision est rarement beaucoup améliorée par le passage d'une puce 50K à une puce 700K.

Un nombre élevé de marqueurs doit donc permettre de capturer la plus grande partie de la variance génétique, avec une valeur seuil après laquelle une augmentation de la densité n'est plus profitable. Cependant il faut aussi que ces marqueurs soient correctement positionnés sur le génome pour capturer les effets des QTL.

D'après Muir *et al*. (2007), si la position des QTL n'est pas connue les marqueurs devraient être choisis de façon à être séparés par des intervalles de même taille. Schaeffer *et al*. (2006) proposent

pour les bovins laitiers qu'une nouvelle puce soit conçue connaissant les SNPs les plus utiles pour l'évaluation, afin d'optimiser la répartition des SNPs pour obtenir des valeurs génétiques plus précises. Moser *et al*. (2010) montrent avec des simulations que quand beaucoup de SNPs sont disponibles leur répartition a peu d'effet sur la qualité des estimations. En revanche quand le nombre de SNPs est limité, ce qui peut arriver dans la pratique quand le coût des puces nécessite de restreindre le nombre de marqueurs, la précision de la sélection génomique est beaucoup plus sensible à leur répartition sur le génome. En effet des SNPs en nombre insuffisant ou mal répartis ne seront pas capables de capturer les effets des QTL faute d'un déséquilibre de liaison suffisant.

L'utilisation d'haplotypes à la place de SNPs a été proposée par Goddard et Hayes (2007). Un haplotype est constitué de plusieurs SNPs consécutifs situés sur le même chromosome et qui ségrégent ensemble au cours de la méiose. Cette méthode serait intéressante dans le cas de QTL en déséquilibre de liaison incomplet avec des SNPs qui auraient par contre un meilleur déséquilibre de liaison avec des haplotypes. Mais à la différence des SNPs les haplotypes ne sont pas nécessairement bi-alléliques, et il existe donc souvent pour un même loci plus de 2 versions de l'haplotype, ce qui requiert une quantité de données suffisante pour estimer correctement les effets des différentes versions des haplotypes présents dans une population. L'utilisation d'haplotypes nécessite tout d'abord de les définir, notamment en choisissant une taille de fenêtre, c'est à dire le nombre de SNPs consécutifs qui seront considérés pour repérer les haplotypes, et aussi le taux de ressemblance entre les marqueurs qui sera utilisé pour définir les limite des haplotypes. Mais d'après Calus *et al*. (2008), ces deux paramètres influent peu sur la précision de la sélection génomique. Calus *et al.* (2008) montrent que l'utilisation d'haplotypes permet d'obtenir une meilleure précision quand la densité des SNPs est faible. Au cours de ma thèse, je n'ai pas utilisé d'haplotypes en sélection génomique, en revanche j'ai réalisé une détection de QTL pour la performance en saut d'obstacles en utilisant des haplotypes afin de mieux capturer d'éventuels QTL en déséquilibre de liaison incomplet avec les SNPs.

Il faut donc un nombre minimum de SNPs suffisamment bien répartis sur le génome pour capturer les effets des QTL et estimer les valeurs génétiques des individus. Plusieurs travaux ont montré qu'au-delà d'un nombre de marqueurs qui dépend de la population, l'augmentation du nombre de SNPs utilisés ne permet plus d'augmenter la précision de la sélection génomique. Ces résultats sont vrais dans le cadre simple où ils ont été obtenus : celui de populations sélectionnées en race pure. Dans des cas plus complexes, une augmentation de la densité des marqueurs peut se révéler intéressante.

Intérêt d'utiliser plus de SNPs : les puces à haute-densité

Dans la partie consacrée aux populations de référence, on a vu que l'augmentation de la population de référence par ajout d'animaux d'autres races améliorait rarement la précision de la sélection génomique à cause d'une densité de SNPs insuffisante (Daetwyler *et al*. 2010). Goddard *et al.* (2006) et Hayes *et al*. (2009b) estiment qu'il faudrait un espacement inférieur à 10kb entre marqueurs consécutifs. Avec suffisamment de marqueurs, on devrait avoir un déséquilibre de liaison correct entre SNPs et QTL quelles que soient les races utilisées dans la population de référence (Wientjes *et al*. 2013). Dans ce cas il devient possible de faire des évaluations avec des races différentes dans la population de référence et dans la validation.

Cependant Muir *et al*. (2007) précisent que si augmenter la densité des SNPs peut effectivement permettre de mieux capturer les effets des QTL, il faut que le nombre de phénotypes soit augmenté également au risque sinon de détériorer la précision par une mauvaise estimation des effets des SNPs. Meuwissen *et al*. (2009) font la même recommandation.

Il serait également possible d'utiliser la séquence complète du génome en sélection génomique. En effet on peut aujourd'hui séquencer des fragments d'ADN suffisamment longs pour être alignés sur la séquence complète déjà connue d'une espèce, donnant ainsi accès à la séquence complète d'un individu. D'après Meuwissen *et al*. (2013) ce type de données pourrait être bien exploité avec les modèles non-linéaires. Les polymorphismes causaux devraient être accessibles, et donc plusieurs dizaines de SNPs ne seront plus nécessaires pour capturer l'effet d'un QTL et certains pourront être mis à 0. Une étude sur données simulées a montré qu'avec les données de séquence il n'y aurait pas de perte en précision 10 générations après la population de référence. Mais les coûts de génotypage sont encore très élevés pour cette technique. Pour l'instant l'application envisagée par Meuwissen *et al.* (2013) consisterait à séquencer les principaux fondateurs d'une population puis à imputer les séquences de leurs descendants. Le projet 1 000 génomes bovins a appliqué cette méthode chez les bovins laitiers en séquençant des ancêtres importants dans plusieurs races. 234 séquences complètes d'individus de race Holstein, Simmental ou Jersey ont permis d'identifier une mutation responsable d'une maladie létale. A partir des séquences imputées, des variants liés à la production laitière ont été identifiés et pourront être utilisés pour sélectionner plus finement les animaux (Daetwyler *et al*. 2014). On voit donc qu'augmenter le nombre de marqueurs sur le génome peut être un moyen d'améliorer la précision de la sélection génomique en augmentant le DL entre SNPs et QTL. Cependant utiliser plus de marqueurs coûte plus cher, et pour que la sélection génomique puisse être mise en place il faut parfois envisager d'utiliser moins de marqueurs afin que le coût soit supportable. Nous allons voir dans la partie suivante que l'imputation peut être utilisée dans cette optique de réduction des coûts de la sélection génomique.

Peut-on estimer les valeurs génétiques précisément en utilisant moins de marqueurs ?

Habier *et al.* (2009) proposent pour réduire le coût de la sélection génomique d'utiliser un panel restreint de SNPs plutôt que d'augmenter la densité des marqueurs. Ils envisagent deux types de sélection pour constituer les panels de marqueurs : soit prendre des SNPs à intervalles réguliers, soit retenir les SNPs qui capturent une part plus importante de la variance génétique. Ces 2 alternatives ont des avantages et des inconvénients. Une puce où les marqueurs sont choisis suivant la part de variance qu'ils expliquent permet d'atteindre une précision proche de celle obtenue avec une puce classique de 50 000 marqueurs, mais elle a l'inconvénient d'être spécifique à chacun des caractères, ce qui réduit son intérêt du point de vue des coûts. Avec une puce où les marqueurs sont pris à intervalle régulier la perte en précision est plus importante, mais cette puce est polyvalente. Comme elle couvre l'ensemble du génome elle permet aussi de prévenir la fixation d'allèles indésirables pour d'autres caractères que ceux sélectionnés (Habier *et al*. 2009). Finalement ils proposent une utilisation du génotypage à basse densité pour faire un tri des animaux. Les individus candidats pourraient être génotypés à basse densité seulement, et leur parents dans la population de référence à la densité habituelle. Une fois retenus pour être reproducteurs, les candidats génotypés à basse densité pourraient être re-génotypés avec une densité supérieure de marqueurs pour être inclus à la population de référence. Weigel *et al*. (2009) s'intéressent eux aussi à la faisabilité d'une sélection génomique basée sur des puces à basse densité. Chez des bovins laitiers, ils choisissent un sous-ensemble de SNPs qui rassemblent les marqueurs apportant le plus d'information sur le caractère. Ils

montrent que 300 SNPs choisis avec la Lasso bayésien expliquent 50% de la variance capturée pour l'index du net merit avec la puce 50K. Ils trouvent eux aussi que la précision est plus élevée quand les SNPs sont choisis en fonction du caractère et non pris au hasard dans le génome. Dans leur étude, les SNPs retenus avec le Lasso bayésien sont en majorité sur 5 chromosomes, ce qui confirme le risque soulevé par Habier *et al.* (2009) qui est de ne plus avoir accès à une partie du génome et donc de ne pas pouvoir prévenir la fixation éventuelle d'allèles indésirables.

Les puces à basse densité sont aussi utilisées pour imputer des génotypes de densité classique, autour de 50 000 marqueurs. L'imputation à partir de puces à basse densité consiste à prédire statistiquement les allèles présents au SNPs manquants. La prédiction est réalisée en combinant deux informations : connaissant une partie des SNPs portés par l'individu qui a été génotypé à basse densité, les SNPs manquant par rapport à la puce de densité intermédiaire sont inférés connaissant les génotypages d'autres individus de la population séquencés à cette densité intermédiaire. Grâce au déséquilibre de liaison entre les SNPs, il est ainsi possible de reconstituer un génotypage de densité haute ou moyenne à partir de génotypages à basse densité.

Pszczola *et al*. (2011) montrent que l'ajout d'individus dont les génotypes sont imputés dans la population de référence n'améliore la précision que si l'imputation est suffisamment précise. Dassonneville *et al*. (2011) testent l'imputation d'une puce 50K à partir d'une puce 3K sur données réelles chez des Holsteins, en utilisant des populations de référence nationales ou bien la population d'Eurogenomix. Ils montrent que le taux d'erreur d'imputation est un peu plus faible quand la population d'Eurogenomix est utilisée, et que dans tous les cas la perte en précision sur les EBVs est faible. Comme Habier *et al*. (2009), ils proposent d'utiliser les puces à basse densité pour réaliser une pré-sélection des jeunes animaux, et comme Habier *et al*. (2009) ils soulignent l'importance de conserver des animaux génotypés à 50K dans la population de référence afin que les imputations restent de qualité. Hayes *et al*. (2012) ont comparé des imputations intra-races à des imputations inter-races chez les ovins. D'après leurs résultats, l'imputation d'une race vers l'autre est meilleure quand la diversité à l'intérieur des races est faible. Quand l'imputation se fait au sein d'une même race, ce sont les génotypes des animaux les plus proches de la population de référence qui sont les mieux imputés. Ils concluent finalement qu'il vaut mieux utiliser une population d'une seule race homogène, même petite, plutôt qu'une population de référence incluant plusieurs races différentes. Berry *et al*. (2014) trouvent eux aussi que l'imputation intra-race donne de meilleurs résultats que l'imputation inter-races chez des bovins laitiers et des bovins allaitants. Ventura *et al*. (2014) font le même constat chez des bovins allaitants. Concernant l'ajout d'animaux d'une autre race dans la population de référence pour l'imputation, Berry *et al*. (2014) trouvent que les résultats dépendent de la combinaison de races qui est faite : ajouter des vaches Holsteins dans la population de référence de Rouges Danoises améliore l'imputation des Rouges Danoises, mais l'ajout de Rouges Danoises dans la population de référence pour les Holsteins diminue la précision de l'imputation des Holsteins. Hayes *et al*. (2012) ont aussi comparé différentes puces basse densité. Dans leur cas, il faut une puce 5K pour imputer la puce 50K avec une précision de 80%, alors que chez les bovins laitiers une puce 3K est suffisante.

VanRaden *et al*. (2013) et Hayes *et al*. (2012) ont testé l'imputation de la puce 50K vers une puce 700K. Chez Hayes *et al*. (2012), les animaux n'étant pas génotypés à haute densité, la précision de l'imputation a été vérifiée en supprimant quelques-uns des SNPs de la puce 50K afin de les imputer. VanRaden *et al*. (2013) et Berry *et al.* (2014) ont aussi vérifié la faisabilité de l'imputation de données

de séquence à partir d'une puce 700K. VanRaden *et al*. (2013) et Hayes *et al*. (2012) montrent qu'il est possible d'imputer une puce à haute densité à partir d'une puce 50K avec une excellente précision. L'imputation de données de séquence à partir d'une puce 700K est aussi possible. Dans leur cas qui est uni-racial le gain en précision obtenu en utilisant les génotypes à haute densité imputés est très faible, ce qui peut être dû au fait que dans leur population les marqueurs de la puce 50K sont en déséquilibre de liaison avec les QTL. L'imputation de la puce 700K pourrait être plus intéressante en population croisée. Berry *et al.* (2014) trouvent en plus qu'il est possible d'imputer d'une race à l'autre des données de séquence à partir de la puce 700K. D'après les résultats de Berry *et al*. (2014), la précision de l'imputation peut varier le long du génome même si elle est globalement la même sur tous les chromosomes. Il y a des régions où la précision est un peu plus faible quelles que soient les combinaisons de puces testées, ce qui pourrait selon eux être dû à la présence de zones où le taux de recombinaison est très élevé (points chauds de recombinaison) ou à des erreurs d'annotation du génome.

Il est donc reconnu qu'en race pure il y a un nombre optimum de marqueurs à utiliser, au-delà duquel la précision n'est plus améliorée par l'ajout de nouveaux SNPs. Il est intéressant d'utiliser plus de marqueurs dans des contextes de sélection multi-raciaux afin de capturer les effets des QTL dans des races différentes où les fréquences et les effets des QTL peuvent varier. Il est aussi possible d'utiliser moins de marqueurs pour estimer les valeurs génétiques sans perdre en précision, à condition que les SNPs retenus soient ceux qui expliquent le mieux la variance génétique additive. Les puces de petite taille peuvent aussi être utilisées pour imputer les puces de densité habituelle et obtenir la même précision. Dans ce cas on peut envisager une amélioration de la précision dans la mesure où le coût par individu plus faible permettra de génotyper plus d'animaux.

1.4. Conclusion de la partie bibliographique sur la sélection génomique

Les facteurs permettant de faire varier la précision de la sélection génomique ont été présentés dans trois sous-parties distinctes : choix de la population de référence (taille, apparentement, enrichissement, composition raciale), choix du modèle (proche de la réalité biologique ou non), choix des marqueurs (densité, répartition, utilisation de sous-ensembles ou de panels plus larges). Il est apparu plusieurs fois que ces trois paramètres interagissent : agrandir la population de référence avec des individus de différentes races est intéressant à condition d'avoir une densité de marqueurs élevée, les modèles bayésiens sont pénalisés quand le nombre de marqueurs est insuffisant, les modèles bayésiens capturent bien le déséquilibre de liaison et peuvent être plus intéressants dans des populations de très grande taille… A cause de ces interactions, il n'existe pas de règle simple pour définir les conditions dans lesquelles la précision de la sélection génomique sera maximale. Plusieurs travaux, non-cités dans cette partie, ont voulu décomposer la précision de la sélection génomique en utilisant quelques paramètres plus ou moins simples à obtenir sur le caractère, la population de référence et les marqueurs. Leur objectif était de proposer des formules permettant d'estimer la précision de la sélection génomique à partir de ces paramètres : le sélectionneur a ainsi un moyen simple d'estimer la précision qu'il peut espérer obtenir avant d'avoir à génotyper, et le cas échéant il peut ajuster les différents paramètres qui peuvent l'être pour obtenir une meilleure précision. L'étude de ces formules fera l'objet du chapitre 3.

Le chapitre suivant présente des disciplines pour lesquelles les chevaux de sport sont sélectionnés en France ainsi que les races correspondantes, et fait un état des lieux de la sélection actuellement réalisée dans ces populations.

2. Le cheval athlète en France

2.1. Introduction : évolution de l'utilisation du cheval

L'usage du cheval a beaucoup évolué. Domestiqué en 5 000 av. JC, sa présence aux côté de l'Homme a participé à sa sédentarisation : d'abord élevé pour sa viande, il est ensuite devenu un partenaire pour le travail et pour la guerre. Son élevage coûteux a longtemps été réservé aux classes les plus aisées de la société. Jusqu'à la fin du XIXe siècle, l'énergie qu'il pouvait fournir a été utilisée pour le travail aux champs, la traction de véhicules et la guerre. La seconde révolution industrielle a entrainé le déclin de son utilisation avec le développement de nouvelles sources d'énergie aussi bien pour l'agriculture que pour les transports, tandis que les guerres de tranchées remplaçaient les guerres de conquête. De 3 à 3,5 millions de chevaux en 1900, la population équine ne comptait plus que 450 000 chevaux dans les années 1970. Dans ce contexte le cheval de traction ou de selle a progressivement été transformé ou remplacé par le cheval de loisir, de course ou de compétition. La consommation de viande de cheval, pourtant très faible en France, a permis le maintien de 9 races de chevaux de trait, initialement élevées pour le travail de la terre (REFErences 2011a).

