Quel cadre théorique et pratique pour l'utilisation de la sélection génomique dans l'amélioration génétique des chevaux ?

Sophie Brard

Quel cadre théorique et pratique pour l'utilisation de la sélection génomique dans l'amélioration génétique des chevaux ?

Sophie Brard

Directeur de thèse : Anne Ricard

Laboratoire d'accueil : UMR 1388 INRA-GenPhySE (Génétique Physiologie et Systèmes d'élevage)

Membres du jury : Sophie Danvy, Mathilde Dupont-Nivet, Steven Janssens, Pascale Le Roy, Anne Ricard

Date de soutenance : 8 octobre 2015

Lieu : Castanet Tolosan

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

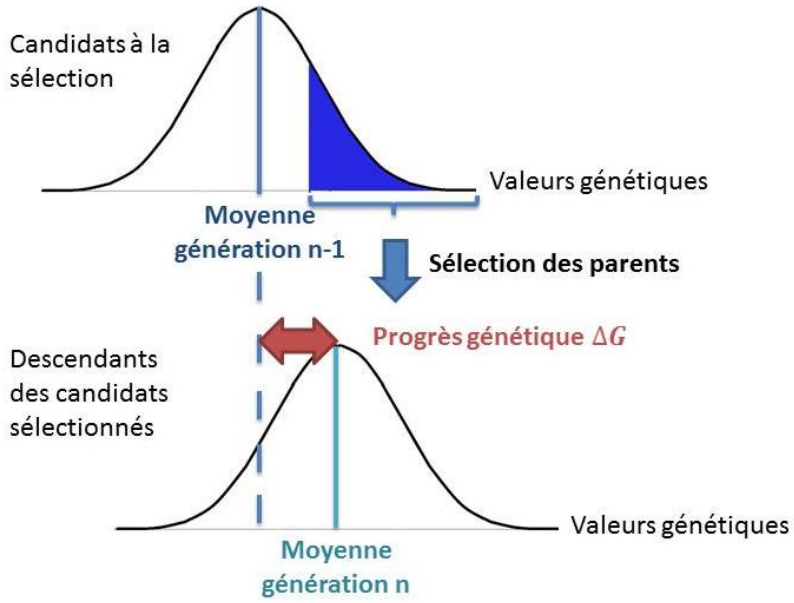
génomique, une méta-analyse reprenant les formules déterministes de calcul de la précision a été réalisée (3). Puis nous avons cherché une solution pour chacune des populations : recherche de marqueurs à effet important (4) et utilisation de l'ensemble de la population dans l'évaluation génomique (5) 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

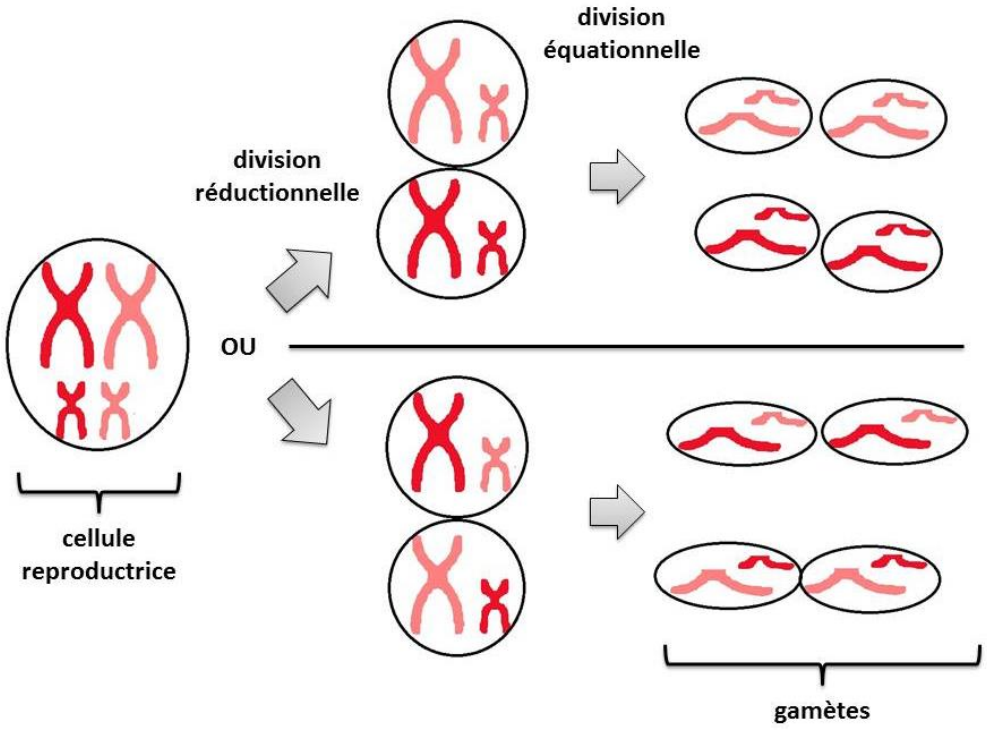
Qu'est-ce qu'un SNP (Single Nucleotid Polymorphism)?

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.

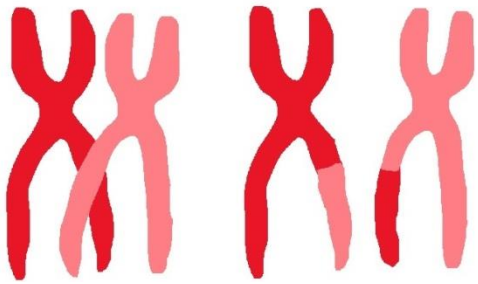
Une mesure courante du DL est r^2. Cette mesure est légèrement différente du D présenté précédemment. r^2 est le coefficient de corrélation entre les génotypes au marqueur et au QTL (ou à un 2ème marqueur), il représente la proportion de la variance expliquée par le QTL qu'on peut observer avec le marqueur (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 (h² 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

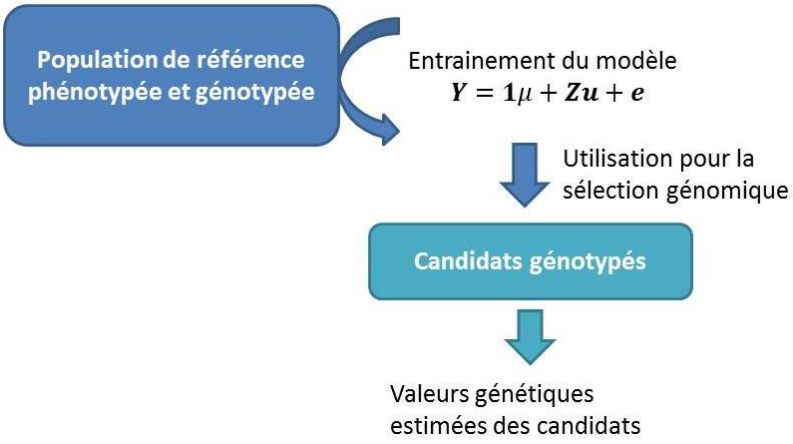
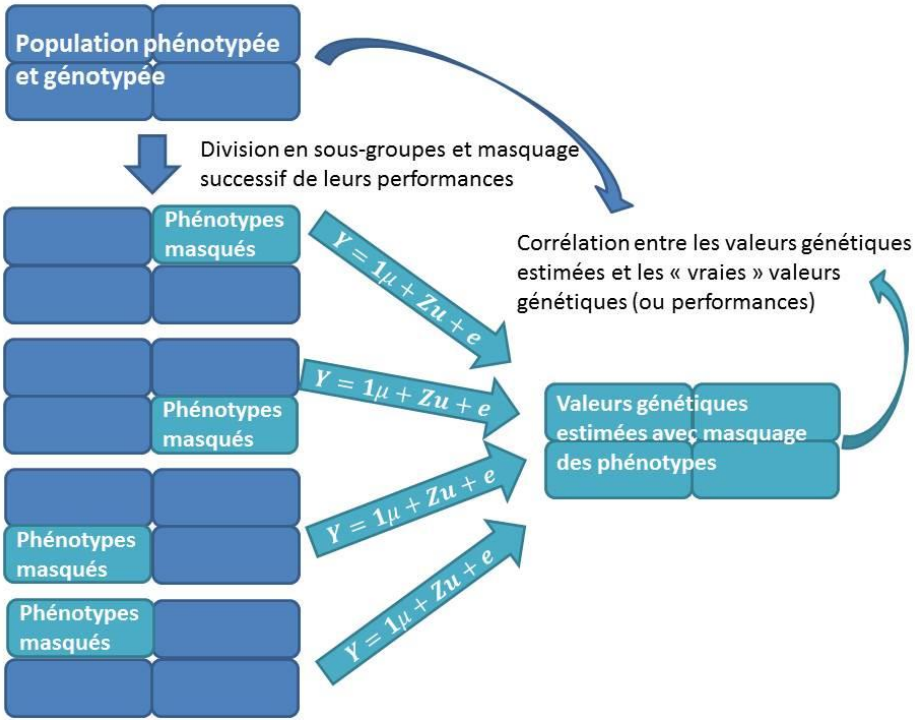


Figure 1.5 : Illustration du principe de la validation croisée



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

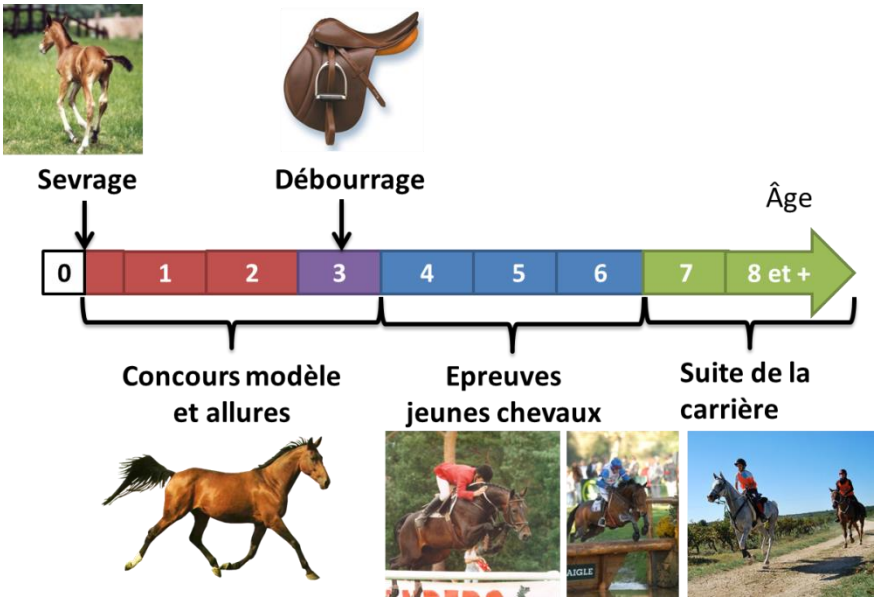
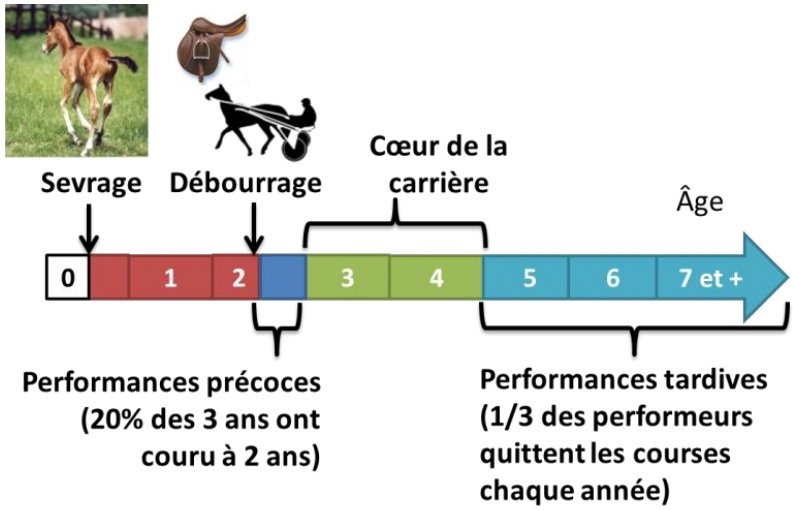


Figure 2.7 : Carrière d'un trotteur



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}

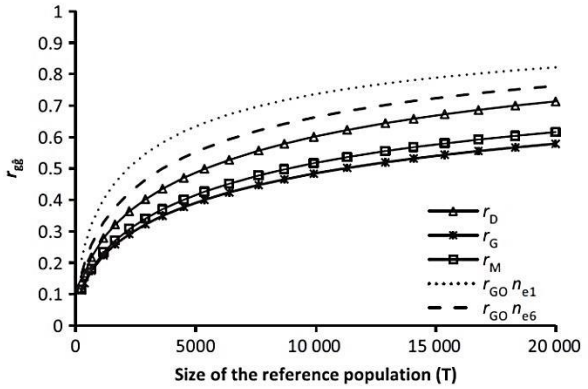


Figure 1 Marginal distribution of accuracy as a function of the size of the reference population *T*.

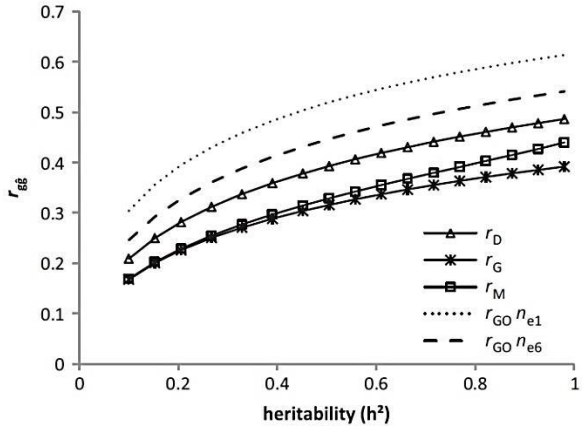


Figure 2 Marginal distribution of accuracy as a function of heritability h^2.

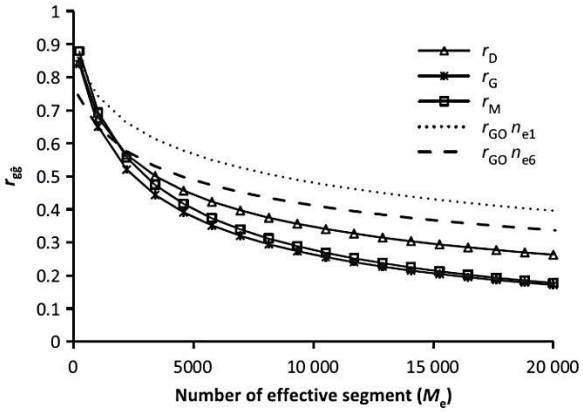


Figure 3 Marginal distribution of accuracy as a function of the number of effective segments M_e.

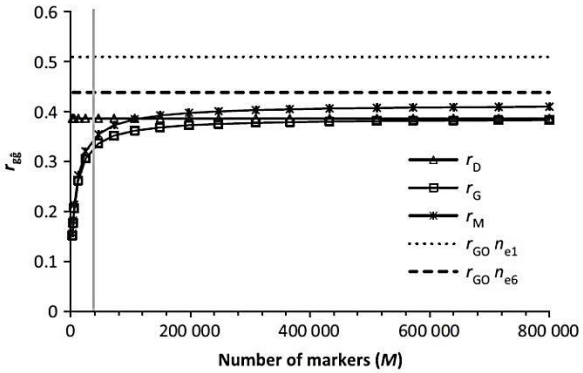


Figure 4 Marginal distribution of accuracy as a function of the number of markers *M*.

(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.

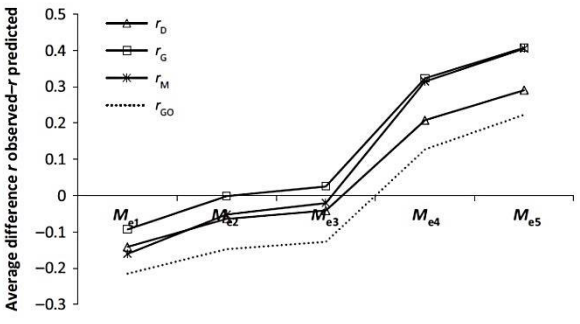


Figure 5 Average differences between the observed and predicted accuracies depending on the formulae for accuracy r and for the number of effective segments M_e.

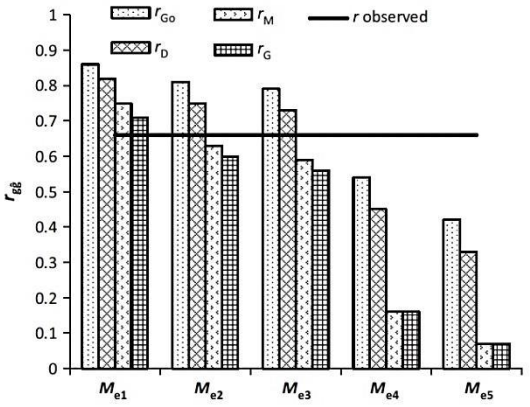


Figure 6 Application of the formulae predicting accuracy using different methods to calculate M_e, and comparison with the accuracy observed by Meuwissen *et al.* (2001). Parameter values are: $h^2 = 0.5$, $T = 1000$, $M = 1010$, $L = 10\ M$, $N_e = 100$ (simulated data).

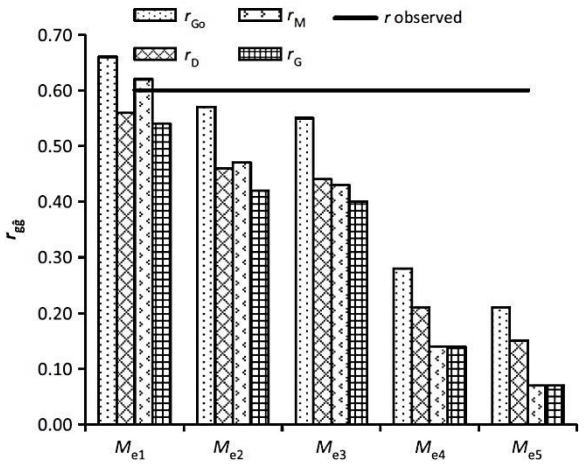


Figure 7 Application of the formulae predicting accuracy using different methods to calculate M_e, and comparison with the accuracy observed by Luan *et al.* (2009). Parameter values: h^2 = CD = 0.97, T = 500, M = 18 991, N_e = 167 (real data).

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

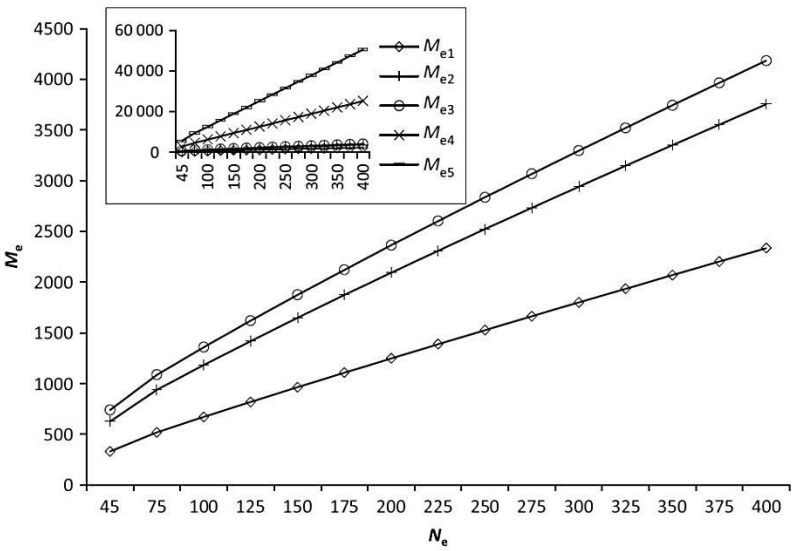


Figure 8 Variation of M_e depending on N_e according to the formulae used for computation.

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.

4. Existe-t-il des gènes à effet majeur pour l'aptitude à la performance en concours de saut d'obstacle et au concours complet d'équitation ?

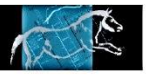
Nous avons vu dans le chapitre bibliographique sur la précision de la sélection génomique que l'architecture génétique d'un caractère peut être prise en compte pour améliorer l'estimation des valeurs génétiques correspondantes. En CSO et CCE, la présence de gènes à effets majeurs chez les chevaux de sport français n'avait pas été vérifiée. Nous avons donc réalisé des détections de QTL pour ces caractères. L'article inclus dans ce chapitre présente les résultats obtenus pour le CSO. Après une introduction de cet article, puis un complément sur les résultats obtenus pour le CSO, nous présenterons les résultats obtenus pour le CCE.

4.1. Analyse d'association pour l'aptitude à la performance en CSO

4.1.1. Introduction de l'article

Pour le CSO, nous avons utilisé les génotypes de 999 chevaux de sport (une partie des 1 101 chevaux génotypés ont été écartés car le génotypage était incomplet ou le CD de leur BSO n'était pas assez élevé). L'échantillon était composé de 68% de Selles Français et de 13% d'Anglo-Arabes, et les autres chevaux étaient des chevaux de sport étrangers. La majorité de ces chevaux étaient des étalons, et ainsi la quasi-totalité des étalons reproducteurs en Selle Français en activité au moment de l'étude étaient présents dans l'échantillon. Les performances utilisées en CSO étaient des pseudo-phénotypes, obtenus en dérégressant l'indice génétique présenté dans la partie bibliographique sur le cheval de sport. Cette approche était intéressante dans la mesure où notre échantillon était constitué quasiment exclusivement d'étalons : des chevaux mis à la reproduction et donc bien connus sur leurs performances propres et sur celles de leurs descendants (le CD moyen était de 0.73). La déregression des BSO a donc permis de tenir compte pour chaque cheval de toute l'information disponible, à l'exception des informations apportées par des individus apparentés génotypés également. En plus de la quantité d'information prise en compte, ces pseudo-phénotypes avaient aussi l'avantage d'être déjà corrigés pour les effets fixes et les effets de harem. La puce utilisée (Illumina Equine SNP50 BeadShip) comptait 54 602 SNPs. 44 424 SNPs ont été retenus sur les critères suivants : respect de l'équilibre d'Hardy-Weinberg, fréquence de l'allèle minimum supérieure à 5%, taux de génotypes manquants inférieur à 20%. Nous avons utilisé deux modèles pour détecter les SNPs : un modèle mixte uni-SNP, et un modèle mixte haplotypique. Le modèle mixte uni-SNP teste l'effet de chaque SNP sur le pseudo-phénotype tour à tour. Dans le modèle mixte haplotypique, les SNPs sont remplacés par des haplotypes. Les haplotypes ont brièvement été décrits dans le chapitre bibliographique sur la sélection génomique. Il s'agit d'une reconstitution statistique de groupes de SNPs adjacents qui ségrégent ensemble au cours de la méiose. On suppose qu'un QTL en déséquilibre de liaison imparfait avec un SNP sera mieux capturé par un haplotype, s'il est inclus dans cet haplotype par exemple. Dans les deux cas, uni-SNP et haplotypique, le modèle comportait un effet polygénique afin de tenir compte de la structure de la population. La méthodologie et les résultats obtenus pour la détection de QTL pour l'aptitude au CSO sont présentés de façon plus détaillée dans l'article suivant et dans ses annexes.

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics



SHORT COMMUNICATION

doi: 10.1111/age.12245

Genome-wide association study for jumping performances in French sport horses

S. Brard[*†‡] **and A. Ricard**[§¶]

*INRA, GenPhySE (Génétique Physiologie et Systèmes d'Elevage), F-31326 Castanet-Tolosan, France. †INP, ENSAT, GenPhySE (Génétique Physiologie et Systèmes d'Elevage), Université de Toulouse, F-31326 Castanet-Tolosan, France. ‡INP, ENVT, GenPhySE (Génétique Physiologie et Systèmes d'Elevage), Université de Toulouse, F-31076 Toulouse, France. §INRA, UMR 1313, 78352 Jouy-en-Josas, France. ¶IFCE, Recherche et Innovation, 61310 Exmes, France.

Summary

A genome-wide association study was performed to identify single nucleotide polymorphisms (SNPs) associated with jumping performances of warmbloods in France. The 999 horses included in the study for jumping performances were sport horses [mostly Selle Français (68%), Anglo-Arabians (13%) and horses from the other European studbooks]. Horses were genotyped using the Illumina EquineSNP50 BeadChip. Of the 54 602 SNPs available on this chip, 44 424 were retained after quality testing. Phenotypes were obtained by deregressing official breeding values for jumping competitions to use all available information, that is, the performances of each horse as well as those of its relatives. Two models were used to test the effects of the genotypes on deregressed phenotypes: a single-marker mixed model and a haplotype-based mixed model (significant: $P < 1E\text{-}05$; suggestive: $P < 1E\text{-}04$). Both models included a polygenic effect to take into account familial structures. For jumping performances, one suggestive quantitative trait locus (QTL) located on chromosome 1 (*BIEC2_31196* and *BIEC2_31198*) was detected with both models. This QTL explains 0.7% of the phenotypic variance. *RYR2*, a gene encoding a major calcium channel in cardiac muscle in humans and mice, is located 0.55 Mb from this potential QTL.

Keywords equine, jumping, quantitative trait loci, whole-genome association study

Nowadays, warmblood horses are bred mainly to produce high level sport horses (Koenen *et al.* 2004). Over the last few years, the horse sequence genome has become available (Wade *et al.* 2009), and genome-wide association studies (GWASs) have been performed in horses for diseases such as osteochondrosis (Corbin *et al.* 2012; Teyssèdre *et al.* 2012a), laryngeal neuropathy (Dupuis *et al.* 2011), hypersensitivity to insect bites (Schurink *et al.* 2012), as well as for morphological traits (Signer-Hasler *et al.* 2012) and reproduction traits (Sieme & Distl 2012). Targets of selection by comparing more than 30 breeds were also found (Petersen *et al.* 2013). Proven mutations were found for the equine *myostatin* gene related to optimum racing distances in thoroughbred horses (Hill *et al.* 2010) and in *DMRT3*, which controls gaits such as pacing and trotting (Andersson *et al.* 2012). For competition-related traits, Schröeder *et al.* (2011b) performed an analysis in Hanoverian horses and identified six QTL for jumping. The objective of this study was to perform a GWAS for jumping performances in France.

For the most part, the population genotyped in this study had already been used for genomic selection testing (Ricard *et al.* 2013). After quality controls, 999 horses were retained. They were mainly sires of jumping competitors (84%) or own performance stallions (14%) during 2009–2013. Of these, 68% were Selle Français (SF), 13% Anglo-Arabians (AA), 4% Koninklijke Vereniging Warmbloed Paardenstamboek Nederland (KWPN), 3% Belgisch Warmbloed Paard (BWP), 3% Holsteiner Warmblut (HOLST) and the remaining 9% from other European studbooks.

Phenotypes were calculated by deregressing the official estimated breeding values (EBVs). EBVs are based on results in official competition. Two criteria are used to measure the performance: the logarithm of annual sum of earnings or points and the rank in each event. Details about EBVs can be found in Ricard (1997) or in Appendix S1. The

Address for correspondence

S. Brard, INRA-GenPhySE, Auzeville BP52627, 31326 Castanet Tolosan Cedex, France.
E-mail: sophie.brard@toulouse.inra.fr

Accepted for publication 26 September 2014

Table 1 Characteristics of the single nucleotide polymorphisms (SNPs) detected for jumping performances from deregressed breeding values (DEBVs) using either a single-SNP mixed model or a haplotype-based mixed model with the genotypes of 999 sport horses (44 424 SNPs).

					Single-SNP mixed model		Haplotype-based mixed model	
SNP	ECA[1]	Pos, Mbp[2]	Alleles	MAF[3]	$-\log_{10}(P)$[4]	Add (SE)[5]	$-\log_{10}(P)$[6]	% of variance explained
BIEC2_31196	1	73.35	A/G	0.12	**4.27**	0.166 (0.043)	**4.16**	0.73
BIEC2_31198	1	73.35	G/A	0.12	**4.27**	0.166 (0.043)	3.71	0.61
BIEC2_864208	4	47.66	G/A	0.10	0.02	−0.003 (0.045)	**4.17**	0.33
BIEC2_864210	4	47.66	A/G	0.10	0.00	0.001 (0.045)	**4.08**	0.31
BIEC2_158739	11	51.49	G/A	0.25	0.04	−0.003 (0.030)	**4.06**	0.31
BIEC2_158740	11	51.49	A/G	0.25	0.04	−0.003 (0.030)	**4.21**	0.42
BIEC2_159643	11	53.08	A/G	0.08	1.13	−0.080 (0.046)	**4.17**	0.43

[1]*Equus Caballus* chromosome.
[2]Position on the genome in Megabase pairs.
[3]Minor allele frequency.
[4]–log10(*P*-value) obtained.
[5]Additive effect of the SNP on DEBVs (SE: Standard error) in phenotypic standard deviations.
[6]–log10(*P*-value) that reached the suggestive threshold are in bold.

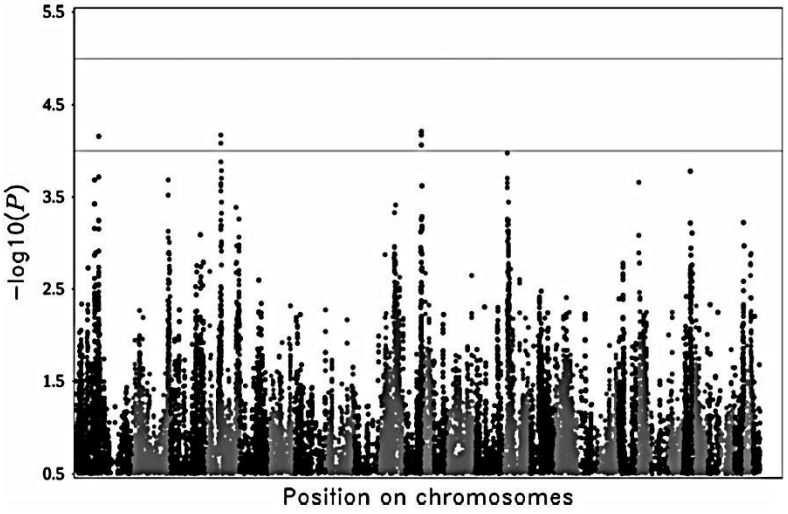


Figure 1 Manhattan plot for association analysis for jumping performances from deregressed breeding values using a haplotype-based mixed model with the genotypes of 999 warmbloods (44 424 SNPs). Alternating colors mark the chromosome limits. Horizontal lines represent suggestive and significance thresholds (Bonferroni correction considering 5000 exclusive tests).

deregression process is described in detail in Ricard *et al.* (2013). These deregressed EBVs (DEBVs) were used as pseudo-phenotypes for the GWAS (Table S1).

Genotyping was performed at Labogena (Jouy-en-Josas, France) with the Illumina Equine SNP50 BeadChip. After quality tests were applied to the 54 602 available SNPs (minor allele frequency >5%, call frequency >80%, *P*-value for the Hardy–Weinberg disequilibrium test >1E-08), 44 424 SNPs were retained for the study.

Two methods were used to perform the GWAS. The first one was a single-marker association study and the second a haplotype analysis. For single-marker analysis, the following mixed model was used:

$$y = xb + Za + e, \quad (1)$$

where y is the vector of DEBVs; b is the coefficient of regression for each SNP; x is the incidence vector for b with the genotypes of all individuals (coded 0, 1 or 2 depending on the number of reference alleles in their genotypes); and a is the vector of random polygenic effects, $V(a) = A\sigma_a^2$, where A is the relationship matrix based on pedigree. Z is an incidence matrix and e is the vector of random residual effects, with $V(e) = D\sigma_e^2$, where D is a diagonal matrix with coefficients $d_{jj} = 1/w_j$, where w_{ii} are the weights measuring the amount of information brought by the DEBVs. The best linear unbiased estimates of b and F-tests were obtained using ASREML software (Gilmour *et al.* 2006). Association analysis was also performed using a model with haplotypes. Haplotypes were obtained using PHASEBOOK (Druet & Georges 2010), a package that runs successive programs to obtain phased haplotypes in a highly interrelated population (Appendix S2), and BEAGLE (Browning & Browning 2007). Haplotypes are based on hidden states given by the Markov model, which are introduced directly in the model. This methodology has already been used by Dupuis *et al.* (2011). In that model, the two hidden states of the two chromosomes replaced the SNP effect in model (1). REMLF90 (Misztal *et al.* 2002) software was used. We chose a significance

threshold of 1E-05 and a suggestive one of 1E-04, according to Teyssèdre *et al.* (2012a).

Q–Q plots were used to compare the theoretical *P*-values to the observed *P*-values (Fig. S1). The distribution of the observed *P*-values was the expected one. No SNPs reached the significance threshold using the single-SNP mixed model (Fig. S2). A suggestive association was found for two SNPs located on chromosome 1 ($P = 5.4E-05$ for both markers *BIEC2_31196* and *BIEC2_31198*) (Table 1). Linkage disequilibrium between these markers is equal to one. Using the haplotype-based mixed model, SNP *BIEC2_31196* exceeded the suggestive threshold ($P = 7.0E-05$). Two suggestive SNPs were detected on chromosome 4, and three suggestive SNPs were detected on chromosome 11 (Fig. 1; Table 1). SNP *BIEC2_31196* was the only SNP that either exceeded or tended toward the thresholds in both models. Details about this SNP and haplotypes estimated effects are summarized in Fig. S3. The SNP and the hidden state with the greatest effect are the rarest.

The power of this protocol was calculated according to Teyssèdre *et al.* (2012b). For a QTL in complete linkage disequilibrium with a SNP in the data, the power was 77% for a QTL, explaining 3% of the variance and assuming a type-1 error of 1.0E-05. Hence, if a QTL explaining a large percentage of the variance (>3%) had existed, we should have been able to localize it. We therefore conclude that no major gene exists for jumping performances in sport horses in France. Nevertheless, we have localized some QTL of moderate effect, not exceeding 0.7% of phenotypic variance. Therefore, the impact on any selection scheme will be low.

Quantitative trail loci affecting sprinting ability, gait or height have been reported previously (Hill *et al.* 2010; Andersson *et al.* 2012; Signer-Hasler *et al.* 2012). Due to possible correlations between these traits and jumping (Ricard 2004; Ducro *et al.* 2007), we checked their *P*-values (or closest SNPs). None of them reached even the suggestive threshold; the lowest was 0.04 for a SNP detected for height. Whatever the correlation between these traits and jumping in competition, these QTL have no effect on the performances we studied.

A GWAS for a similar trait was reported recently by Schröeder *et al.* (2011b). The phenotypes used in that study were the breeding values for style and ability of free jumping scored by judging commissions during mares' performance tests or during inspections of males and females before auctions. QTL were detected on chromosomes 1, 8, 9 and 26, and nearby potential candidate genes were identified. None of these QTL was detected in our study. In spite of the correlation with performances (Luhrs-Behnke *et al.* 2006a,b), it seems that the phenotypes of both studies would not depend on the same QTL, but they used only 115 sires. In another publication, Schröeder *et al.* (2011a) stated that 0.05% of the genes on chromosome 1 could be candidate genes for performance. We therefore investigated potential candidate genes close to the QTL we detected for jumping performances on chromosome 1. Using the horse genome assembly EquCab2.0, three potential candidate genes were identified. The gene closest to the QTL is *ryanodine receptor 2 (cardiac)* (*RYR2*) at 0.55 Mb of the SNP *BIEC2_31196*. This gene encodes a ryanodine receptor found in cardiac muscle which is part of the calcium channel. Wehrens *et al.* (2003) showed that, in mice, the phosphorylation of the protein encoded by this gene increases intracellular calcium release and cardiac contractility, which causes cardiac ventricular arrhythmia during exercise and may lead to sudden cardiac death. SNP *BIEC2_31196* is also close to two other genes coding for components of the skeletal and cardiac muscle contractile apparatus, *actinin, alpha 2* (*ACTN2*) and *Actin, alpha 1, skeletal muscle* (*ACTA1*), which are at 1.49 Mb and 4.94 Mb from *BIEC2_31196* respectively.

Acknowledgements

We would like to thank the owners of the genotyped horses and the Fédération Nationale du Cheval, the Association Nationale du Selle Français and the Association Nationale de l'Anglo-Arabe, who supported the project. We would also like to thank the staff at Labogena where analyses were performed. This project was funded by the Institut Français du Cheval et de l'Equitation, the Institut National de la Recherche Agronomique and the Fond Eperon.

Conflict of interest

The authors declare no conflict of interest.

References

Andersson L.S., Larhammar M., Memic F. *et al.* (2012) Mutations in *DMRT3* affect locomotion in horses and spinal circuit in mice. *Nature* **488**, 642–6.

Browning B.L. & Browning S.R. (2007) Efficient multilocus association mapping for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology* **31**, 365–75.

Corbin L.J., Blott S.C., Swinburne J.E. *et al.* (2012) A genome-wide association study of osteochondritis dissecans in the Thoroughbred. *Mammalian Genome* **23**, 294–303.

Druet T. & Georges M. (2010) A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* **184**, 789–98.

Ducro B.J., Koenen E.P.C., van Tartwijk J.M.F.M. & Bovenhuis H. (2007) Genetic relations of movement and free-jumping traits with dressage and show-jumping performance in competition of Dutch Warmblood horses. *Livestock Science* **107**, 227–34.

Dupuis M.C., Zhang Z., Druet T., Denoix J.M., Charlier C., Lekeux P. & Georges M. (2011) Results of a haplotype-based GWAS for recurrent laryngeal neuropathy in the horse. *Mammalian Genome* **22**, 613–20.

Gilmour A.R., Gojel B.J., Cullis B.R. & Thompson R. (2006) *ASREML User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, UK.

Hill E.W., Gu J., Eivers S.S., Fonseca R.G., McGivney B.A., Govindarajan P., Orr N., Katz L.M. & MacHugh D. (2010) A sequence polymorphism in *MSTN* predicts sprinting ability and racing stamina in Thoroughbred horses. *PLoS One* **5**, e8645.

Koenen E.P.C., Aldridge L.I. & Philipsson J. (2004) An overview of breeding objectives for warmblood sport horses. *Livestock Production Science* **88**, 77–84.

Luhrs-Behnke H., Rohe R. & Kalm E. (2006a) Estimation of genetic parameters for tournament sports performances within examination classes and their relations to traits from mare and stallion performance tests. *Zuchtungskunde* **78**, 173–83.

Luhrs-Behnke H., Rohe R. & Kalm E. (2006b) Genetic analyses of riding test and their connections with traits of stallion performance and breeding mare tests. *Zuchtungskunde* **78**, 119–28.

Misztal I., Tsuruta S., Strabel T., Auvray B., Druet T. & Lee D.H. (2002) BLUPF90 and related programs (BGF90). In: *Proceedings of the 7th World Congress of Genetics Applied to Livestock Production*, Montpellier, France.

Petersen J.L., Mickelson J.R., Rendahl A.K. *et al.* (2013) Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genetics* **9**, e1003211.

Ricard A. (1997) Breeding evaluations and breeding programs in France. In: *48th Annual Meeting of EAAP*, Vienna, Austria.

Ricard A. (2004) Heritability of jumping ability and height of pony breeds in France. *Livestock Production Science* **89**, 243–51.

Ricard A., Danvy S. & Legarra A. (2013) Computation of deregressed proofs for genomic selection when own phenotypes exist with an application in French show-jumping horses. *Journal of Animal Science* **91**, 1076–85.

Schröeder W., Klostermann A. & Distl O. (2011a) Candidate genes for physical performance in the horse. *The Veterinary Journal* **190**, 39–48.

Schröeder W., Klostermann A., Stock K.F. & Distl O. (2011b) A genome wide association study of quantitative trait loci of show-jumping in Hanoverian warmblood horses. *Animal Genetics* **43**, 392–400.

Schurink A., Wolc A., Ducro B.J., Frankena K., Garrick D.J., Dekkers J.C.M. & Arendonk van J.A.M. (2012) Genome-wide association study of insect bite hypersensitivity in two horse populations in the Netherlands. *Genetics Selection Evolution* **44**, 31.

Sieme H. & Distl O. (2012) Genomics and fertility in stallions. *Journal of Equine Veterinary Science* **32**, 467–70.

Signer-Hasler H., Flury C., Haase B., Burger D., Simianer H., Leeb T. & Rieder S. (2012) A genome-wide association study reveals loci influencing height and other conformation traits in horses. *PLoS One* **7**, e37282.

Teyssèdre S., Dupuis M.C., Guérin G., Schibler L., Denoix J.M., Elsen J.M. & Ricard A. (2012a) Genome-wide association studies for osteochondrosis in French Trotter horses. *Journal of Animal Science* **90**, 45–53.

Teyssèdre S., Elsen J.M. & Ricard A. (2012b) Statistical distributions of test statistics used for quantitative trait association mapping in structured populations. *Genetics Selection Evolution* **44**, 32.

Wade C.M., Giulotto E., Sigurdsson S. *et al.*, Broad Institute Genome Sequencing Platform, Broad Institute Whole Genome Assembly Team, Lander E.S., Lindblad-Toh K. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865–7.

Wehrens X.H.T., Lehnart S.E., Huang F. *et al.* (2003) FKBP12.6 deficiency and defective calcium release channel (ryanodine receptor) function linked to exercise-indiced sudden cardiac death. *Cell* **113**, 829–40.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Details about official breeding values.

Appendix S2 Details about phases.

Figure S1 Quantile–quantile plot for association analysis for jumping performances from deregressed breeding values using a single-SNP mixed model (a) or a haplotype-based mixed model (b) with the genotypes of 999 warmbloods (44 424 SNPs).

Figure S2 Manhattan plot for association analysis for jumping performances from deregressed breeding values (DEBVs) using a single-SNP mixed model with the genotypes of 999 sport horses (44 424 SNPs).

Figure S3 (a) Effect and number of copies of hidden states in the sample for SNP BIEC2_31196; (b) Distribution of estimated breeding values (EBVs) depending on the genotype at SNP BIEC2_31196.

Table S1 Descriptive statistics of estimated breeding values (EBVs) and their reliabilities, and deregressed EBVs (DEBVs) and their weights for jumping performances.

Supporting information S1. Details about official breeding values.

In one competition, according to the rules (number of faults or time), the horses are ranked. All breeds compete together without distinction. According to this rank and to the technical difficulty of the event, earnings (before 2009) and points (after 2009) are distributed in an exponential way: the second earns 25% of the earning of the first, the third earns 25% of the earning of the second…Official EBVs are based on a bivariate animal models with two traits: Log(annual earnings) and ranks in each competition.

The model used for Log(annual earnings) is:

$y_1=X_1b_1+Z_1u_1+W_1c_1+M_1m_1+e_1$

$\mathbf{b}_1$ is a vector of fixed effects including sex with two classes (females, and male and geldings in the same one as geldings are not always correctly declared) and combination of age class and year. 5 classes of age were considered: 4, 5, 6-7, 8-9-10, and 11 and older, and years from 1974 to 2012. $\mathbf{u}_1$ is a vector of additive genetic value for annual earnings. $\mathbf{c}_1$ is a vector of permanent environmental effect common to the different annual performances of the same horse,$\mathbf{m}_1$ is a vector of herd-maternal environmental effect common to the performances of the different descent of one mare. The maternal effect is partly confounded with herd effect since most of the breeders have registered only one mare (73% in 2013), and so no relationship matrix is added to the variance covariance matrix of this effect. Z_1, W_1, M_1 are design matrices.

For the rank, an underlying model is assumed with an unobservable liability which ranking explains the observable rank. The model used is:

$y_2=X_2b_2+Z_2u_2+W_2c_2+M_2m_2+e_2$

$\mathbf{b}_2$ is a vector of fixed effects including sex with two classes (as previously) and age with 9 classes: 4 to 10 by 1, 11-12, and 13 and older. The year effect is not estimable because all performances measured in the same event have the same level of year effect, and only contrasts inside one event are estimable. $\mathbf{u}_2$ is a vector of additive genetic value for underlying performance. $\mathbf{c}_2$ is a vector of permanent environmental effect common to the different performances of the same horse in each event. $\mathbf{m}_2$ is a vector of herd-maternal environmental effect common to the performances of the different descent of one mare. Z_2, W_2, M_2 are design matrices.

Variances covariances matrices between the two traits are built according to the following parameters and with a relationship matrix for the genetic values. For jumping, heritability is 0.27 for Log(annual earnings/points) and 0.16 for underlying performance responsible for ranks in every events. Repeatability is 0.47 (between years) and 0.29 (between events) respectively. Herd-maternal effects represent 5% and 3% of phenotypic variance respectively. The genetic correlation between the two traits analyzed is 0.90, as well as the correlation between the permanent environmental effects and between herd-maternal environmental effects of both traits. In these models, nor breed effect nor genetic groups are added.

Finally, the official EBV is the mean of the EBVs of the two traits with a weight of 0.75 for the logarithm of annual earnings and of 0.25 for the ranking. EBVs are standardized using a reference population in which the mean EBV is set to zero. For jumping performances, all Selle Français and

Anglo-Arabians born five years before the calculation of EBV are included in the reference population. So as the EBVs used were published in 2012, for jumping the reference population includes all Selle Français and Anglo-Arabians that were born in 2007.

Supporting information Table S2. Descriptive statistics of estimated breeding values (EBVs) and their reliabilities, and deregressed EBVs (DEBVs) and their weights for jumping performances.

	Mean	Standard deviation	Minimum	Maximum
EBV[1]	0.73	0.31	-0.68	1.5
Reliability	0.73	0.16	0.34	0.99
DEBV[1]	0.96	0.56	-0.93	4.1
Weight[2]	21	38	1.1	368

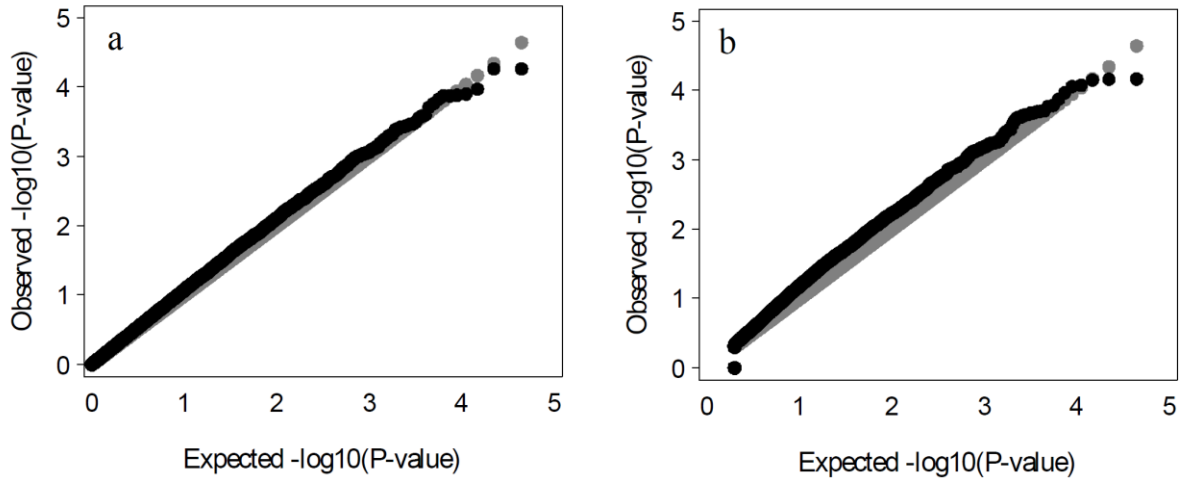
[1] in phenotypic standard deviations

[2] weights are the inverse of residual variance of the DEBVs

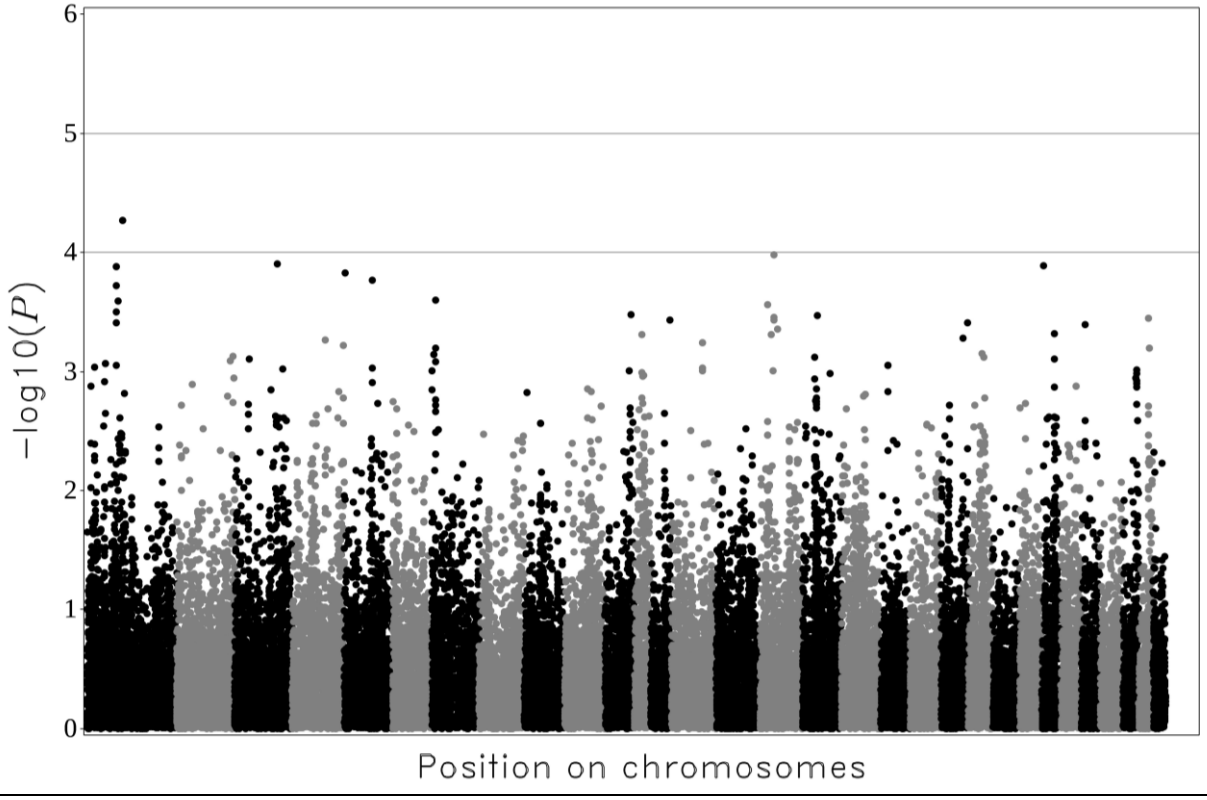
Supporting information S3. Details about phases.

First step was to reconstruct the haplotypes using the pedigree information: LinkPHASE used the Mendelian segregation rules to determine at each heterozygous marker which allele was from the sire or from the dam. Then, DAGPHASE was run to randomly attribute missing alleles. Next these remaining alleles were really attributed with BEAGLE: the linkage information was used in a Hidden Markov Model to infer the origin of alleles. BEAGLE constructed an optimal directed acyclic graph (DAGs) and then the haplotypes from these DAGs were sampled with DAGPHASE. BEAGLE and DAGPHASE were used iteratively to improve the haplotypes. The last output produced by BEAGLE consisted of the haplotypes but also the hidden states used to construct them. These hidden states for each SNP were used for association analysis. Hidden states are clusters of haplotypes, so very similar haplotypes were considered as the same ones (sharing the same allele of QTL), avoiding the difficulty to estimate very rare haplotype effects.

Supporting Information Figure S4. Quantile–quantile plot for association analysis for jumping performances from deregressed breeding values using a single-SNP mixed model (a) or a haplotype-based mixed model (b) with the genotypes of 999 warmbloods (44424 SNPs). Grey dots: expected –$\log_{10}(P)$; black dots: observed –$\log_{10}(P)$.



Supporting Information Figure S5. Manhattan plot for association analysis for jumping performances from deregressed breeding values (DEBVs) using a single-SNP mixed model with the genotypes of 999 warmbloods (44 424 SNPs). Alternating colors mark the chromosome limits. Horizontal lines represent the suggestive and significance thresholds (Bonferroni correction considering 5000 exclusive tests).



Supporting information Figures S6a and b.

Figure S6a. Effect and number of copies of hidden states in the sample for SNP *BIEC2_31196*.

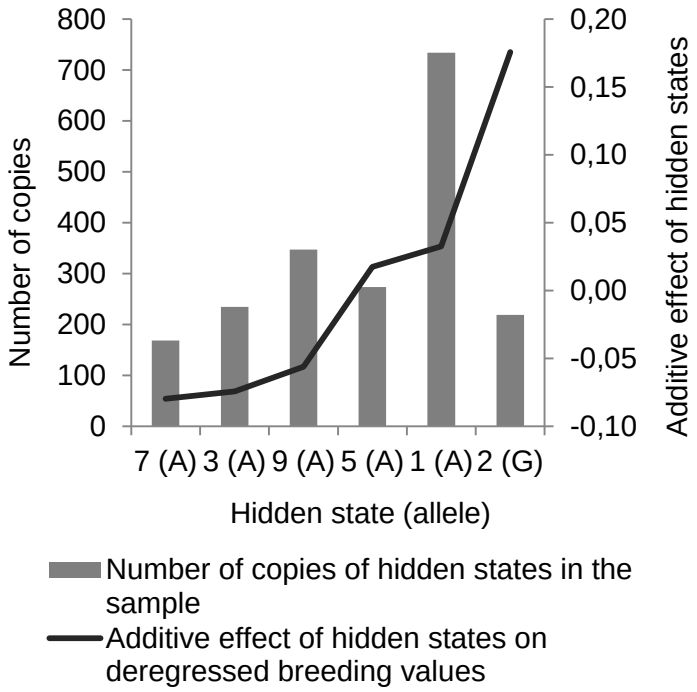
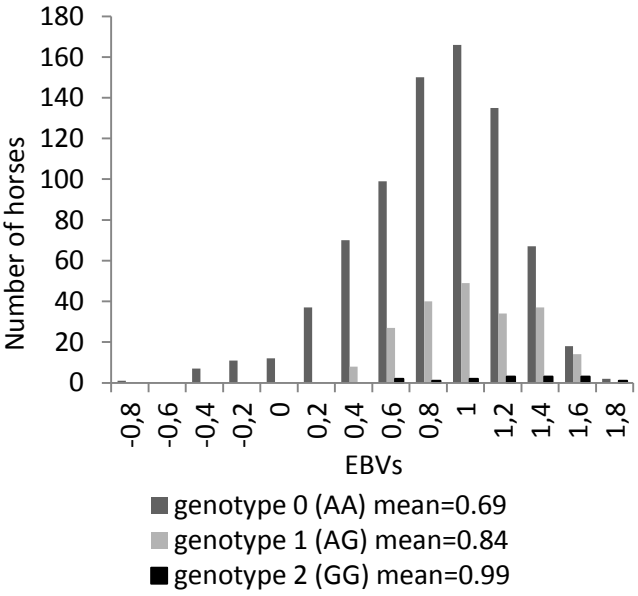


Figure S6b. Distribution of estimated breeding values (EBVs) depending on the genotype at SNP *BIEC2_31196*.



4.1.2. Bilan partiel pour l'aptitude à la performance en CSO basé sur les résultats de l'article

La détection de QTL pour le CSO n'a pas permis de mettre en évidence un gène ayant un effet majeur sur les performances, dans la mesure où aucun SNP n'a dépassé le seuil de significativité fixé (10^{E}-05). La performance en CSO est donc un caractère complexe polygénique. Cependant, un QTL potentiel a été identifié: le SNP *BIEC2-31196*. Ce SNP a atteint le seuil de tendance (10^{E}-04) dans le modèle mixte uni-SNP (*P-value* = 5.4^{E}-05) et dans le modèle mixte haplotypique (*P-value* = 7.0^{E}-05). L'échantillon était composé en majorité d'étalons, qui sont des chevaux sélectionnés pour la reproduction sur la base de leurs bonnes performances. Il est donc possible que les *P-values* n'atteignent pas le seuil de significativité car la variabilité effective pour l'aptitude au saut d'obstacles n'est pas entièrement représentée dans l'échantillon utilisé pour la détection de QTL.

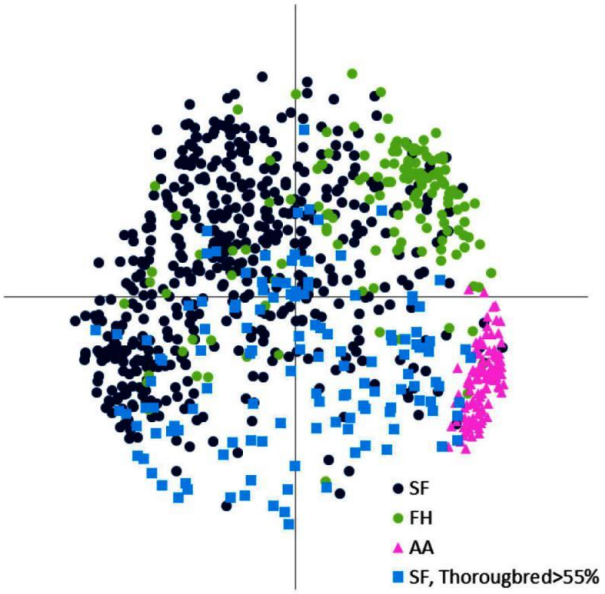
Sa longueur étant limitée à 4 pages, l'article ne présente que des résultats obtenus sur la totalité de l'échantillon. En réalité, nous avons réalisé une analyse d'association supplémentaire sur un sous-échantillon de chevaux afin de montrer que le QTL potentiel détecté n'était pas le marqueur de différences entre les races présentes dans l'échantillon, mais était bien associé à la performance en CSO pour toutes les races confondues. Cette contrainte est en grande partie la conséquence des délais de réponse des relecteurs inhabituellement longs, partiellement dus au retrait du premier éditeur associé pour raisons personnelles (7 mois pour le 1^{er} avis et 3 mois pour le $2^{ème}$, avant d'arriver à un rythme d'échange normal) : le sujet de l'article initialement considéré comme « fitt(ing) very well with Animal Genetics » s'est retrouvé noyé dans une arrivée massive d'articles portant sur des détections de QTL dans d'autres espèces. On nous a finalement demandé de transformer l'article original d'une dizaine de pages en communication courte, malgré notre utilisation peu commune des haplotypes pour l'analyse d'association. Les résultats de l'analyse d'association sans Anglo-Arabes sont donc présentés ici en complément.

4.1.3. Complément : résultats obtenus avec un échantillon sans Anglo-Arabes

Pourquoi exclure les Anglo-Arabes de l'échantillon ?

En complément des analyses présentées dans l'article, nous avons réalisé la détection de QTL dans un échantillon sans Anglo-Arabes. Ce choix a été fait en conséquence des résultats de Ricard *et al*. (2013) qui ont réalisé une analyse en composante principale sur la matrice d'apparentement génomique de l'échantillon. Ils ont mis en évidence l'existence de deux sous-populations : une sous-population incluant les Anglo-Arabes, et une autre sous-population incluant tous les autres chevaux (Figure 4.1). Le fait que les Selle Français et les chevaux de sport étrangers forment une seule sous-population est dû à l'ouverture aux autres races du stud-book Selle Français. Comme décrit dans le chapitre bibliographique consacré aux chevaux, des chevaux de sport étrangers peuvent être admis à la reproduction dans ce stud-book, et un poulain Selle Français n'est pas nécessairement le produit de deux chevaux inscrits au stud-book Selle Français : 63% des Selles Français concourant en France actuellement sont nés de deux parents Selle Français, et c'est aussi le cas pour les KWPN (53%) ou encore le BWP (24%) (A. Ricard, communication personnelle). Il s'agit d'une situation plutôt originale comparée au stud-book Anglo-Arabe par exemple, et la race Selle Français doit donc être considérée plus comme une appellation que comme une race. Le but de l'analyse réalisée sur l'échantillon sans Anglo-Arabes était de vérifier si le QTL potentiel détecté sur le chromosome 1 était bien lié à la performance quelle que soit la race des chevaux, et non le marqueur de différences entre sous-populations.