La pratique de l'équitation s'est démocratisée : aujourd'hui, la Fédération Française d'Equitation est la 3ème de France derrière celles du football et du tennis, avec plus de 700 000 licenciés en 2013, et un nombre de cavaliers total d'environ 2,2 millions (IFCE-OESC 2015b). Le nombre d'équidés estimé (identifiés ou non) fin 2012 serait de 1 million (IFCE-OESC 2015a). En marge de la pratique de l'équitation, le cheval se trouve aussi de nouveaux usages: il devient territorial dans les collectivités en assurant le ramassage des déchets ou le transport de personnes, médiateur par son utilisation en milieu carcéral, ou encore thérapeute via l'équitation thérapeutique, l'hippothérapie, l'équitation adaptée… (REFErences 2011b)

Parmi toutes ces pratiques plus ou moins récentes, celles qui ont fait l'objet d'études au cours de ma thèse sont le concours de saut d'obstacles, le concours complet d'équitation, les courses d'endurance et les courses au trot.

2.2. Usages du cheval athlète : compétitions équestres et courses hippiques

Deux des disciplines étudiées au cours de ma thèse sont présentes aux Jeux Olympiques : le Concours de Saut d'Obstacles (CSO) et le Concours Complet d'Equitation (CCE). Le dressage fait également partie des disciplines olympiques mais ne sera pas étudié en raison de la faible quantité de données disponibles.

La première compétition de CSO, alors appelé « concours hippique », a été organisée en 1870 par la Société Hippique Française (SHF). Trente ans plus tard, cette discipline entre aux Jeux Olympiques. C'est aujourd'hui le type de compétition le plus répandu en France : le CSO représente 90% des sorties en compétitions organisées chaque année. Le CCE a des origines militaires. Démocratisé à la fin des années 1980, il s'agissait à l'origine d'un ensemble d'épreuves visant à vérifier les qualités des chevaux de l'armée. Le CCE est une discipline olympique depuis 1912. Aujourd'hui il se compose de trois épreuves : une épreuve de dressage, une épreuve de cross aussi appelée « épreuve de fond » et une épreuve de saut d'obstacles aussi appelée « hippique ». Les courses d'endurance sont organisées depuis le 19ème siècle. L'objectif d'une épreuve est d'effectuer un parcours de plusieurs kilomètres le plus rapidement possible en maintenant l'intégrité physique du cheval. Les courses au trot ont une

origine rurale. En France la première course a été organisée en 1836. Le cheval est attelé (ou monté dans certaines courses) et doit courir une courte distance le plus rapidement possible au trot, sans changer d'allure. Contrairement aux disciplines équestres décrites précédemment, les courses au trot font l'objet de paris. Les autres grands types de courses hippiques sont les courses de plat, qui se courent au galop et montées, et les courses d'obstacles qui sont des courses de galop incluant le franchissement de haies. Le nombre d'épreuves organisées et le nombre d'engagements en 2013 sont indiqués dans le Tableau 2.1 et le Tableau 2.2 pour une partie des disciplines équestres (FFE 2015a) et pour les courses hippiques (IFCE-OESC 2014) respectivement.

Tableau 2.1 : Nombre d'épreuves et nombre d'engagements dans les trois disciplines olympiques et en courses d'endurance pour l'année 2013.

Discipline équestre	CSO	CCE	Dressage	Endurance
Nombre d'épreuves	75 000	5 000	1 500	2 500
Nombre d'engagements	1 300 000	62 000	82 000	22 000

Tableau 2.2 : Nombre de courses organisées et nombre de partants pour l'année 2013.

Discipline hippique	Course de plat	Course d'obstacles	Course au trot
Nombre de courses	4 900	2 200	11 000
Nombre de partants	55 000	23 000	149 000

La suite de cette partie présente plus précisément les disciplines étudiées au cours de ma thèse. L'indexation pour le CSO ayant été revue au cours de ma thèse, plus de détails seront donnés sur cette discipline.

2.2.1. Qu'est-ce que le CSO ?

Une épreuve de CSO consiste à réaliser un parcours de 10 à 12 obstacles dans un ordre déterminé et dans un temps imparti. Les obstacles sont mobiles (Figure 2.1), ce qui signifie que les barres qui les constituent peuvent tomber, et construits sur un terrain plat généralement rectangulaire (minimum 40m par 80m). Une épreuve de CSO est précédée d'une reconnaissance de quelques minutes pendant laquelle les cavaliers peuvent accéder à pied à la carrière dans laquelle a lieu l'épreuve afin de mémoriser le parcours et d'en appréhender les difficultés. Les chevaux sont classés d'abord suivant leur nombre de points de pénalité, et en fonction de leur temps de parcours en cas d'une égalité de pénalités. Les points de pénalité sanctionnent une chute de la barre la plus haute de l'obstacle, une erreur de parcours, une volte (le cavalier ajoute un cercle à son parcours), un refus de sauter (le cheval pile devant l'obstacle et recule d'au moins un pas) ou une dérobade (le cheval contourne l'obstacle). Des points de pénalité sont aussi donnés en cas de dépassement du temps imparti pour réaliser le parcours. Sont éliminatoires une chute du cavalier ou trois désobéissances. Ces règles de classement constituent le barème A. Dans le barème C les chevaux sont classés uniquement suivant le chronomètre, et les points de pénalité sont attribués sous la forme de secondes ajoutées à leur temps de parcours (FFE 2014a).

Il existe une grande variété d'épreuves en terme de niveaux, des épreuves destinées aux jeunes cavaliers à poney jusqu'au Jeux Olympiques ou aux Jeux Equestres Mondiaux. Des épreuves sont réservées aux cavaliers amateurs et d'autres aux cavaliers professionnels. Au sein de ces deux

catégories, la hauteur des obstacles augmente avec le niveau de l'épreuve. Pour les cavaliers amateurs, la hauteur varie entre 95cm en niveau 3 et 125cm en niveau élite. Pour les cavaliers professionnels les hauteurs vont de 120cm en niveau 3 à 150cm en niveau Elite. Ces épreuves incluent un obstacle double, c'est-à-dire une combinaison de deux obstacles placés sur une ligne et séparés d'une à trois foulées. Les épreuves de catégorie 1 et élite incluent en plus un obstacle triple, c'est-à-dire une combinaison de trois obstacles alignés et séparés d'une à trois foulées. La forme des épreuves peut varier : se jouer en deux manches, se jouer au barrage (les cavaliers sans-faute doivent réaliser un 2nd parcours plus court soit directement à la fin de leur passage, soit après le passage de tous les cavaliers), avoir un temps différé (le chronomètre se déclenche en cours de parcours), ou se jouer en manches successives avec augmentation progressive de la hauteur des obstacles pour les cavaliers sans fautes (puissance). Il existe aussi des épreuves réservées aux jeunes cavaliers à poney, dans ce cas la hauteur des obstacles tient compte de la catégorie de poney (A, B, C ou D suivant leur hauteur au garrot) pouvant participer aux épreuves. On verra plus tard que seules les épreuves dites « chevaux » sont traitées pour l'indexation des chevaux de sport.

Figure 2.1 : Franchissement d'un obstacle de CSO

Les épreuves de CSO peuvent être de type préparatoire (sans chronomètre, les ex aequo sont départagés par tirage au sort), vitesse (au barème C ou avec un temps différé), spéciale (une puissance par exemple) ou grand prix (en deux manches ou avec un barrage). Il existe également des épreuves dites « jeunes chevaux » qui se jouent sans chronomètre et sont ouvertes aux chevaux suivant leur âge (4 ans, 5 ans ou 6 ans). Ces épreuves ont pour but de former les jeunes chevaux aux compétitions de CSO dans un contexte où finir sans-faute prime sur la vitesse, car seuls les sans-fautes sont classés. Ces concours sont organisés à l'échelle des régions, et les jeunes chevaux ayant eu les meilleurs résultats peuvent s'affronter au cours d'une finale nationale. Dans ces concours les chevaux sont classés sur leurs résultats à l'obstacle mais aussi grâce à des notes de modèle et allures attribuées par des juges de la Société Hippique Française. Ces concours jeunes chevaux sont répartis en deux catégories : le cycle libre est plutôt destiné aux amateurs, et les chevaux peuvent participer à d'autres types d'épreuves. Le cycle classique est plutôt destiné aux professionnels : les obstacles sont un peu plus hauts, les chevaux inscrits dans ce cycle ne peuvent participer à des compétitions en dehors des épreuves d'élevage, et les chevaux de race Selle Français et Anglo-Arabe, qui seront présentés dans la partie 2.4, bénéficient d'une dotation spécifique.

Au niveau international, les couples chevaux-cavaliers s'affrontent dans des Concours de Saut Internationaux (CSI) dont le niveau de difficulté est indiqué par un nombre d'étoiles, de 1 à 5 étoiles. Les Concours de Saut Internationaux Officiels sont les compétitions de plus haut niveau, avec les Jeux Olympiques et les Jeux Equestres Mondiaux. Les CSIO sont eux aussi classés de 1 à 5 étoiles suivant la difficulté. Comparé à un CSI, un CSIO ayant le même nombre d'étoiles sera plus difficile. Un CSIO doit en plus inclure une épreuve de deux manches se jouant en équipe nationale. Les Jeux Olympiques et les Jeux Equestres Mondiaux ont lieu tous les 4 ans, et chaque pays ne peut organiser que 2 CSIO (un indoor et un outdoor) par an.

Dans cette discipline, les chevaux sont évalués sur leur franchise (pas de dérobade ni de refus), leur puissance (capacité à sauter haut et large), leur adresse (sur des courbes serrées, des enchainements nécessitant de varier l'amplitude des foulées), leur rapidité (chronomètre) et leur respect de l'obstacle. On verra plus tard que la grande diversité des épreuves au niveau national et international nécessite pour la sélection un critère d'évaluation qui soit représentatif des performances des chevaux et tienne compte de la difficulté des épreuves dans lesquelles ces performances sont réalisées.

2.2.2. Le CCE combine dressage, saut d'obstacles et cross

Le CCE se compose de trois épreuves réalisées par un même couple cheval-cavalier (FFE 2015b). L'épreuve de dressage a lieu en premier, ensuite l'ordre du saut d'obstacles et du cross peut varier. En général un CCE est organisé sur deux journées consécutives, mais pour les compétitions de plus haut niveau les épreuves sont organisées sur 3 jours.

L'épreuve de dressage consiste à présenter un programme de figures appelé reprise, en respectant l'ordre d'exécution imposé. L'épreuve se déroule sur un terrain rectangulaire (60m x 20m). Des lettres disposées à intervalles réguliers autour du terrain servent à marquer l'endroit où les figures ou changements d'allure devront être réalisés. Les reprises de dressage sont connues à l'avance et sont donc préparées par le couple cavalier-cheval. La qualité de l'exécution de chaque figure est notée sur 10 par 2 à 5 juges. Des notes sont en plus attribuées pour évaluer sur l'ensemble de la reprise la technique du cavalier, les allures du cheval, son impulsion (désir de se porter en avant) et sa soumission au cavalier (Figure 2.2). Les notes attribuées sont transformées en pénalités : les points obtenus par un couple cavalier-cheval sont soustraits à la note maximale qu'il est possible d'obtenir (10 à toutes les figures), ce qui donne une note négative à laquelle est appliquée un coefficient.

Figure 2.2 : Photo d'un couple cheval-cavalier en épreuve de dressage.

L'épreuve de cross est un parcours d'obstacles « naturels » fixes en terrain varié, qui doit être réalisé sans erreurs dans le parcours et en s'approchant d'un temps idéal. Le temps de parcours ne doit donc être ni trop rapide, ni trop lent, et les cavaliers sont généralement équipés de chronomètres sonnant les minutes afin de vérifier leur allure pendant le parcours. Les obstacles peuvent être des constructions comme des coffres ou des stères généralement en bois, mais aussi des dénivelés, des gués ou des trous (Figure 2.3). Contrairement au CSO où les obstacles s'enchainent sur une petite surface et le plus rapidement possible, dans une épreuve de cross les obstacles sont répartis sur un tracé en terrain dégagé ou en forêt. Le parcours comprend des tronçons sans obstacles permettant l'échauffement et la récupération des chevaux. Le dernier obstacle du parcours est volontairement imposant afin d'obliger le cavalier à ménager son cheval et de vérifier les capacités du cheval en fin d'épreuve. Comme en CSO, les cavaliers font une reconnaissance du parcours à pieds. Si en CSO les couples chevaux-cavaliers effectuent leurs parcours tour à tour, en cross la longueur du parcours et le nombre de participants nécessitent souvent que plusieurs cavaliers soient sur le parcours en même temps. Les cavaliers partent donc successivement. Des pénalités sont appliquées au couple cheval-cavalier en cas de refus : 20 points pour le 1er refus et 40 points pour le 2nd. Trois refus sur le même obstacle sont éliminatoires, tout comme 4 refus sur l'ensemble du parcours ou bien une chute du cavalier. De lourdes pénalités sont prévues pour les refus car le cross est une épreuve qui teste la franchise des chevaux face à des obstacles imposants. Des pénalités sont données pour chaque seconde de temps dépassé (faire le double du temps idéal est éliminatoire), et le couple est éliminé si le parcours a été réalisé trop rapidement, en dessous d'un temps minimum de parcours fixé. En cas d'égalité c'est le couple le plus proche du temps idéal qui est favorisé.

Figure 2.3 : Franchissement d'un gué et d'une haie sur un parcours de cross

L'épreuve de saut d'obstacles est similaire à une épreuve de CSO, mais avec des distances entre les obstacles un peu plus longues et un tracé comportant moins de difficultés techniques. Cette épreuve se déroule généralement après le cross, ce qui permet de vérifier l'état des chevaux. Placer le saut d'obstacles après le cross constitue une difficulté pour le couple cheval-cavalier, car pendant l'épreuve de cross le cheval peut toucher certains obstacles sans les faire tomber, alors qu'en saut d'obstacles cette erreur entraine des chutes de barres sanctionnées par des pénalités. De plus le profil et l'espacement des obstacles d'un parcours de saut nécessitent une vitesse et une posture du cheval différentes de celle du cross, qui se court plus rapidement et avec un cheval moins redressé à l'abord des obstacles. Les pénalités appliquées sont proches de celles du CSO. Un temps imparti est

défini au-delà duquel les secondes supplémentaires coûteront des pénalités, mais les cavaliers respectant le temps imparti ne seront pas départagés sur leur vitesse.

Les couples sont finalement classés en additionnant les pénalités reçues dans les 3 épreuves. Tout comme en CSO, une grande variété de niveaux de compétition existe. La SHF organise des CCE réservés aux jeunes chevaux de 4 à 6 ans : dans ces compétitions le temps minimum en saut d'obstacles est plus élevé. Les tracés en cross et en saut d'obstacle sont plus simples, et en début de saison les obstacles doivent être moins hauts que les cotes prévues pour un niveau d'épreuve donné afin de ne pas mettre les chevaux en difficulté. La FFE organise également des compétitions pour différentes catégories de niveaux (pro, amateur). Des compétitions internationales sont aussi organisées, et le CCE est présent aux Jeux Equestres Mondiaux.

2.2.3. L'endurance : des courses en pleine nature dans le respect de l'intégrité du cheval

Les courses d'endurance sont des courses de fond sans obstacles organisées en pleine nature (Figure 2.4). Le tracé est balisé. La longueur des courses d'endurance varie entre 10 et 160km (FFE 2014b). Sur les épreuves les plus importantes la distance peu atteindre 240km, mais dans ce cas la course a lieu sur 2 ou 3 jours consécutifs. Le but est de parcourir la distance le plus rapidement possible en respectant l'intégrité physique du cheval. Une course inclut plusieurs étapes au cours desquelles le cavalier et le cheval doivent s'arrêter afin de procéder à des contrôles vétérinaires (fréquence cardiaque, état des muqueuses, état de déshydratation, examen des allures) qui vérifient si le cheval est apte à continuer la course. La vitesse peut être imposée ou libre. Les courses à vitesse imposée se font sur de petites distances (60 km et moins), tandis que la vitesse est laissée libre pour les grandes distances (à partir de 90 km). Dans les épreuves à vitesse imposée la vitesse doit être comprise entre une vitesse minimale et une vitesse maximale (par exemple, entre 12 et 15km/h pour une course départementale en une étape de 30km). Dans ces courses le chronomètre est arrêté quand le cavalier et le cheval arrivent à l'étape, et le cheval a 30 minutes pour récupérer avant de passer les contrôles vétérinaires, qui l'autoriseront ou non à reprendre la course 30 minutes plus tard. Dans les courses à vitesse libre seule la vitesse minimale est imposée. Le chronomètre n'est arrêté que quand le cheval se trouve dans la zone de contrôle vétérinaire, et le délai pour se présenter au contrôle une fois arrivé à l'étape n'est que de 20 minutes (le délai effectif étant reporté sur la fiche de suivi des vétérinaires). Dans ce type d'épreuve d'autres critères sont observés, comme la fréquence respiratoire ou la récupération de la fréquence cardiaque. Dans tous les cas un dernier contrôle vétérinaire a lieu 30 minutes après la fin de la course. Les cavaliers sont autorisés à mettre pied à terre pendant la course, mais ils doivent franchir les lignes d'arrivée et de départ en selle. Des points d'assistance auxquels les cavaliers peuvent abreuver et rafraichir leurs montures sont répartis le long du parcours. Les modalités de classement dépendent du type d'épreuve : en vitesse imposée le classement est fait en comparant la vitesse réalisée à la vitesse imposée sur le parcours, et en tenant compte de la fréquence cardiaque du cheval à la fin de l'épreuve. Quand la vitesse est libre le classement est établi en fonction de l'ordre d'arrivée des cavaliers. La difficulté des courses réside dans la distance à parcourir : les petites distances sont celles des épreuves départementales ou régionales, tandis que les grandes distances (plus de 100km) sont rencontrées au niveau national et international. Des épreuves sont organisées par la SHF pour les jeunes chevaux. Cette discipline est également présente aux Jeux Equestres Mondiaux.