Figure 4.1 : Résultats de l'analyse en composantes principales (deux composantes principales) réalisée sur la matrice de relations génomique de 908 chevaux (SF=Selle Français, FH=cheval de sport étranger, AA=Anglo-Arabe), d'après Ricard *et al*. (2013)



Matériel & méthodes

Pour cette analyse les mêmes performances et les mêmes modèles que pour l'échantillon complet ont été utilisés. Une fois les Anglo-Arabes enlevés, l'échantillon comptait 866 chevaux génotypés. Les statistiques sur les BSO et les pseudo-phénotypes de ces chevaux sont présentés dans le Tableau 4.1. On peut constater que les valeurs génétiques sont un peu plus élevées pour cet échantillon que dans l'échantillon incluant les Anglo-Arabes (tableau en annexe 2 de l'article), et que le CD moyen est le même dans les 2 cas.

Tableau 4.1 : Statistiques sur les valeurs génétiques (BSO : BLUP Saut d'Obstacle) et leur précision, et sur les pseudo-phénotypes (BSO dérégressé) et leur poids dans l'échantillon de Selles Français et de chevaux de sport étranger (866 individus génotypés)

	Moyenne	Ecart-type	Minimum	Maximum
BSO[1]	0.81	0.29	-0.71	1.60
CD	0.73	0.16	0.34	0.99
BSO dérégressé[1]	1.10	0.56	-0.97	4.30
Poids[2]	22	40	1.20	368

[1] en écart-type phénotypique
[2] le poids est l'inverse de la variance résiduelle des BSO dérégressés

Résultats

La détection de QTL dans l'échantillon sans Anglo-Arabe a permis de retrouver le QTL potentiel détecté sur le chromosome 1. Les *P-values* obtenues pour le SNP *BICE2-31196* étaient de 3.14^{E}-04

avec le modèle mixte uni-SNP, et de 1.72^{E}-05 avec le modèle mixte haplotypique. Par ailleurs, tous les autres QTL atteignant les seuils dans l'échantillon complet atteignent les seuils également ou s'en approchent dans l'échantillon sans Anglo-Arabes (Tableau 4.2). En plus de ces SNPs, 4 SNPs supplémentaires sont détectés sur le chromosome 16, mais avec le modèle mixte haplotypique seulement.

Tableau 4.2 : Caractéristiques des SNPs détectés pour l'aptitude à la performance en CSO à partir des indices génétiques dérégressés en utilisant soit un modèle mixte uni-SNP, soit un modèle mixte haplotypique avec les génotypes de 866 chevaux Selle Français ou chevaux de sport étrangers (44 424 SNPs).

					Modèle mixte uni-SNP		Modèle mixte haplotypique	
SNP	**ECA[1]**	**Pos, Mbp[2]**	**Allèles**	**MAF[3]**	**-log$_{10}$(*P*)**	**Add (SE)[4]**	**-log$_{10}$(*P*)**	**% de variance expliquée**
BIEC2-31196	1	73.35	A/G	0.12	**4.16**	0.158 (0.045)	**4.43**	0.84
BIEC2-864208	4	47.66	G/A	0.10	**4.17**	-0.011 (0.047	**4.04**	0.36
BIEC2-864210	4	47.66	A/G	0.10	**4.08**	-0.008 (0.047)	3.95	0.34
BIEC2-158739	11	51.49	G/A	0.25	**4.06**	-0.017 (0.032)	3.16	0.27
BIEC2-158740	11	51.49	A/G	0.25	**4.21**	-0.017 (0.032)	3.06	0.33
BIEC2-159643	11	53.08	A/G	0.08	**4.17**	-0.102 (0.049)	3.76	0.47
BIEC2-329707	16	11.47	G/A	0.40	3.60	0.076 (0.030)	**4.14**	0.39
BIEC2-329730	16	11.70	A/C	0.21	3.97	0.063 (0.035)	**4.47**	0.63
BIEC2-330264	16	14.95	A/G	0.45	3.20	0.072 (0.030)	**4.10**	0.41
BIEC2-330266	16	14.96	A/G	0.45	3.20	0.072 (0.030)	**4.10**	0.41

[1] *Equus Caballus* chromosome

[2] Position sur le génome en Méga paires de bases

[3] Fréquence de l'Allèle Minimum

[4] Effet additif du SNP sur le BSO dérégressé (SE : Erreur-type) en écart-type phénotypique

Conclusion de l'analyse d'association pour l'aptitude à la performance en CSO

Le SNP *BIEC2-31196* étant détecté avec et sans les Anglo-Arabes dans l'échantillon, quel que soit le modèle utilisé, sa détection dans l'échantillon complet n'est pas due à une différence entre sous-populations mais est bien liée à la performance en CSO. La figure en annexe numéro 3 de l'article présente la distribution des BSO pour chacun des génotypes au SNP *BICE2-31196*. Nous avons constaté que les Anglo-Arabes ne sont pas porteurs du génotype GG, qui chez les Selle Français et chevaux de sport étrangers est porté par les individus ayant les BSO les plus élevés.

J'ai recherché les gènes d'intérêts proches du SNP détecté qui pourraient être des gènes candidats en utilisant la séquence annotée du génome équin EquCab2.0. Le gène le plus proche est le *ryanodine receptor 2 (cardiac)* (*RYR2*), situé à 0.55mb du SNP *BIEC2-31196*. Ce gène code pour un récepteur à la ryanodine, qui se trouve dans les cellules du muscle cardiaque et qui est impliqué dans la chaine d'échange de calcium. Ce gène est suspecté d'être impliqué dans la mort subite de l'athlète : des expérimentations chez la souris ont montré qu'une phosphorylation de la protéine codée par ce gène augmente le relâchement de calcium extracellulaire et les contractions cardiaques, ce qui cause

durant l'exercice des arythmies ventriculaires pouvant entrainer la mort. Le SNP *BIEC2-31196* peut donc être considéré comme un gène candidat potentiel.

4.2. Détection de QTL pour la performance en CCE

4.2.1. Introduction

Après la première analyse d'association réalisée pour l'aptitude à la performance en CSO, il était possible de réaliser la même étude pour la performance en CCE, car les chevaux indicés en CSO l'étaient aussi pour le CCE. D'un point de vue méthodologique, les calculs à mettre en œuvre pour obtenir les pseudo-phénotypes étaient les mêmes que pour le CSO. Il était possible également d'utiliser un modèle mixte uni-SNP et un modèle mixte haplotypique, en utilisant les SNPs retenus pour l'analyse d'association en CSO. Il n'était pas prévu au cours de ma thèse de tester la sélection génomique sur les chevaux de concours complet. L'intérêt de cette étude résidait dans le fait que le CCE comporte une épreuve de saut d'obstacle. Il était envisageable que le QTL potentiel détecté pour l'aptitude à la performance en CSO soit détecté également dans l'analyse d'association pour le CCE.

4.2.2. Matériel & Méthodes

La taille de l'échantillon était plus restreinte en CCE (289 individus) qu'en CSO car beaucoup de chevaux avaient un CD trop faible (inférieur à 0.13). Dans cet échantillon la proportion de Selle Français était plus faible comparée à l'échantillon utilisé pour le CSO, et la proportion d'Anglo-Arabes était plus importante : il y avait 57% de Selle Français, 26% d'Anglo-Arabes, et 17% de chevaux de sport étrangers.

L'indice génétique (le BCC) a été dérégressé pour obtenir les pseudo-performances. Les caractéristiques de l'indice génétique et des pseudo-performances sont résumées dans le Tableau 4.3.

Tableau 4.3 : Statistiques sur les valeurs génétiques (BCC : BLUP Concours Complet) et leur précision, et sur les pseudo-phénotypes (BCC dérégressé) et leur poids pour l'ensemble des chevaux génotypés (289 individus)

	Moyenne	Ecart-type	Minimum	Maximum
BCC[1]	0.42	0.25	-0.37	1.32
CD	0.54	0.13	0.24	0.85
BCC dérégressé[1]	0.57	0.58	-1.30	2.34
Poids[2]	4.49	3.63	1.12	21.5

[1] en écart-type phénotypique
[2] le poids est l'inverse de la variance résiduelle des BCC dérégressés

Le même modèle mixte uni-SNP qu'en CSO a été utilisé :

$$\boldsymbol{y} = \boldsymbol{x}b + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e},$$

avec $\boldsymbol{y}$ le vecteur des pseudo-phénotypes, b le coefficient de régression de chaque SNP, $\boldsymbol{x}$ un vecteur d'incidence pour b avec les génotypes des chevaux codés 0, 1 ou 2 en fonction du nombre d'allèles de référence portés. $\boldsymbol{a}$ est un effet aléatoire polygénique tel que $V(\boldsymbol{a}) = \boldsymbol{A}\sigma_a^2$ ($\boldsymbol{A}$ étant la matrice d'apparentement basée sur le pédigrée). $\boldsymbol{Z}$ est une matrice d'incidence, et $\boldsymbol{e}$ est un vecteur

contenant les effets aléatoires résiduels tels que $V(\boldsymbol{e}) = \boldsymbol{D}\sigma_e^2$, où $\boldsymbol{D}$ est la matrice diagonale dont les coefficients sont les inverses des poids associés aux pseudo-performances. Le modèle mixte haplotypique s'écrit de la même façon, mais en remplaçant b par les effets des états cachés des haplotypes, et le vecteur $\boldsymbol{x}$ par une matrice $\boldsymbol{X}$ qui contient des 0 et des 1 suivant les états cachés portés par les chevaux (pour un SNP auquel correspondent n états cachés, la ligne correspondant à un cheval compte deux 1 et $n-2$ zéros).

Comme en CSO, un seuil de détection significative a été fixé à 10^{E}-05, ainsi qu'un seuil de tendance à 10^{E}-04 (Teyssèdre *et al*. 2012).

4.2.3. Résultats

Les QQ-plots obtenus avec ces deux modèles sont présentés dans la Figure 4.2. La distribution des *P-values* obtenues correspond bien à la distribution des *P-values* attendues. Pour le modèle mixte haplotypique, la distribution des *P-values* montrait une sous-estimation des effets des SNPs. En conséquence un contrôle génomique a été appliqué afin de retrouver la distribution attendue des *P-values* (Devlin *et al.* 2001, Bacanu *et al.* 2002). Les résultats présentés sur la Figure 4.2b sont ceux avec application du contrôle génomique. Les résultats obtenus avec les modèles mixtes uni-SNP et haplotypique sont présentés dans la Figure 4.3 et la Figure 4.4. Avec le modèle mixte uni-SNP, deux SNPs situés sur les chromosomes 2 et 29 atteignent le seuil de tendance, avec des *P-values* respectives de 5.5^{E}-05 (*BIEC2-506494*) et de 8.2^{E}-05 (*BIEC2-754864*). Avec le modèle mixte haplotypique, deux SNPs localisés sur les chromosomes 9 et 29 sont détectés entre les deux seuils. Il s'agit du SNP *BIEC2-1102825* (*P-value*=7.42^{E}-05) et du SNP *BIEC2-748434* (*P-value*=5.80^{E}-05). Les caractéristiques de ces SNPs sont rassemblées dans le Tableau 4.4.

Figure 4.2 : Quantile-quantile plot de l'analyse d'association réalisée à partir des indices génétiques pour le CCE dérégressés en utilisant un modèle mixte uni-SNP (a) et un modèle mixte haplotypique (b), avec les génotypes de 289 chevaux (44 424 SNPs). Les points en gris représentent les valeurs attendues et les points en noir les valeurs observées.

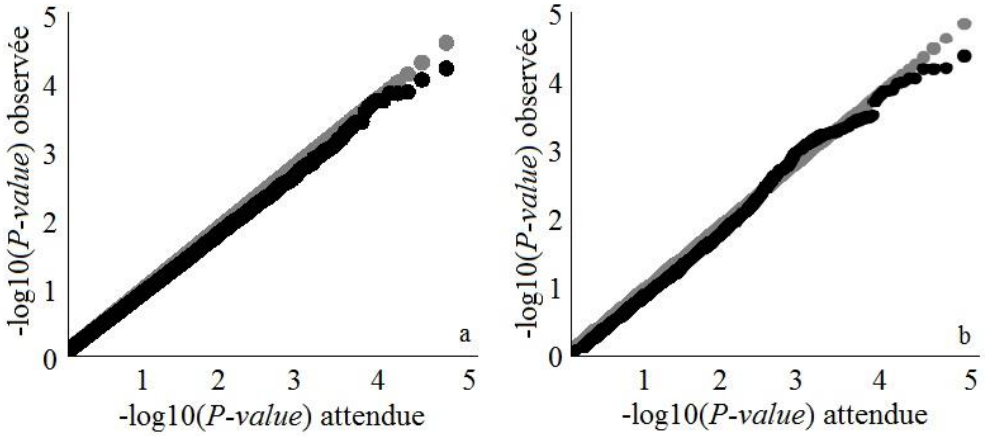


Figure 4.3 : Manhattan plot de l'analyse d'association réalisée à partir des indices génétiques pour le CCE dérégressés utilisés dans un modèle mixte uni-SNP, avec les génotypes de 289 chevaux (44 424 SNPs). L'alternance du noir et du gris marque les différents chromosomes. Les lignes horizontales indiquent les seuils de tendance et de significativité.

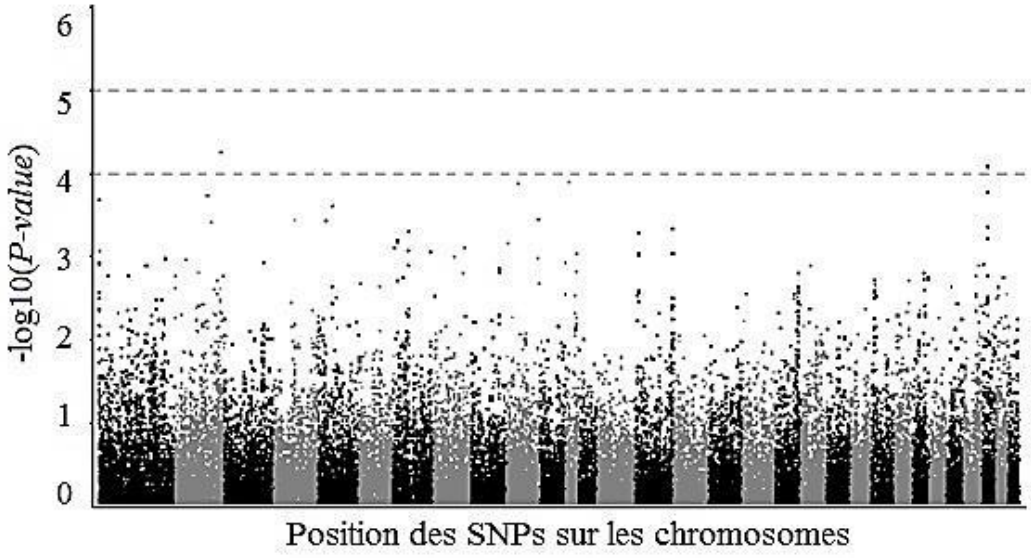


Figure 4.4 : Manhattan plot de l'analyse d'association réalisée à partir des indices génétiques pour le CCE dérégressés utilisés dans un modèle mixte haplotypique, avec les génotypes de 289 chevaux (44 424 SNPs). L'alternance du noir et du gris marque les différents chromosomes. Les lignes horizontales indiquent les seuils de tendance et de significativité.

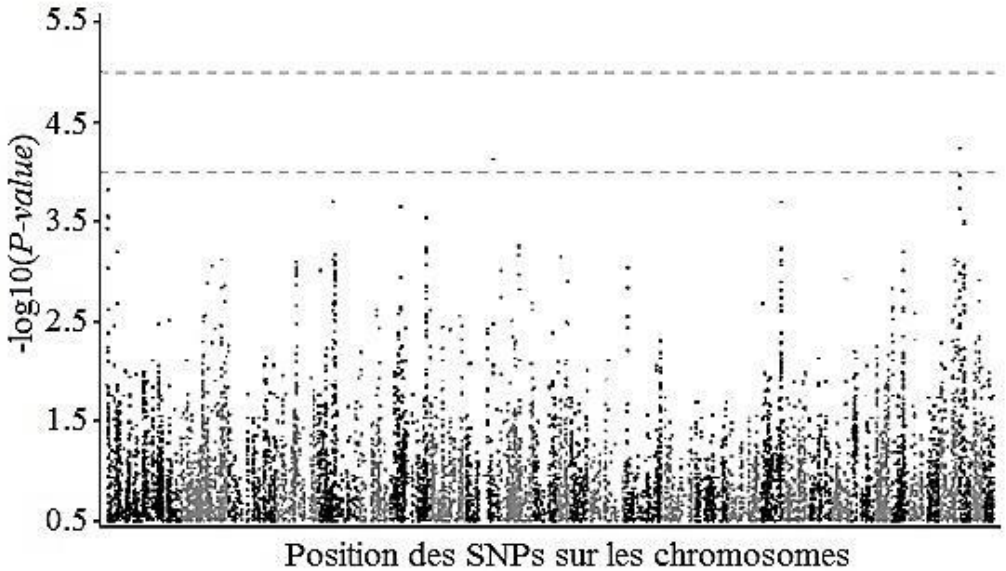


Tableau 4.4 : Caractéristiques des SNPs détectés pour l'aptitude à la performance en CCE à partir des indices génétiques dérégressés utilisés soit dans un modèle mixte uni-SNP, soit dans un modèle mixte haplotypique avec les génotypes de 289 chevaux (44 424 SNPs).

					Modèle mixte uni-SNP		Modèle mixte haplotypique	
SNP	**ECA[1]**	**Pos, Mbp[2]**	**Allèles**	**MAF[3]**	**-log$_{10}$(*P*)**	**Add (SE)[4]**	**-log$_{10}$(*P*)**	**% de variance expliquée**
BIEC2_506494	2	111.64	A/G	0.34	4.26	0.198 (0,048)	2.70	0,93
BIEC2_1102825	9	67.05	A/G	0.36	2.19	0.141 (0,051)	4.13	2,37
BIEC2_748434	29	1.57	G/A	0.32	1.57	0.108 (0,049)	4.23	2,45
BIEC2_754684	29	14.73	A/G	0.20	4.09	0.233 (0,058)	3.07	1,24

[1] *Equus Caballus* chromosome

[2] Position sur le génome en Méga paires de bases

[3] Fréquence de l'Allèle Minimum

[4] Effet additif du SNP sur le BCC dérégressé (SE : Erreur-type) en écart-type phénotypique

4.2.4. Conclusion de l'analyse d'association pour le CCE

Contrairement à l'analyse d'association réalisée pour le CSO, aucun SNP n'a été détecté avec les deux modèles. Un sous-échantillon sans Anglo-Arabes a été utilisé également, mais le trop faible nombre de chevaux (un peu plus de 200) ne permet pas de mettre en évidence des QTL potentiels. Nous concluons donc comme pour le CSO que la performance en CCE est un caractère complexe polygénique. Cependant vu le faible nombre de chevaux utilisés, il est envisageable d'obtenir de meilleurs résultats en utilisant plus de données. Par ailleurs, le QTL potentiel détecté sur le chromosome 1 pour la performance en CSO obtient des *P-values* élevées dans l'analyse d'association pour le CCE, et plus généralement les SNPs détectés pour un type de compétition ne sont pas détectés dans l'autre type. Malgré la présence d'une épreuve de saut d'obstacles dans le CCE, les SNPs impliqués dans la variabilité de ces performances ne sont donc pas les mêmes.

4.3. Conclusion du chapitre

Cette partie de ma thèse a donc permis de vérifier l'architecture génétique de l'aptitude à la performance en CSO. Un QTL potentiel a été détecté sur le chromosome 1. Il a un effet sur les performances en CSO aussi bien chez le Selle Français que chez les Anglo-Arabes, mais la part de la variance additive expliquée est trop faible pour justifier de le prendre en compte séparément des autres SNPs dans le cadre d'une évaluation génomique. Par ailleurs, nous avons montré que ce SNP n'a pas d'effet sur les performances en CCE, malgré la présence d'une épreuve de saut d'obstacles dans ces compétitions.

Le chapitre suivant est consacré à l'étude de l'évaluation génomique pour le CSO chez les chevaux de sport français, avec l'utilisation d'une méthode permettant d'utiliser les performances de chevaux non-génotypés.

5. Le single-step permet-il d'améliorer la précision de l'évaluation génomique pour la performance en CSO ?

5.1. Introduction

Dans le cadre de la production de chevaux de sport, l'utilisation de la sélection génomique pourrait modifier plusieurs paramètres du progrès génétique. La sélection qui est optimale à 5 ans pourrait être réalisée dès 3 ans, réduisant de deux années l'intervalle de génération. Pour cela, il faudrait que l'évaluation génomique permette d'atteindre plut tôt la précision dont on dispose actuellement à 5 ans. La sélection génomique pourrait aussi permettre d'augmenter l'intensité de la sélection, par exemple en génotypant les juments car actuellement la voie femelle est peu sélectionnée, ou encore en génotypant les poulains afin de réaliser une pré-sélection. La séquence du génome équin est disponible depuis 2009 (Wade *et al*. 2009). Un essai de sélection génomique a été réalisé chez les chevaux de sport Français (Ricard *et al*. 2013). L'échantillon de 908 chevaux génotypés comptait en majorité des Selle Français (71%), mais aussi des chevaux de sport étrangers (17%) et des Anglo-Arabes (13%). 95% étaient des étalons. Les performances utilisées ont été obtenues par dérégression des BSO. Cette méthode a permis d'obtenir des pseudo-phénotypes corrigés pour les effets fixes, les effets d'environnement permanent et les effets maternels. Pour un cheval génotypé donné, la pseudo-performance a été calculée en pondérant les BSO de tous ses apparentés connus de telle sorte que les informations entrant dans la pseudo-performance dépendent uniquement des relations de parenté extérieures à l'échantillon génotypé. Chaque pseudo-performance a été accompagnée d'un poids résumant la quantité d'information incluse dans l'indice dérégressé (dBSO). Les chevaux génotypés avaient été sélectionnés suivant la disponibilité de leur ADN et la qualité de l'échantillon (sang ou sperme), sur la précision de leur CD et sur la taille de leur famille. Les marqueurs moléculaires utilisés étaient les SNPs de la puce 54k fournie par Illumina. Ils avaient eux aussi été triés pour répondre aux tests de qualité suivants : équilibre d'Hardy-Weinberg respecté, fréquence de l'allèle minimum de 5%, typage correct chez au moins 98% des chevaux. Finalement 44 444 SNPs avaient été retenus. Le test de la sélection génomique par validation croisée avec différents modèles (GBLUP, Bayes Cπ) et différents échantillons d'apprentissage et de validation a montré que la précision de la prédiction de la valeur génétique est peu améliorée lors du passage de l'évaluation classique basée sur le pédigrée (BLUP) à l'évaluation génomique (GBLUP) : la précision de 0.36 passait à 0.39. Pour que la mise en œuvre de la sélection génomique constitue un réel apport par rapport à la sélection classique, il faudrait améliorer la précision de la prédiction des valeurs génétiques.

Une solution envisagée au cours de la thèse pour améliorer la précision est l'utilisation d'une évaluation génomique en une étape. Cette méthode abordée dans le 1er chapitre bibliographique permet d'utiliser la totalité des performances propres disponibles directement dans l'évaluation génomique, que les individus performeurs soient génotypés ou non, et de s'affranchir des approximations nécessaires au calcul de performances dé-régressées. L'évaluation génomique en une étape a donc été testée chez les chevaux de sport. A cette fin, un critère d'évaluation de la performance homogène pour l'ensemble des années de performances disponibles a été recalculé. Les paramètres génétiques pour la performance en CSO avaient été estimés sur des critères hétérogènes et avec des modèles incomplets il y a plusieurs dizaines d'années : nous les avons donc ré-estimés pour le critère d'évaluation calculé. Enfin l'évaluation classique et l'évaluation génomique en une étape ont été comparées.

5.2. Calcul d'un critère homogène pour l'ensemble de la population depuis 1985

L'évaluation en une étape nécessite d'avoir le même critère de performance pour tous les chevaux inclus dans l'analyse. Or, comme expliqué dans la partie bibliographique sur le cheval, depuis 2009 le gain annuel (critère d'évaluation originel et ayant le poids le plus important dans l'index) a été remplacé par des points. Ce changement de critère est dû à une dérégularisation des dotations qui a rendu le critère du gain annuel obsolète, les dotations des épreuves n'étant plus représentatives de leur difficulté. Il était donc nécessaire en premier lieu de choisir et de calculer un critère unique pour tous les chevaux utilisés dans l'évaluation.

5.2.1. Choix du critère de performance

La partie bibliographique consacrée au cheval et à l'évaluation génétique pour l'aptitude au CSO a présenté à quel point la performance en saut d'obstacles est difficile à évaluer, du fait de la diversité des types d'épreuves, de leurs niveaux et des règlements associés qui sont modifiés chaque année, entrainant l'apparition de nouvelles catégories d'épreuves et la disparition d'autres épreuves. Face à cette complexité, un critère simple et représentatif des performances des chevaux a été mis en place : le gain annuel, puis les points. Un temps mentionné dans le stud-book, le BSO (BLUP Saut d'Obstacle) n'est plus utilisé officiellement mais est toujours publié. Il n'est pas possible de vérifier si les éleveurs choisissent les étalons en tenant compte de leur indice génétique, mais dans la pratique ce sont les étalons ayant les indices les plus élevés qui réalisent le plus de saillies (A. Ricard, communication personnelle), et il y avait un réel progrès génétique dans la population quand le BSO était utilisé (Dubois et Ricard 2007). Il n'était donc pas question au cours de ma thèse de partir en quête d'un nouveau critère de mesure de la performance, mais plutôt d'arriver à une homogénéisation du critère existant pour l'ensemble des fichiers de performances disponibles (années 1985 à 2012), avec pour objectif final son utilisation dans une évaluation génomique en une étape.

La méthode retenue est à mi-chemin entre l'utilisation des gains annuels et l'utilisation des points. Compte-tenu des données disponibles, nous avons choisi de recalculer un gain annuel pour chacun des chevaux en fonction de leurs classements. Les modalités de ce calcul ont varié en fonction du type de compétition et des fichiers où sont enregistrées les performances.

5.2.2. Particularités des performances brutes et des fichiers de données

Les fichiers de performance étaient disponibles pour les années 1985 à 2012. A chaque année correspondait au moins un fichier de performance. Il est arrivé, en 1997 et en 1998, que les résultats des épreuves d'élevage organisées par les associations de races soient enregistrés dans des fichiers différents de ceux de la FFE rassemblant les compétitions fédérales. A peu de choses près, la structure des fichiers était la même d'une année sur l'autre et contenait le même type d'informations. Il a cependant fallu tenir compte de modifications de codage des informations, par exemple la discipline codée avec des chiffres (1, 2 ou 3) puis avec des caractères (DR, CC, SO), de changement d'ordre des colonnes d'informations ou de leur longueur (Annexe 1 : Format des fichiers de compétitions équestres des chevaux de sport), ou bien encore de changement d'unités pour les enregistrements des gains.

Les fichiers de 2008 à 2012 contenaient les gains reçus par les chevaux ainsi que leur classement. Les chevaux dits « classés » sont ceux qui ont obtenus les meilleurs résultats lors de l'épreuve (le

classement compte les 8 premiers chevaux ou bien le premier ¼). Les chevaux classés ont reçu un gain indiqué dans le fichier, et les chevaux non-classés ont tout le même un rang, dépendant comme pour les chevaux classés des pénalités reçues et du temps de parcours.

En revanche, de 1985 à 2007 les gains étaient bien indiqués, mais seuls les chevaux ayant eu un gain avaient aussi un rang. Tous les chevaux non-classés étaient indiqués comme « derniers » quel que soit leur classement réel : pour ¾ des partants on ne disposait donc ni du gain puisqu'ils n'en n'avaient pas reçu, ni du classement réel.

Il existait également une particularité due au type d'épreuve : épreuves fédérale d'une part, ouvertes à tous les chevaux, ou épreuves réservées aux jeunes chevaux par classe d'âge d'autre part. Dans la partie bibliographique ces épreuves « jeunes chevaux » ont été présentées : il s'agit d'un circuit de compétition dont le but est de préparer les jeunes chevaux à leur future carrière. Sur ces épreuves seuls les chevaux sans fautes sont classés, et le chronomètre n'est pas utilisé pour départager les chevaux ayant le même nombre de pénalités. Pour les enregistrements des résultats des épreuves jeunes chevaux il y avait donc 2 cas de figure possibles : dans les fichiers où seuls les chevaux classés avaient un rang, les chevaux classés étaient tous premiers ex aequo, et les chevaux non-classés étaient tous derniers. Dans les fichiers où tous les rangs étaient enregistrés, les chevaux sans fautes étaient tous ex aequo, puis les chevaux ayant eu 4 points de pénalités étaient tous ex aequo, etc.

On avait donc dans tous les cas les gains des chevaux classés. Suivant le types d'épreuve et le fichier contenant les enregistrements, on disposait ou non des rangs des chevaux non-classés (Tableau 5.1). Nous allons voir comment ces différentes quantités d'informations ont été prises en compte dans le calcul du critère.

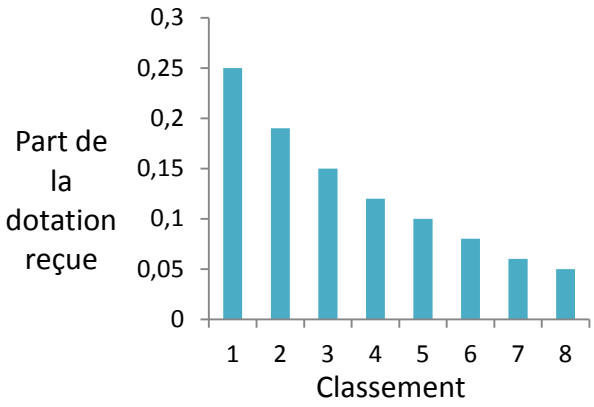
Tableau 5.1 : Récapitulatif des informations enregistrées en fonction de l'année de performance et du type de compétition.

Année de performance	Compétition fédérale ou internationale	Particularité « jeunes chevaux »
1985-2007	Gain + rang des chevaux classés	Classés : 1ers ex aequo, non-classés : derniers ex aequo
2008-2012	Gain des chevaux classés + rang de tous les chevaux partants	Classés par groupes d'ex aequo suivant le nombre de pénalités

5.2.3. Calcul du critère : un gain annuel basé sur des gains fictifs

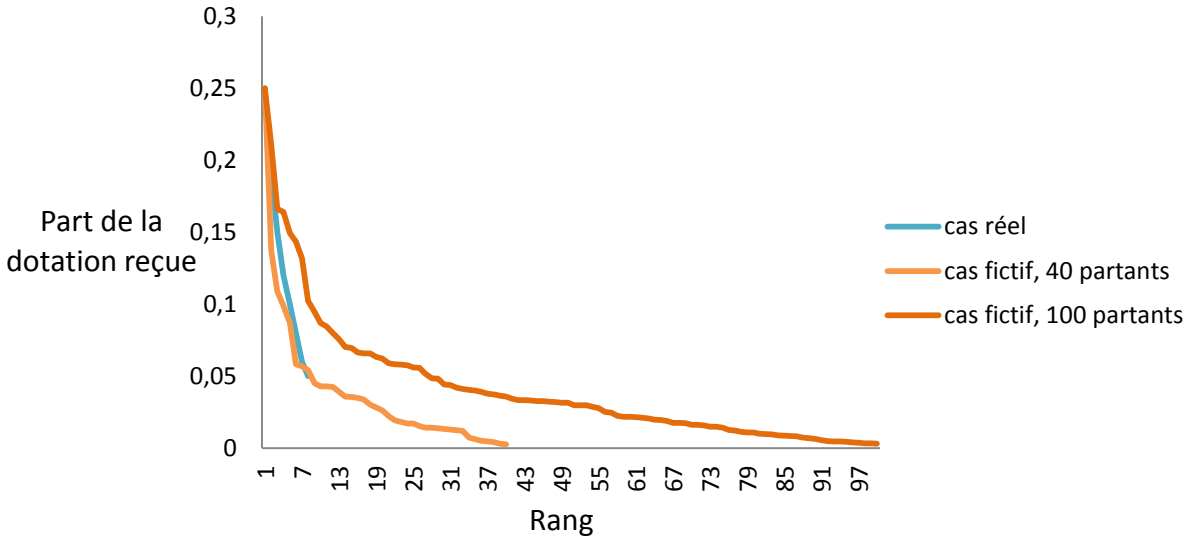
Les fichiers détaillés des performances de 2008 à 2012 contenaient en plus des gains les rangs des chevaux non-classés. Le calcul d'un nouveau critère d'évaluation de la performance ayant pour but de prendre en compte le plus d'informations possible dans le single-step, il aurait été dommage d'utiliser seulement ¼ des informations disponibles en se contentant du gain réel. Nous ne souhaitions cependant pas calculer un deuxième indice spécifique au classement comme c'est le cas dans l'indexation actuelle. Pour les performances réalisées en compétitions fédérales, nous avons donc envisagé de prolonger la décroissance des gains en tenant compte du classement. En effet, la distribution des gains est quasiment exponentielle (Figure 5.1). Il est donc tout à fait envisageable de prolonger cette décroissance des gains, et d'affecter un gain fictif aux chevaux non-classés mais dont on connait le rang. Cette solution a permis de tenir compte de tous les rangs connus en utilisant un seul critère, celui du gain.

Figure 5.1 : Décroissance de la part de la dotation reçue en fonction du rang de classement du couple cheval-cavalier dans une compétition de CSO



La méthode suivante a été utilisée pour calculer les gains manquants : connaissant le nombre de partants n dans une épreuve, on tire n valeurs dans une loi normale centrée réduite. On ordonne ensuite les n valeurs tirées de la plus grande à la plus petite. On obtient ainsi le *normal score*, qui est l'espérance de la statistique d'ordre sur les n valeurs. On calcule ensuite l'exponentielle de ces valeurs afin de respecter la distribution réelle des gains. Finalement, on standardise de façon à ce que le gain du 1er représente 25% de la dotation totale, comme dans la réalité. La Figure 5.2 représente la part de la dotation touchée dans le cas réel d'une épreuve avec une trentaine de partants, et dans deux cas fictifs où la décroissance des gains a été calculée avec la méthode décrite précédemment. On peut voir sur la Figure 5.2 qu'à partir d'un grand nombre de partants, le fait de donner des gains fictifs seulement aux chevaux non-classés en réalité risque de perturber le classement : il est possible que le gain fictif calculé pour le premier cheval non-classé soit supérieur au gain réel du dernier cheval classé. L'idée initiale était de prolonger la décroissance des gains en calculant des gains fictifs pour les chevaux non-classés. Au vu des risques de modification des classements, nous avons finalement décidé de recalculer un gain fictif pour tous les chevaux, y compris ceux qui ont reçu un gain en réalité. La méthode utilisée est la même que précédemment, mais les gains fictifs remplacent complètement les gains réels.

Figure 5.2 : Part de la dotation reçue en fonction du classement, en réalité ou avec calcul d'un gain fictif (40 partants ou 100 partants).



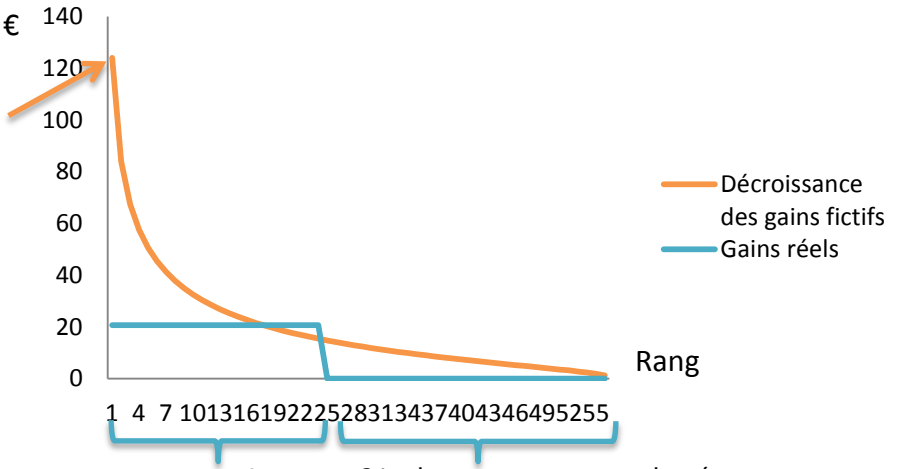
Un autre choix a été fait concernant l'attribution des gains fictifs : celui de ne pas limiter les gains fictifs distribués dans une épreuve à la dotation réelle. En effet, si plus de chevaux reçoivent un gain à partir d'une dotation totale identique, les chevaux classés parmi les premiers voient leur gain diminuer. Ceci est problématique, par exemple quand des épreuves de difficulté similaire ont la même dotation totale, mais que le nombre de partants est plus faible dans une épreuve que dans l'autre. Les chevaux classés dans l'épreuve avec beaucoup de partants recevront des gains fictifs inférieurs à ceux des chevaux de l'épreuve avec moins de partants, et seront donc injustement pénalisés. Nous avons donc choisi d'utiliser la dotation totale comme référence pour initier la décroissance des gains : le 1er reçoit 25% de la dotation réelle, puis la loi de décroissance calculée est appliquée. Les gains ont ainsi été recalculés pour l'ensemble des épreuves dont les classements complets étaient disponibles, de 2008 à 2012.

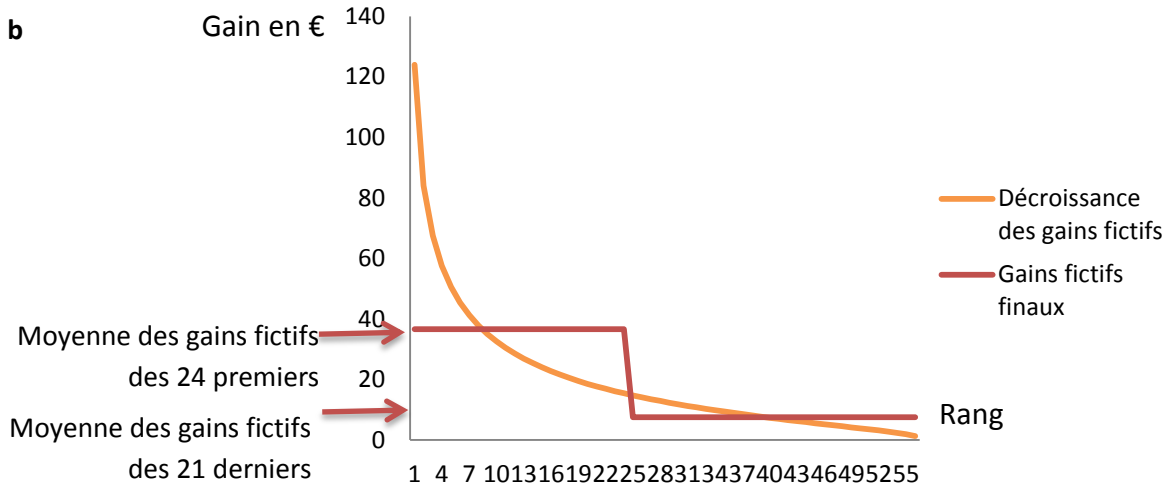
Un traitement particulier a été appliqué aux épreuves pour lesquelles les seuls rangs connus sont ceux des chevaux classés. La décroissance des gains a été calculée pour le nombre de partants de l'épreuve, et les gains fictifs ont été attribués aux chevaux classés dont on connaissait le rang. Les chevaux non-classés ont reçu la moyenne des gains restants. Un procédé similaire a été utilisé pour les épreuves jeunes chevaux. La méthode a été la même quel que soit le type d'enregistrement (que des 1ers ex æquo et des derniers ex æquo, ou bien un classement plus détaillé en fonction des pénalités des non-classés). La décroissance des gains a été calculée pour le nombre de partants (Figure 5.3 a), puis les x chevaux classés 1er ex æquo ont reçu la moyenne des gains fictifs calculés suivant la décroissance des gains des x 1er. Les chevaux non sans-faute ont reçu la moyenne des gains restants (Figure 5.3 b). Nous avons choisi de ne pas tenir compte du classement détaillé des épreuves jeunes chevaux quand il était disponible car dans la réalité ce classement n'est pas valorisé. Une fois que le couple-cheval cavalier a reçu une pénalité, il est certain qu'il ne sera pas classé. Il n'y a donc plus d'enjeu compétitif sur la suite du parcours, et il est possible que le cavalier se déconcentre ou perde en motivation, ou bien qu'il choisisse d'économiser son cheval sur la fin de l'épreuve. Ceci peut conduire à des pénalités supplémentaires non-représentatives des qualités du cheval, c'est pourquoi nous ne prenons pas en compte les rangs des chevaux non-classés des épreuves jeunes chevaux dans notre attribution des gains fictifs.

Figure 5.3 : Exemple d'un calcul de gains fictifs dans le cas d'une épreuve jeunes chevaux dotée à 480€ avec 55 partants et 24 chevaux sans fautes, 1er ex æquo. Les chevaux restants sont 25èmes ex æquo.

a

Dans une épreuve fédérale dotée à 480€, le 1er devrait recevoir 120€





Les gains antérieurs à 2001 ont été convertis en €. Ensuite, pour chaque année de compétition, les gains des chevaux ont été sommés afin d'obtenir un gain annuel. Le nombre de gains fictifs donnés par année de compétition est représenté dans la Figure 5.4. Entre 1985 et 2012 le nombre de gains annuels fictifs a augmenté régulièrement, passant de près de 12 500 à plus de 51 000.

Un exemple de distribution des gains annuel est présenté dans la Figure 5.5. Sur les 44 800 chevaux ayant eu un gain fictif en 2008, près de la moitié ont un gain annuel fictif inférieur à 150€. Un peu plus de 6 700 chevaux atteignent un gain fictif annuel supérieur à 1000€, et à peine 200 chevaux ont un gain annuel de plus de 10 000€. En conséquence, une transformation logarithmique a été appliquée aux gains fictifs afin d'approcher une distribution normale du critère d'évaluation de la performance. Le résultat de cette transformation pour l'année 2008 est présenté dans la Figure 5.6.

Dans la suite on désignera le critère de performance comme le **gain annuel**, mais il s'agira bien du **logarithme du gain annuel fictif**.

Finalement, les performances décrites par le gain annuel ont été rassemblées dans un seul et même fichier. Le nombre total de performances était de près de 942 000 gains annuels, réalisés sur 28 années de compétitions par un peu plus de 224 000 chevaux, inclus dans un pédigrée de 413 000 chevaux. A partir de ces données, les paramètres génétiques pour le gain annuel ont été estimés.

Figure 5.4 : Nombre de gains fictifs annuels distribués par année de compétition

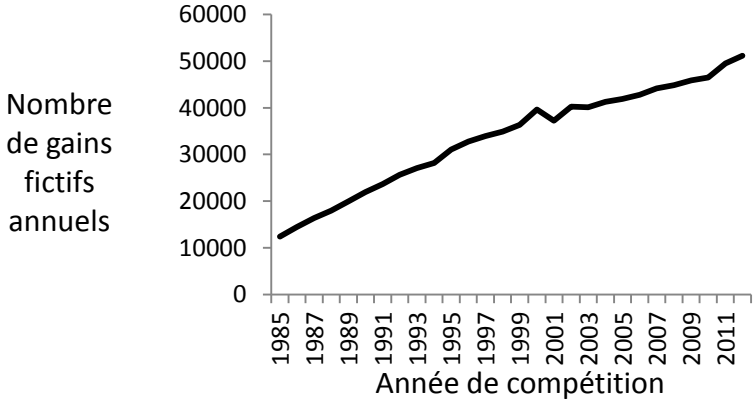


Figure 5.5 : Distribution du gain annuel fictif par cheval en 2008, pour les gains annuels inférieurs à 1000€ (a) et pour les gains annuels supérieurs à 1000€ (b) avant transformation logarithmique.

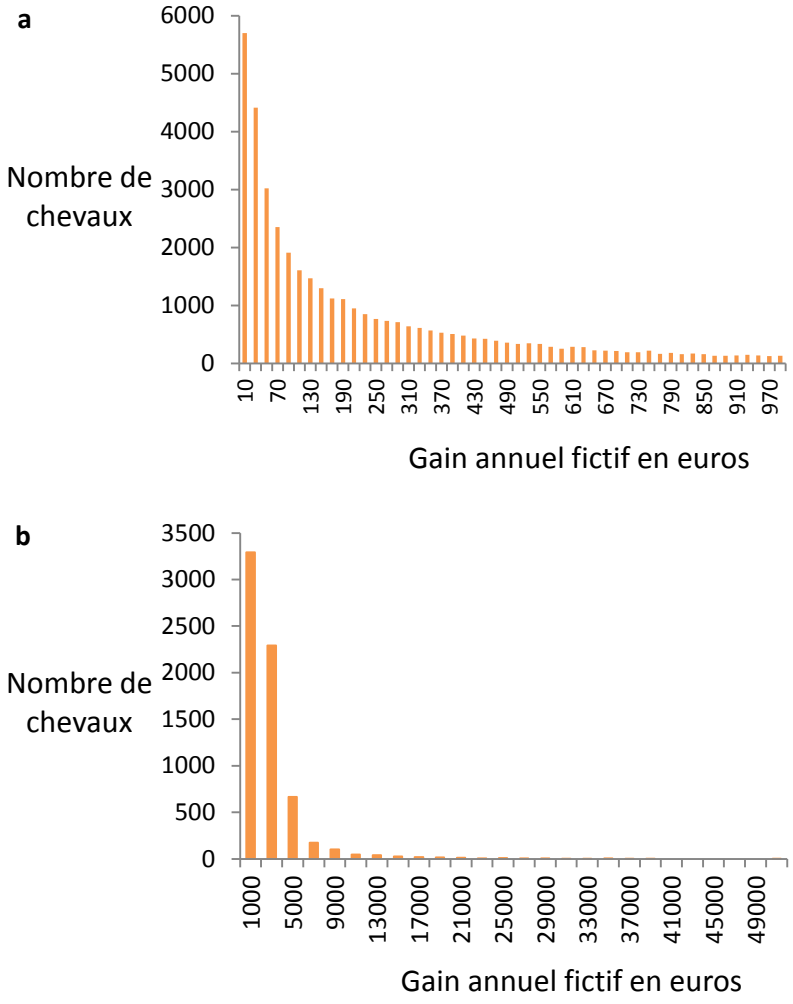
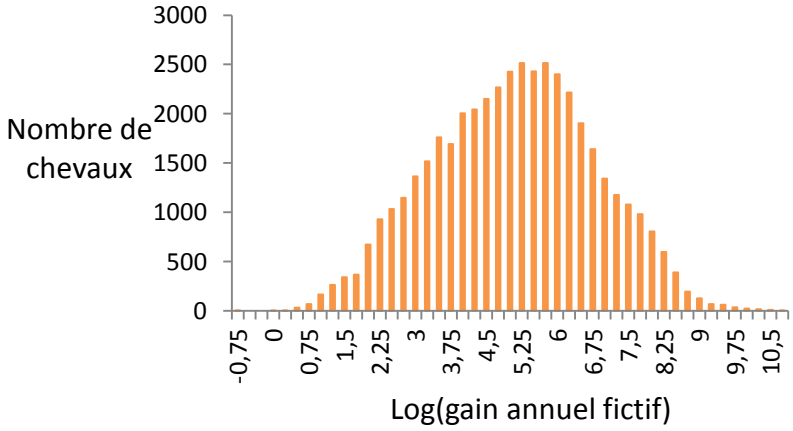


Figure 5.6 : Distribution du logarithme du gain annuel fictif pour l'année 2008



5.3. Estimation des paramètres génétiques avec différents effets fixes dans le modèle

Les fichiers de performances incluaient des informations susceptibles d'être prises en compte en effets fixes dans l'estimation des paramètres génétiques du gain annuel tels que le cavalier, le lieu de naissance du cheval, sa race, son âge au moment de la compétition, son sexe, et l'année de la compétition. L'estimation finale des paramètres génétiques a été obtenue en testant plusieurs modèles et différentes classes d'effets fixes. J'ai choisi de ne pas présenter la totalité des résultats intermédiaires qui m'ont conduit à revoir progressivement les classes des effets fixes.

5.3.1. Deux effets écartés : le cavalier et la région de naissance

La prise en compte de l'effet du cavalier n'a pas été envisagée pour les mêmes raisons que dans l'indexation : un cavalier monte trop peu de chevaux différents pour que cet effet puisse être estimé correctement. Par exemple pour l'année 2012, sur 35 059 cavaliers ayant pris part à des compétitions, 23 191 ont monté un seul cheval, 6 379 cavaliers ont monté 2 chevaux, 2 227 3 chevaux, et 1 082 ont monté 4 chevaux. Soixante-cinq cavaliers ont monté plus de 20 chevaux.

La grande majorité des performeurs en CSO sont nés en France. La prise en compte d'un effet du pays de naissance avait été envisagée dans l'indexation classique. L'étude menée à cette fin a montré un effet non-négligeable du pays de naissance sur les performances, mais la mise en œuvre de la prise en compte de cet effet est pour l'instant impossible. En effet, pour corriger correctement les performances pour l'effet du pays de naissance, il faudrait que les étalons aient suffisamment de produits nés en France et à l'étranger pour déterminer quelle part de la performance est due à la génétique et quelle part est due à l'environnement (sélection des importations, meilleur environnement des chevaux étrangers). Or, les effectifs des performeurs ayant pour père le même étalon et nés en France et à l'étranger sont tels que l'effet ne peut pas être estimé correctement avant 2001 (A. Ricard, communication personnelle). Comme les gains annuels ont été calculés en utilisant des performances remontant jusqu'à 1985, nous avons choisi de ne pas prendre en compte cet effet, et de l'écarter dès l'estimation des paramètres génétiques.

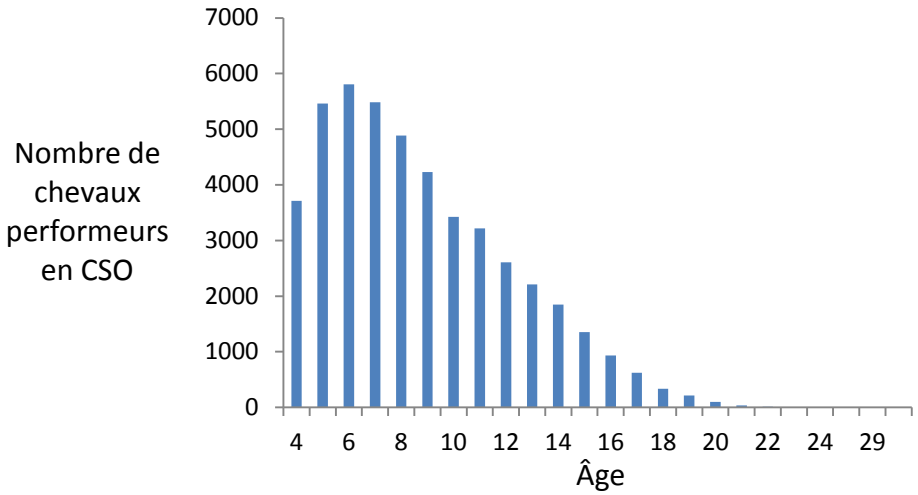
5.3.2. Prise en compte de l'âge, du sexe et de l'année de la compétition : le trio indispensable

Connaissant l'année de la compétition et l'année de naissance des chevaux, il était possible de calculer leur âge au moment de la compétition. La distribution des âges des chevaux au moment de la réalisation de leurs performances est similaire d'une année de compétition à l'autre. L'exemple de l'année 2010 est visible dans la Figure 5.7. 20% des chevaux ont 4 ou 5 ans, et la moitié des performeurs ont 8 ans ou moins. A partir de 6 ans, l'âge des chevaux participant aux compétitions décroit régulièrement, et seuls 15% des performeurs ont 15 ans ou plus. Les chevaux âgés seront donc regroupés par classes : les 7-8 ans, les 9-10 ans, les 11-12 ans, les 13-14 ans, et les chevaux de 15 ans et plus. Dans le calcul du BSO, l'objectif de l'effet âge est de tenir compte de l'expérience acquise par le cheval puis de son vieillissement, mais aussi de prendre en compte les politiques de dotation spécifiques des épreuves jeunes chevaux.

Les mâles et les femelles concourent dans les mêmes épreuves, et les mâles peuvent être entiers ou bien castrés. Le pourcentage d'entiers est stable autour de 15-16%. Les pourcentages de juments et de hongres varient tous les deux entre un peu moins et un peu plus de 40% suivant les années. Afin

de prévenir les erreurs d'enregistrement du statut des mâles en tant que hongre ou entier, ils ont été rassemblés dans la même catégorie.

Figure 5.7 : Distribution de l'âge des chevaux performeurs en CSO en 2010.



L'effet de l'année de la compétition doit aussi être pris en compte afin de tenir compte de l'inflation et de la variabilité des dotations allouées d'une année sur l'autre, notamment par cheval en fonction de l'évolution des effectifs.

Ces trois effets âge, sexe et année interagissent : suivant l'année les politiques de dotation peuvent changer, en favorisant les jeunes chevaux par exemple, ce qui se répercute sur l'effet de l'âge. Ces relations entre les effets ont donc été prises en compte en les intégrant dans le modèle d'estimation des paramètres génétiques comme des effets fixes d'interaction.

Le modèle testé pour l'estimation des paramètres génétiques est donc le suivant :

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{Zp} + \boldsymbol{e},$$

avec $\boldsymbol{y}$ le vecteur des performances, $\boldsymbol{b}$ le vecteur des effets fixes (interaction de l'âge, du sexe et de l'année de la compétition), $\boldsymbol{a}$ le vecteur de l'effet animal aléatoire, $\boldsymbol{p}$ le vecteur de l'effet d'environnement permanent du cheval, et $\boldsymbol{e}$ la résiduelle. $\boldsymbol{X}$ et $\boldsymbol{Z}$ sont des matrices d'incidence. Les analyses ont été réalisées avec le logiciel ASReml (Gilmour *et al.* 2006).

L'héritabilité obtenue était de 0.28, et la répétabilité de 0.51. Les effets moyens de l'âge et de l'année de la compétition sont représentés dans la Figure 5.8 et la Figure 5.9. On constate un effet favorable de l'âge sur les performances pour les très jeunes chevaux (4 ans) qui décline légèrement avant de remonter à 7-8 ans, et qui diminue assez rapidement passé 10 ans. L'effet de l'année de compétition est de plus en plus négatif au fil des années. Ceci est dû à l'augmentation du nombre de chevaux participants aux compétitions qui n'a pas été accompagnée d'une augmentation proportionnelle des dotations. L'effet du sexe est très proche entre les juments et les mâles.

Figure 5.8 : Effet moyen de l'âge sur le gain annuel

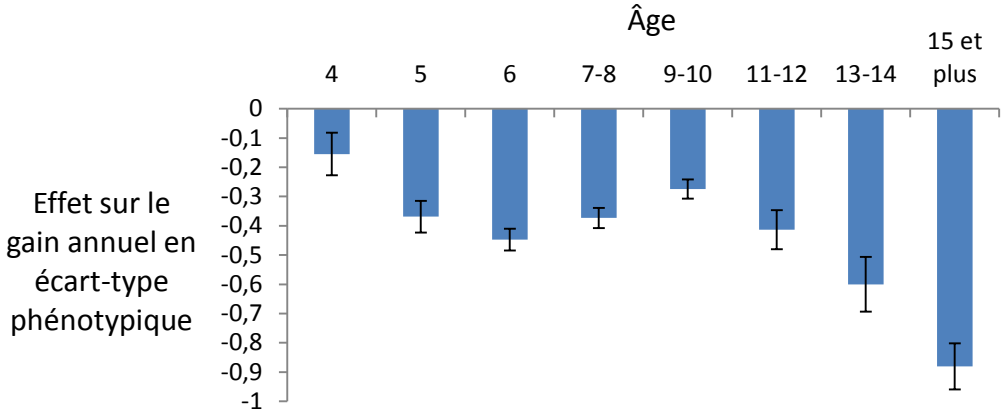
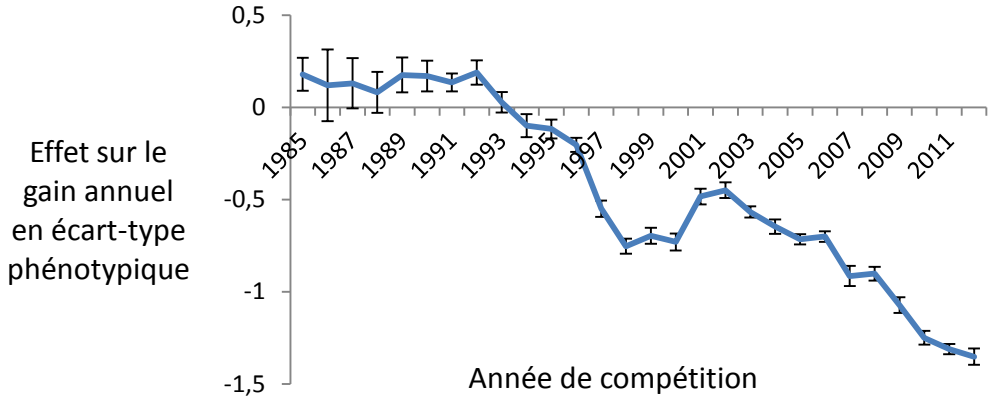


Figure 5.9 : Effet moyen de l'année de compétition sur le gain annuel



5.3.3. Utilisation de groupes de parents inconnus : pallier aux informations manquantes dans le pédigrée

Dans le pédigrée, près de 28 500 chevaux sont nés d'un ou de deux parents inconnus. Dans l'estimation des paramètres génétiques qui vient d'être présentée, ces parents sont supposés appartenir tous à la même population de base. On ne prend pas en compte le progrès génétique, qui fait que le parent inconnu d'un cheval né en 1950 sera sûrement d'un moins bon niveau génétique que le parent inconnu d'un cheval né en 2000. Nous avons donc tenté d'utiliser des groupes de parents inconnus pour pallier ce problème. Différents groupes de parents inconnus permettent de relier les chevaux à différentes populations d'origine, soit en fonction de leur année de naissance soit en fonction de leur race.