Figure 2.4 : Les épreuves d'endurance se courent en pleine nature

2.2.4. Les courses au trot

Les courses au trot se courent autour d'une piste. Les chevaux prennent le départ en même temps et doivent atteindre la ligne d'arrivée le plus vite possible, sans changer d'allure (galop, amble) sous peine d'être disqualifiés (SECF 2015). Avant d'être autorisés à courir, les chevaux doivent passer un test de qualification, qui consiste à courir une distance de 2 000m en dessous d'un temps imposé. Ce temps dépend de l'âge du cheval et est actualisé chaque année en tenant compte du progrès dans la population. Il s'agit d'un test réellement sélectif car 40% d'une génération le réussit. La totalité des chevaux nés une même année ne seront pas présentés aux mêmes âges aux courses de qualification : par exemple sur les 11 000 chevaux nés en 2010, un peu plus de la moitié ont été présentés à 2 ans, avec un taux de qualification de 37%. A 3 ans un peu plus de trotteurs ont été présentés (7 200), avec un taux de qualification plus faible (24%). Le nombre de chevaux présentés à 4 ans est faible (1 200 chevaux, dont 14% de qualifiés) (SECF 2013, 2014 et 2015). La majorité des courses sont attelées : le cheval tracte un sulky (voiture légère à 2 roues, Figure 2.5), le meneur qui le conduit est appelé driver. Les courses montées, où le cheval n'est pas attelé mais porte un jockey, sont une spécificité de la France. Les distances en courses attelées peuvent varier de 1 600 à 4 100m, mais en général elles sont comprises entre 2 100 et 2 800m. En courses montées la distance est comprise entre 1 800 et 3 000m. Chaque course a une allocation, c'est-à-dire une somme d'argent qui sera répartie entre les premiers arrivés. L'allocation moyenne en France est de 22 000 euros. Le 1er reçoit la moitié de l'allocation, le second reçoit la moitié restante, etc. Le gain est réparti de la façon suivante : 80% pour le propriétaire du cheval, 15% pour son entraineur et 5% pour le driver ou le jockey. Indépendamment de cette distribution des gains, l'éleveur du cheval touche également une prime dont la valeur est de 12.5% de l'allocation. Deux cent cinquante mille euros sont ainsi distribués chaque année. Il existe différents types de courses. Dans les courses à réclamer, des enchères sur les concurrents sont faites à bulletin secret avant la course. Les chevaux sont vendus à l'issu de la course aux personnes ayant fait les offres les plus élevées. Il s'agit du niveau de course le plus faible. Les courses sont ensuite réparties en catégories de niveau croissant de H à A. Les chevaux sont autorisés ou non à courir dans une course en fonction des gains cumulés au cours de leur carrière : le cheval ne doit pas dépasser un certain montant de gains pour être autorisé à courir. D'un niveau supérieur, les courses de groupes III, II ou I requièrent un cumul de gains minimum. Les courses de groupe III sont en général des Grands prix de province. Les courses de groupe II et de groupe I excluent les hongres car elles servent à sélectionner les reproducteurs. Les courses de groupe I sont des grandes épreuves internationales et intergénérationnelles. A la différence des courses nationales réservées au Trotteur Français, les courses internationales sont ouvertes à toutes les races de trotteurs. Le prix d'Amérique

est une course de groupe I dont le montant d'allocation est supérieur à 200 000 euros, et à laquelle seuls les chevaux ayant cumulé plus de 800 000 euros de gains peuvent participer. Le départ des courses peut se faire de 2 façons. En France le plus fréquent est le départ volté, qui est une entrée simultanée de tous les chevaux sur la piste, après leur alignement et leur coordination sur une aire de départ adjacente à la piste. Plus rare, le départ à l'autostart nécessite un véhicule doté d'une structure latérale rétractable derrière laquelle les chevaux s'alignent. Le véhicule contient les chevaux sur quelques centaines de mètres avant d'accélérer et de replier sa structure pour les laisser partir.

Une particularité des courses par rapport aux autres disciplines présentées est l'importance des paris. La prise des paris est organisée par le PMU, qui réalise 10 milliards d'euros de recette par an pour les courses au trot et au galop. 70% des sommes reviennent aux joueurs, 14% sont prélevées par l'Etat, et 16% sont utilisées pour les allocations et le fonctionnement des sociétés mères (SECF et France Galop). Pour garantir aux parieurs des résultats non-truqués, les contrôles anti-dopage de chevaux sont très fréquents : 18 000 prélèvements sont effectués par an, soit le double des prélèvements réalisés dans les principales disciplines sportives en France (cyclisme, natation, football…) (A. Duluard, communication personnelle). Dans une course, tous les chevaux gagnants ainsi qu'un cheval pris au hasard sont prélevés. Les chevaux à l'entrainement, au repos ou encore à l'élevage peuvent aussi être contrôlés. Les 25 meilleurs d'une année sont contrôlés également.

Figure 2.5 : Cheval en course au trot attelé

2.3. Carrières des chevaux athlètes

Les chevaux sont sevrés à 6 mois. Pour les chevaux destinés au CSO ou au CCE le débourrage a lieu à partir de 3 ans, et plutôt vers 4 ans pour les chevaux d'endurance. Avant 6 mois, les poulains peuvent participer avec leur mère à des concours de modèle et allures réservés aux poulinières suitées. De 6 mois à 3 ans, ils peuvent participer à des concours de modèle et allures seuls. A partir de 4 ans, les compétitions jeunes chevaux deviennent accessibles. A sept ans et plus, les autres compétitions leurs sont ouvertes, sans limite d'âge (Figure 2.6). Il est cependant rare qu'un cheval continue la compétition passé 20 ans.

Les chevaux élevés pour les courses au trot ont une carrière différente. Les poulains sont sevrés à 6 mois, mais ils sont débourrés et mis à l'entrainement à 18 mois. Sous réserve de se qualifier, ils peuvent courir dès leurs 2 ans, mais une très faible part d'une génération court effectivement à cet âge-là (Figure 2.7). La plus importante part de leur carrière a lieu à 3 et 4 ans, et seuls les très bons chevaux continueront à courir à 5 ans ou plus (un gain cumulé sur la carrière minimum est requis,

plus élevé pour chaque année de course supplémentaire). Leur carrière est donc plus précoce et souvent plus courte que celle des chevaux de sport.

Figure 2.6 : Carrière d'un cheval de sport

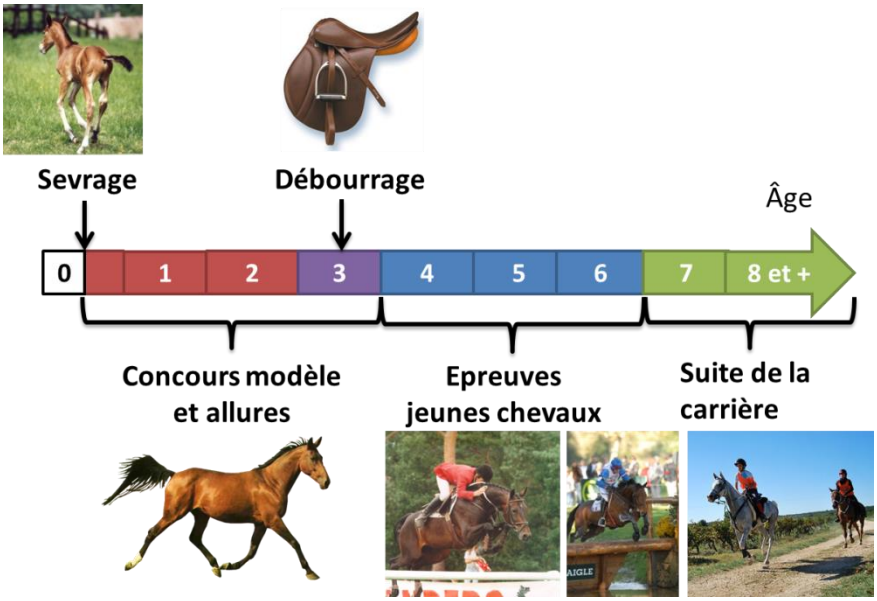
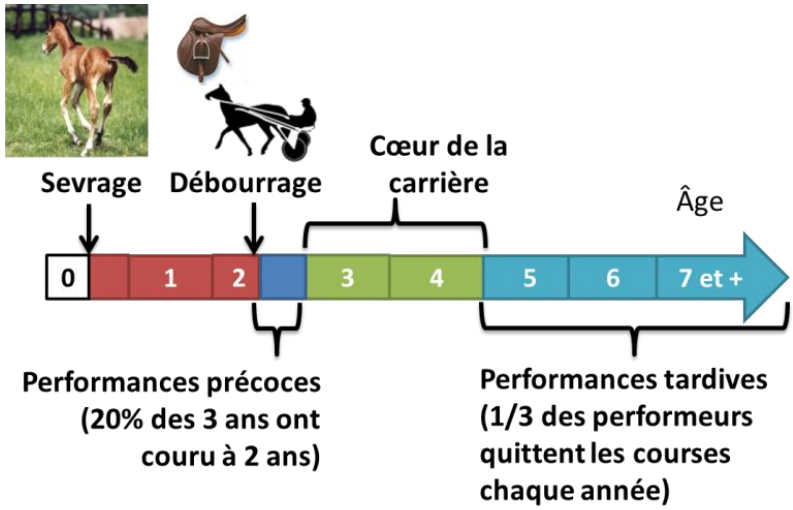


Figure 2.7 : Carrière d'un trotteur



2.4. Races françaises sélectionnées pour le sport ou la course

En France, deux races de chevaux dominent dans l'élevage de chevaux de sport : le Selle Français et l'Anglo-Arabe. L'utilisation du Selle Français est plus tournée vers le CSO, tandis que les Anglo-Arabes sont plutôt élevés pour le CCE (mais des Anglo-Arabes peuvent participer à des CSO et des Selles Français à des CCE). Des chevaux de sport étrangers participent également à ces compétitions, ils sont soit issus de l'importation, soit nés sur le territoire avec des effectif de naissances totaux proches de celui des Anglo-Arabes. Pour les courses d'endurance, des chevaux très robustes à l'effort sont nécessaires, particulièrement pour les courses de haut niveau. Les chevaux élevés à cette fin sont principalement des Pur-Sang Arabes et des croisés Arabes. Enfin la France élève sa propre race de chevaux pour les courses au trot : le Trotteur Français.

2.4.1. Le Selle Français

Le stud-book du Selle Français a été fondé en 1958 (Stud-book Selle Français 2012). Il a à l'époque regroupé des chevaux demi-sang produits dans trois berceaux distincts : le demi-sang normand, le demi-sang vendéen et le demi-sang charolais. Ces trois rameaux avaient eux-mêmes été obtenus par le croisement de Pur-Sang Anglais avec la jumenterie autochtone. Ce rassemblement de chevaux pour la fondation d'un stud-book a conduit à une grande diversité dans la race, qui n'a pas de standard bien défini. Cependant, du fait de leur usage exclusivement sportif, les chevaux Selle Français sont de grande taille : entre 1,65m et 1,70m au garrot.

Pour être inscrit au stud-book Selle Français sur ascendance, le cheval doit être né de 2 reproducteurs Selle Français, ou bien d'un reproducteur Selle Français et d'un facteur de Selle Français, ou encore d'un étalon approuvé Selle Français et d'une jument labellisée Selle Français (Stud-book Selle Français, 2015). Les juments facteurs de Selle Français peuvent être de race Pur-Sang, Autre Que Pur-Sang, Anglo-Arabe, Demi-Sang Anglo-Arabe, Trotteur (Français ou étranger). Des juments de races étrangères peuvent être utilisées à condition d'être reconnues dans l'Union Européenne ou par le WBFSH (World Breeding Federation for Sport Horses) et d'avoir un numéro SIRE (Système d'Information Relatif aux Equidés). Les juments d'autres races qui ne remplissent pas ces conditions peuvent être reconnues facteurs de Selle Français si leur indice individuel en compétition est supérieur ou égal à 110 (cet indice sera présenté dans la partie 2.5). La labellisation des juments se fait à la demande des propriétaires par une commission qui évalue la jument sur son modèle, ses allures, ses performances et sa généalogie. L'approbation des mâles se fait à la demande des propriétaires et sera décrite dans le paragraphe consacré au schéma de sélection. Du fait de l'autorisation pour la reproduction en Selle Français de chevaux d'autres races françaises ou étrangères labellisés ou reconnus comme facteur de Selle Français, une grande variété de croisements de races peut donner naissance à un cheval inscriptible au stud-book Selle Français. Ainsi, pour l'année 2012, les Selle Français ayant un indice en CSO étaient issus de 1 094 croisements différents (Anne Ricard, communication personnelle). Les croisements les plus représentés sont ceux réalisés avec des chevaux de sport inscrits dans d'autres stud-books européens : Holsteiner, KWPN (Koninklijke Vereniging Warmbloed Paardenstamboek Nederland), Belgian Warmblood, Hannoveraner, Oldenburger, Cheval de sport Belge, Zangersheide et Rheinisches Warmblut. On verra par la suite que l'existence de croisements aussi divers rendra délicate l'utilisation d'un effet race dans l'estimation des valeurs génétiques. Par ailleurs le règlement du stud-book a évolué au cours du temps, et la population actuelle Selle Français est le résultat de plusieurs règlements successifs.

L'insémination artificielle est autorisée par le stud-book : le Selle Français peut donc aisément être produit partout dans le monde. Dans ce cas un contrôle de filiation doit être réalisé avant l'inscription au stud-book. Le transfert d'embryon peut également être utilisé.

Pour l'année 2014, l'ANSF a répertorié un peu plus de 5 700 éleveurs de Selle Français. Environ 4 400 détiennent une poulinière (140 élevages comptent plus de 5 poulinières). Neuf mille juments Selle Français ont été saillies en 2014. En 2013, le nombre de naissances de Selles Français était de quasiment 6 500. Il y a actuellement 700 étalons approuvés pour la reproduction en Selle Français, dont 500 de race Selle Français (IFCE 2015).

2.4.2. L'Anglo-Arabe

Le stud-book de la race Anglo-Arabe (ANAA, Association Nationale Anglo-Arabe) a été fondé en 1833 (ANAA 2014a). Le berceau de la race se situe dans le Sud-Ouest où ont eu lieu les croisements de

chevaux Pur-Sang Anglais et Pur-Sang Arabe, avec un apport de sang local amené par la jumenterie autochtone. Les qualités recherchées sont d'une part celles du cheval Pur-Sang Arabe, comme l'endurance et élégance des allures, et d'autre part celles du Pur-Sang Anglais: taille et vitesse. Comme pour le Selle Français, le croisement de différentes races a conduit à un standard peu défini. Ces chevaux sont un peu moins grand que les Selle Français : entre 1,58m et 1,65m au garrot.

Les chevaux sont inscriptibles au stud-book Anglo-Arabe sur leur ascendance. Il faut que les ascendants soient tous Pur-Sang Anglais ou Pur-Sang Arabe et soient chacun inscrits dans leurs stud-books respectifs (ANAA 2014b). Si le pédigrée comporte un ascendant n'étant pas Pur-Sang Anglais ou Pur-Sang Arabe, il faut qu'à la 4ème génération 15 des 16 ancêtres du cheval à inscrire soient Pur-Sang Anglais ou Pur-Sang Arabe. Les chevaux ne répondant pas à ces critères peuvent être inscrit à condition d'avoir un ascendant Pur-Sang issu d'un croisement Anglo-Arabe, Pur-Sang Anglais ou Pur-Sang Arabe croisé avec demi-sang Arabe ou Anglo-Arabe ou avec un stud-book reconnu par la WBFSH et s'ils ont au moins 25% de sang Pur-Sang Arabe. Les chevaux ne peuvent pas être inscrits si à la 4ème génération ils ont un ou des ascendants poneys, traits, cobs ou d'origine non constatée.

Les Anglo-Arabes sont produits principalement dans le Sud-Ouest, mais aussi dans le reste de la France. Là aussi l'insémination artificielle est autorisée, ainsi que le transfert d'embryons.

En 2014, un peu plus de 1 100 juments Anglo-Arabes ont été saillies. Il y avait une centaine étalons Anglo-Arabes en activité. En 2013, un peu plus de 550 poulains Anglo-Arabes sont nés (IFCE 2015).