J'ai choisis un découpage des groupes de parents inconnus avec un pas de temps variable : un seul groupe pour les chevaux nés avant 1920, puis un groupe tous les 10 ans jusqu'en 1980. A partir de 1980, j'ai abaissé le pas de temps à 5 ans, car cette année marque celle du début du progrès génétique dans la population des Selles Français. Le découpage des groupes de parents inconnus par pas de temps n'a pas été croisé avec un découpage par effet race en raison des effectifs trop faibles de certaines races ou certains types de races dans les groupes les plus anciens, mais aussi dans les groupes les plus récents avec lesquels on veut prendre en compte le progrès génétique (10 à 60 chevaux de sport français dans les 4 derniers groupes, moins de 100 poneys et chevaux Arabes dans les 2 derniers groupes). Pour les chevaux ayant un seul parent inconnu, le groupe de parents inconnus a été attribué en se basant sur l'année de naissance du parent connu. Dans le cas de deux

parents inconnus, les groupes de parents inconnus étaient attribués en se basant sur l'année de naissance de l'individu. Parmi les 28 423 chevaux nés d'un ou de deux parents inconnus, l'année de naissance était inconnue pour 9 955 d'entre eux. Il a donc fallu l'estimer. Cette estimation a été faite en remontant le pédigrée. Soit une jument d'année de naissance inconnue, ayant donné naissance à plusieurs poulains. Comme la jument devait avoir au moins 3 ans l'année où son premier poulain est né (mise à la reproduction à 2 ans et gestation de 11 mois), on suppose que son année de naissance est celle de son 1er poulain moins 3 ans, et de même pour les étalons. On calcule ainsi de proche en proche une estimation de l'année de naissance. Ce calcul a été fait pour un total de 29 275 chevaux dans le pédigrée (9 955 chevaux nés de parents inconnus et 19 320 chevaux dont on connaissait les parents mais pas l'année de naissance). La distribution des années de naissance avant et après estimation des années manquantes est représentée dans la Figure 5.10 pour les chevaux nés de parents inconnus. On peut voir que le nombre de chevaux nés de parents inconnus dans le pédigrée a longtemps été de plusieurs centaines chaque année, et a diminué à partir des années 1990 pour passer à environ une centaine par an dans les années 2000. La répartition des chevaux dans les groupes de parents inconnus est présentée dans la Figure 5.11. Les 2 groupes les plus anciens et les 2 groupes les plus récents comptent un peu moins de 1 000 chevaux, contre 2 000 à 6 000 pour les groupes correspondant aux chevaux nés entre 1930 et 1995.

Figure 5.10 : Années de naissance connues et estimées pour les chevaux nés de parents inconnus.

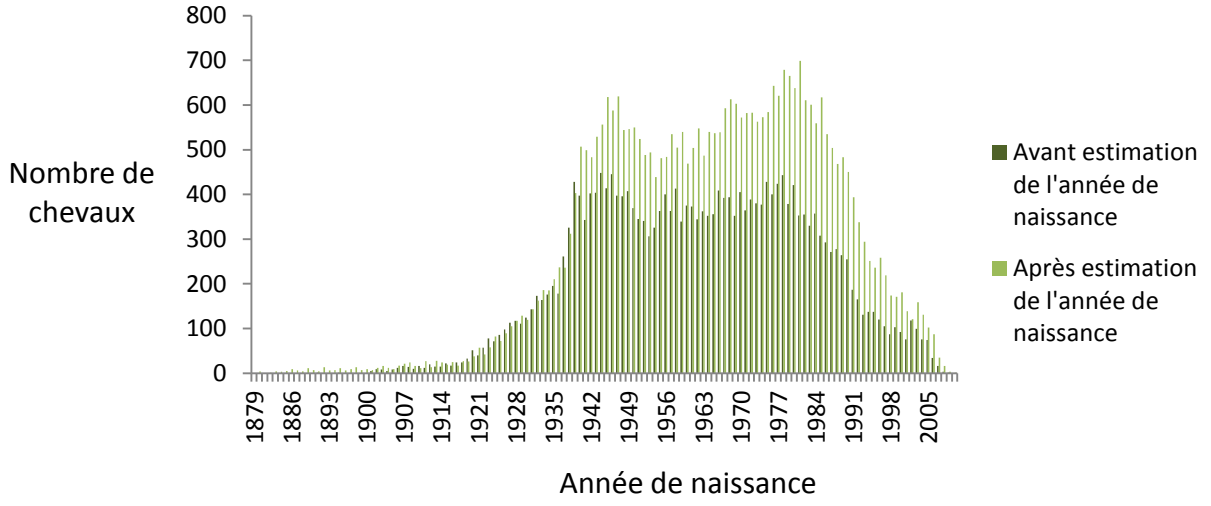
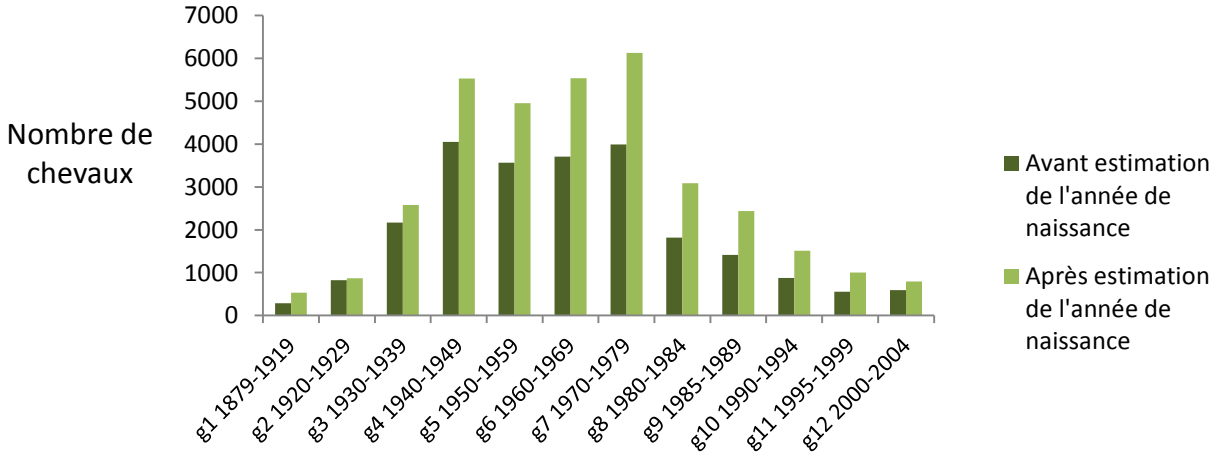


Figure 5.11 : Répartition des chevaux en bout de pédigrée dans les groupes de parents inconnus (effectifs avant et après estimation des années de naissance manquantes)



L'estimation des paramètres génétiques avec ces groupes de parents inconnus n'a pas modifié significativement les estimations des variances obtenues dans le 1er modèle, l'héritabilité reste de 0.28 et la répétabilité 0.51. En revanche, les valeurs génétiques des chevaux nés de parents inconnus ont été modifiées : elles augmentent (Figure 5.12), ce qui peut faire supposer qu'en l'absence de groupes de parents inconnus les chevaux nés de parents inconnus étaient pénalisés par une valeur génétique de la population de base trop faible. Les effets fixes âge*sexe*année ne sont pas affectés par la prise en compte des groupes de parents inconnus dans le modèle. Les solutions obtenues pour les groupes de parents inconnus sont représentées dans la Figure 5.13. Les groupes les plus anciens (1, 2 et 3, chevaux nés de parents inconnus avant 1940) ont des solutions plus élevées que les 4 groupes suivants (chevaux nés de parents inconnus entre 1940 et 1979), mais les erreurs d'estimation sont plus grandes également. On observe à partir du groupe 8 (chevaux nés de parents inconnus entre 1980 et 1984) une augmentation des solutions des groupes, qui est cohérente avec l'augmentation connue du niveau génétique de la population à partir de la fin des années 1970. La variabilité des solutions des groupes de parents inconnus est plus faible que celle des effets fixes (environ 0.5 écart-type phénotypique d'écart au lieu d'environ un écart-type phénotypique).

Figure 5.12 : Valeurs génétiques estimées des chevaux pour le gain annuel en CSO, avec ou sans l'utilisation de groupes de parents inconnus.

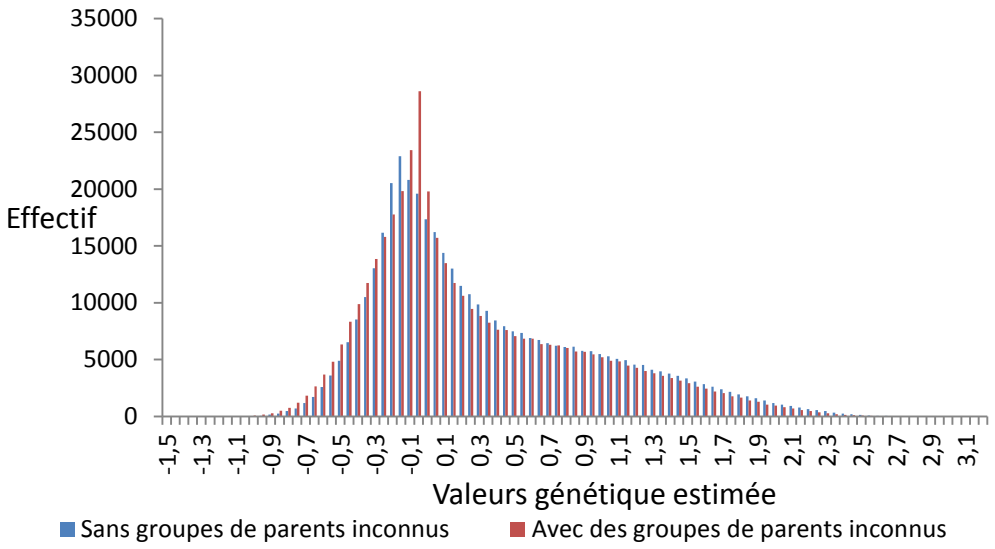
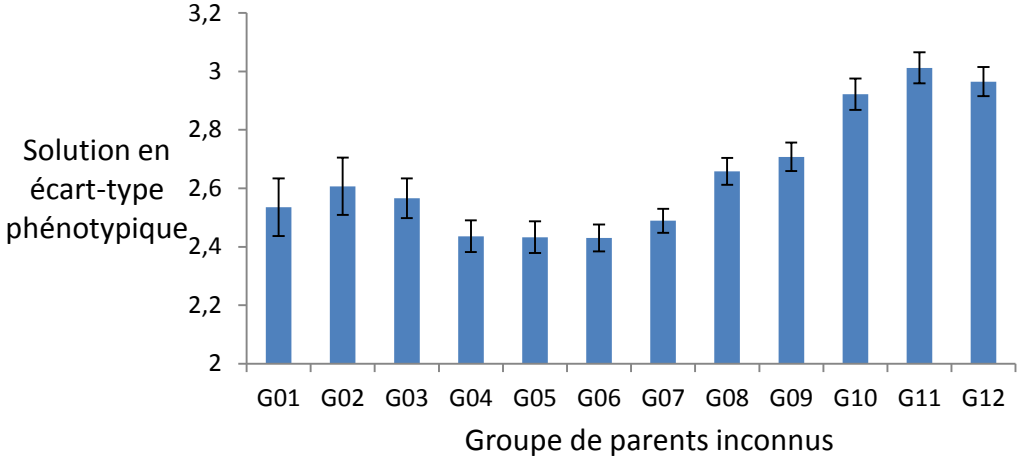


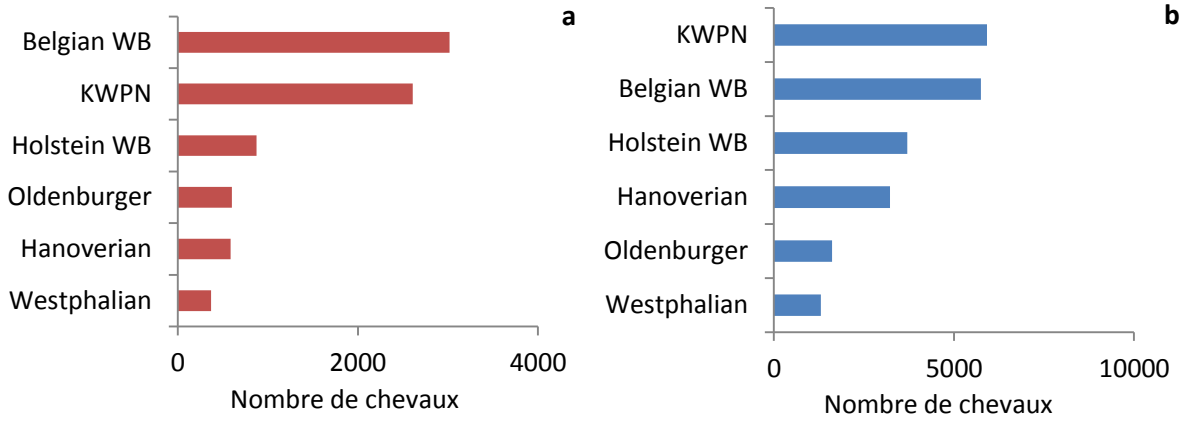
Figure 5.13 : Effet des groupes de parents inconnus sur le gain annuel



5.3.4. Prise en compte de la race, regroupement suivant la discipline de prédilection des chevaux

Près d'une centaine de races différentes sont représentées parmi les performeurs. Il peut s'agir de races réellement distinctes, ou bien de plusieurs sous-catégories d'un même stud-book. Les races ont été regroupées de la façon suivante : les Selle Français ont été rassemblés dans une catégorie, et les différents types d'Anglo-Arabes dans une autre, afin d'avoir un groupe pour chacune des 2 grandes races de chevaux de sport français. Les chevaux de sport européens les plus représentés parmi les performeurs et dans le pédigrée ont été considérés séparément également (Hanoverian, Oldenburger, Westphalian, Holstein Warmblood, Koninklijke Vereniging Warmbloed Paardenstamboek Nederland, Belgian Warmblood), leurs effectifs sont représentés dans la Figure 5.14. Les autres chevaux de sport étrangers ont été regroupés dans une seule catégorie.

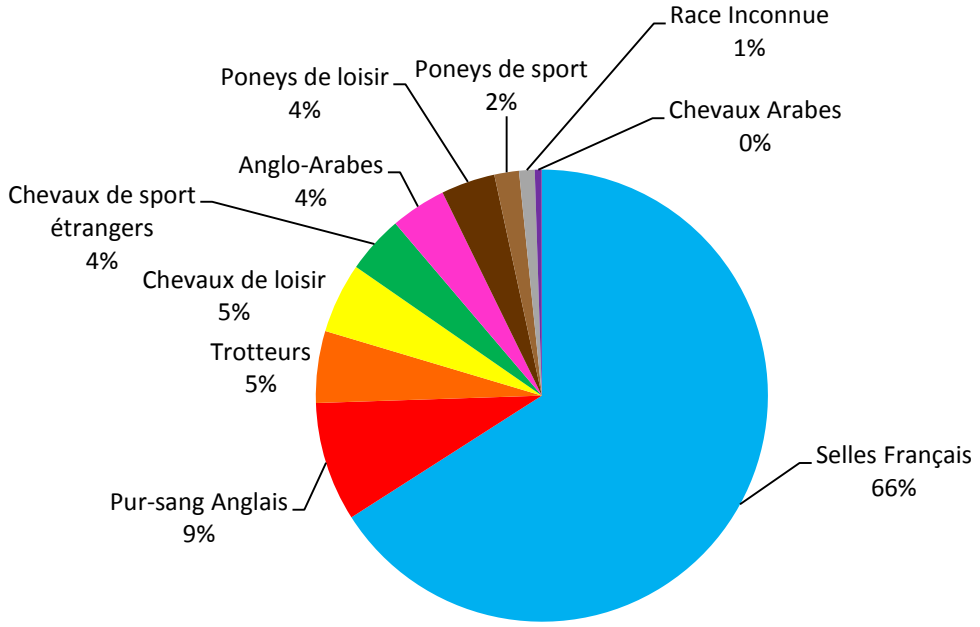
Figure 5.14 : Chevaux de sport étrangers européens présents parmi les performeurs (a) et dans le pédigrée (b) (WB=Warmblood)



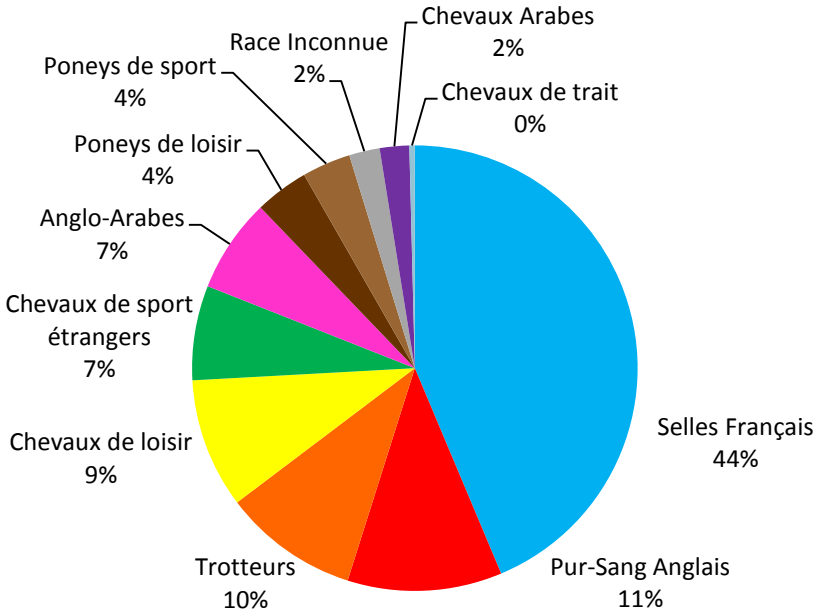
Les chevaux qui ne sont pas élevés pour leurs performances en CSO ont été répartis dans des groupes en fonction des utilisations auxquelles ils sont habituellement destinés. Il y a donc eu un groupe rassemblant les Trotteurs Français, élevés pour les courses au trot, un groupe pour les Pur-Sang Anglais, élevés pour les courses au galop, un groupe pour les chevaux Pur-Sang Arabes ou de type Arabe, plutôt sélectionnés pour l'endurance. Un groupe rassemble toutes les races de chevaux de loisirs, principalement des races françaises régionales, mais aussi des chevaux de travail américains comme le Quarter Horse, ou encore des races de chevaux élevés pour le dressage ou le spectacle comme le Pure Race Espagnol. Deux autres groupes ont été créés pour les poneys de sport et les poneys de loisir. Les parts respectives de ces groupes de chevaux parmi les performeurs et dans le pédigrée complet sont présentées dans la Figure 5.15 (a et b respectivement). Le Selle Français est la race la plus représentée parmi les performeurs en CSO et dans le pédigrée (66% et 44% respectivement). L'Anglo-Arabe représente 4% des performeurs en CSO, tout comme les chevaux de sport étrangers (les 6 races européennes isolées pour l'estimation des paramètres génétiques sont inclus aux chevaux de sport étrangers dans la figure). Même si une grande partie des performeurs sont des chevaux de sport, la présence parmi les performeurs de chevaux de loisir ou de course m'a conduit à envisager la prise en compte de l'effet de la race dans l'estimation des paramètres génétiques. La prise en compte de cet effet pouvait également être intéressante dans la mesure où au sein des chevaux de sport des orientations différentes existent, avec notamment l'Anglo-Arabe qui est plus élevé pour le CCE que pour le CSO.

Figure 5.15 : Types de chevaux performeurs en CSO (a) et présents dans le pédigrée (b)

a



b



Avec la prise en compte de la race en effet fixe, le modèle d'estimation des paramètres génétiques devient :

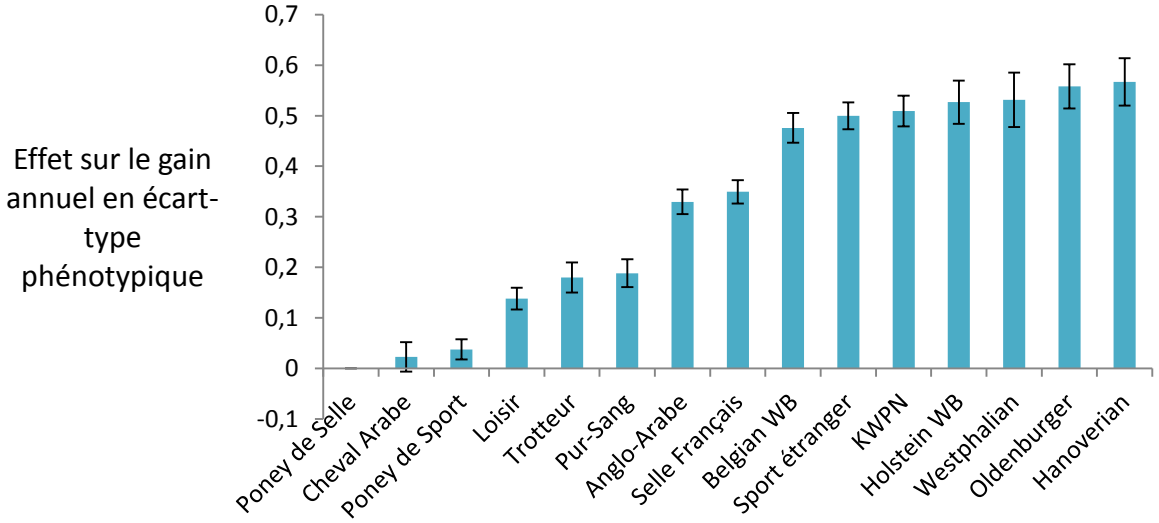
$$\boldsymbol{y = Xb + Wr + Za + Zp + e},$$

avec $\boldsymbol{r}$ le vecteur contenant les effets races et $\boldsymbol{W}$ la matrice d'incidence correspondante.

Ce modèle a été testé avec et sans l'utilisation de groupes de parents inconnus. Sans groupes de parents inconnus, l'héritabilité est de 0.27 et la répétabilité est de 0.51. Quand les groupes de parents inconnus sont utilisés, l'héritabilité est de 0.28 et la répétabilité de 0.51. Les effets fixes estimés restent similaires quel que soit le modèle, et les effets races estimés ne sont pas significativement différents en fonction de l'utilisation ou non de groupes de parents inconnus. L'effet race estimé en utilisant des groupes de parents inconnus est présenté dans la Figure 5.16. Les

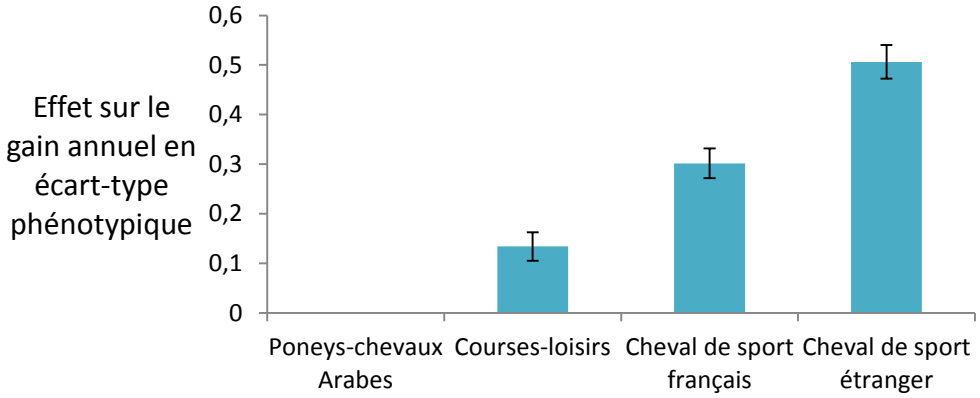
effets races obtenus n'étant pas significativement différents pour plusieurs catégories, j'ai voulu tester un modèle où les races seraient rassemblées de la façon suivante : un groupe incluant les poneys et les chevaux Arabes, un groupe regroupant les chevaux de loisir et de course, un groupe regroupant les 2 races françaises de chevaux de sport, et un groupe regroupant tous les chevaux de sport étrangers.

Figure 5.16 : Effet de la race sur le gain annuel estimé en séparant les races de chevaux de sport européens les plus représentées



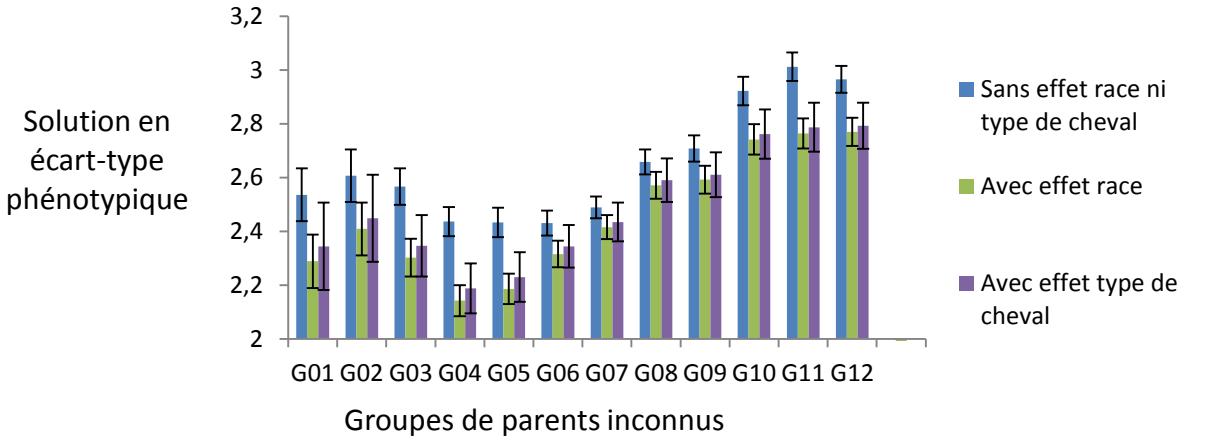
Avec cette modification des classes pour l'effet race, qui devient plutôt un effet « type de cheval », l'héritabilité obtenue est de 0.27 et la répétabilité de 0.51, aussi bien avec que sans les groupes de parents inconnus. L'ajout d'un effet « type de cheval » ne modifie pas les effets fixes déjà présents dans le modèle (interaction année de la compétition, âge et sexe). Les effets des types de chevaux sont présentés dans la Figure 5.17.

Figure 5.17 : Effet du type de cheval sur le gain annuel



Les solutions des groupes de parents inconnus estimées avec un effet race ou type de cheval dans le modèle sont présentées dans la Figure 5.18. Les solutions estimées pour les groupes de parents inconnus sont similaires pour les modèles avec effet race ou type de cheval. Cependant on peut constater que ces solutions ne sont pas tout à fait les mêmes que celles présentées dans la Figure 5.13, obtenues dans un modèle où la race ou le type de cheval ne sont pas pris en compte.

Figure 5.18 : Comparaison des solutions des groupes de parents inconnus avec utilisation ou non d'un effet race ou type de cheval dans le modèle



5.3.5. Interaction entre l'effet race ou type de cheval et les solutions estimées pour les groupes de parents inconnus.

Les solutions estimées pour les groupes de parents inconnus dans différents modèles (avec effet race, avec effet type de cheval, sans effet race ni type de cheval) sont rassemblées dans la Figure 5.18.

Si on compare d'abord les solutions estimées dans le modèle sans effet race avec les solutions obtenues dans le modèle avec effet race (respectivement en bleu et en vert dans la Figure 5.18), on constate que les solutions estimées sont différentes pour les groupes 1 à 6, puis pour les groupes 9 à 12. La Figure 5.19 représente la proportion des chevaux nés de différentes races dans chacun des groupes de parents inconnus. On peut voir une évolution des proportions des races au fil des groupes. Les 3 premiers groupes de parents inconnus ont 5 % de fils ou de filles Selle Français. Ceci pourrait expliquer le fait que ces groupes obtiennent des solutions un peu plus élevées que les groupes 4 et 5 bien qu'ils soient plus anciens. Les solutions estimées des groupes 4 et 5 sont identiques (quel que soit le modèle) et plus faibles que celles des groupes 1, 2 et 3, ce qui peut s'expliquer par de faibles proportions de Selle Français dans le groupe 4 et de chevaux de sport étrangers dans le groupe 5. On peut noter qu'à partir du groupe 5, l'augmentation des solutions des groupes de parents inconnus semble correspondre à l'augmentation de la part de fils ou filles chevaux de sport étranger. On peut remarquer aussi que les solutions des groupes de parents inconnus sont plus faibles quand l'effet race est dans le modèle. Ceci pourrait signifier qu'une partie de la solution du groupe de parents inconnus due aux races des chevaux nés du groupe est enlevée à la solution totale quand l'effet race est pris en compte dans le modèle. Cet effet ne pouvait être pris en compte au travers des groupes de parents inconnus en raison des effectifs trop faibles de certaines races ou de certains types de chevaux dans les groupes de parents inconnus les plus anciens et les plus récents.

La même comparaison entre le modèle sans effet race ni type de cheval et le modèle avec type de cheval peut être effectuée (Figure 5.18 et Figure 5.20). Là aussi les solutions des groupes de parents inconnus sont plus faibles quand l'effet type de cheval est ajouté au modèle, ce qui peut signifier qu'une part des solutions était due aux types de chevaux issus des groupes de parents inconnus. L'écart-type d'erreur d'estimation des solutions étant plus grands pour les 3 1er groupes de parents inconnus, on ne trouve cette fois pas de différence entre les solutions estimées avec ou sans prise en

compte de l'effet type de cheval. En revanche on observe à nouveau l'augmentation des solutions des groupes de parents inconnus parallèle à l'augmentation de la proportion de chevaux de sport étrangers nés de ceux-ci.