2.4.3. Le Pur-Sang Arabe

Le Pur-Sang Arabe a une origine ancienne. Il est élevé en France en race pure depuis le règne de Napoléon, et son introduction sur le territoire remonterait aux premières croisades. C'est un cheval plus petit que le Selle Français et l'Anglo-arabe : il mesure entre 1,48m et 1,56m au garrot (ACA 2014a). Il est principalement élevé en France pour les courses d'endurance. Des chevaux de cette race sont aussi élevés pour les shows (des concours de modèle et allures où les chevaux sont jugés sur leur esthétique). Cette race est beaucoup utilisée en croisement, notamment pour la production d'Anglo-Arabes. Les Pur-Sang Arabes peuvent aussi être croisés avec des races de loisir, de poney ou encore de trait quand l'objectif est d'affiner le modèle de la race ou d'améliorer son endurance.

Un cheval est inscriptible au stud-book du cheval Arabe sur ascendance s'il est issu d'une jument inscrite à ce stud-book (et âgée d'au moins 2 ans l'année de la saillie) et d'un père approuvé pour la production de cheval Arabe (ACA 2014b). Le stud-book comporte une annexe pour les chevaux dits demi-sang Arabe. Le cheval doit avoir au moins 50% de sang Arabe. Il doit avoir pour parent un cheval inscrit au stud-book du cheval Arabe ou du demi-sang Arabe, et un parent inscrit à un stud-book de chevaux de sang, de trait ou de poney. Le second parent peut aussi être inscrit à un registre d'origine constatée ou non. L'insémination artificielle et le transfert d'embryons sont autorisés, mais pas le clonage.

En 2014 le nombre d'étalons en activité était de près de 700, et environ 2 300 juments Arabes ont été saillies. En 2013 le nombre de naissance dans cette race était d'un peu plus de 1 400.

2.4.4. Le Trotteur Français

Le Trotteur Français est élevé depuis le début des courses au trot en France. Il est le fruit d'un croisement entre des juments normandes et des étalons Pur-Sang Anglais ou avec des trotteurs de Grande-Bretagne (plus rarement, avec des trotteurs américains). Ces chevaux mesurent en général

entre 1,60m et 1,70m au garrot. Ils sont élevés dans le Nord-Ouest de la France, en majorité en Basse-Normandie.

Un cheval peut être inscrit au stud-book Trotteur Français sur ascendance à condition que le père soit approuvé pour la reproduction en Trotteur Français et que la mère soit inscrite au stud-book et admise à la reproduction. Le contrôle de filiation est obligatoire. L'insémination artificielle en sperme frais est autorisée, en revanche les produits issus du clonage ne peuvent être inscrits au stud-book. L'utilisation de le technique de transfert d'embryon n'est autorisée que dans de rares cas : jument âgée, excellente performeuse ou mère d'excellents performeurs (3 victoires dans des courses de groupe I), ou bien jument ayant été saillie sans succès 2 années consécutives (SECF 2011).

En 2014 il y avait 501 étalons Trotteur Français en activité, et le nombre de juments Trotteur Français saillies était d'environ 15 700. En 2013, un peu plus de 11 300 poulains ont été inscrits au stud-book Trotteur Français.

2.5. Quels critères pour évaluer et comparer les performances des chevaux ?

2.5.1. En CSO et CCE les index reposent sur deux critères

Les gains et le classement pour mesurer la performance

Jusqu'en 2008, les épreuves de CSO, de CCE et de dressage étaient dotées en fonction de leur prestige et de leur niveau de difficulté. Les cavaliers les mieux classés étaient récompensés par un gain financier. La dotation de l'épreuve était distribuée au 8 premiers cavaliers classés d'une épreuve, et à tout le 1er ¼ des cavaliers quand l'épreuve compte plus de 32 partants. L'attribution des gains était telle que le 1er gagnait 25% de la dotation, puis le 2ème gagnait 25% de la somme gagnée par le 1er, etc. La distribution obtenue était une exponentielle décroissante, et les écarts de gains en fonction du classement assuraient que les cavaliers voudraient être aussi hauts dans le classement que possible. Les gains totalisés par un cheval sur une année de compétition étaient alors un bon critère pour évaluer sa capacité à se classer dans le 1er ¼.

Comme il n'existe pas de mesure simple de la difficulté technique d'une épreuve, les gains ont longtemps été exploités sous la forme d'un gain annuel pour l'indexation. Pour chaque cheval, les gains étaient sommés sur l'année de compétition. Il était ainsi possible que des chevaux jamais classés dans le 1er quart n'aient pas de gain. Depuis 2009, les gains ont été remplacés par des points pour le calcul des indices. Cette décision a été prise pour palier à la libéralisation des dotations des compétitions. Une grille indicative des prix est fournie aux organisateurs de concours, mais son application n'est plus obligatoire. Des épreuves pouvant être sur-dotées ou bien ne plus être dotées du tout, les gains ne sont plus représentatifs de la difficulté de l'épreuve et sont devenus un critère obsolète. Ils ont été remplacés par des points. En CSO les points qui remplacent la dotation des épreuves sont définis en fonction de la difficulté physique de l'épreuve : volume des obstacles, longueur du parcours, vitesse imposée. En CCE, les classes d'épreuves existantes sont suffisantes pour définir la difficulté technique des épreuves (Ricard *et al*. 2010). Les points ont l'avantage par rapport aux gains réellement perçus de pouvoir être distribués à l'ensemble des partants, qu'ils se soient classés dans le 1er ¼ ou non. Ils conservent une information détaillée sur le classement, contrairement aux gains avec lesquels les chevaux hors du 1er ¼ étaient considérés comme dernier ex aequo. Une distribution exponentielle décroissante a été conservée pour la distribution de ces

points, elle dépend du nombre de partants dans l'épreuve. Dans la suite nous continuerons de parler de gain annuel, il faut cependant noter que c'est le logarithme du gain annuel qui est utilisé, afin d'avoir une distribution normale des performances.

Les classements des chevaux d'une épreuve apportent de l'information sur plus d'animaux, dans la mesure où le classement de chaque cheval est enregistré, qu'il se soit classé dans le 1er ¼ ou non. Cependant le classement en lui-même donne moins de poids à la capacité du cheval à être le meilleur parmi les 1ers. Les classements des chevaux dans chaque épreuve d'une année de compétition sont utilisés pour obtenir un classement général des chevaux. Contrairement aux gains ou aux points, les classements sont utilisés en tant que tels sans ajout d'information sur la difficulté de l'épreuve. Les chevaux concourant en France ne se rencontrent pas tous sur une année de compétition, mais la connaissance du classement des chevaux dans une épreuve donnée sachant leurs classements dans d'autres épreuves où ils ont rencontrés d'autres concurrents permet d'obtenir un classement général, qui repose sur une variable sous-jacente maximisant la probabilité d'observer les classements réalisés pendant l'année de compétition. Le niveau de l'épreuve est ainsi estimé par le niveau du plateau des chevaux présents plutôt que par une mesure subjective de la difficulté technique.

Ces deux types d'informations sont utilisés pour calculer un indice de performances et un indice génétique.

Calcul d'un indice de performance annuel

L'indice de performance permet d'évaluer un cheval par rapport à ses contemporains. La performance est corrigée pour les effets d'environnement, mais la valeur génétique de l'animal n'est pas estimée.

Les indices de performance sont l'ISO (Indice Saut d'Obstacles) et l'ICC (Indice Concours Complet). L'ISO et l'ICC sont calculés chaque année pour chaque cheval sur les performances qu'il a réalisées pendant l'année précédente (Ricard 2008). La période de calcul, du premier week-end d'octobre de l'année n-1 au dernier week-end de septembre de l'année n, a été choisie afin que la publication des indices puisse avoir lieu en décembre pour le choix des reproducteurs. Les épreuves donnant lieu à un double classement ou incluant des notes de modèle et allures ne sont pas utilisées. La formule d'un indice de performance s'écrit :

Indice de performance = performance mesurée – effets environnementaux.

Un indice de performance est calculé pour les deux critères gain annuel et classement dans chaque épreuve. Les effets pris en comptes sont l'année de compétition, l'âge et le sexe, ils sont estimés par une analyse de variance. L'effet cavalier n'est pas pris en compte car un même cavalier monte peu de chevaux et un cheval est monté par trop peu de cavaliers différents au cours d'une année de compétition pour que l'effet puisse être correctement estimé. L'ISO et l'ICC sont ensuite obtenus en sommant les 2 indices, avec une pondération ajustée suivant la différence entre les variances des deux critères. Ces indices sont publiés depuis 1972. Depuis 1998, ils sont publiés avec des coefficients de précision (CP) qui indiquent la fiabilité des indices. Le CP augmente quand le nombre de sorties en compétition du cheval augmente, et également quand il a participé à des épreuves comptant beaucoup de partants par rapport auxquels il peut être comparé.

Pour faciliter leur usage, l'ISO et l'ICC sont standardisés. Les chevaux de la population de référence sont choisis en fonction du CP qu'ils ont obtenu pour l'année considérée : leur CP doit être d'au moins 0.60 pour l'ISO et de 0.40 pour l'ICC. Les indices sont ensuite ajustés de façon à ce que dans la population 3% des chevaux aient un indice de performance supérieur à 140, et 40% des chevaux aient un indice de performance d'au moins 110.

Les indices génétiques : BSO et BCC

Contrairement à l'ISO et à l'ICC qui sont réservés aux performeurs, tous les chevaux ayant un apparenté performeur en CSO ou CCE ont un indice génétique appelé BSO (BLUP Saut d'Obstacle) ou BCC (BLUP Concours Complet) respectivement, même s'ils n'ont pas de performances propres dans l'une ou l'autre de ces disciplines. Les indices génétiques permettent d'estimer la valeur génétique d'un cheval, c'est-à-dire sa capacité à transmettre ses qualités à ses produits. Ils permettent d'aider à sélectionner des reproducteurs en utilisant toutes les informations disponibles sur les performances de leurs apparentés en plus de leurs performances propres. Les gains sont remontés jusqu'en 1974, et les classements jusqu'en 1985 (Ricard 2008). Les relations de parenté sont prises en compte en incluant tous les chevaux nés après 1945. Les performances des apparentés utilisées pour le calcul des indices d'un cheval sont pondérées en fonction de l'apparentement avec le cheval et de l'héritabilité du caractère. En plus des corrections pour le sexe, l'âge et l'année, une correction pour le harem rencontré par l'étalon est prise en compte afin de ne pas surestimer l'indice d'un étalon qui n'aurait rencontré que de bonnes juments, et inversement. Les effets sont estimés simultanément avec un modèle animal en utilisant la méthode du BLUP (Best Linear Unbiased Predictor). Le BSO et le BCC sont publiés depuis 1986 et 1997 respectivement. Les paramètres génétiques pour le gain et le classement en CSO et en CCE sont présentés dans le Tableau 2.3.

Tableau 2.3 : Paramètres génétiques des indices pour le CSO et le CCE, d'après Ricard (2008)

Discipline	**CSO**		**CCE**	
Critère	Gain	Classement	Gain	Classement
Héritabilité	0.27	0.16	0.14	0.07
Répétabilité	0.47	0.29	0.45	0.33
Composante maternelle	0.05	0.03	0.03	0.03

Le BSO et le BCC sont obtenus par la somme pondérée de l'indice génétique gain annuel et de l'indice génétique classement correspondant avec des poids respectifs de 0.75 et 0.25. Il s'agit d'un modèle animal bi-varié, la corrélation entre les critères gain et classement étant de 0.90. Les indices sont standardisés par rapport à une population de référence. Au sein de cette population la moyenne des indices génétiques est mise à 0 et sert ainsi de base mobile. Cette standardisation a l'avantage de conserver les indices des chevaux actuels dans une même échelle. On peut noter qu'au cours de la vie d'un cheval, l'information utilisée pour l'évaluer va évoluer : à sa naissance, seules les performances réalisées par ses apparentés seront disponibles. Au cours de sa carrière, ses performances propres vont progressivement prendre plus de poids dans le calcul de son indice, et en fin de carrières seront complétées par les performances de ses descendants si le cheval a été mis à la reproduction. Des outils sont donc disponibles pour évaluer les chevaux, que ce soit par rapport à ses contemporains via l'ISO et l'ICC, ou sur ses qualités génétiques via le BSO et le BCC: l'utilisation (ou non) de ces outils sera abordée dans la partie 2.6.

2.5.2. Trois critères mesurent les performances en courses d'endurance

Les indices de performance et les indices génétiques sont respectivement publiés depuis 2006 et 2012.

Trois informations sont utilisées pour évaluer les performances.

Les épreuves prises en compte sont les épreuves à vitesse libre (≥90 km). La vitesse (en km/h) est disponible pour tous les chevaux ayant fini la course. Elle n'est pas utilisée en tant que telle mais standardisée par rapport à l'épreuve. Le niveau de la concurrence rencontrée dans la course est pris en compte en introduisant dans le modèle l'effet de la course. La vitesse est manquante pour tous les chevaux qui n'ont pas fini la course.

Le classement est mesuré pour tous les partants. Trois classes sont considérées : finissant, abandon, éliminé. Là aussi, l'effet de la course est introduit dans le modèle et permet de corriger à la fois pour les effets environnementaux induisant un taux de réussite plus ou moins important mais aussi le niveau de compétitivité par la qualité des chevaux rencontrés.

Le dernier critère est la distance de la course. Le cheval reçoit comme performance la distance réelle de la course en km (Ricard 2008).

Calcul de l'indice de performance

Contrairement à l'ISO et à l'ICC, l'IRE (Indice en Raid d'Endurance) est calculé en utilisant les performances réalisées sur toute la carrière du cheval. Les critères vitesse, classement et distance constituent en eux-mêmes des indices accompagnés d'un Coefficient de Précision (CP), qui sont utilisés dans un modèle multi-caractères pour obtenir l'indice de performance global. Une pondération donne un léger avantage aux critères vitesse et distance (35%) par rapport au classement (30%), et le CP dépend du nombre de courses courues. Les corrélations génétiques entre les caractères sont aussi prises en compte depuis 2012. Pour l'instant, le seul effet d'environnement pris en compte est l'âge, ainsi qu'un effet course pour les caractères vitesse et classement en plus d'un effet d'environnement permanent (qui inclut la valeur génétique).

L'IRE est standardisé de façon à ce que 50% des chevaux ayant couru aient un indice supérieur 100, 17% aient un indice supérieur à 120, et seulement 2.9% un indice supérieur ou égal à 140 (Ricard 2008).

L'indice génétique : BRE

L'indice génétique BRE (BLUP Raid d'Endurance) est calculé avec un modèle multi-caractères à partir des 3 critères décrits. Les corrélations génétiques entre ces 3 caractères sont de 0.50. Le calcul est fait chaque année, et les valeurs obtenues sont centrées sur 0. Un cheval avec un BRE positif aura donc un indice génétique supérieur à la moyenne dans la population. Comme le BSO et le BCC, le BRE est publié avec un CD. Les paramètres génétiques pour la vitesse, la distance et le classement évalués en uni-caractère et en multi-caractère sont dans le Tableau 2.4.

Tableau 2.4 : Paramètres génétiques des critères vitesse, distance et classement.

Critère	héritabilité	Répétabilité
Vitesse	0.20	0.42
Distance	0.10	0.19
Classement	0.10	0.25

2.5.3. Un critère unique pour l'évaluation des trotteurs

Les performances des trotteurs sont mesurées avec le logarithme du gain annuel divisé par le nombre de départs.

L'indice de performance pour le trot est l'ITR (Indice Trot). L'indice d'un cheval est calculé chaque année, en utilisant toutes les courses courues entre mi-octobre de l'année de calcul et mi-octobre de l'année précédente. Les courses sont prises en compte quel que soit leur type (attelées ou montées) ou les conditions pour y participer. La performance est corrigée pour l'âge (une classe par âge jusqu'à 5 ans, puis une seule classe pour tous les chevaux de 6 ans ou plus) et pour le sexe. Avec ce mode de calcul, seuls les chevaux ayant eu un gain sont indicés. Les indices sont standardisés de façon à avoir une moyenne de 100 et un écart-type de 20.

Pour l'indice génétique, toutes les performances réalisées depuis 1966 sont utilisées. L'évaluation est réalisée avec un modèle animal similaire aux modèles déjà décrits. L'indice est le BTR (BLUP Trot), et contrairement aux indices génétiques déjà décrits il n'y a pas de standardisation. L'héritabilité du BTR est de 0.26, sa répétabilité de 0.36, et la composante maternelle est de 0.04 (Ricard 2008).

2.6. Sélection du cheval athlète en France

2.6.1. Les acteurs

La sélection est assurée par des Organismes de Sélection (OS), qui peuvent sous-traiter des tâches comme l'identification des chevaux ou les évaluations génétiques à d'autres organismes. Si aucun organisme de sélection n'est agréé pour une race, c'est l'IFCE (Institut Français du Cheval et de l'Equitation) qui assure les missions d'un organisme de sélection.

La définition des objectifs

Les objectifs de sélection sont définis par les OS comme le stud-book Selle Français, l'Association Nationale de l'Anglo-Arabe (ANAA), l'Association nationale française du Cheval Arabe pur-sang et demi-sang (ACA) ou bien la Société d'Encouragement à l'Elevage du Cheval Français (SECF). Il s'agit d'une tâche particulière dans la mesure où, contrairement à la plupart des autres productions animales, beaucoup d'éleveurs sont des particuliers amateurs. Les élevages sont de petite taille, et dans la plupart des cas aucune rentabilité économique n'est attendue : c'est plutôt l'équilibre budgétaire qui est visé. Il y a donc peu ou pas de référence technico-économiques sur lesquelles s'appuyer pour fixer les objectifs de sélection. De plus, les résultats économiques sont peu représentatifs du sérieux de la conduite d'un élevage, car le prix d'un cheval varie de façon plus exponentielle que linéaire en fonction de ses qualités. Seuls de rares chevaux d'élite feront faire de réels bénéfices à leurs éleveurs.

L'identification

Le SIRE (Système d'Information Relatif aux Equidés) assure depuis 1975 l'identification des chevaux et l'enregistrement des généalogies. Chaque cheval est identifié par son numéro SIRE composé de 8 chiffres suivis d'une lettre clé. Plusieurs pays européens sont impliqués dans le projet du Universal Equine Life Number. Ce système d'identification partagé par plusieurs pays prévoit un code pour le pays de naissance, un code pour l'organisme ayant enregistré le cheval, suivi de l'identifiant national du cheval. Cette nomenclature uniformisée permettrait une mise en commun des bases de données et un suivi des chevaux à l'étranger plus facile qu'à l'heure actuelle.

L'enregistrement des performances

Les performances de CSO, CCE et d'endurance sont enregistrées par la Fédération Française d'Equitation (FFE). Certains stud-books faisant appel aux données issues de ces performances, la gestion des épreuves et l'enregistrement des résultats tiennent compte des impératifs liés au calcul des indices. Pour les courses au trot les performances sont enregistrées par la SECF, qui a la particularité d'avoir la charge du stud-book du Trotteur Français et de l'organisation des courses au trot.

2.6.2. Les schémas de sélection

Le Selle Français : un schéma basé sur les performances propres proposé par le stud-book mais peu suivi

Le schéma de sélection du Selle Français repose sur le cycle de vie du cheval de sport en France présenté dans la partie 2.3 (Figure 2.6).

Le stud-book du Selle Français prévoit une sélection qui doit être principalement réalisée sur les mâles à 3 niveaux : les jeunes mâles de 2 et 3 ans, les jeunes performeurs de 4 à 7 ans, et la sélection et approbation confirmée sur performances internationales ou sur descendance (Stud-book Selle Français, 2015).

A 2 ans, les jeunes mâles peuvent être présentés à des tests de modèle, de locomotion et d'aptitude à l'obstacle en liberté dans des concours ayant lieu partout sur le territoire. Sur les 300 jeunes présentés chaque année, 80 sont sélectionnés pour participer à la finale nationale, et 10 à 20% sont approuvés pour la reproduction dès 2 ans. A 3 ans, 500 jeunes sont présentés à ces concours, qui incluent en plus un test à l'obstacle monté. 100 chevaux sont sélectionnés, et 25 à 40% sont approuvés. Pour conserver leur agrément, les jeunes mâles approuvés dès 2 ans doivent être confirmés à 3 ans.

De 4 ans à 6 ans, les chevaux peuvent participer aux épreuves jeunes chevaux. L'approbation a lieu lors des finales des championnats de la race. Le modèle et la locomotion sont toujours évalués, mais l'aptitude sportive prend plus de poids dans l'évaluation. Un jeune cheval peut aussi être approuvé à partir d'un indice sur performances propres minimum. Ces étalons sont approuvés pour 7 ans (sauf ceux approuvés à 2 ans qui doivent confirmer leurs résultats à 3 ans), et sont confirmés définitivement suivant leurs performances et celles de leurs descendants quand les données deviennent disponibles.

A partir de 7 ans, l'approbation nécessite un indice de performance minimum pour les étalons concourant en France. Les étalons qui sont à l'étranger doivent avoir au moins 3 classements parmi les 8 premiers d'un Grand Prix CSI 3*** ou plus. Un étalon peut aussi être approuvé sur descendance s'il a au moins 2 produits classés parmi les 200 premiers du classement mondial.

Il n'y a pas à proprement parler de sélection sur la voie femelle. Cependant, pour les approbations à 2-3 ans, des points bonus sont attribués suivant la qualité de la lignée maternelle remontée sur 5 générations. Des concours sont organisés pour caractériser les poulinières et leurs foals, et des labels ont été mis en place pour valoriser différentes qualités des juments : Sport, Reproductrice, Meilleure lignée maternelle, Modèle et allures. Il existe aussi une Prime d'Aptitude à la Compétition Equestre (PACE) dont l'attribution dépend des performances de la jument, de celles de ses descendants et de

ses apparentés. Ces mesures ont pour but d'inciter les éleveurs à mettre à la reproduction leurs meilleures jeunes juments afin de réduire l'intervalle de génération.

Si le schéma de sélection pose clairement des étapes de sélection, dans la pratique son application est peu observée. L'intensité de sélection est très faible, puisque comme indiqué dans la présentation du Selle Français on compte en 2013 près de 6 500 naissances en Selle Français pour 700 étalons en activité. L'insémination artificielle étant autorisée, retenir 30 à 40 étalons devrait suffire pour assurer la production actuelle de Selles Français. De plus leur sélection se fait sur performances propres, sans chercher à dissocier les effets génétiques des effets d'environnement, alors que l'indice génétique est disponible. Cette sélection sur performances propres a lieu dans l'idéal à 5 ans, ce qui peut sembler optimal en l'absence d'utilisation de sélection génomique, mais en réalité les jeunes étalons produisent peu de poulains au profit des vieux étalons : l'intervalle de génération est donc proche de 10 ans. Quand le BSO était utilisé pour sélectionner, le progrès génétique était important : environ 9% d'écart-type génétique par an (Dubois et Ricard 2007).

Les stud-books des chevaux de sport élevés en Europe utilisent eux aussi les performances propres des individus pour l'approbation des étalons. Par ailleurs, certains stud-books Hollandais (KWPN), Danois (Danois Sang-Chaud), et Allemands (Holsteiner, Westphalien, Oldenbourg, Hanovrien) prévoient également une phase de test des étalons en stations. Une première sélection est réalisée avant l'entrée des chevaux en station. Les tests durent en général 70 jours de façon à uniformiser les effets d'environnement et minimiser leur importance pour la comparaison des chevaux (Koenen et Aldridge, 2002).

L'Anglo-Arabe et le Pur-sang Arabe: une approbation des reproducteurs basée sur leur race

Les stud-books de l'Anglo-Arabe et du Pur-Sang Arabe ne proposent pas un schéma basé sur les performances comme le stud-book du Selle Français. L'approbation des chevaux repose sur leur race, et les éleveurs sont ensuite libres dans leurs choix de reproducteurs.

Tous les chevaux entiers de race Anglo-Arabe, Pur-Sang Anglais ou Pur-Sang Arabe peuvent être approuvés automatiquement pour produire dans le stud-book Anglo-Arabe (ANAA 2014b) à partir de 2 ans. Les entiers d'autres races doivent être approuvés par une commission. Pour cela, il faut qu'ils soient autorisés à reproduire dans leur stud-book d'origine (le stud-book étant reconnu par la WBSH). Leur approbation dépendra ensuite de leurs pédigrées remontés sur 4 générations, de leurs modèle et allures, et de performances sportives (les leurs ou celles de leurs descendants). Comme en Selle Français, il n'y a pas de sélection sur la voie femelle. Les juments peuvent être saillies à partir de 2 ans. Les éleveurs peuvent inscrire leurs chevaux au programme d'élevage de la race Anglo-Arabe, ce qui leur donne le droit de faire participer leurs chevaux aux concours d'élevage. L'inscription à ce programme est obligatoire pour recevoir des primes d'élevage distribuées par l'ANAA ou les primes PACE destinées aux meilleures poulinières. Le montant de ces primes dépend de l'ICC de la jument et de ses descendants.

En Pur-Sang Arabe, un étalon est approuvé pour la reproduction dès lors qu'il est inscrit au stud-book (ACA 2014). Etalons et juments peuvent être mis à la reproduction à partir de 2 ans. L'ACA verse des primes « qualité » aux naisseurs des meilleurs chevaux de l'année. Il existe un programme d'élevage spécifique à l'endurance, qui donne droit à la participation aux concours d'élevage et à l'obtention de primes PACE pour les juments pour cette discipline. Pour obtenir une prime PACE, la jument doit avoir participé à au moins un concours d'élevage avec une note minimale de 10 sur 20, à moins

d'avoir un indice de performance très élevé. Le montant de la prime versée dépend d'un nombre de points qui peut être calculé à partir du meilleur IRE obtenu par la jument sur sa carrière, et tient compte des indices de ses descendants également.

Le Trotteur Français : une approbation sur performances

Les étalons sont approuvés sur leurs performances (SECF 2011): ils doivent avoir été classés dans les 3 premiers d'une course de groupe I, ou être arrivés premier dans 6 courses avec un temps de parcours au km inférieur à un seuil qui varie en fonction de l'âge au moment de la course (il existe des équivalences, par exemple être arrivé premier dans une course de groupe II équivaut à 2 victoires avec le temps de parcours au km requis). Pour les étalons de 5 ans ou plus, le nombre de records à obtenir pour être approuvé peut être diminué si le cheval est le frère d'un grand gagnant, s'il a eu des gains élevés au cours de sa carrière ou s'il a reçu d'excellentes notes lors du concours national de sélection des chevaux entiers. En fonction du critère utilisé pour approuver l'étalon, le nombre de saillies annuelles autorisées varie de 20 à 100. Les étalons autorisés à moins de 100 saillies pourront voir leur nombre de saillies augmenter s'ils ont des produits classés dans les courses les plus prestigieuses. A l'inverse les étalons dont le niveau des produits serait insuffisant peuvent se voir retirer leur approbation.

Il y a de plus une sélection sur la voie femelle. L'appartenance à une catégorie dépend du record de vitesse obtenu, les valeurs seuils dépendant de l'âge de la jument et des conditions d'obtention du record (type de course, de départ, distance). Les juments nées jusqu'en 2004 inclus doivent avoir réussi la qualification (ou un test équivalent dans un autre pays), ou être la sœur d'un cheval classé dans les 3 premiers d'une course de groupe I, ou être classée en 1ère catégorie, ou être la fille d'une jument de cette catégorie (pour les juments nées entre 1997 et 2004 être fille d'une jument de 2ème catégorie suffit). Les juments nées après 2005 doivent avoir obtenu une victoire dans une course publique (en France ou à l'étranger), ou être classée en 1ère ou 2ème catégorie sur ses performances, ou être la fille d'une jument répondant elle-même à ce critère. Une jument doit avoir au moins 5 ans pour être mise à la reproduction. La reproduction dès 4 ans peut être autorisée pour les juments de 1ère ou 2ème catégorie ou les juments ayant leur mère en 1ère catégorie. Les juments approuvées peuvent aussi être suspendues si leurs produits n'ont pas d'assez bonnes performances.

2.6.3. Comment utiliser les indices génétiques?

Les indices génétiques sont publiés chaque année, accompagnés de leur CD (coefficient de détermination), qui définit un intervalle autour de la valeur génétique estimée dans lequel il y a 95% de chances que la vraie génétique se trouve. Evaluer un cheval sur la base de la borne basse de son intervalle de confiance minimise le risque de surestimer sa valeur génétique. L'intervalle de confiance se rétrécit au fur et à mesure de l'ajout de nouvelles informations dans le calcul de l'indice. C'est ce qui est recommandé avec le BSO et le BCC. Avec la moyenne des BSO de la base mobile mise à 0, la classification des chevaux suivant la borne basse de leur BSO se fait de la façon suivante : élite si supérieur à 15, très bon si compris entre 7.5 et 15, améliorateur entre 0 et 7.5, acceptable entre -7.5 et 0, médiocre entre -7.7 et -15, et déconseillé si inférieur à -15. L'augmentation du CD qui permet une estimation précise de la valeur génétique nécessite d'accumuler suffisamment d'information. A la naissance, le CD sur ascendance d'un cheval est d'environ 0.20-0.30. Avec 32 descendants, un étalon n'a qu'un CD moyen (0.70). Il faut 54 descendants avec des performances propres pour que la précision de la valeur génétique soit élevée (CD de 0.80).

Pour l'endurance, les chevaux ayant un CD trop faible dû à un apparentement lointain avec des chevaux participant à des courses d'endurance n'ont pas de BRE publié.

Le BSO et le BTR ont été pendant un temps mentionnés dans les stud-books du Selle Français et du Trotteur Français respectivement, mais ils ne sont à l'heure actuelle plus utilisés officiellement.

2.6.4. Quelles perspectives pour l'utilisation de la sélection génomique dans l'amélioration génétique des chevaux ?

Actuellement, l'efficacité des schémas de sélection des chevaux repose sur l'intensité de la sélection : en CSO par exemple, 70% d'une génération sort en compétition et est donc testée. L'âge optimum pour sélectionner conseillé par Dubois *et al*. (2008) est 5 ans. A cet âge les chevaux sont connus sur ascendance et sur performances propres, mais n'ont pas encore de descendants performeurs : la précision de la sélection est donc moyenne. En revanche les schémas ne sont pas optimums concernant l'intervalle de génération : ce paramètre est détérioré par une utilisation trop longue d'étalons bien connus mais dont le niveau génétique est rattrapé par les jeunes étalons.

Le génome du cheval est séquencé depuis 2009 (Wade *et al*.), et des puces SNPs sont disponibles (50K et 74K, Illumina). Compte-tenu des résultats déjà obtenus dans d'autres espèces, il est possible de réfléchir à l'intérêt d'une mise en place d'évaluations génomiques chez les chevaux.

L'intérêt de la sélection génomique chez les chevaux reposerait sur l'obtention de valeurs génétiques aussi précises qu'à l'âge actuel de sélection (5 ans) dès la naissance ou la maturité sexuelle du cheval. L'obtention de valeurs génétiques plus précocement devrait permettre de réduire l'âge de mise à la reproduction à 2 ans, soit un gain de 3 ans sur l'intervalle de génération. Le fait d'avoir une information précise suffisamment tôt devrait aussi permettre d'améliorer le tri réalisé parmi les reproducteurs potentiels au moment de la castration, améliorant ainsi l'intensité de la sélection. L'utilisation de la sélection génomique dans l'amélioration génétique des chevaux pourrait donc permettre d'améliorer deux paramètres du progrès génétique : l'intervalle de génération et l'intensité de la sélection. L'amélioration de ces deux paramètres repose sur l'obtention de valeurs génétiques suffisamment précises plus précocement : l'objectif de la thèse a donc consisté à vérifier les précisions qu'il était possible d'obtenir avec des évaluations génomique dans plusieurs populations de chevaux.

Le chapitre suivant présente les aspects théoriques de l'utilisation de la sélection génomique étudiés au cours de la thèse : cette partie théorique a consisté en l'étude des formules déterministes pour la prédiction de la précision de la sélection génomique.

3. Les formules pour la prédiction de la précision de la sélection génomique à l'épreuve de la méta-analyse

3.1. Introduction de l'article

Tout comme pour la sélection classique, la sélection génomique requiert de penser les schémas de sélection de façon à obtenir le meilleur progrès génétique possible. Le progrès génétique dépend de l'intervalle de génération, de la précision des évaluations génétiques et de l'intensité de la sélection. Suivant les schémas de sélection existants, il faut identifier le ou les paramètres qui seront influencés par le passage à la sélection génomique, et les possibles interactions entre les paramètres modifiés afin d'optimiser les schémas en conséquence.

L'optimisation d'un schéma classique consiste à déterminer les phénotypes à mesurer, les individus sur qui réaliser les mesures, et la quantité de données nécessaires. Dans le cas de la sélection génomique, une question supplémentaire se pose : celle du génotypage. Là aussi les interrogations concernent le choix des individus : faut-il génotyper tous les performeurs d'une génération, peut-on inclure des individus dont on ne connait pas le pédigrée ; leur nombre : effectif nécessaire pour avoir un échantillon représentatif de la population ; et la méthode : quelle densité de marqueurs utiliser.

Des formules ont été développées afin d'estimer la précision de la sélection génomique attendue en fonction de quelques paramètres: l'héritabilité du caractère, le nombre de SNPs, le nombre d'individus dans la population de référence, et le nombre de segments indépendants dans le génome (Daetwyler *et al*. 2008, Goddard 2009, Goddard *et al*. 2011, Meuwissen *et al*. 2013). Ces formules proposent de vérifier simplement la précision attendue en intégrant la sélection génomique dans des plans de sélection complexes. On peut par exemple calculer le nombre d'individus à génotyper pour atteindre la précision visée, ou calculer la précision attendue connaissant le nombre d'individus que l'on peut génotyper. Ces formules semblent donc constituer un bon outil pour étudier l'intérêt de l'utilisation de la sélection génomique dans un schéma de sélection.