Figure 5.19 : Proportion des races des chevaux nés des différents groupes de parents inconnus

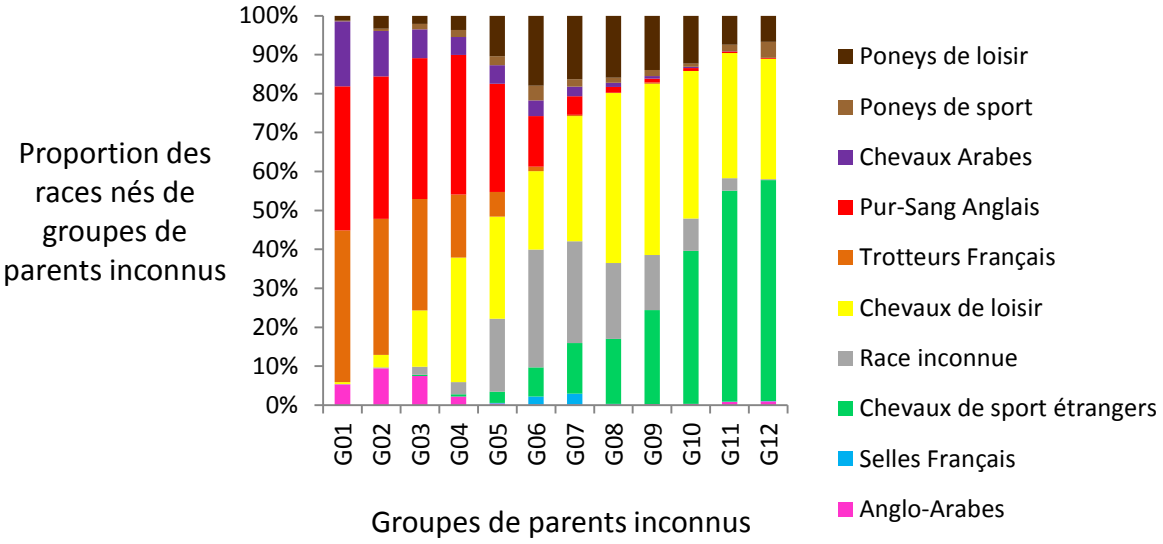
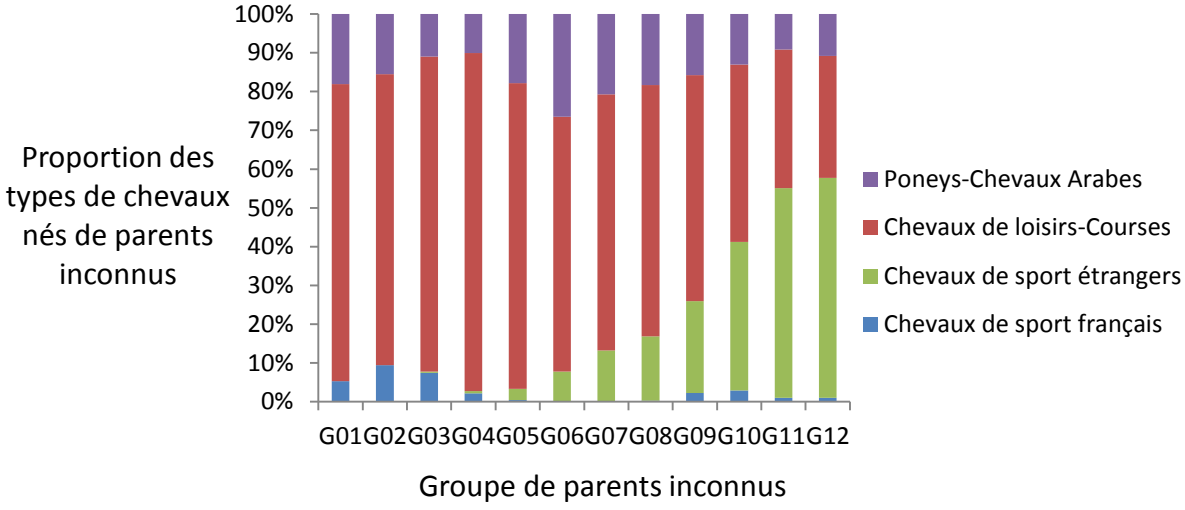


Figure 5.20 : Proportion des types de chevaux nés des différents groupes de parents inconnus



Les solutions obtenues pour les groupes de parents inconnus sont donc un peu plus faibles quand un effet type de cheval ou race est ajouté au modèle. L'augmentation des solutions obtenues par les groupes de parents inconnus à partir du 6ème groupe est cohérente avec l'augmentation des fils et filles de parents inconnus qui sont des chevaux de sport étrangers, ce type de chevaux étant ceux qui ont l'effet le plus élevé sur le gain annuel. De même les solutions élevées obtenues par les groupes anciens pourraient être dues à la présence de quelques chevaux Selle Français parmi les fils et filles issus de ces groupes. Ces tendances observées dans le modèle sans effet race ni type de chevaux restent observables quand un effet race ou type de cheval est inclus dans le modèle : il est difficile de conclure si les solutions des groupes de parents inconnus obtenues dans ces modèle reflètent seulement le progrès génétique dans la population, ou si l'effet race ou type de cheval n'a pas été

complètement pris en compte à cause d'une mauvaise transmissibilité de cette caractéristique d'une génération à l'autre. En effet les chevaux de sport français et étrangers ont dans leur pédigrée des chevaux de loisir, des Trotteurs Français et des Pur-Sang Anglais dès les 1ères générations remontées.

5.3.6. Conclusion de l'estimation des paramètres génétiques

Les paramètres génétiques calculés avec les différents modèles sont résumés dans le Tableau 5.2. Quel que soit le modèle la répétabilité est de 0.51. L'héritabilité obtenue est de 0.27 ou 0.28. Le choix du modèle n'a donc pas eu un fort effet sur l'estimation des paramètres génétiques. Pour cette raison, et aussi à cause des incertitudes sur ce qui est réellement mesuré avec l'effet race et les groupes de parents inconnus, les résultats de l'évaluation classique et génomique présentés dans la partie suivante seront uniquement ceux obtenus sur le modèle le plus simple qui prend en compte l'âge et le sexe du cheval ainsi que l'année de la performance.

Tableau 5.2: Paramètres génétiques estimés pour le gain annuel en CSO avec différents modèles

	Utilisation de groupes de parents inconnus	Prise en compte de la race	héritabilité	Répétabilité
Interaction année âge sexe	Non	Non	0.28	0.51
		Oui	0.27	0.51
		Oui, regroupées en types de chevaux	0.27	0.51
	Oui	Non	0.28	0.51
		Oui	0.28	0.51
		Oui, regroupées en types de chevaux	0.27	0.51

5.4. Comparaison de l'évaluation classique et de l'évaluation génomique en une étape

5.4.1. Matériel & méthodes

Choix des candidats et de la population de référence

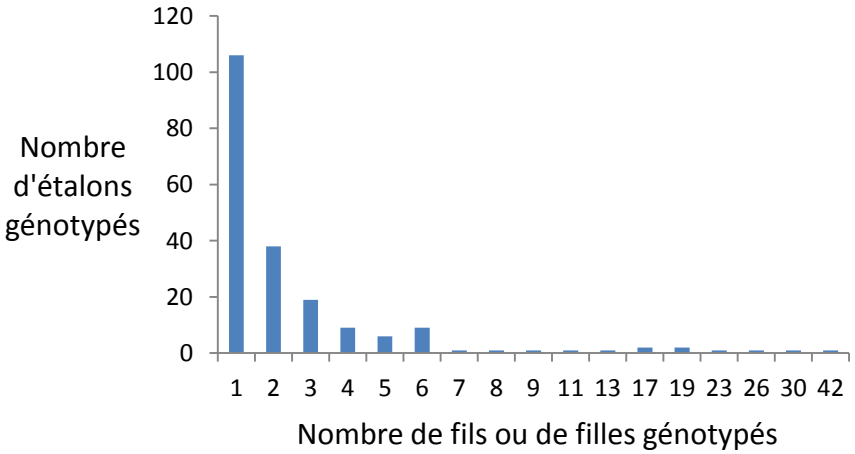
L'échantillon de chevaux génotypés était composé de 1 008 individus qui étaient tous des étalons à l'exception de 3 juments. La sélection des étalons étant principalement basée sur leurs performances propres, une grande partie des chevaux génotypés (888) étaient performeurs. Parmi les étalons génotypés, 200 avaient au moins un fils génotypé et performeur. La Figure 5.21 représente le nombre de descendants génotypés des étalons génotypés. Parmi les étalons qui ont des produits génotypés, la majorité (106) ont un seul descendant génotypé. Un peu moins de 40 ont deux descendants génotypés. Moins de 15 étalons génotypés ont plus de 6 descendants génotypés. Le fait que tous les étalons n'aient pas de produits génotypés et performeurs est dû à leur sélection récente : ils n'ont pas encore de produits en âge de participer à des compétitions.

Lors du premier essai de sélection génomique réalisé par Ricard *et al*. (2013), les phénotypes utilisés étaient des indices dé-régressés, qui avaient l'avantage d'être pré-corrigés pour les effets fixes, et qui incluaient les performances propres des chevaux génotypés, mais aussi les performances de tous leurs apparentés en dehors de l'échantillon génotypé. La validation croisée avait été réalisée sur des

chevaux dont les valeurs génétiques étaient suffisamment précises (CD d'au moins 0.52), et dont les pères étaient génotypés, avec comme condition sur les pères qu'ils aient un CD élevé ou bien 3 fils en validation.

Ici, nous avons choisi de nous placer dans une situation proche de la réalité d'un processus de sélection, où les valeurs génétiques seraient estimées pour les dernières générations de chevaux nés, et où les candidats auraient leur père génotypé dans la population de référence. Suivant ces critères, pour être considérés comme des candidats dans le cadre de notre validation, il fallait que les chevaux soit nés récemment, génotypés, performeurs (l'évaluation devant avoir lieu avant la sélection pour la reproduction, ils n'ont pas encore de descendants performeurs), et fils d'étalons génotypés. La Figure 5.22 présente les années de naissance des chevaux répondant à ces conditions. Compte-tenu des effectifs, les chevaux performeurs génotypés et fils d'étalons génotypés nés en 2003, 2004 et 2005 ont été retenus comme candidats. Afin de nous placer dans la situation où les valeurs génétiques des chevaux seraient estimées l'année de leur naissance, leurs performances et celles de tous les chevaux réalisées après l'année de naissance considérée ont été supprimées. Ainsi, pour les candidats nés en 2003, les performances réalisées à partir de 2004 n'ont pas été utilisées pour estimer les valeurs génétiques des candidats, et de même pour les candidats nés en 2004 puis pour ceux de 2005.

Figure 5.21 : Nombre de descendants génotypés par étalon génotypé.



Les races des chevaux candidats pour chacune des années sont représentées dans la Figure 5.23. La part des chevaux de sport étrangers est assez stable sur les 3 années, par contre entre 2003 et 2005 la proportion d'Anglo-Arabes diminue au profit des Selles Français. Le nombre de chevaux génotypés restant dans la population de référence varie d'une centaine sur les 3 années de naissance considérées : 877 pour les candidats nés en 2003, 921 pour les candidats nés en 2004, et 977 pour les candidats nés en 2005.

Une évaluation classique et une évaluation génomique ont donc été réalisées sur les candidats retenus. L'évaluation classique et l'évaluation génomique ont également été réalisées sur un sous-échantillon de ces candidats, en ne conservant que ceux dont les pères avaient un CD minimum de 0.60.

Figure 5.22 : Année de naissance des chevaux candidats potentiels

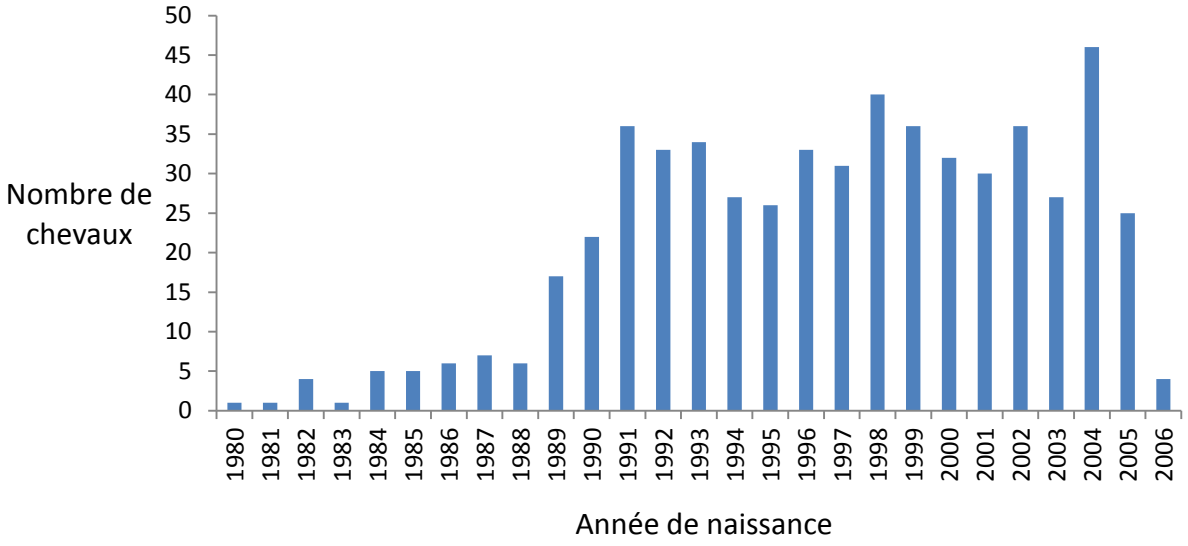
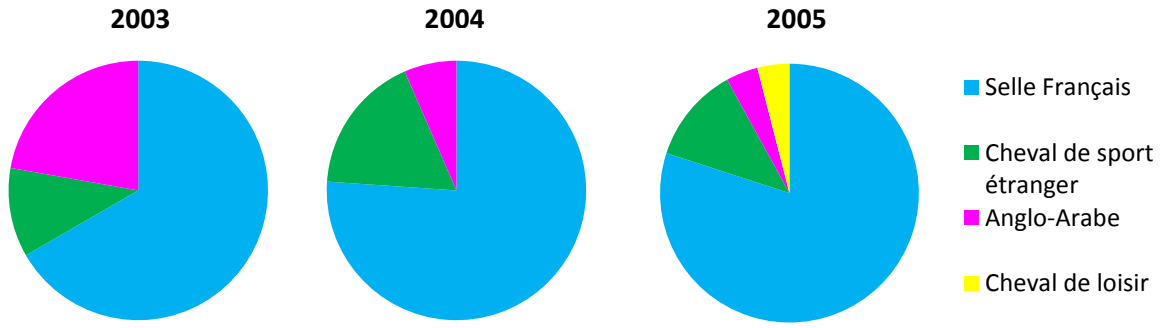


Figure 5.23 : Races des candidats nés en 2003, 2004 et 2005



Modèles utilisés

Le modèle utilisé pour l'évaluation classique est un modèle animal :

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{Wc} + \boldsymbol{e},$$

avec $\boldsymbol{y}$ le vecteur des gains annuels, μ la moyenne des performances, $\boldsymbol{b}$ le vecteur des effets fixes : interaction de l'année de la compétition, de l'âge et du sexe du cheval. Pour rappel, les années de compétition vont de 1985 à 2012, les classes d'âges sont 4 ans, 5 ans, 6 ans, 7-9 ans, 9-10 ans, 11-12 ans, 13-14 ans, et 15 ans et plus, et pour le sexe les hongres et les étalons sont rassemblés dans le même groupe. $\boldsymbol{a}$ est le vecteur des valeurs génétiques additives telles que $V(\boldsymbol{a}) = \boldsymbol{A}\sigma_a^2$, $\boldsymbol{c}$ est le vecteur des effets de l'environnement permanent dans lesquelles sont répétées les performances, et $\boldsymbol{e}$ est la résiduelle telle que $V(\boldsymbol{e}) = \boldsymbol{I}\sigma_e^2$. $\boldsymbol{X}$, $\boldsymbol{Z}$ et $\boldsymbol{W}$ sont des matrices d'incidences.

Pour le single-step, le même modèle animal est utilisé, mais la matrice d'apparentement $\boldsymbol{A}$ est remplacée par la matrice $\boldsymbol{H}$, qui inclut à la fois l'apparentement observé dans le pédigrée et reporté dans la matrice $\boldsymbol{A}$, et l'apparentement observé grâce aux génotypes et reporté dans la matrice $\boldsymbol{G}$. L'inverse de la matrice $\boldsymbol{H}$ s'écrit de la façon suivante (Legarra *et al*. 2009, Christensen et Lund 2010):

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

La matrice $\boldsymbol{A}$ est en effet scindée de la façon suivante :

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} \end{pmatrix}$$

Les chevaux 2 sont les chevaux génotypés, et les chevaux 1 sont les chevaux non-génotypés. L'estimation des valeurs génétiques a été réalisée avec le logiciel BLUPf90 (Mistzal *et al*. 2002).

Critère de validation

Le critère utilisé pour comparer la sélection classique et la sélection génomique en une étape est la corrélation entre les valeurs génétiques des candidats (estimées sans utiliser les performances réalisées après leur année de naissance) et leurs performances réalisées. Les chevaux candidats ayant pour la majorité une carrière d'au moins 2 ans, ils ont plusieurs gains annuels. Les résultats seront donc présentés suivant différents angles (année de la compétition ou âge lors de la performance). Les gains annuels ont été préalablement corrigés pour les effets fixes et les effets d'environnement permanent.

5.4.2. Résultats : comparaison de l'évaluation classique et de l'évaluation génomique en une étape

Le Tableau 5.3 présente les coefficients de corrélation entre les valeurs génétiques et les logarithmes des gains annuels obtenus par année de compétition, ainsi que les coefficients de régression associés. Le coefficient de corrélation le plus faible est obtenu pour l'année 2008, et les coefficients les plus élevés pour les années 2009 et 2010. L'avantage de l'évaluation génomique en une étape (single-step) est faible pour la plupart des années de performances : l'écart entre les coefficients de corrélation est de 0.02 à 0.03 en faveur de la sélection génomique en une étape pour toutes les années de performance, à l'exception de l'année 2011 où l'écart est de 0.09 en faveur de l'évaluation génomique en une étape. Les différences entre les coefficients de régression obtenus par les 2 modèles sont un peu plus variables (entre 0.06 et 0.28), et les coefficients de régression sont plus proches de 1 avec l'évaluation classique. Ces résultats ne mettent pas en évidence un avantage clair de l'évaluation génomique en une étape par rapport à l'évaluation classique.

Tableau 5.3 : Coefficients de corrélation et de régression entre les valeurs génétiques des candidats estimées avec un modèle animal classique (BLUP) ou une évaluation génomique en une étape (single-step) et le logarithme de leurs gains annuels corrigés, présentés par année de performance. Pour chaque ligne les meilleurs résultats sont en gras.

		Coefficient de corrélation		Coefficient de régression	
Année de compétition	**Effectif**	**BLUP**	**Single-step**	**BLUP**	**Single-step**
2008	62	0.19	**0.23**	0.39	**0.45**
2009	93	0.51	**0.53**	**1.22**	1.29
2010	83	0.49	**0.51**	**1.16**	1.22
2011	77	0.31	**0.40**	**0.89**	1.17
2012	73	0.38	**0.41**	**1.00**	1.10

Le Tableau 5.4 présente les coefficients de corrélation et de régression des deux modèles en fonction de l'âge auquel les gains annuels ont été touchés. Quand les résultats sont présentés de cette

manière, l'évaluation classique obtient des résultats un peu meilleurs que l'évaluation génomique pour les gains obtenus à 4 ans (+0.05 pour le coefficient de corrélation). A 6 ans les deux méthodes donnent les mêmes résultats. A 5 ans et à 7 ans, les résultats de l'évaluation génomique en une étape sont un peu meilleurs mais l'avantage par rapport à l'évaluation classique reste faible (+0.02 et +0.03 de corrélation). Pour les gains obtenus à 8 ans, l'avantage de l'évaluation génomique est plus net : +0.13 pour le coefficient de corrélation comparé à l'évaluation classique. Pour tous les âges de performance à l'exception des gains à 8 ans, le coefficient de régression est plus proche de 1 quand l'évaluation classique est utilisée.

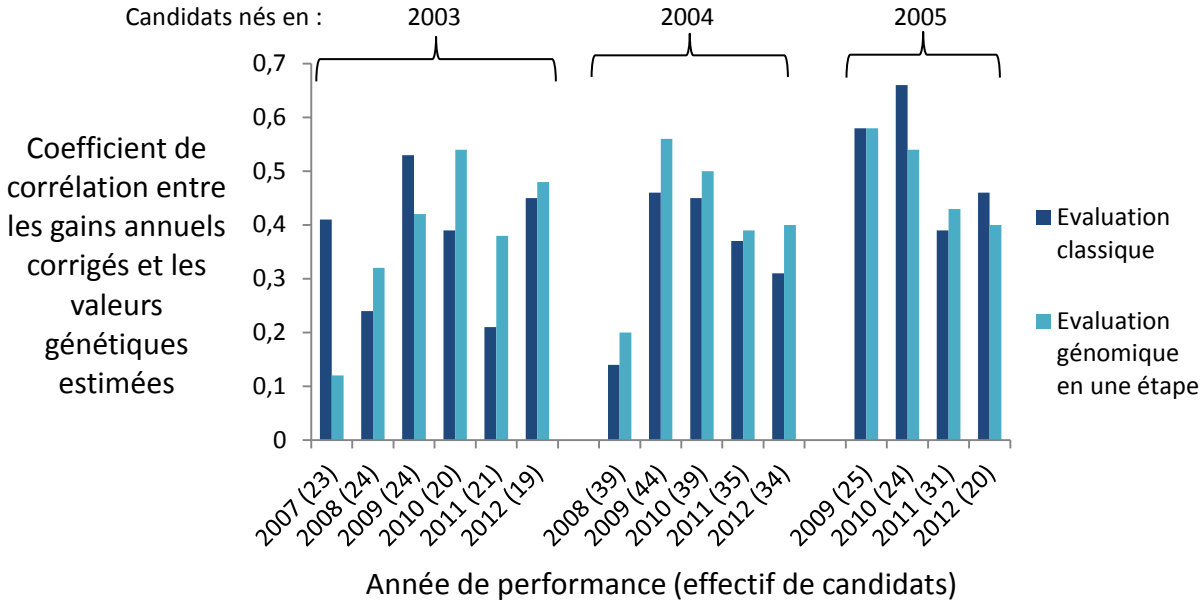
Tableau 5.4 : Coefficients de corrélation et de régression entre les valeurs génétiques des candidats estimées avec un modèle animal classique (BLUP) ou une évaluation génomique en une étape (single-step) et leurs gains annuels corrigés, présentés par âge au moment de la performance. Pour chaque ligne les meilleurs résultats sont en gras.

		Coefficient de corrélation		Coefficient de régression	
Âge auquel le gain a été reçu	**Effectif**	**BLUP**	**Single-step**	**BLUP**	**Single-step**
4 ans	87	**0.33**	0.28	**0.65**	0.55
5 ans	92	0.46	**0.49**	**1.05**	1.14
6 ans	84	**0.46**	**0.46**	1.06	**1.05**
7 ans	75	0.40	**0.42**	**1.08**	1.22
8 ans	55	0.26	**0.39**	**0.77**	**1.13**

La Figure 5.24 présente les coefficients de corrélation et de régression détaillés par année de naissance des candidats et par année d'obtention du gain annuel. On peut voir que pour les candidats nés en 2004, les résultats de l'évaluation génomique en une étape sont toujours meilleurs que ceux de l'évaluation classique. Il n'y a pas de lien entre l'âge à laquelle la performance est réalisée et la corrélation de cette performance avec la valeur génétique : par exemple à 4 ans le coefficient de corrélation est élevé pour les candidats nés en 2005, faible pour les candidats nés en 2004, et dépend de la méthode pour les candidats nés en 2003. On peut remarquer que les coefficients de corrélations obtenus pour les candidats nés en 2005 sont globalement plus élevés que ceux des candidats des autres groupes. Ces candidats sont ceux qui ont la plus grande population de référence (977 chevaux génotypés), cependant l'effet de la taille de la population de référence sur le coefficient de corrélation ne se vérifie pas clairement en comparant les candidats nés en 2003 et en 2004, alors qu'il y a 46 individus en plus dans la population de référence des candidats nés en 2004.

Les caractéristiques des pères des candidats des trois échantillons (nés en 2003, 2004 et 2005) ont été comparées. Les étalons avaient un nombre moyen de fils autour de 350, dont 5 à 6 étaient génotypés. La durée de carrière en compétition moyenne des étalons était de 8 ans pour les candidats nés en 2003, de 10 ans pour ceux de 2004 et de 9 ans pour ceux de 2005, et le CD moyen des pères était de 0.77, 0.71 et 0.78 respectivement pour ces 3 groupes. Par ailleurs, les candidats nés en 2003, 2004 et 2005 avaient respectivement des nombres moyens de demi-frères génotypés dans la population de référence de 3.6, 2.7 et 3.6. Ces informations ne permettent pas d'expliquer les différences de résultats observées entre les groupes de candidats.

Figure 5.24 : Comparaison des coefficients de corrélation entre les valeurs génétiques des candidats estimées avec un modèle animal classique (BLUP) ou une évaluation génomique en une étape (single-step) et leurs gains annuels corrigés, présentés par année de naissance des candidats et par année de performance.



Les résultats obtenus en ne gardant que les candidats dont les pères ont un CD supérieur à 0.60 sont présentés dans le Tableau 5.5. Les coefficients de corrélation et de régression sont améliorés de la même façon pour l'évaluation classique et l'évaluation génomique : le fait de n'utiliser que les candidats issus des étalons les mieux connus améliore la précision globalement, sans agrandir l'écart entre la sélection génomique et la sélection classique.

Tableau 5.5 : Coefficients de corrélation entre les valeurs génétiques des candidats estimées avec un modèle animal classique (BLUP) ou une évaluation génomique en une étape (single-step) et leurs gains annuels corrigés, présentés par année de performance. Les pères des candidats ont un CD minimum de 0.60. Pour chaque ligne les meilleurs résultats sont en gras.

		Coefficient de corrélation		Coefficient de régression	
Année de compétition	**Effectif**	**BLUP**	**Single-step**	**BLUP**	**Single-step**
2008	46	0.31	**0.34**	0.54	**0.55**
2009	72	0.56	**0.57**	**1.33**	1.39
2010	64	0.56	**0.58**	**1.34**	1.42
2011	55	0.32	**0.47**	**0.82**	1.24
2012	56	0.41	**0.46**	**1.05**	1.20

5.5. Conclusion de la comparaison de l'évaluation classique et de l'évaluation génomique pour les performances des chevaux de CSO

Malgré l'utilisation d'une évaluation génomique en une étape permettant d'utiliser tous les gains annuels réalisés depuis 1985, les corrélations obtenues entre les performances et les valeurs génétiques estimées ne sont pas de beaucoup supérieures à celles obtenues avec la sélection classique. Il arrive que les résultats de l'évaluation basée sur le pédigrée soient meilleurs que ceux

utilisant les génotypes, mais les causes de ce résultat n'ont pas été identifiées. Pour un même groupe de candidats, les corrélations entre leurs valeurs génétiques et leurs performances variaient suivant leur âge au moment de la performance. Ces variations étaient différentes d'un groupe à l'autre, et nous n'avons pas non plus d'explication pour cette variabilité. Une limite de cette étude est certainement la taille des échantillons de validation, limitée à quelques dizaines de chevaux.

Dans ce chapitre, nous avions signalé lors de l'estimation des paramètres génétiques nos doutes concernant les solutions estimées pour les groupes de parents inconnus, et les effets estimés pour la race ou le type de cheval. En effet, il était vérifié que la race ou le type de cheval n'étaient pas des caractéristique héritables à cause des nombreux croisements autorisés par les stud-books du Selle Français et des chevaux de sport étrangers. Pourtant, la modification des solutions des groupes de parents inconnus suivant l'utilisation ou non de ces effets races ou type de cheval avais mis en évidence que ces solutions dépendaient en partie des races/types des chevaux issus des groupes de parents inconnus. On peut donc supposer qu'utiliser des groupes de parents inconnus et prendre en compte la race dans le modèle pourrait améliorer la précision de l'évaluation.

En utilisant les groupes de parents inconnus dans l'évaluation génomique en une étape (résultats non présentés), nous avons obtenu des valeurs génétiques visiblement fausses. Notre cas n'est pas isolé : Mistzal *et al*. (2013) posent la question de l'utilisation de groupes de parents inconnus dans l'évaluation génomique en une étape. En effet, la construction du pédigrée avec des groupes de parents inconnus est telle que ces groupes sont considérés indépendants. Cette hypothèse peut être vraie quand par exemple deux groupes de parents inconnus sont utilisés pour des animaux nés de parents inconnus dans deux races bien distinctes, non-apparentées pour les années de naissance où les groupes sont utilisés. Pour les chevaux de CSO ce n'est clairement pas le cas : le but des différents groupes était de tenir compte du progrès génétique dans la population des chevaux de sport. Nous avons vu que les groupes de parents inconnus rassemblaient tous plusieurs races ou types de chevaux, et que 2 groupes successifs pouvaient donner naissance à des proportions similaires des différentes races ou des différents types de chevaux. Or, la matrice d'apparentement génomique $\boldsymbol{G}$ est capable de capturer cet apparentement entre les groupes qui se répercute sur leurs descendants. A l'inverse, dans la matrice $\boldsymbol{A}$ les groupes de parents inconnus sont supposés non-apparentés. L'information apportée par les marqueurs sur l'apparentement en bout de pédigrée est mal exploitée, ce qui peut conduire à estimer des valeurs génétiques biaisées. Nous nous sommes trouvées dans ce cas de figure et avons choisi une des solutions proposée par Mistzal *et al*. (2013) qui consiste simplement à ne pas utiliser de groupes de parents inconnus dans l'évaluation génomique en une étape.

Une amélioration des estimations des valeurs génétiques sera peut-être possible en utilisant des méta-fondateurs (Legarra *et al*. 2015). Cette méthode développée récemment consiste à ajouter au pédigrée un ou plusieurs individus fictifs constituant un ou des pools de gamètes. Ce ou ces pools de gamètes sont ceux de la (ou des) population(s) dont sont issus les groupes de parents inconnus. Il est ainsi possible de prendre en compte dans la matrice $\boldsymbol{A}$ le fait que les groupes de parents inconnus peuvent être apparentés.

Enfin, une dernière limite de cette étude est que les chevaux génotypés étaient en majorité des étalons, c'est-à-dire des individus choisis pour la reproduction. Dans le chapitre suivant, la sélection génomique a été testée dans une population de chevaux moins sélectionnée.