Cependant, si ces formules prédisent mal la précision de la sélection génomique, les conséquences seront fâcheuses. Si la précision attendue connaissant le nombre d'individus pouvant constituer la population de référence est sous-estimée ou surestimée, l'utilisation de la sélection génomique risque d'être écartée à tort, ou bien les dépenses de génotypages seront réalisées mais les résultats obtenus seront décevants par rapport aux résultats attendus. De même, si le nombre d'individus à génotyper pour atteindre une certaine précision est mal estimé, la précision de la sélection génomique risquera d'être plus faible que celle attendue, ou atteindra la valeur prévue mais en ayant réalisé des génotypages inutiles.

Ricard *et al*. (2013) ont testé la sélection génomique chez des chevaux de CSO par validation croisée, une méthode qui permet d'estimer empiriquement la précision de la sélection génomique. Ricard *et al*. (2013) concluent qu'il y a un faible intérêt à utiliser la sélection génomique dans les conditions actuelles au moment de l'essai, car la précision de la sélection classique passe de 0.36 à 0.39 avec la sélection génomique. Disposant des données, j'ai pu utiliser les paramètres nécessaires dans les formules de prédiction de la précision de la sélection génomique, ce qui m'a amené à deux constats. D'une part, la précision prédite varie suivant la formule utilisée. Il existe même des résultats différents pour chaque formule, qui dépendent de la méthode utilisée pour calculer le nombre de segments indépendants M_e, paramètre qui peut se calculer à partir de la taille efficace de la

population N_e, elle-même estimable de plusieurs façons. D'autre part, une seule des précisions prédites approchait la précision observée (0.32). Quelques-unes des précisions prédites sous-estimaient la précision obtenue avec la sélection génomique, et beaucoup de prédictions étaient beaucoup trop optimistes, prévoyant une précision de 0.60 à 0.70 ! Or, lors de leur publication ces formules avaient été mises à l'épreuve par leurs auteurs, et les résultats présentés ne montraient pas de si grands écarts entre les précisions prédites et les précisions réalisées.

En conséquence, il nous a semblé nécessaire d'étudier plus en détail ces formules. Pour cela, je me suis basée sur 13 publications portant sur la sélection génomique contenant 145 valeurs de précision. Dans un premier temps, j'ai réalisé une analyse de sensibilité des formules à leurs paramètres. Pour cela, j'ai relevé dans les 13 publications les valeurs prises par les paramètres utilisés dans les formules : l'héritabilité, le nombre d'individus dans la population de référence, le nombre de marqueurs, la taille efficace de la population (nécessaire pour calculer le nombre de segments indépendants). Les publications étaient basées sur des données réelles ou simulées. Il est apparu que les paramètres prenaient des valeurs plus faibles dans les simulations, mais qu'il s'agissait seulement d'une diminution de l'échelle visant à réduire le temps des calculs, donc pour l'analyse de sensibilité nous avons utilisé des intervalles de variation des paramètres correspondant aux données réelles. Nous avons aussi tenu compte de la fréquence des valeurs prises par les paramètres. L'héritabilité et la taille efficace de la population peuvent prendre toutes les valeurs de l'intervalle défini. En revanche, la taille de la population de référence et le nombre de marqueurs n'ont pas des distributions continues : on peut avoir de très petites populations de référence, des populations de taille moyenne ou encore de très grandes populations incluant des animaux de plusieurs pays, et de même les puces utilisées comptent en général environ 3 000, 50 000 ou 700 000 marqueurs. L'analyse de sensibilité a été réalisée en calculant pour chacun des paramètres la densité marginale de la précision en fonction du paramètre, en intégrant les autres paramètres sur leur intervalle défini, avec une distribution continue pour l'héritabilité et le nombre de segments indépendants, et une distribution logarithmique pour la taille de la population de référence et le nombre de marqueurs. Dans un second temps, j'ai réalisé une méta-analyse basée sur les 145 valeurs de précision collectées dans les 13 publications, afin de vérifier si les formules sont réellement capables de prédire la précision de la sélection génomique. Pour cela, j'ai calculé les précisions prédites par les formules en utilisant les paramètres donnés dans les publications, et j'ai comparé les précisions prédites aux précisions réellement obtenues dans les publications. La méthodologie et les résultats de cette étude sont présentés en détail dans l'article inclus à ce chapitre.

Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Is the use of formulae a reliable way to predict the accuracy of genomic selection?

S. Brard[1,2,3] & A. Ricard[4,5]

1 INRA, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), Castanet-Tolosan, France
2 Université de Toulouse, INP, ENSAT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), Castanet-Tolosan, France
3 Université de Toulouse, INP, ENVT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), Toulouse, France
4 INRA, UMR 1313, Jouy-en-Josas, France
5 IFCE, Recherche et Innovation, Exmes, France

Keywords
Effective number of segments; genomic selection; Reliability.

Correspondence
S. Brard, INRA-GenPhySE, Auzeville BP52627, 31326 Castanet-Tolosan Cedex, France.
Tel: +33 561 285 182;
Fax: +33 561 285 353;
E-mail: sophie.brard@toulouse.inra.fr

Received: 19 February 2014;
accepted: 16 September 2014

Summary

We studied four formulae used to predict the accuracy of genomic selection prior to genotyping. The objectives of our study were to investigate the impact of the parameters of each formula on the values of accuracy calculated using these formulae, and to check whether the accuracies reported in the literature are in agreement with the formulae. First, we computed the marginal distribution of accuracy (by integration) for each parameter of all four formulae: heritability h^2, reference population size T, number of markers M and number of effective segments in the genome M_e. Then, we collected 145 accuracies and corresponding parameters reported in 13 publications on genomic selection (mainly in dairy cattle), and performed analysis of variance to test the differences between observed and predicted accuracy with effects of formulae and parameters. The variation of accuracy for different values of each parameter indicated that two parameters, T and M_e, had a significant impact and that considerable differences existed between the formulae (mean accuracies differed by up to 0.20 point). The results of our meta-analysis showed a big formula effect on the accuracies predicted using each formula, and also a significant effect of the value obtained for M_e calculated from N_e (effective population size). Each formula can therefore be demonstrated to be optimal depending on the assumption used for M_e. In conclusion, no rules can be applied to predict the reliability of genomic selection using these formulae.

Introduction

Ever since the very first publication by Meuwissen *et al.* (2001) which exposed the principles of genomic selection, it has been widely used in dairy cattle production. Genomic evaluation uses genotypes to estimate breeding values instead of or in addition to pedigree data. To design an efficient breeding plan, it is essential to know the accuracy of the breeding values predicted for the candidates to selection. In classical genetic evaluation methods based on pedigree, the accuracy of a breeding value depends on heritability and the number of performances recorded for the animal itself and its relatives. The accuracy can then be predicted based on these parameters, and a breeding plan elaborated before any phenotypes is recorded. With the advent of genomic evaluation methods, the need to predict accuracy knowing the breeding plan design arises for the same reasons, with in addition the necessity of deciding which animals should be genotyped and how many. Over the past few years, various formulae have been developed to predict the

  doi:10.1111/jbg.12123

accuracy of genomic evaluation (Daetwyler *et al.* 2008; Goddard 2009; Goddard *et al.* 2011; Meuwissen *et al.* 2013). These formulae use parameters that describe the data available for genomic selection (animals, traits and markers). Such formulae are intended to be used to provide a general picture of the possible interest of a genomic selection project before actually starting it, so the decision to implement or reject a selection project can be dependent on the accuracy predicted with these formulae.

However, to our knowledge, the importance of the effect of the parameters on the accuracy predicted using such formulae has never been explored in detail. Moreover, up to now, the reliability of the accuracies predicted using the formulae has been ascertained only using data chosen (real data) or generated (simulated data) specifically for the purpose of testing the formulae.

Our objectives were (i) to study to what extent variations of the parameters (heritability, reference population size, number of markers and number of independent segments) have an effect on the accuracy calculated using the formulae and (ii) to investigate whether the accuracy predicted using the formulae is exact. For this second part of the study, accuracies and the corresponding parameters were collected from published reports (mainly describing the accuracy of genomic selection in dairy cattle), and the observed accuracies were compared to accuracies calculated with the formulae using the parameters.

Material and methods

Four formulae for predicting the accuracy of genomic selection

This work focused on four recently developed formulae used to predict the accuracy of genomic selection.

Daetwyler formula

Daetwyler *et al.* (2008) reported a formula intended to predict the accuracy of genomic selection. The formula was $r_{g\hat{g}} = \sqrt{Th^2/(Th^2 + n_g)}$, $r_{g\hat{g}}$, being the accuracy of genomic selection, T the number of animals in the reference population, h^2 the heritability and n_g the number of independent loci affecting the trait. This formula was derived by considering the regression of phenotypes at one locus at a time, using a fixed model, in what can be called a 'marker model'. In 2010, Daetwyler *et al.* proposed a slightly different version of the formula based on the recent findings of Goddard (2009). Because of linkage disequilibrium, all loci are not independent, and the number of independent loci that have an additive and independent effect on a trait is inferior to the total number of loci. Therefore, they replaced n_g by the effective number of independent chromosome segments M_e. The formula therefore became as follows:

$$r_{g\hat{g}} = \sqrt{\frac{Th^2}{Th^2 + M_e}}. \quad (1)$$

Goddard 2009 formula

A different formula, also based on a marker model, was reported by Goddard (2009). In this case, the random normal marker effects are estimated using BLUP

$$r^2_{g\hat{g}} = \frac{\sum_{j=1}^{M_e} \left(\frac{\mathrm{TV}(m_j)}{\mathrm{TV}(M_j)+\lambda} V(m_j) \right)}{\sum_{j=1}^{M_e} V(m_j)}$$

where $V(m_j)$ is the variance of markers (m_j = 0,1,2 depending on the number of copies of each SNP allele) and $\lambda = \sigma^2_e/\sigma^2_\beta$ where σ^2_β is the variance of random marker effects (equal for all markers). Daetwyler *et al.* (2008) had previously calculated the explained total genetic variance, that is $h^2 = \sigma^2_g = \sum_{j=1}^{M_e} V(m_j)\sigma^2_\beta$. But here the summation had to be carried out over the distribution of markers for complex terms and required a hypothesis for the density of marker allele frequency. The distribution of marker frequencies under neutral mutation model was assumed to be $f(p) = 1/(\mathrm{Log}(2N_e)2p(1-p))$ (Hill *et al.* 2008) where N_e is the effective size of the population, so summations were approximated by integrating over this distribution. The final formula was therefore (with $\sigma^2_e = 1$)

$$r_{g\hat{g}} = \sqrt{1 - \frac{\lambda}{2T\sqrt{a}} \mathrm{Log}\left(\frac{1 + a + 2\sqrt{a}}{1 + a - 2\sqrt{a}} \right)} \quad (2)$$

where $\lambda = M_e/(h^2 \mathrm{Log}(2N_e))$ and $a = 1 + 2(M_e/Th^2 \mathrm{Log}(2N_e))$.

Goddard 2011 formula

A similar formula was developed by Goddard *et al.* (2011) using an 'animal model', where the phenotype is explained by a genomic value that is the sum of marker effects. Moreover, the variance–covariance between genomic animal values is expressed in a relationship matrix calculated using the genotypes instead of the pedigree:

$$r_{g\hat{g}} = \sqrt{b \frac{Tbh^2/M_e}{1 + Tbh^2/M_e}} \quad (3)$$

where $b = M/(M + M_e)$ is the proportion of genetic variance explained by the markers. Although derived

from a different but equivalent model (VanRaden *et al.* 2009), this formula differs from the Daetwyler formula only by the addition of the coefficient b for the regression between markers and QTL.

Meuwissen formula

Meuwissen *et al.* (2013) reviewed the recent advances of genomic selection and discussed its accuracy. Based on the formulae developed by Daetwyler *et al.* (2008) and Goddard (2009), the authors derived a new formula:

$$r_{g\hat{g}} = \sqrt{b\frac{Tbh^2/M_e}{1 + Tbh^2/M_e - h^2 r^2_{g\hat{g}}}}$$

in which a new term $-h^2 r^2_{g\hat{g}}$ appears to take into account the fact that when the accuracy of the predicted breeding values increases the error variance in the model decreases. Previously, in formulae (1) and (2), the error variance was assumed to be equal to the phenotypic variance because only one locus was taken into account at a time. However, when multiple loci are used, the error variance decreases. The implementation of this correction was first suggested by Daetwyler *et al.* (2008) in the Appendix of their report. The formula involves $r^2_{g\hat{g}}$ on both sides and may be solved for $r^2_{g\hat{g}}$ giving

$$r_{g\hat{g}} = \sqrt{\frac{\theta + 1 + \sqrt{(\theta+1)^2 - 4h^2\theta b}}{2h^2}}, \qquad (4)$$

where $\theta = Tbh^2/M_e$.

The accuracies computed with formulae of Daetwyler *et al.* (2008), Goddard (2009), Goddard *et al.* (2011) and Meuwissen *et al.* (2013) will hereafter be called r_D (1), r_{Go} (2) r_G (3) and r_M (4), respectively.

Formulae for effective number of segments M_e

The effective number of chromosome segments M_e (also called 'effective number of loci' or 'number of independent chromosome segments') was introduced in the formula used to predict accuracy by Goddard in 2009. Goddard assumed that every potential QTL (whatever its position in the genome) is tagged by a marker. Linkage disequilibrium reduces the number of markers needed to tag every QTL to a value called M_e that is less than the total number of loci. In that article, he proposed two formulae to link M_e to the effective population size N_e:

$$M_e = \frac{2N_eL}{\mathrm{Log}(4N_eL)} \quad (M_{e1})$$

$M_e = 4N_eL$ (M_{e5}) as described by Stam (1980), where L is the size (in Morgan) of the genome.

In the article published in 2011, Goddard *et al.* proposed two new formulae

$$M_e = \frac{2N_eL}{\mathrm{Log}(2N_el)} \quad (M_{e2})$$

$$M_e = \frac{2N_eL}{\mathrm{Log}(N_el)} \quad (M_{e3})$$

where l is the average length of a chromosome ($n_{chro}l = L$ where n_{chro} is the number of chromosomes).

M_{e2} is also used by Meuwissen *et al.* (2013) assuming $l = 1$.

The formula proposed by Hayes *et al.* (2009b), $M_e = 2N_eL$ (M_{e4}), results in a M_e value comprised between that of Stam (1980) and those obtained with the other formulae.

Finally, no less than five possibilities were used to compute M_e. These formulae are ranked from the lowest value for M_e: $M_e = 2N_eL/(\mathrm{Log}(4N_eL))$ (M_{e1}) to the highest value for M_e: $M_e = 4N_eL$ (M_{e5}) for a given value of N_e; the definition of N_e being also subject to several interpretations.

Data from thirteen distinct publications

The data from thirteen articles, published between 2001 and 2012 and investigating various issues pertaining to the accuracy of genomic selection (mainly in dairy cattle), were used to study the reliability of the formulae. The 13 publications were based either on simulated data (Meuwissen *et al.* 2001; Habier *et al.* 2007, 2009; Calus *et al.* 2008, 2009; Brito *et al.* 2011; Pszczola *et al.* 2011; Bastiaansen *et al.* 2012) or on real data (Hayes *et al.* 2009a; Luan *et al.* 2009; Verbyla *et al.* 2009; Habier *et al.* 2010; Moser *et al.* 2010). The ranges of the values collected are reported in Table 1.

Accuracy values gathered from the publications

The selected studies analysed the accuracy of genomic selection in different situations when the parameters (T, M, h^2, M_e, N_e), methods used and type of data (simulated or real) varied. The accuracies from the articles will hereafter be called 'observed accuracies', whereas the accuracies calculated later in this paper using the various formulae will be called 'predicted accuracies'.

The observed accuracies were of two types. In publications using simulated data, accuracy was the

Table 1 Range of values found in publications for accuracy of genomic selection and parameters, in real data or simulated data

	Real data (76 cases)				Simulated data (69 cases)			
	Mean	Standard deviation	Minimum	Maximum	Mean	Standard deviation	Minimum	Maximum
Observed accuracy	0.57	0.13	0.17	0.78	0.50	0.18	0.11	0.90
Heritability (h^2)	0.88	0.10	0.58	0.97	0.45	0.25	0.10	0.94
Size of the reference population (T)	812	551	250	2096	880	520	480	2200
Number of markers (M)	29 011	10 377	18 991	42 576	41 236	163 476	100	800 000
Effective size of population (N_e)	127	44	45	167	184	91	95	400
Length of the genome (L)	31.60	–	–	–	8.77	7.20	3.00	23.33
Effective number of segments 1 (M_{e1})	822	262	329	1060	432	498	81	1646
Effective number of segments 2 (M_{e2})	1264	381	542	1610	601	754	95	2440
Effective number of segments 3 (M_{e3})	1421	419	625	1800	669	836	108	2707
Effective number of segments 4 (M_{e4})	1621	464	737	2042	756	938	124	3038
Effective number of segments 5 (M_{e5})	7998	2794	2844	10 554	4097	5335	570	17 190

correlation between true breeding values and genomic breeding values, so it could be compared directly to r_D, r_G r_M and r_{Go}. But in publications based on real data, accuracy was computed as the correlation between the daughter yield deviation (DYD) and the genomic breeding value in a validation sample. In the validation sample, genomic breeding values were estimated without individual phenotypes from estimates obtained from a training sample combining phenotypes and genotypes. In this case, two corrections were needed. First, the observed accuracy was divided by $\sqrt{CD}$ (when not already done in the publication) to determine the true breeding value from the DYD (coefficient of determination (CD), squared correlation between the breeding values and the true genetic values in pedigree indexes). Second, because the phenotype was the mean of progeny results, h^2 was replaced in the formulae by CD. The ranges of values for observed accuracies are shown in Table 1.

Values for the parameters gathered from the publications

The values for the size of the training population T, the number of markers M, the heritability h^2, the effective size of population N_e and the effective number of chromosome segments M_e were collected from the various publications to define a range of values for each parameter. T and M are easily observable parameters. Heritability is fixed in simulated data and is a well-established parameter in studies using real data. Therefore, these three parameters were easy to find in publications on both simulated and real data. M_e is the number of independent loci that results in the same variance of realized relationship matrix as that obtained in realistic situations where an unknown number of QTL act together. In publications using simulated data, QTL are introduced at the start of the simulation of the population: M_e results from recombinations along the chromosomes and mating during the simulation of generations. Therefore, this parameter differs from other parameters by often not being mentioned directly in simulations and by being unknown in real data sets. However, the above-described formulae could be used to compute M_e from N_e and L (length of the genome). In most of the publications using simulated data, N_e was one of the simulation parameters, so it could be found in the Materials and Methods section. But for studies based on real data, the effective size of the population could be estimated in various ways using either pedigree, demographic or molecular data. N_e was most often not given in publications on genomic selection with real data; we therefore searched for it in other publications on the same breeds raised in the same countries (for example, De Roos *et al.* 2008). Finally, M_e was computed from N_e using the five formulae. The ranges of the parameters are reported in Table 1.

Parameter-dependent variation of predicted accuracy

The five formulae depend on four parameters: T, h^2, M_e and M (except for r_D and r_{Go} that do not depend on M). In addition, r_{Go} varies with N_e but, as N_e and M_e are related, N_e was computed using this relationship in five different ways according to the five formulae. For formulae (1–3), the zero searching routine C05AYF from NAG (Numerical Algorithms Group Ltd., Oxford, UK) library was used to find N_e as a function of M_e.

To analyse how variation of the parameters affects the predicted accuracy, variation ranges were defined for each parameter based on the values observed in the 13 articles (Table 2). It should be noted that a

single range was chosen for each parameter whatever the data type, either simulated or real. The choice of variation ranges took into account the fact that the very low values for parameters found in some simulated data studies were intended to mimic real data but at a smaller scale. Hence, for example, the minimum number of markers used in a simulation study (100) was not retained as no one at present would begin genomic selection with such low density of markers. A minimum of 3000 markers corresponding to a low-density beadchip was used. The range for the size of the reference population was the same range in both simulated and real data. Nevertheless, consistent with the report by Lund *et al.* (2011) who used a reference population consisting of 20 000 animals, a higher maximum was chosen for this parameter. We chose to retain this value as maximum with the objective of dealing with all possible situations. Nevertheless, because some of the parameters were missing for that study, it was not used in the meta-analysis. The range of values used for M_e was reduced slightly to avoid extreme values for N_e (<0) with some formulae.

The marginal probability density function of accuracy was computed for each parameter, by integration over the other parameters, for example for T,

$$f(r_{g\hat{g}}|T) = \iiint_{M,M_e,h^2_{e\Delta}} f(r_{g\hat{g}}|T,M,M_e,h_2)p(M)p(M_e)\,p(h^2)\mathrm{d}M\mathrm{d}M_e\mathrm{d}h^2,$$

with $f(r_{g\hat{g}}|T,M,M_e,h^2)$ for the four previous formulae, and $p(M)$, $p(M_e)$, $p(h^2)$ and $p(T)$ the density of each parameter. Similar formulae may be built for the other parameters. This integration was performed using the D01FCF routine of NAG (Numerical Algorithms Group Ltd.) library. For each parameter, the density function was chosen to attribute similar probabilities to the most common values. To do so, a uniform distribution over the range of values taken by parameters h^2 and M_e was chosen. Low (0.1) and medium (0.5) heritabilities reflected a reference population with phenotypic records, and high heritability (up to 0.98) reflected a reference population with progeny testing, each situation being possible. For M and T, a uniform log distribution was chosen to attribute equal probabilities to the three main ranges: low-density beadchip (3K), medium (50K) or high density (800K). For T, the most common cases were a small reference population (simulation studies, 250), a conventional reference population (few thousands) or a large international population (Eurogenomics, 20 000).

Table 2 Range of values of parameters chosen for the study of the marginal distribution of accuracy

Parameter	Minimum	Maximum
Heritability	0.10	0.98
Size of the reference population	250	20 000
Number of markers	3000	800 000
Effective population size	45	400
Effective number of segments	250	20 000

$$p(T) = \frac{1}{\max(T) - \min(T)},$$
$$\max(T) < T < \min(T) \text{ for } h^2, M_e.$$

$$p(\mathrm{Log}(T)) = \frac{1}{\max(\mathrm{Log}(T)) - \min(\mathrm{Log}(T))},$$
$$\max(T) < T < \min(T) \text{ for } M, T.$$

Correspondence between observed and predicted accuracies

One hundred and forty-five values of observed accuracies were gathered from the 13 publications and could be compared to the predicted accuracies. An analysis of variance was performed on the differences between observed and predicted accuracies. The sources of variation were as follows: combination of formula (four levels) and method used to calculate M_e (five levels), so in all 20 levels, with type (simulated or real data) and T, h^2, M, L (genome length) as covariates.

Results

Marginal distribution of accuracy

Variation of accuracy

Figures 1–4 display the variation of accuracy for the four formulae as a function of the four parameters. For r_{Go}, only the curves for the two extreme assumptions about the relationship between M_e and N_e are shown (all the others are fall within). The curves differed depending on the parameters. The accuracy increased when T, h^2 or M increased, and decreased when M_e increased. For M, curves reached a plateau, but it can be noted that it was not reached with the 50 000 markers of the most common beadchip. T and M_e induced more important variations of $r_{g\hat{g}}$ than did h^2 or M: depending on the formula, the higher variations of accuracy varied by up to 0.70, 0.61, 0.31 and 0.28 for M_e, T, h^2 and M, respectively.

The average accuracy was different for each formula: r_G (0.31) < r_M (0.33) < r_D (0.39) < r_{Go} N_{e6}

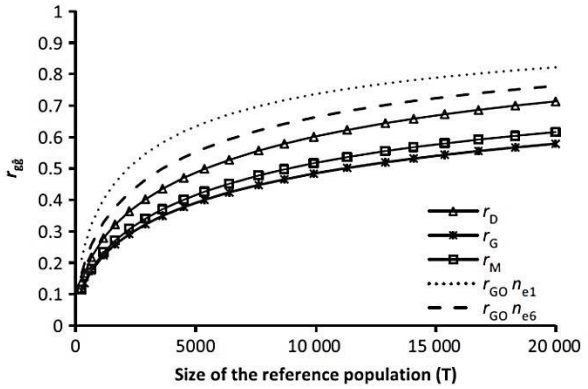


Figure 1 Marginal distribution of accuracy as a function of the size of the reference population *T*.

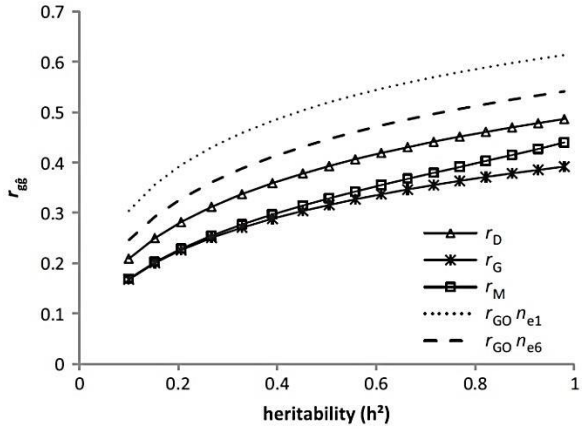


Figure 2 Marginal distribution of accuracy as a function of heritability h^2.

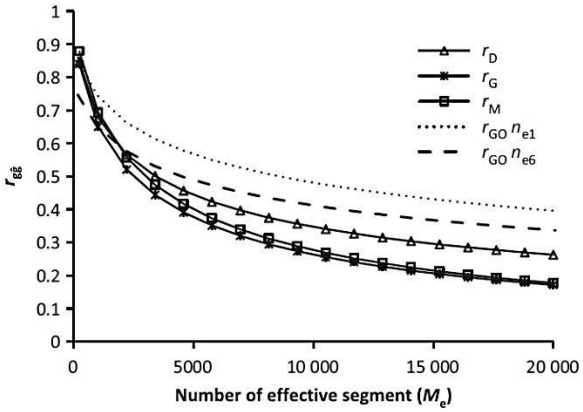


Figure 3 Marginal distribution of accuracy as a function of the number of effective segments M_e.

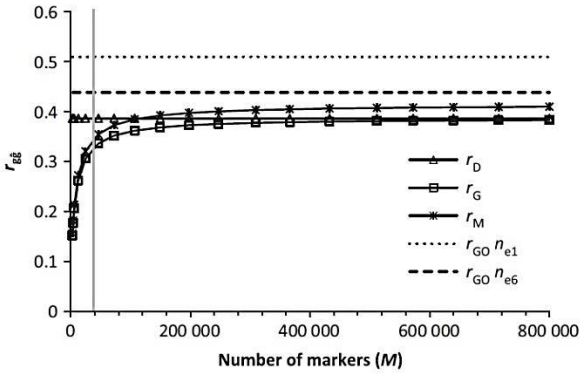


Figure 4 Marginal distribution of accuracy as a function of the number of markers *M*.

(0.44) < r_{Go} N_{e1} (0.51). In comparison with the accuracy values obtained with pedigree evaluation in a standard breeding scheme, these accuracies were higher than those obtained for the prediction of individual performances with an intermediate heritability, but lower than those obtained by progeny testing. However, depending on the value of the parameters, the accuracy calculated using the formulae could be much lower and unfavourable (<0.20) or much higher and favourable (>0.70).

Comparison of formulae

According to the results shown in Figures 1–4, r_D provides higher values of accuracy than r_G. This lower accuracy calculated using r_G is due to the regression between markers and QTL that Goddard *et al.* (2011) took into account compared with Daetwyler *et al.* (2008). The difference between r_G and r_D varied between 0.08 and 0.10 which proved the impact of such a term on the results and may favour one of the formulae over the other when comparing observed accuracies.

The accuracy calculated using r_M was higher than that of r_G because Meuwissen *et al.* (2013) improved the formula by introducing the decrease of residual variance, and this leads to a further increase of the accuracy by 0.02–0.05.

Whatever the value used for N_e, the accuracy calculated using r_{Go} was greater than that of r_D when the parameters *T*, *M* or h^2 varied. For low values of M_e (<1600), the accuracy calculated using r_{Go} was lower than of r_D, but higher for higher values of M_e.

Comparison of observed and predicted accuracies

Variance analysis showed that the type of data (real or simulated) was not significant, whereas all other effects were significant ($p < 0.0001$). Predicted

accuracy overestimated observed accuracy for high values of T (+0.06 for every additional 1000 animals), M (+0.01 for every additional 100 000 markers) and L (+0.008 for every additional Morgan). Predicted accuracy underestimated observed accuracy for high heritabilities (−0.057 per 0.1 h^2).

Figure 5 displays the average differences between observed and predicted accuracies depending on the formulae used to compute $r_{g\hat{g}}$ and M_e. The ranking of the formulae for accuracy was identical to that obtained when investigating the effect of the parameters. The differences between observed and predicted accuracies increased as M_e increased and depended on the method used to get this parameter. Basically, accuracy was overestimated when using M_{e1} and underestimated when using M_{e4} or M_{e5}. So, according to these results, the best formula to predict accuracy depends on the formula used to compute M_e. With M_{e2}, M_{e3} or M_{e4}, the formulae that give the best predictions are r_G, r_M and r_{Go}, respectively.

Some detailed results on simulated data and real data

Figure 6 shows the predicted accuracies and observed accuracy for the simulated data set used by Meuwissen *et al.* (2001). Parameters values were $h^2 = 0.5$, $M = 1010$, $T = 1000$, $L = 10\ M$ and $N_e = 100$. The observed accuracy was 0.66. When M_{e4} or M_{e5} were used, all formulae greatly underestimated the accuracy. The best prediction was obtained using r_M with M_{e2} (relative difference = 5%), followed by r_G with M_{e1} or M_{e2} (relative difference ≤10%).

Figure 7 shows the predicted accuracies and observed accuracy for the real data set used by Luan *et al.* (2009). Parameter values were h^2 = CD = 0.97, M = 18 991, T = 500 and N_e = 167. Accuracy was well predicted (relative difference ≤5%) when r_M was used with M_{e1} and when r_{Go} was used with M_{e2}. Accuracy was fairly well predicted with r_{Go}, r_D and r_G used with M_{e1} (relative difference ≤10%).

Similar results were obtained when the accuracies of other data sets were compared. Globally, the predicted accuracy could either be close to the observed value, or in other cases very far from the observed accuracy, depending on the formulae used to calculate the accuracy and M_e. Generally, accuracy was underestimated when M_{e4} or M_{e5} were used.

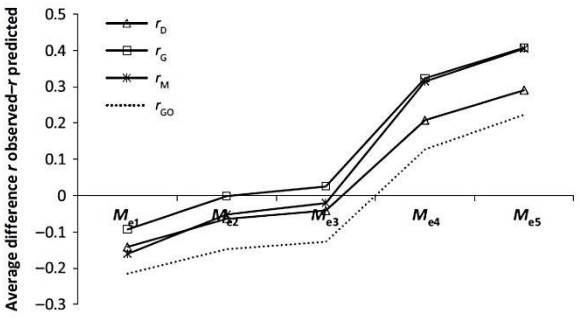


Figure 5 Average differences between the observed and predicted accuracies depending on the formulae for accuracy r and for the number of effective segments M_e.

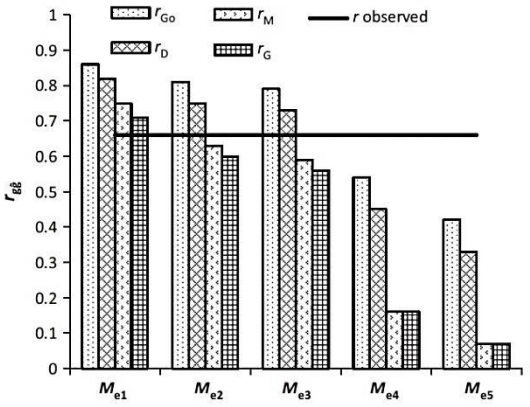


Figure 6 Application of the formulae predicting accuracy using different methods to calculate M_e, and comparison with the accuracy observed by Meuwissen *et al.* (2001). Parameter values are: $h^2 = 0.5$, $T = 1000$, $M = 1010$, $L = 10\ M$, $N_e = 100$ (simulated data).

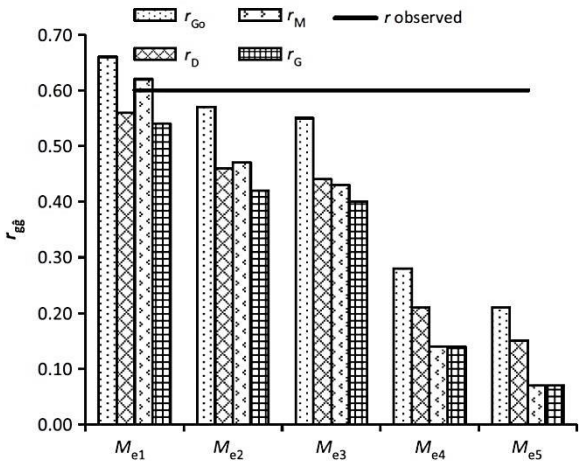


Figure 7 Application of the formulae predicting accuracy using different methods to calculate M_e, and comparison with the accuracy observed by Luan *et al.* (2009). Parameter values: h^2 = CD = 0.97, T = 500, M = 18 991, N_e = 167 (real data).

Discussion

The first objective of this study was to investigate how the accuracy predicted for genomic selection varied depending on the values of the parameters involved in the formulae. Our results demonstrated that two parameters had a significant impact: the number of animals in the reference population T and the effective number of segments M_e. Moreover, depending on the formula used, the values computed for M_e from N_e were quite different (Figure 8). This is considerable importance as we showed that the weight of this parameter (M_e) was significant in the accuracy calculated using the formulae.

When comparing the reliability of the predicted accuracies, the relative performances of the formulae were as could be expected. Ever since the first formula was developed by Daetwyler *et al.* (2008), improvements have aimed at obtaining a better fit with real data and enhancing the prediction of accuracy. Nevertheless, the results of our meta-analysis did not establish the superiority of one formula over the others.