6. Test de l'évaluation génomique chez les chevaux d'endurance

6.1. Introduction

Des chevaux d'endurance ont été génotypés dans le cadre du projet GENENDURANCE. L'objectif de ce projet est de déterminer les caractéristiques phénotypiques (biochimie, métabolisme, morphologie et allures) et génétiques favorables ou défavorables à la performance en course d'endurance. Dans le chapitre précédent, la sélection génomique a été testée dans une population de chevaux sélectionnée, les individus génotypés étant en grande majorité des étalons. Or Bijma *et al*. (2012) ont montré que la précision de l'évaluation génomique peut être plus faible dans une population sélectionnée comparée au résultat qu'on obtiendrait dans une population peu sélectionnée. Les chevaux du projet GENENDURANCE étaient en majorité des performeurs, et non des reproducteurs déjà sélectionnés sur leurs performances. Dans le cadre de ma thèse, j'ai donc utilisé les génotypes des chevaux d'endurance pour tester la sélection génomique sur cette population, moins sélectionnée que la population des chevaux de CSO. Après une présentation des données disponibles et de leurs particularités, je décrirai la méthode utilisée et les résultats obtenus.

6.2. Matériel & Méthodes

6.2.1. Candidats potentiels

Pour tester l'intérêt de l'évaluation génomique pour la sélection des chevaux d'endurance, je disposais des génotypes de 783 chevaux, parmi lesquels 597 étaient performeurs en endurance. 72% des chevaux étaient de race Pur-Sang Arabe. Ces 597 chevaux étaient inclus dans un pédigrée de 9 481 chevaux. Parmi les chevaux performeurs et génotypés, seulement 45 avaient leur père génotypé et performeur également. En effet il y avait 17 pères performeurs dans le pédigrée, et seulement 12 qui étaient également génotypés. Comme il est préférable que les pères des candidats soient génotypés, le faible nombre de pères génotypés et performeurs limitait le nombre de candidats pour la validation dans cette étude.

6.2.2. Marqueurs

Les marqueurs utilisés sont issus de la puce Illumina 74K. Un tri a été effectué sur les SNPs de façon à avoir un taux de typage parmi les chevaux génotypés d'au moins 80%, une fréquence de l'allèle minimum supérieure à 1%, et le respect de l'équilibre d'Hardy-Weinberg. Sur ces critères, 56 200 SNPs ont été retenus.

6.2.3. Phénotypes : des moyennes de performances corrigées pour les effets fixes

L'indexation classique est réalisée à partir de trois critères décrits dans le chapitre bibliographique sur les chevaux. Les indices sont obtenus en utilisant un modèle multi-caractères combinant la vitesse, le résultat en fin de course (finissant, abandon ou élimination) et la distance courue. Le résultat en fin de course est codé par une variable discrète, et son utilisation dans un modèle multi-caractères nécessite d'utiliser un logiciel adapté au modèle complexe actuellement utilisé, qui inclut notamment des effets courses. La sélection génomique aurait été compliquée à implémenter dans ce logiciel. Nous avons donc choisi d'utiliser les indices de performances disponibles : ils peuvent être assimilés à des moyennes de performances déjà corrigées pour les effets fixes. Les indices de performance prennent en compte toutes les performances réalisées pendant la carrière du cheval, et

représentent la somme de l'effet génétique et d'environnement permanent aux différentes performances d'un même cheval. Le modèle peut s'écrire de la façon suivante :

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zc} + \boldsymbol{e},$$

avec $\boldsymbol{y}$ les performances, $\boldsymbol{b}$ les effets fixes, $\boldsymbol{c}$ les effets « chevaux » (génétique et environnement permanent) et $\boldsymbol{e}$ la résiduelle. $\boldsymbol{X}$ et $\boldsymbol{Z}$ sont des matrices d'incidence. Ce modèle permet de calculer les effets chevaux estimés $\hat{c}$, qui sont accompagnés d'un coefficient de précision (CP). La répétabilité des performances est calculée par $r = \frac{\sigma_c^2}{\sigma_c^2+\sigma_e^2}$. Soit y^* les performances corrigées pour les effets fixes. On peut alors écrire le modèle équivalent :

$$\boldsymbol{y}^* = \boldsymbol{Zc} + \boldsymbol{\varepsilon},$$

avec $\boldsymbol{\varepsilon}$ tel que $V(\boldsymbol{\varepsilon}) = \Delta\sigma_e^2$, où Δ est une matrice diagonale dont les termes diagonaux valent $1/w_i$, avec $w_i = CP_i(1-r)/(r(1-CP_i))$. J'ai donc estimé les valeurs génétiques des candidats à partir des indices de performance corrigés pour les effets fixes $\boldsymbol{y}^*$ et pondérés par les poids ρ_i tels que $\rho_i = CP_i(1-r)/(r-CP_i h^2)$. Les indices de performance avaient été évalués soit en uni-caractère, soit en multi-caractères en tenant compte des corrélations phénotypiques entre les 3 indices. Les paramètres génétiques des trois indices de performances sont indiqués dans le Tableau 6.1.

Tableau 6.1 : Paramètres génétiques des pseudo-phénotypes utilisés pour l'estimation des valeurs génétiques (vitesse, code d'état en fin de course et distance)

Caractère	Evaluation uni-caractère		Evaluation multi-caractères	
	héritabilité	Répétabilité	héritabilité	Répétabilité
Distance	0.19	0.36	0.22	0.41
Code d'état en fin de course	0.10	0.18	0.10	0.19
Vitesse	0.10	0.18	0.10	0.18

6.2.4. Modèles utilisés

Le modèle utilisé pour l'estimation des valeurs génétiques est :

$$\boldsymbol{y}^* = \mathbf{1}\mu + \boldsymbol{Za} + \boldsymbol{e},$$

avec $\boldsymbol{y}^*$ les indices de performance corrigés pour les effets fixes, μ la moyenne, $\mathbf{1}$ un vecteur de 1, $\boldsymbol{a}$ les valeurs génétiques, et $V(\boldsymbol{\delta}) = \Delta\sigma_e^2 + \boldsymbol{I}\sigma_p^2$ une matrice diagonale de termes diagonaux σ_e^2/ρ_i. Les ρ_i sont les poids des performances de chaque cheval i, ils permettent de pondérer les performances en fonction du nombre de courses courues. $\boldsymbol{Z}$ est une matrice d'incidence.

J'ai comparé l'évaluation classique et l'évaluation génomique en appliquant un BLUP et un GBLUP à ces indices de performance. Dans le BLUP, $\boldsymbol{a}$ est tel que $Var(\boldsymbol{a}) = \boldsymbol{A}\sigma_a^2$, où $\boldsymbol{A}$ est la matrice d'apparentement basée sur le pédigrée. Dans le GBLUP, $\boldsymbol{a}$ est tel que $V(\boldsymbol{a}) = \boldsymbol{G}\sigma_a^2$, avec $\boldsymbol{G}$ la matrice d'apparentement basée sur les marqueurs. L'estimation des valeurs génétiques a été réalisée avec le logiciel BLUPf90 (Misztal *et al*. 2002).

6.2.5. Critère de validation

Les critères utilisés pour comparer la sélection classique et la sélection génomique sont, pour chacun des caractères pris séparément, les coefficients de corrélation et de régression entre les valeurs

génétiques estimées et les pseudo-performances correspondantes, c'est-à-dire les pseudo-phénotypes obtenus en pondérant les indices de performance. Pour la validation, les valeurs génétiques des candidats ont été estimées en supprimant au préalable leurs pseudo-performances.

6.3. Résultats

Les coefficients de corrélation entre les valeurs génétiques estimées avec le BLUP et le GBLUP et les pseudo-performances (indices de performances estimés en uni-caractère ou en multi-caractères et pondérés) sont présentés dans le Tableau 6.2. Les coefficients de corrélation obtenus en utilisant les indices calculés avec des modèles uni-caractère sont très mauvais (proches de 0.10 ou négatif), alors que les coefficients de corrélation sont d'environ 0.30 à 0.40 quand les indices ont été calculés avec un modèle multi-caractères. Ce résultat est observé aussi bien pour l'évaluation classique que pour l'évaluation génomique : l'intérêt de calculer un indice de performance multi-caractères est donc confirmé. Quand la pseudo-performance est un indice multi-caractères, le coefficient de corrélation le plus élevé est toujours obtenu avec l'évaluation génomique. Cependant la supériorité du GBLUP sur le BLUP est faible : les écarts sont de 0.01 à 0.03. Les coefficients de régression correspondants sont proches de 1, indiquant des estimations non-biaisées.

Tableau 6.2 : Pour les caractères vitesse, code d'état en fin de course et distance : coefficients de corrélation et de régression obtenus entre les valeurs génétiques estimées avec un BLUP ou un GBLUP et les indices de performance pondérés des 45 candidats.

Caractère	Modèle pour l'indice de performance	Coefficient de corrélation		Coefficient de régression	
		BLUP	GBLUP	BLUP	GBLUP
Vitesse	Uni-caractère	0.11	0.10	0.41	0.35
	Multi-caractères	0.28	**0.29**	1.06	**0.96**
Code d'état en fin de course	Uni-caractère	-0.04	0.10	-0.26	0.46
	Multi-caractères	0.35	**0.38**	**0.95**	0.84
Distance courue	Uni-caractère	0.07	0.02	0.21	0.06
	Multi-caractères	0.40	**0.43**	**0.94**	0.85

6.4. Conclusion

La comparaison de l'évaluation classique et de l'évaluation génomique pour les performances des chevaux d'endurance a montré que la précision était un peu plus élevée avec le GBLUP qu'avec le BLUP. Cependant, les différences entre les coefficients de corrélation mesurés entre les performances et les valeurs génétiques estimées des candidats sont faibles. Les corrélations sont meilleures quand les indices de performances ont été évalués en multi-caractères.

Une limite de cette étude était la faible taille de l'échantillon, et surtout le faible nombre de pères à la fois génotypés et performeurs dont dépendait le nombre de candidats. L'importance d'avoir des pères génotypés avait été évoquée dans le chapitre bibliographique portant sur la sélection génomique.

Le chapitre suivant aborde une application différente de la sélection génomique comparée aux cas déjà testé. Il s'agit pour l'évaluation du Trotteur Français de vérifier l'intérêt de la prise en compte d'un gène à effet majeur identifié.

7. Comparaison de l'évaluation classique et de l'évaluation génomique chez le Trotteur Français en présence d'un gène à effet majeur

7.1. Introduction de l'article

La plupart des chevaux n'ont que trois allures naturelles : le pas, le trot, et le galop. Le pas est une allure marchée dans laquelle le cheval a toujours trois pieds au sol. L'avancée et le posé d'un postérieur entraine l'avancée et le posé de l'antérieur du même côté. Au trot, allure sautée, le cheval se déplace par bipède diagonal, avec une phase de projection entre le posé de chaque bipède. Le galop est une allure sautée non-symétrique : le cheval pose le postérieur droit, puis le bipède diagonal droit, puis l'antérieur gauche avec ensuite une phase de projection (ou postérieur gauche, bipède diagonal gauche et antérieur droit puis projection suivant le pied sur lequel le cheval galope). Deux allures supplémentaires existent notamment chez le cheval Islandais : l'amble et le tölt. A l'amble, le cheval se déplace par bipède latéral avec une phase de projection entre le posé de chaque bipède, comme au trot. Au tölt, le cheval se déplace également par bipède latéral mais a toujours au moins un pied en contact avec le sol. Andersson *et al*. (2012) ont étudié le déterminisme génétique du nombre d'allures chez des chevaux Islandais ayant les trois allures naturelles et le tölt, ou bien les trois allures naturelles et le tölt et l'amble. Leur étude leur a permis de mettre en évidence une mutation dans le gène *DMRT3*, qui conduit à une modification de la locomotion des chevaux. Chez les chevaux Islandais ayant 5 allures la fréquence du génotype muté (A) est de 99%, et chez les chevaux Islandais ayant 4 allures cette fréquence est de 65%. Plus généralement, les races chevaux ayant des allures supplémentaires ont une fréquence de l'allèle A dans la population proche de 1, tandis que chez les races de chevaux n'ayant que les 3 allures naturelles cette fréquence est nulle : tous les chevaux sont porteurs du génotype CC, C étant l'allèle non-muté (Promerová *et al*. 2014). Les grandes races de trotteurs ont également été étudiées. Il est apparu dans leurs résultats que chez le Standardbred élevé aux USA, qui court au trot et à l'amble, la fréquence de l'allèle A est de 1. Chez le Standardbred suédois, quelques sujets portent encore l'allèle non-muté, avec un effet négatif sur leurs performances en courses (Jäderkvist *et al*. 2014). Or, chez le Trotteur Français, la fréquence de l'allèle A n'est que de 77%. Compte-tenu de la sélection réalisée dans cette race depuis plusieurs générations, il était étrange que l'allèle C n'ait pas été éliminé.

Ricard (2015) a étudié plus précisément l'effet de la mutation de *DMRT3* chez le Trotteur Français, en utilisant le SNP *BIEC2-620109* dont le déséquilibre de liaison avec la mutation est de 0.90 (aux allèles C et A de *DMRT3* correspondent respectivement les allèles C et T au SNP utilisé). Cette étude a montré que l'allèle C a un effet négatif sue la capacité des chevaux à réussir le test de qualification : les chevaux de génotype CC ont 20% de chance de réussir, contre 48% pour les TT. Cependant les résultats obtenus sur les gains étaient moins tranchés. Comparé au génotype TT, le génotype CC avait toujours un effet défavorable sur les gains obtenus à 2 ans, 3 ans, 4 ans et 5 ans et plus. En revanche, pour le génotype CT, à 4 ans le génotype CT avait le même effet sur les gains que le génotype TT, et à 5 ans et plus l'effet de CT sur le gain était très significativement supérieur à l'effet de TT. Cette supériorité du génotype CT pour les gains tardifs n'est pas un artefact, car la même étude a montré par simulation qu'en l'absence de cet effet positif sur les gains tardifs du génotype CT la fréquence de l'allèle C devrait être plus faible que celle observée actuellement. L'effet du génotype au SNP *BIEC2-620109* étant démontré sur les performances du Trotteur Français, il nous a semblé

intéressant de tester l'effet de sa prise en compte dans l'estimation des valeurs génétiques des Trotteurs Français.

Pour cette étude nous avons utilisé 684 Trotteurs Français, génotypés avec la puce Illumina 50K et en majorité issu du projet GENEQUIN. Les SNPs ont été soumis à des tests de qualité : fréquence de l'allèle minimum d'au moins 1%, taux de typage des SNPs d'au moins 80%, et respect de l'équilibre d'Hardy-Weinberg. Finalement, 41 711 SNPs ont été retenus. Dans cet échantillon, la fréquence des génotypes au SNP *BIEC2-620109* était de 56% pour les TT, 39% pour les TC et 5% pour les CC. Les performances étudiées étaient la capacité à se qualifier ainsi que les gains à différents âges (2 ans, 3 ans, 4 ans et 5 ans et plus). Ces performances étaient déjà corrigées pour les effets fixes. L'héritabilité de la qualification était de 0.56, et pour les gains l'héritabilité variait entre 0.25 et 0.32 suivant l'âge. Le modèle utilisé pour l'évaluation des valeurs génétiques était un modèle multi-caractères incluant la qualification et les gains, permettant de prendre en compte les corrélations génétiques entre ces caractères. Les valeurs génétiques ont été estimées avec un BLUP classique basé sur le pédigrée (6 599 chevaux), et un BLUP génomique utilisant les génotypages. L'utilisation du génotype au SNP *BIEC2-620109* en effet fixe a été testée dans le modèle classique et dans le modèle génomique. Dans ce dernier cas, le génotype au SNP *BIEC2-620109* était supprimé du fichier des génotypes. Trois cent chevaux étaient performeurs génotypés et fils de chevaux performeurs génotypés, ils pouvaient être candidats pour la validation. Afin de garder un nombre suffisant d'individus dans la population de référence, la validation a été réalisée en trois fois avec trois échantillons disjoints de 100 chevaux, en supprimant du fichier de performances les phénotypes des candidats. Les coefficients de corrélation et de régression entre les valeurs génétiques estimées pour chacun des caractères et les performances réalisées ont ensuite été calculés pour l'ensemble des 300 chevaux. Nous avons également réalisée la validation sur 50 groupes de 126 chevaux candidats tirés au hasard, afin de vérifier la variabilité des coefficients de corrélation et de régression due à l'échantillonnage. Plus de détails sur les données et les méthodes sont donnés dans l'article suivant, accepté dans Journal of Animal Science.

Should we use the SNP linked to *DMRT3* in genomic evaluation of French trotter?[1]

S. Brard[*†‡2], A. Ricard[§ #]

[*]INRA, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), 31326 Castanet-Tolosan, France, [†]Université de Toulouse, INP, ENSAT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), 31326 Castanet-Tolosan, France, [‡]Université de Toulouse, INP, ENVT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), 31076 Toulouse, France, [§]INRA, UMR 1313, 78352 Jouy-en-Josas, France, [#]IFCE, Recherche et Innovation, 61310 Exmes, France.

[1]Acknowledgments: This work was funded by the SelGen metaprogramme and by the French Institute for Horses and Riding (Institut Français du Cheval et de l'Equitation, IFCE).

[2]Corresponding author: sophie.brard@toulouse.inra.fr

ABSTRACT: An A/C mutation responsible for the ability to pace in horses was recently discovered in the *DMRT3* gene. It has also been proved that allele C has a negative effect on trotter performances. However, in French trotter (**FT**), the frequency of allele A is only 77% due to an unexpected positive effect of allele C in late-career FT performances. Here we set out to ascertain whether the genotype at SNP *BIEC2-620109* (linked to *DMRT3*) should be used to compute estimated breeding values (**EBVs**) for FT. We used the genotypes of 630 horses, with 47,711 SNPs retained. The pedigree counted 6,599 horses. Qualification status (trotters need to complete a 2000-m race within a limited time to begin their career) and earnings at different ages were pre-corrected for fixed effects and evaluated with a multi-trait model. Estimated breeding values were computed with and without the genotype at SNP *BIEC2-620109* as a fixed effect in the model. The analyses were performed using pedigree only via BLUP and using the genotypes via GBLUP. The genotype at SNP *BIEC2-620109* was removed from the file of genotypes when already taken into account as a fixed effect. Alternatively, three groups of 100 candidates were used for validation. Validations were also performed on 50 random-clustered groups of 126 candidates and compared against the results of the 3 disjoint sets. For performances on which *DMRT3* has a minor effect, the coefficients of correlation were not improved when the genotype at SNP *BIEC2-620109* was a fixed effect in the model (earnings at 3- and 4-years). However, for traits proven as strongly related to *DMRT3*, the accuracy of evaluation was improved, increasing +0.17 for earnings at 2-years, +0.04 for earnings at 5-years and older, and +0.09 for qualification status (with the GBLUP method). For all traits, the bias was reduced when the SNP linked to *DMRT3* was a fixed effect in the model. This work finds a clear rationale for using the genotype at *DMRT3* for this multi-trait evaluation. Genomic selection seemed to achieve better results than classic selection.

Key words: *DMRT3*, genomic selection, horse, major gene, single nucleotide polymorphism, trotter

INTRODUCTION

Andersson et al. (2012) discovered a major gene affecting locomotion in horses. A stop mutation in *DMRT3* is strongly associated with ambling gaits, which are very comfortable gaits that some breeds naturally have or are easily able to learn due to a genetic predisposition, in addition to the

usual gaits (walk, trot and canter). This mutation is caused by a single base change: the wild-type allele C is replaced by the mutant allele A. Promerová et al. (2014) found that the mutated allele was also fixed in many breeds dedicated to trot races but missing in breeds selected for gallop races. A feature of trot races is that horses that break stride are disqualified, and so trotters have been selected on their ability to trot easily at high speed. The mutated allele was proved to have a positive effect on racing performances in Swedish standardbred trotter and is fixed in American standardbred trotter. Nevertheless, in French trotter (**FT**), the frequency of the mutation is only 77%. Ricard (2015) studied the effect of the genotype at SNP *BIEC2-620109* (linked to *DMRT3,* C-C, A-T) and confirmed positive effects of the mutation on ability to trot easily and on earnings through most of the career of FT. Nevertheless, the greater earnings are obtained in prestigious events that are mainly raced at 5-years and older by only few horses. The wild-type allele in heterozygotes had a positive and highly significant effect on these late earnings ($P<0.001$), which justified its frequency in a long-term-selected breeds like FT. Our objective is to ascertain whether using the genotype at this same SNP in high linkage disequilibrium with *DMRT3* would enable a better estimation of breeding values for performances by FT in harness races. We also considered the effect of the method used to obtain the relationship matrix: expected relationships based on pedigree or realized relationships based on genotypes. This work therefore reports the results of a first attempt at genomic selection in FT.

MATERIALS AND METHODS

Phenotypes

The races studied in this work are harness races, in which the horse pulls an ultra-light roadster called a sulky. Performances in these races may be analyzed through different traits. The first step in the career of a trotter is to pass a qualification test to gain the right to compete in races. The qualification test consists in a 2000-m race that has to be completed within a limited time allocation, which can change every year depending on improvement of the racing performances of the whole population, and which is also dependent on the age of the horse. About 40% of a given generation passes the test. Qualification is therefore an important trait for two reasons: first, it is a relatively highly heritable (Table 1) and early trait, and second, it means that horses that will race are a selected sample of their generation. The subsequent career of a trotter can then be considered at three stages. The first stage is racing as a 2-year-old. It is an early stage as only 20% of the horses racing at 3-years started at 2-years. The second stage, racing as a 3 or 4-year-old, is the crux of a trotter's career. Few of them will go on to make the third stage, i.e. racing at 5-years and older: a ⅓ of the horses stop racing after 4 years old, and another ⅓ will stop each year that follows. Horses win money prizes depending on their ranking. Most of the time, only the first 7 horses receive a price. The first horse earns the half of the total prize, then the second receives the half of the remaining money, and so on down to the seventh horse. The next 9 horses are ranked but do not earn money. According to Thiruvenkadan et al. (2009), performances in trot racing can be studied using the logarithm of annual earnings divided by the annual number of finished races (**LnE**), assuming a horse can be disqualified if it breaks stride. Here we study LnE based on these three stages: early earnings at 2-years, peak of career at 3-years and 4-years separately, and late earnings which will include all prizes at 5-years and older. Heritability of earnings is moderate (around 0.30, Table 1). The records for all these traits were provided by the Society for the Promotion of French Horse Breeding (**SECF**–*Société d'encouragement à l'élevage du cheval français*) and by the French Institute for Horses and Riding (**IFCE**–*Institut français du cheval et de l'équitation*). Although records were available for all horses that took part in French races from

1996 to 2011, the data were truncated: records were only kept for horses born between 1994 and 2008 to have a sufficient amount of information on all horses.

Horses

The blood samples of 623 horses had been previously collected for a genome-wide association study on osteochondrosis in FT (Teyssèdre et al. 2012). The horses were recruited between 2008 and 2010 at French Center for Imaging and Research on Equine Locomotor Disorders (**CIRALE**) and at a few veterinary clinics. These horses are not exactly a random sample of the population, and they have globally better performances than other trotters (79% of them were qualified whereas only about 40% of a generation usually makes the cut). Another 59 horses were genotyped, giving a total of 682 genotyped horses available. Finally 630 were retained, based on the availability of their records for racing performances: they were born between 1996 and 2008 and their performances were recorded between 1996 and 2011. Of the retained sample, 41% were females and 61 horses were sires. Looking in from an alternative perspective, 300 genotyped horses had their sire genotyped. As sires are selected on own performances, their presence in the sample also explains the better performances recorded. The 630 trotters were included in a pedigree of 5699 horses.

Genotypes

The genotypes were obtained using the Illumina Equine SNP50 BeadChip. A quality control of SNP genotypes based on minimum allele frequency (≥1%), genotype assessment rate (≥80%) and Hardy-Weinberg equilibrium (*P*-value for the test >10E-08) was performed, and 41,711 SNPs were retained. The chip marker that had the strongest linkage disequilibrium (**LD**) with the mutation identified in *DMRT3* (Andersson et al. 2012) was the SNP *BIEC2-620109*. Promerová et al. (2014) estimated a LD of 0.91 between this SNP and the *DMRT3* mutation in a population of 2,749 horses including 59 FT. The C allele at SNP *BIEC2-620109* is associated with the C allele in the mutation, whereas the T allele of the SNP is associated with the A allele. The frequencies for the 3 genotypes among the retained horses were 56% for TT, 39% for TC and 5% for CC.

Statistical models

Qualification status and LnE at different ages were studied by Ricard (2015) in the same population with the objective of assessing the effect of SNP *BIEC2-620109* on performances in trot racing. Ricard et al. (2015) used the following model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} ,$$

where **y** is the performance vector, **b** the fixed effect vector that combined gender and year of birth, **a** the vector of random polygenic values, and **e** the vector of residuals. $V(\mathbf{a})=\mathbf{A}\sigma^2_a$, where **A** is the relationship matrix, and $V(\mathbf{e})=\mathbf{I}\sigma^2_e$. **X** and **Z** are incidence matrices. Heritability, genetic correlation and residual correlation between traits had been estimated in a multi-trait model using more than 173,000 FT, with about 64,000 of them qualified. Here we used performances pre-corrected for fixed effects according to the estimations obtained with this model. This same approach has already been used by Pribyl et al. (2010). We realized a multi-trait estimation of breeding values for qualification status and LnE at different ages to exploit the genetic correlations between traits (Table 1). Note that qualification status was first a binary variable (0: unqualified, 1: qualified) but the correction for fixed effects turned it into a continuous trait. Qualification status is important in the multi-trait evaluation because it allows the use of horses that are not qualified and have not yet posted earnings. It has been demonstrated that it is important to use those horses without earnings in the estimation of breeding values to reduce the bias (Klemetsdal 1992, Árnason 1999).

Our first objective was to assess whether the genotype of the SNP linked to *DMRT3* should be used to compute the EBVs, as it has been shown that QTL are a useful source of information for animal selection (Soller & Beckmann, 1983). Our second objective was to check whether genomic selection should be preferred over pedigree-based selection. For comparison of classic vs. genomic selection, two methods were used to compute the EBVs: a BLUP and a genomic BLUP (GBLUP). On the one hand, EBVs were calculated using the pedigree, and on the other hand genomic EBVs were calculated using the relationship matrix deduced from the horse genotypes. The corresponding statistical model is:

$$\mathbf{y^*} = \mathbf{1}\mu + \mathbf{Zg} + \boldsymbol{\varepsilon},$$

where **y*** is the vector of pre-corrected performances of the 630 horses, **1** is a vector of ones, μ is the overall mean, **g** is a random vector of additive genetic values, and **ε** is a vector of residuals. In BLUP, $\text{Var}(\mathbf{g})=\mathbf{A}\sigma_g^2$,where **A** is the pedigree-based relationship matrix, and σ^2_g is the additive genetic variance. In GBLUP, $\text{Var}(\mathbf{g})=\mathbf{G}\sigma_g^2$, and **G** is the genomic relationship matrix as defined by VanRaden (2008). **Z** is an incidence matrix. This model allows the comparison of pedigree-based and marker-based evaluations. To test the value of using the SNP linked to *DMRT3* for the estimation of breeding values, we modified the model by adding a fixed effect. The model therefore became:

$$\mathbf{y^*} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where **β** is the vector of the fixed effect of the 3 genotypes, and **W** is the incidence matrix for the horse's genotype at this SNP. When this model was coupled to the GBLUP method, the SNP *BIEC2-620109* was removed from the file of genotypes that was used to compute **G**, therefore the genotype at this SNP was used only once in the model. EBVs were computed using BLUPF90 (Misztal *et al*. 2002).

Validation

Comparison between methods was based on cross-validation. The file was divided into two populations: reference population and candidate population. Estimated breeding values were calculated including performances of the reference population only, and the validation criteria were then based on relationships between EBVs and performances of the candidate population. For the accuracy, we used the correlation coefficient between the candidates' EBVs and their realized performances. For the bias, we used the regression of the realized performances of candidates on their EBVs. When the model included the SNP *BIEC2-620109*, the total EBV of a horse was the solution for the animal effect summed to the solution for its genotype at *SNP BIEC2-620109*.

The candidates had to meet the following requirements:

- have at least one record (not necessarily a record for each of the traits studied in the multi-trait analysis),
- be genotyped,
- be the son of a genotyped sire, itself having at least one record,
- not have any progeny with records.

Because the performances were pre-corrected for fixed effects including year of birth, this information was not used to design the reference and validation samples.

A total 300 horses were potential candidates. If EBVs had been computed simultaneously for all of them, the reference population would have been reduced to 330 individuals. To use enough information to compute candidates' EBVs, a 3-fold cross-validation was used. Candidates were randomly divided into 3 non-overlapping training datasets of equal size (100 horses), and each group of candidates was then evaluated one by one, with the two other groups included in the reference population. This method has already been used in dairy cattle (Luan et al. 2009), beef cattle, (Saatchi et al. 2011), and pine (Resende et al. 2012). The advantage of this method was to guarantee that EBVs of the trotters in the validation sample would be computed using a reference population counting at least five times more individuals, in line with recommended practice (Legarra et al. 2008). Accuracy and bias were then computed for the three groups of candidates together.

To quantify the standard errors due to sampling of the results of the 3-fold cross-validation, EBVs were also computed for 50 random groups of 126 candidates, yielding 4/5 genotyped horses with performances in the reference population and 1/5 in the validation population. In this case, the groups were obviously overlapping, and candidates had several EBVs as they could be picked from in several datasets for validation. We therefore computed the accuracy and the bias separately for each of the 50 validation datasets. The accuracy and bias obtained with the 3-fold cross-validation were compared against the distributions of the correlation and regression coefficients obtained for the 50 overlapping datasets.