The main interest of formulae intended to predict accuracy is for designing selection plans and estimating the required training population size before starting genotyping. The minimal suitable accuracy values vary according to species, breeds and traits. By way of example, let us consider what the expected training population size would be if the required accuracy is 0.5. According to Figure 2, the number of animals needed to reach this level of accuracy is comprised between 2140 and 11 300 depending on which formulae are used to compute M_e and $r_{g\hat{g}}$. Such a range of values is much too large to be helpful; one must therefore be able to choose the appropriate formula with certainty before using it for predictions.

Our second objective was to investigate whether the formulae used to predict the accuracy of genomic selection actually work. Unfortunately, variance analysis testing the differences between observed and predicted accuracies did not evidence the superiority of any one of the formula over the others. This is due to the uncertainty introduced by the number of methods used to estimate M_e. In the present report, M_e was computed using formulae with two parameters: the length of the genome and the effective population size. For one particular N_e, very different values of M_e could be obtained, especially for high N_e values. Moreover, the five formulae used do not take into account that M_e does not depend only on the species and the population, but also on the relationship between the reference population and the population to be estimated. In addition, N_e is also a source of uncertainty because it can be calculated using at least six different methods each based on their own hypothesis and leading to different values.

Both Daetwyler *et al.* (2010) and Goddard *et al.* (2011) compared the accuracies predicted with their formulae to observed accuracies to ascertain their formulae. In both papers, populations were simulated over several generations from a base population to reach mutational drift equilibrium and achieve linkage disequilibrium between markers. Simulations were performed for different heritabilities and effective sizes of population. They calculated the breeding values using GBLUP in both publications, as well as

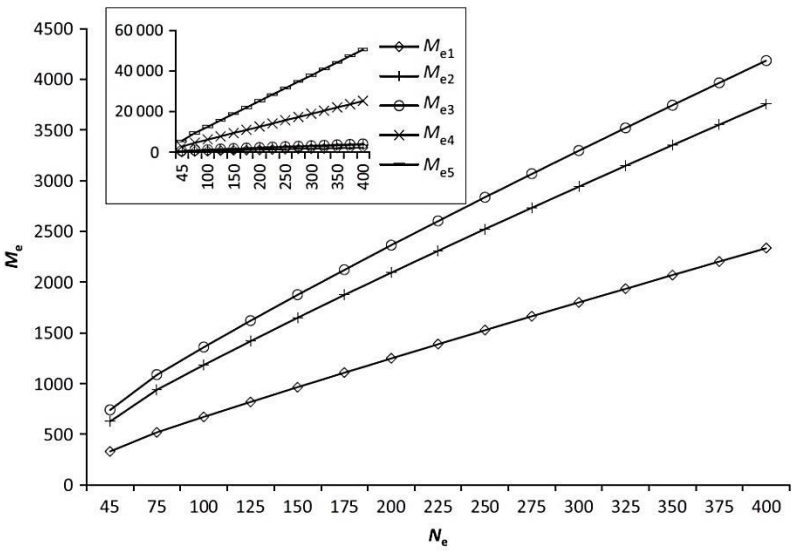


Figure 8 Variation of M_e depending on N_e according to the formulae used for computation.

Bayes B in Daetwyler *et al.* (2010). In Daetwyler *et al.* (2010), the formula used to calculate M_e was M_{e1}. When $N_e = 200$, the formula correctly predicted the accuracy obtained with GBLUP. In other cases, for example when $N_e = 1000$, the accuracy was underestimated, meaning that the results of a genomic selection plan would actually be better than those predicted with the formula. This underestimation could be due to the fact that the chosen N_e value was much higher than actual values (between 50 and 120 in most dairy cattle species). When N_e is high, M_e is also high so there are more effective segments to be estimated, and thus the accuracy predicted by the formula decreases. On the contrary, when N_e is low (as is the case for dairy cattle), then M_e is also low and the number of effective segments to be estimated is smaller, so the result of the formula is much more optimistic (we observed this situation when applying the formula). In the same way, a value $N_e = 1000$ was used for the population simulated in Goddard *et al.* (2011). M_e was calculated with M_{e3}, but the resulting effect was the same: in real populations, N_e and M_e are smaller, so the risk of getting over-optimistic results with the different formulae is higher.

Goddard *et al.* (2011) also worked on real data from dairy cattle, the fat percentage in Australian Holstein bulls, and used M_{e3}, with $N_e = 100$. Their results showed that the agreement between predicted and observed accuracies was good. Nevertheless, Hayes *et al.* (2009a) reported observed accuracies of genomic selection in the same breed, albeit for different traits (milk yield, protein, fat, protein percentage, fat percentage), and found that the observed accuracies depended on the trait whereas the accuracy predicted with Goddard's formula and M_{e3} was the same for all traits. When using the formulae, all the predicted accuracies were identical because the parameters were the same for all these traits (same population and same markers). However, the heritability of the traits was different, but as the phenotypes were DYDs, the reliabilities of DYDs were used instead of the heritabilities. As the reliabilities were the same for all traits, the predicted accuracies were identical for all traits. Therefore, the differences between the observed accuracies and the predicted accuracies (same for all traits) could be due to a different genetic architecture for the various traits. Hence, it might not be possible to generalize the good results found by Goddard *et al.* (2011).

Results obtained with the formulae developed by Goddard (2009) and Daetwyler *et al.* (2010) were compared by Hayes *et al.* (2009c) and shown to be very close. For this comparison, Hayes *et al.* used M_{e4} and the Daetwyler formula was corrected for the decrease of residual variance (from 1 to $1-h^2$), but without solving the second-degree equation, only by the approximation given in the appendix of Daetwyler *et al.* (2008). Without this correction, the similarity would not have been so pronounced, around 0.09 (especially for high heritabilities). To be fair, the comparison should also have included the decrease of residual variance in the Goddard formula, which would have resulted in an increase of reliability and a greater difference with the Daetwyler formula.

Some publications have evidenced that the accuracy of genomic selection depends on other parameters that are not directly used in the formulae such as the genetic architecture of the trait (Bastiaansen *et al.* 2012), the proportion of genetic variance truly captured with markers, or the source of information such as cosegregation or genetic relationships (Habier *et al.* 2007, 2013; Hayes *et al.* 2010; Pszczola *et al.* 2012). However, the results we obtained suggest that the problems encountered when comparing predicted and observed accuracies are more likely to be due to the uncertainty on the estimation of M_e than to defaults of the formulae. Although the formulae have already been improved, solving the problem of properly estimating M_e seems to be the next important step to take. Recently, Erbe *et al.* (2013) tested the validity of r_D and r_G by estimating b and M_e from accuracies obtained in data using different randomly chosen replicated training set sizes. They proved that the proportion of genetic variance captured by markers (b) follows a function based on the logarithm of marker density rather than the simple formula ($b = M/(M + M_e)$) proposed by Goddard *et al.* (2011). They found very different M_e values for the two breeds studied, Brown Swiss and Holstein, without any link to effective population sizes. Using a M_e value estimated from a portion of the data to predict the accuracy of the full data set led to an overestimation of the accuracy. However, these results should perhaps be taken with some caution because of the very high accuracy obtained (0.70 and higher) and the fact that the authors did not discuss how relationships might be taken into account. This was one of the first attempts to consider b and M_e as parameters to be estimated before used in other sets of the same breed and same trait. Although the authors did not solve the problem of the theoretical prediction of such parameters without knowing the data, they proposed a practical way to extend the first results of genotyping programme to larger population. As for classical genetics, the heritability is always estimated at the beginning, perhaps M_e

should also be estimated before designing selection plans.

Additional parameters could be introduced to further enhance these formulae that still do not predict sufficiently reliable accuracy values. On the other hand, a good estimation of M_e might suffice as it could take into account the parameters suggested to be missing such as genetic architecture, cosegregation, genetic relationship.

So far, the main problem we evidenced is that the uncertainty on the appropriate method to use to estimate M_e prevents proper testing of the formulae for their prediction of accuracy, and determining whether one of the formulae is superior over the others. There is no evidence from population history or structure demonstrating that a formula is more suitable than another. Our only recommendation to people aiming to plan genomic selection using these formulae is to pay attention to the parameters in general, because we have proved that the formulae can both overestimate and underestimate accuracy for extreme values of parameters, and to be very careful with M_e because this parameter has a huge weight and its estimation is completely different depending on the method used.

For the moment, the only advice we can give is of opposite nature. In effect, in a population where genomic selection already works, the formulae could be reversed to compute M_e from the accuracy and the other parameters, as proposed by Daetwyler *et al.* (2010). The value obtained for M_e could thereafter be used to predict the accuracy of genomic selection for other traits for which genomic selection has not yet been performed. Nevertheless, the applicability of this approach is limited because the population would have to have the very same structure as that used to estimate M_e from $r_{g\hat{g}}$, otherwise M_e would not be the same. Further work to improve the estimation of M_e may be a solution to ensure the use of these formulae to predict accuracy with a limited risk of error.

References

Bastiaansen J.W.M., Coster A., Calus M.P.L., van Arendonk J.A.M., Bovenhuis H. (2012) Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genet. Sel. Evol.*, **44**, 3.

Brito F.V., Neto J.B., Sargolzaei M., Cobuci J.A., Schenkel F.S. (2011) Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. *BMC Genet.*, **12**, 80.

Calus M.P.L., Meuwissen T.H.E., de Roos A.P.W., Veerkamp R.F. (2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, **178**, 553–561.

Calus M.P.L., Meuwissen T.H.E., Windig J.J., Knol E.F., Schrooten C., Vereijken A.L.J., Veerkamp R.F. (2009) Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genet. Sel. Evol.*, **41**, 11.

Daetwyler H.D., Villanueva B., Woolliams J.A. (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE*, **3**, e3395.

Daetwyler H.D., Pong-Wong R., Villanueva B., Woolliams J.A. (2010) The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, **185**, 1021–1031.

De Roos A.P.W., Hayes B.J., Spelman R.J., Goddard M.E. (2008) Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics*, **179**, 1503–1512.

Erbe M., Gredler B., Seefried F.R., Bapst B., Simianer H. (2013) A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS ONE*, **8**, e81046.

Goddard M. (2009) Genomic selection: prediction of accuracy and maximization of long term response. *Genetica*, **136**, 245–257.

Goddard M.E., Hayes B.J., Meuwissen T.H.E. (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.*, **128**, 409–421.

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

Habier D., Fernando R.L., Dekkers J.C.M. (2009) Genomic selection using low-density marker panels. *Genetics*, **182**, 343–353.

Habier D., Tetens J., Seefried F.R., Lichtner P., Thaller G. (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.*, **42**, 5.

Habier D., Fernando R.L., Garrick D.J. (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics*, **194**, 597–607.

Hayes B.J., Bowman P.J., Chamberlain A.C., Verbyla K., Goddard M.E. (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.*, **41**, 51.

Hayes B.J., Visscher P.M., Goddard M.E. (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb)*, **91**, 47–60.

Hayes B.J., Daetwyler H.D., Bowman P., Moser G., Tier B., Crump R., Khatkar M., Raadsma H.W., Goddard M.E. (2009c) Accuracy of genomic selection: comparing theory and results. *Proc. Assoc. Advmt. Anim. Breed. Genet.*, **18**, 34–37.

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J., Goddard M.E. (2010) Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milkfat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet.*, **6**, e1001139.

Hill W.G., Goddard M.E., Visscher P.M. (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.*, **4**, e1000008.

Luan T., Woolliams J.A., Lien S., Kent M., Svendsen M., Meuwissen T.H.E. (2009) The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics*, **183**, 1119–1126.

Lund M.S., de Roos S.P.W., de Vries A.G., Druet T., Ducrocq V., Fritz S., Guillaume F., Guldbrandsten B., Liu Z., Reents R., Schrooten C., Seefrid F., Su G. (2011) A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet. Sel. Evol.*, **43**, 43.

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

Meuwissen T., Hayes B., Goddard M. (2013) Accelerating improvement of livestock with genomic selection. *Annu. Rev. Anim. Biosci.*, **1**, 221–237.

Moser G., Khatkar M.S., Hayes B.J., Raadsma H.W. (2010) Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genet. Sel. Evol.*, **42**, 37.

Pszczola M., Mulder H.A., Calus M.P.L. (2011) Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle. *J. Dairy Sci.*, **94**, 431–441.

Pszczola M., Strabel T., Mulder H.A., Calus M.P.L. (2012) Reliability of direct genomic breeding values for animals with different relationships within and to the reference population. *J. Dairy Sci.*, **95**, 389–400.

Stam P. (1980) The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genet. Res.*, **35**, 131–155.

The NAG library, The Numerical Algorithm Group (NAG), Oxford, UK (available at: www.nag.com; last accessed 15 January 2014).

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

Verbyla K.L., Hayes B.J., Bowman P.J., Goddard M.E. (2009) Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet. Res. (Camb)*, **91**, 307–311.

3.2. Conclusions de l'article

L'analyse de sensibilité des formules pour la prédiction de la précision de la sélection génomique à leurs paramètres a permis de classer ceux-ci par ordre d'importance. Les paramètres qui engendrent le plus de variation dans la précision prédite sont la taille de la population de référence et le nombre de segments indépendants M_e, la précision étant en comparaison moins sensible à une variation de l'héritabilité ou du nombre de marqueurs.

La méta-analyse a confirmé le résultat observé chez les chevaux : les formules prédisent des précisions très différentes de celles attendues et souvent surestimées. Les causes des écarts entre précisions observées et précisions prédites ont été recherchées par une analyse de variance combinant la formule de prédiction de la précision et la formule utilisée pour estimer M_e. Cette analyse de variance a montré que plus la valeur de M_e estimée est grande et plus les précisions prédites diffèrent d'une formule à l'autre. Nous avons aussi montré que sur les 5 formules testées pour le calcul de M_e, deux sont à écarter, celles de Stam (1980) et de Hayes *et al.* (2009), car la précision sera ensuite sous-estimée avec les formules actuelles de prédiction de la précision. Une autre formule d'estimation de M_e, celle de Goddard (2009), doit aussi être évitée car les précisions prédites avec le M_e correspondant seront surestimées.

Il reste donc deux formules qui semblent acceptables pour calculer M_e. Cependant, même si les formules de Daetwyler *et al*. (2008) et Meuwissen *et al*. (2013) prédisent avec une erreur plutôt faible en espérance (écart inférieur à 0.08), les résultats varient beaucoup suivant les cas, et une formule qui prédit bien la précision dans une situation peut la surestimer dans une autre situation. Ces résultats peuvent sembler surprenants dans la mesure où les auteurs des formules les avaient testées. D'après leurs résultats, les formules de Goddard et de Daetwyler prédisaient bien la précision ou bien la sous-estimaient, ce qui peut être considéré comme un moindre mal comparé à une précision surestimée conduisant à lancer la sélection génomique dans une population avec finalement un résultat décevant par rapport à celui attendu. Cependant, ayant vérifié les plages de variation des paramètres des formules afin de réaliser l'analyse de sensibilité, j'ai constaté que les formules pour prédire la précision ont été testées dans des conditions particulières ou ne permettant pas de mettre en lumière leurs dysfonctionnements. Les auteurs ont utilisé des populations simulées, dans lesquelles la taille efficace N_e servant à calculer M_e est fixé à 200 ou 1000, alors que chez les bovins laitiers le maximum observé pour cette valeur est plutôt de 120. En utilisant un N_e trop grand ils estiment un M_e trop grand également. Plus il y a de segments indépendants dans le génome dont les effets sont à estimer, plus la précision est faible, et donc la surestimation de M_e masque la tendance des formules à surestimer la précision de la sélection génomique quand des valeurs plus proches de la réalité sont utilisées en paramètres. Un autre point n'apparaît pas dans leur vérifications : avec ces formules des cas de sélection génomique caractérisés par les mêmes paramètres (par exemple deux caractères évalués dans la même population avec les mêmes marqueurs et ayant la même héritabilité) auront la même précision prédite, alors qu'en pratique les précisions obtenues peuvent être différentes.

Nous concluons donc de cette étude que les formules pour la prédiction de la précision de la sélection génomique doivent être utilisées avec prudence car pour l'instant aucune formule ne peut être préférée aux autres. Le majeur problème identifié est celui de l'estimation de M_e, car suivant la méthode utilisée la formule optimale pour estimer la précision change. Les formules intègrent simplement les paramètres influant sur la précision de la sélection génomique: l'héritabilité, le

nombre de marqueurs, le nombre d'individus dans la population de référence. Le paramètre M_e dépend de l'apparentement, de la variabilité dans la population de référence, de l'étendue du déséquilibre de liaison. Compte-tenu de la nature de ces facteurs et de leur importance dans la précision de la sélection génomique, on pourrait envisager d'estimer M_e préalablement à la sélection, dans chaque population et pour chacun des caractères, au même titre que l'héritabilité. Par ailleurs, une amélioration de son estimation devrait permettre l'utilisation des formules pour la prédiction de la précision de la sélection génomique.

Après cet aspect théorique de l'utilisation de la sélection génomique, les chapitres 5, 6 et 7 présenteront les résultats obtenus en testant la sélection génomique dans différentes populations de chevaux. Le chapitre numéro 4 qui est le suivant présente les résultats d'une analyse d'association pour la performance en CSO, réalisée afin de vérifier l'architecture génétique de ce caractère avant de tester la sélection génomique.