RESULTS

Validation on three non-overlapping datasets

The accuracy and the bias computed for the 300 candidates evaluated in the 3-fold cross validation are shown in Tables 2 and 3. All the horses of the validation population had a record for qualification status but may have had earnings for only some of the years of their career or no earnings at all. Therefore, the correlation and regression coefficients were computed for different numbers of candidates depending on the number of trotters that truly had a record for the trait. Early and late earnings (LnE at 2-years and LnE at 5-years and older) were the traits that had lower number of candidates, at 38 and 83, respectively. Logarithm of annual earnings divided by the annual number of finished races at 3- and 4-years had higher numbers of candidates (at 134 and 171, respectively).

Early and late earnings and qualification status achieved greater accuracies when the SNP *BIEC2-620109* was included as a fixed effect in the model (Table 2). The superiority of this model was more obvious for LnE at 2-years (+0.21 for BLUP and +0.17 for GBLUP) and qualification status (+0.14 for BLUP and +0.09 for GBLUP) than for LnE at 5-years and older (+0.01 for BLUP and +0.04 for GBLUP). This result was consistent with the significant effect of the SNP linked to *DMRT3* on those traits as originally evidenced by Ricard (2015). With this model, GBLUP provided slightly greater accuracies for LnE at 2-years (+0.04) and 5-years and older (+0.01), whereas for the qualification status BLUP gave marginally better results than GBLUP (+0.02).

For LnE at 3- and 4-years, greater accuracies were achieved when the SNP *BIEC2-620109* was not a fixed effect of the model, particularly for LnE at 4-years (+0.12). This is consistent with previous results of Ricard et al. (2015): the SNP linked to *DMRT3* is thought to have a less significant effect on these traits, so the accuracy is not improved when the SNP *BIEC2-620109* is added in the model. For these two traits, the superiority of using GBLUP over BLUP was ascertained (+0.08 for accuracy for LnE at 3-years, +0.06 at 4-years).

For traits that achieved better accuracy when the SNP linked to *DMRT3* was in the model (early and late earnings, qualification status), the regression coefficients closest to 1 were also obtained

with this model (Table 3). For LnE at 3- and 4-years, the regression coefficients were nearly unbiased when the SNP linked to *DMRT3* was in the model, although their coefficient of correlation was not improved.

Validation on fifty overlapping groups of candidates

Figure 1 shows the distribution of the accuracy achieved by the 50 random-clustered validation datasets. Like for the non-overlapping groups, the number of effective candidates for each trait in one group was different depending on availability of records. The numbers of effective candidates are shown in Table 4. The distributions of accuracy for LnE at 2-years and qualification status clearly showed the superiority of the model including the SNP linked to *DMRT3* as a fixed effect. For LnE at 5-years and older, the superiority of this model was less patent, although the distributions for the model including the SNP *BIEC2-620109* visibly achieved slightly better accuracies. For LnE at 3-years, the distributions of accuracies for both models were very similar, whereas the GBLUP approach achieved greater accuracies than the BLUP method. For LnE at 4-years, the shapes of the distributions of accuracy did not single out a better model or a better method. Note in Fig. 1 that the average values of accuracies among the 50 validation datasets were very close to the accuracies computed on the 3 non-overlapping validation datasets.

Figure 2 charts the distributions of bias. For all traits, the distributions were closest to 1 when the model included the SNP linked to *DMRT3* as a fixed effect. This result was consistent with observations on the validation based on the 3 non-overlapping groups. For these models including SNP *BIEC2-620109*, GBLUP seemed to more often achieve unbiased estimations than BLUP for LnE at 2-years, LnE at 5-years and older, and qualification status. The BLUP method seemed to achieve regression coefficients closer to 1 for LnE at 3-years. For LnE at 4-years, neither BLUP nor GBLUP emerged as superior in terms of the bias. Once again, these results were consistent with the difference between methods evidenced with the 3 non-overlapping groups.

DISCUSSION

The results of our study were consistent with the results of Ricard (2015). On one hand, accuracy was improved for LnE at 2-years, LnE at 5-years and older, and qualification status when the genotype at SNP *BIEC2-620109* was included in the model. Ricard (2015) had found that the genotype with the SNP linked to *DMRT3* had a very highly significantly different effect ($P<0.001$) on these three traits. In our estimation of genotype effects with the GBLUP method based on the 50 validation groups, the difference of effect of genotype CT compared to TT was very highly significant for LnE at 2-years ($P<0.001$, -0.84 phenotypic standard deviations) and highly significant for LnE at 5-years and older ($P<0.01$, +0.44 phenotypic standard deviations) and the difference of effect of genotype CC compared to TT was highly significant for qualification status ($P<0.01$, -0.70 phenotypic standard deviations). On the other hand, the accuracy for LnE at 3- and 4-years was slightly lower when the SNP linked to *DMRT3* was included in the model. This could be due to a weak effect of SNP *BIEC2-620109* on these traits: no useful information is added for the computation of the corresponding EBVs.

For qualification status, the accuracy reached with the GBLUP method coupled with the model with the SNP linked to *DMRT3* may be considered low (r=0.25) as this trait has the greatest h^2 (0.56) and the greatest number of horses with recorded performances. Qualification status is a discrete trait, so we had to use averaged performances for the multi-trait evaluation, and the correlation coefficient that we used might not be the most suitable way of measuring accuracy for this kind of phenotype. However, it nevertheless made it possible to observe an increase in accuracy due to the addition of the effect of SNP *BIEC2-620109* in the model, even if we can suppose that accuracies for this trait are generally underestimated with the coefficient of correlation used.

Logarithm of annual earnings divided by the annual number of finished races at 5-years and older obtained quite high accuracies in both models with and without the SNP linked to *DMRT3*. This result was unexpected according to h^2 and the number of horses that do have performances for this trait. These high values may be due to selection.

Even if adding qualification status in the multi-trait evaluation is supposed to reduce bias, high regression coefficients were obtained for LnE at 2- and 5-years and older. This may be due to relative overselection of those horses that do have performances for these traits.

The distributions obtained for accuracy and bias with 50 randomly-selected validation datasets showed that very different values could be obtained for both accuracy and bias depending on the group of candidates, which illustrates the standard error of accuracy and bias of our cross-validation. A genetic evaluation associating genomic EBVs and the effect of a major gene has already been performed in dairy cattle by Hayr et al. (2013) via a single-trait evaluation, whereas we worked on a multi-trait evaluation. Hayr et al. computed EBVs for fat yield with the genotype at *DGAT1* in the model, and found that the better results were obtained when the major gene was considered as a fixed effect, with no improvement of accuracy when the genotype of the major gene was imputed for all animals. Zhang et al. (2010) developed a different method: they realized a weighted genomic evaluation with a trait-specific marker-derived relationship matrix, and achieved better accuracy for traits depending on a major gene compared with BLUP and GBLUP. In their method, the marker-derived relationship matrix is different from the realized relationship matrix used in GBLUP, because a greater weight is attributed to loci depending on the genetic variance they explain. Zhang et al. did not compare their weighted GBLUP to the method used here (a GBLUP with the major gene as a fixed effect in the model), but there is every reason to believe that this method could be very valuable in trotters given how *DMRT3* has a strong effect on performances. Nevertheless, as discussed earlier, the evaluation of breeding values of trotters is a multiple-trait model that should include qualification status in addition to LnE, and it considers the earnings obtained in each year of a horse's career. As the effect of SNP *BIEC2-620109* is different for each of the traits that we evaluated simultaneously, this would imply deriving a weighted relationship matrix for each of the traits and to implement a method able to take into account these different matrices, likely resulting in longer computation times.

For now, we recommend using a model with the SNP linked to *DMRT3* as a fixed effect. This model achieved better accuracies for LnE at 2-years, LnE at 5-years and older, and qualification status. The drop in accuracy when the SNP linked to *DMRT3* was added in the model was low for LnE at 3

years but higher for LnE at 4-years. However, the model including the SNP linked to *DMRT3* as a fixed effect had the advantage of being less biased than the model without the SNP. With the model including the SNP linked to *DMRT3*, we recommend using GBLUP, as it yields greater accuracies than BLUP for all traits except qualification status which was slightly more accurate with BLUP. A combination of a GBLUP approach with a model including the effect of the major gene thus looks a good compromise for estimating breeding values in FT. The way to use these breeding values remains open to discussion. Synthetic indexes may be produced with different weights for each trait depending on the breeder's objectives. It would be possible to select horses with a large weight on qualification status and LnE at 3- and 4- years old to produce horses that would be easy to qualify. We could also imagine an index with a higher weight on performances at 5-years and older to select horses that would be harder to qualify but expected to perform better later in their career in prestigious high-pay-off races. This type of selection would entail breeding for the C allele. As the CC genotype has a negative effect on all traits, a strategy should be defined to keep this allele in heterozygotes horses while minimizing the frequency of the homozygotes in the population.

LITERATURE CITED

Andersson, S. L., M. Larhammar, F. Memic, H. Wootz, D. Schwochow, C. J. Rubin, K. Patra, T. Arnason, L. Wellbring. G. Hjälm, F. Imsland, J. L. Petersen, M. E. McCue, J. R. Mickelson, G. Cothran, N. Ahituv, L. Roepstorff, S. Mikko, A. Vallstedt, G. Lindgren, L. Andersson, and K. Kullander. 2012. Mutations in *DMRT3* affect locomotion in horses and spinal circuit function in mice. Nature 488(7413):642-646. doi:10.1038/nature11399

Árnason, T. 1999. Genetic evaluation of Swedish standard-bred trotters for racing performance traits and racing status. J. Anim. Breed. Genet. 116:387–398. doi:10.1046/j.1439-0388.1999.00202.x

Hayr, M. K., M. Saatchi, D. L. Johnson and D. J. Garrick. 2013. Increasing the accuracy of genomic predictions of fat yield in New Zealand Holstein Friesians using *DGAT1* genotypes. J. Dairy Sci. 96(Suppl. 1):618-619 (Abstr.)

Klemetsdal, G. 1992. Estimation of genetic trend in racehorse breeding. Acta Agric. Scand., Sect. A, Anim. Sci. 42:226–231. doi:10.1080/09064709209410133

Legarra, A., C. Robert-Granié, E. Manfredi and J. M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611-618. doi:10.1534/genetics.108.088575

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen and T. H. E. Meuwissen. 2009. The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. Genetics 183:1119-1126. doi:10.1534/genetics.109.107391

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet and D. H. Lee. 2002. BLUPF90 and related programs (BGF90). In: Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France, XXVIII:1–2.

Pribyl, J., V. Rehout, J. Citek and J. Pribylova. 2010. Genetic evaluation of dairy cattle using a simple heritable genetic ground. J. Sci. Food Agric. 90:1765-1773. doi: 10.1002/jsfa.4041

Promerová, M., L. S. Andersson, R. Juras, M. C. T. Penedo, M. Reissmann, T. Tozaki, R. Bellone, S. Dunner, P. Hořín, F. Imsland, P. Imsland, S. Mikko, D. Modrý, K. H. Roed, D. Schwochow, J. L. Vega-Pla, H. Mehrabani-Yeganeh, N. Yousefi-Mashouf, E. G. Cothran, G.Lindgren and L. Andersson.2014. Worldwide frequency distribution of the '*Gait keeper*' mutation in the *DMRT3* gene. Anim. Genet. 45:274-282. doi:10.1111/age.12120

Resende, M. F. R., Jr., P. Muñoz, M. D. V. Resende, D. J. Garrick, R. L. Fernando, J. M. Davis, E. J. Jokela, T. A. Martin, G. F. Peter and M. Kirst. 2012. Accuracy of genomic selection methods in a standard data set of Loblolly Pine (*Pinus taeda* L.). Genetics 190:1503-1510. doi:10.1534/genetics.111.137026

Ricard, A. 2015. Does heterozygosity at the *DMRT3* gene make French trotters better racers? Genet. Sel. Evol. 47:10. doi:10.1186/s12711-015-0095-7

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. W. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, S. Bauck, B. Woodward, J. C. M. Dekkers, R. L. Fernando, R. D. Schnabel, D. J. Garrick and J. F. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet. Sel. Evol. 43:40. doi:10.1186/1297-9686-43-40

Soller, M. and J. S. Beckmann. 1983. Genetic polymorphism in varietal identification and genetic improvement. Theor. Appl. Genet. 67:25-33. doi:10.1007/BF00303917

Teyssèdre, S., M. C. Dupuis, G. Guérin, L. Schibler, J. M. Denoix, J. M. Elsen and A. Ricard. 2012. Genome-wide association studies for osteochondrosis in French Trotter horses. J. Anim. Sci. 90:45-53. doi:10.2527/jas.2011-4031

Thiruvenkadan, A. K., N. Kandasamy and S. Panneerselvam. 2009. Inheritance of racing performance of trotter horses: an overview. Livest. Sci. 124:163-181. doi:10.1016/j.livsci.2009.01.010

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423. doi:10.3168/jds.2009-2061

Zhang, Z., J. Liu, X. Ding, P. Bijma, D. J. de Koning and Q. Zhang. 2010. Best Linear Unbiased Prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLos ONE 5(9): e12648. doi:10.1371/journal.pone.0012648

Table 1. Heritability (diagonal), genetic correlation (upper triangle) and residual correlation (lower triangle) for logarithm of annual earnings divided by the annual number of finished races (**LnE**) at different ages, and qualification status as per Ricard et al. (2015)

	LnE at				Qualification
Trait	2-years	3-years	4-years	≥ 5-years	Status
LnE at 2-years	0.28	0.85	0.76	0.56	0.48
LnE at 3-years	0.29	0.32	0.91	0.81	0.61
LnE at 4-years	0.12	0.27	0.25	0.92	0.47
LnE at 5-years and older	0.14	0.23	0.41	0.26	0.44
Qualification status	0.00[1]	0.00[1]	0.00[1]	0.00[1]	0.56

[1] residual correlations between qualification status and LnE were fixed to 0

Table 2. Correlation coefficients between estimated breeding values and performances for the corresponding traits: logarithm of annual earnings divided by the annual number of finished races (**LnE**) at different ages or qualification status

Trait[1]	Effective number of candidates[2]	The SNP linked to *DMRT3* is not used		The SNP linked to *DMRT3* is a fixed effect	
		BLUP	GBLUP	BLUP	GBLUP
LnE at 2-years	38	0.19	0.27	0.40	0.44
LnE at 3-years	171	0.21	0.29	0.19	0.27
LnE at 4-years	134	0.28	0.34	0.22	0.26
LnE at 5-years and older	83	0.43	0.41	0.44	0.45
Qualification status	300	0.13	0.16	0.27	0.25

[1] Estimated breeding values were computed in a multi-trait analysis, with 4 combinations of models and methods, and for 3 non-overlapping validation datasets of 100 candidates each.

[2]Results are presented for all candidates pooled together. Qualification status was the only trait for which all candidates had a performance record. For LnE, the non-qualified horses had missing values and could not therefore be candidates.

Table 3. Regression coefficients of the performances on estimated breeding values for the corresponding traits: logarithm of annual earnings divided by the annual number of finished races (**LnE**) at different ages or qualification status

Trait	Effective number of candidates[1]	The SNP linked to *DMRT3* is not used		The SNP linked to *DMRT3* is a fixed effect	
		BLUP	GBLUP	BLUP	GBLUP
LnE at 2-years	38	1.70	1.55	1.39	1.44
LnE at 3-years	171	1.28	1.33	1.02	1.16
LnE at 4-years	134	1.67	1.65	1.04	1.06
LnE at 5-years and older	83	2.29	1.83	1.35	1.25
Qualification status	300	0.74	0.60	1.05	0.78

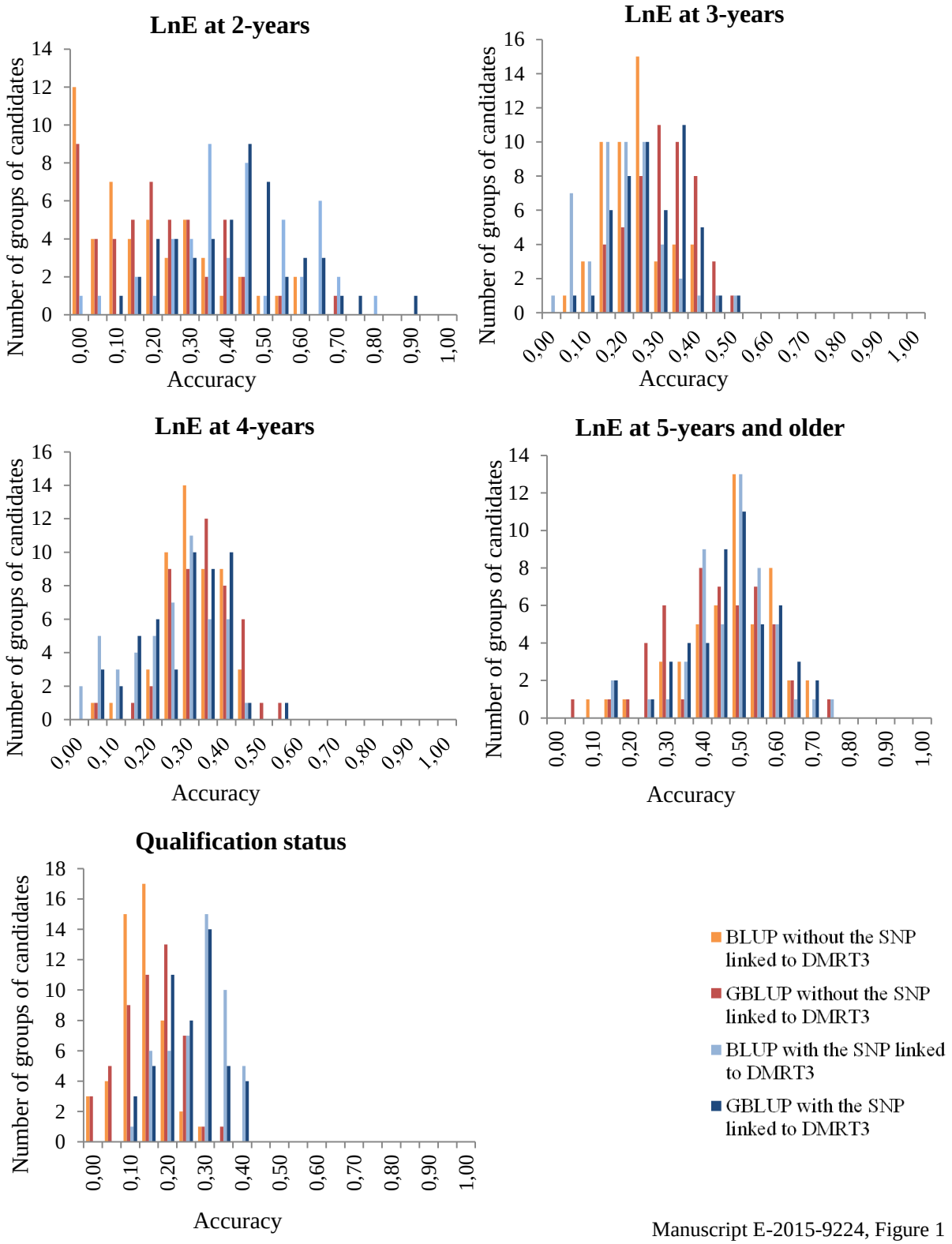
[1] Estimated breeding values were computed in a multi-trait analysis, with 4 combinations of models and methods, and for 3 non-overlapping validation datasets of 100 candidates each.

[2] Results are presented for all 300 candidates pooled together. Qualification status was the only trait for which all candidates had a performance record. For LnE, the non-qualified horses had missing values and could not therefore be candidates.

Table 4. Effective number of candidates in the 50 random-clustered validation datasets for the logarithm of annual earnings divided by the annual number of finished races (**LnE**) at different ages, and qualification status.

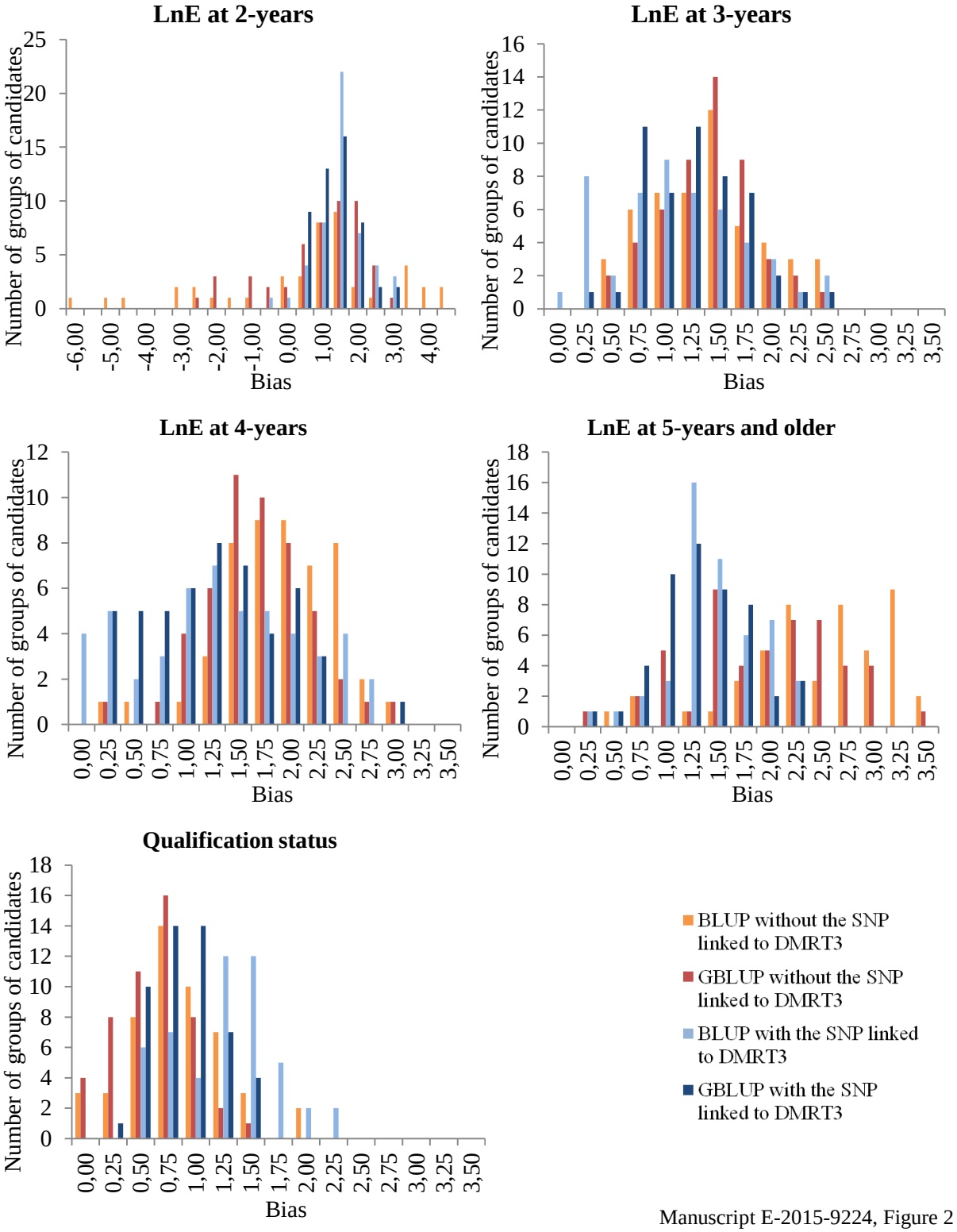
	Effective number of candidates		
Trait	Mean	Minimum	Maximum
LnE at 2-years	15.6	9	21
LnE at 3-years	71.9	64	78
LnE at 4-years	56.2	47	64
LnE at 5-years and older	34.6	28	43
Qualification status	126	126	126

Figure 1. Distributions of correlation coefficients between the estimated breeding values and the corresponding performances for 50 randomly-clustered validation datasets of 126 candidates each. Estimated breeding values were computed in a multi-trait evaluation including logarithm of annual earnings divided by the annual number of finished races (**LnE**) at different ages and qualification status, with four combinations of models and methods.



Manuscript E-2015-9224, Figure 1

Figure 2. Distributions of regression coefficients of the performances on estimated breeding values for the corresponding traits for 50 randomly-clustered validation datasets of 126 candidates each. Estimated breeding values were computed in a multi-trait evaluation including logarithm of annual earnings divided by the annual number of finished races (**LnE**) at different ages and qualification status, with four combinations of models and methods.



Manuscript E-2015-9224, Figure 2

7.2. Résumé des résultats et conclusion

Pour la qualification et les gains obtenus à 2 ans et à 5 ans et plus, les coefficients de corrélation les plus élevés ont été obtenus avec le GBLUP et le SNP *BIEC2-620109* en effet fixe dans le modèle. Les coefficients de corrélation pour ces caractères sont respectivement 0.25, 0.44, et 0.45. Ces résultats sont cohérents avec le fait que ces performances sont celles pour lesquelles le génotype CT a un effet significativement différent de celui du génotype TT, positif ou négatif. En effet pour les gains à 3 ans et à 4 ans l'effet de CT est peu ou non significativement différent de celui de TT. Pour le gain à 3 ans, le meilleur coefficient de corrélation est obtenu avec un GBLUP sans mettre le SNP lié à *DMRT3* en effet fixe (coefficient de 0.29, contre 0.27 quand le SNP lié à *DMRT3* est en effet fixe dans le GBLUP). Pour le gain à 4 ans, le meilleur coefficient de corrélation est également obtenu avec un GBLUP sans isoler le SNP lié à *DMRT3* (0.34, contre 0.26 quand le SNP est en effet fixe dans le GBLUP). Pour tous les caractères, et que la méthode soit un BLUP ou un GBLUP, l'ajout du SNP lié à *DMRT3* en effet fixe diminue le biais. La comparaison des résultats du BLUP et du GBLUP montre un avantage de l'évaluation génomique quasi systématique (de +0.01 à +0.08) à deux exceptions près : les gains tardifs quand le SNP lié à *DMRT3* n'est pas en effet fixe dans le modèle, et la qualification quand le SNP lié à *DMRT3* est en effet fixe dans le modèle. Cette étude montre donc l'intérêt de l'utilisation de la sélection génomique avec la prise en compte d'un gène majeur pour l'évaluation des Trotteurs Français. Nous avons pu vérifier par ailleurs grâce aux évaluations répétées 50 fois sur des candidats tirés au hasard que nos résultats obtenus sur 3 groupes de candidats disjoints étaient valides. Une limite de cette étude est peut-être le fait que l'échantillon ne soit pas complètement représentatif de la population. En effet, 40% des chevaux réussissent le test de qualification dans une génération, et notre échantillon comportait 79% de chevaux qualifiés. Nous avons cependant tenté de prendre en compte la sélection des chevaux en incluant des chevaux non-qualifiés dans l'évaluation. La collecte de données en cours pour le projet GENOTROT, visant à caractériser finement l'effet de *DMRT3* chez le Trotteur Français, devrait permettre de disposer d'un échantillon plus large et complet pour confirmer ces résultats.

La sélection génomique a donc été testée dans trois populations de chevaux. Le chapitre suivant compare les résultats obtenus au sein de chaque population et d'une population à l'autre en fonction des paramètres influant sur la précision de la sélection génomique. Pour cela, le nombre de segments indépendants a été calculé.

8. Estimation de M_e dans les populations de chevaux, comparaison de la précision des évaluations génomiques au regard de M_e et des autres paramètres identifiés

8.1. Introduction

Au cours de cette thèse, des évaluations génomiques ont été réalisées pour l'aptitude de chevaux pour le CSO, les courses d'endurance ou les courses au trot. Pour chacune de ces évaluations, la précision observée a été calculée comme la corrélation entre les valeurs génétiques estimées de candidats dont on supprimait les performances et leurs performances réalisées (performances propres ou moyennes corrigées de performances). Ces essais de sélection génomique ont été réalisés dans des conditions différentes dépendant du nombre de SNPs disponibles et retenus, de la taille de la population de référence, de l'héritabilité des caractères, et des modèles utilisés. Nous avons vu dans le chapitre 3 qu'un paramètre majeur de la précision des évaluations est le nombre de segments indépendants dans le génome M_e, paramètre que nous n'avons jusqu'à présent pas évoqué dans nos populations de chevaux. L'objet de cette partie étant de comparer les résultats des évaluations génomiques obtenus au regard des paramètres connus pour leur effet sur la précision de l'évaluation génomique, nous avons estimé M_e dans les différentes populations de chevaux utilisées à l'aide de deux méthodes. La première méthode est basée uniquement sur les génotypes et estime M_e à partir de moyennes de DL entre les SNPs, et la seconde méthode estime M_e à partir des différences entre les matrices d'apparentement classique et génomique. Après avoir présenté les méthodes d'estimation de M_e et discuté les résultats, nous comparons les précisions obtenues pour les évaluations génomiques réalisées dans différentes populations au regard des paramètres influant sur la précision et des modèles utilisés.

8.2. Matériel et méthodes

8.2.1. Données

Chevaux

Les données utilisées dans le cadre de cette étude sont les mêmes que celles qui ont servi au test de la sélection génomique pour différentes disciplines dans ma thèse. Les caractéristiques des échantillons (nombre d'individus génotypés et taille du pédigrée correspondant) sont rappelées dans le Tableau 8.1. Le nombre de chevaux génotypés variait d'environ 700 à 1 000 suivant la discipline, et la taille du pédigrée complet était comprise entre 6 700 et 10 400 chevaux. Pour l'estimation de M_e utilisant la matrice d'apparentement classique, plusieurs profondeurs de pédigrée ont été testées. On peut remarquer que chez le Trotteur Français, limiter le pédigrée à 5 générations remontées diminue les effectifs de 10%, alors que la diminution est de 50% pour les chevaux de CSO, et de 58% pour les Pur-Sang Arabes et croisés Arabes. Quand le pédigrée est limité à une profondeur de 3 générations, la diminution des effectifs par rapport au pédigrée complet est de 14% pour les Trotteurs Français, de 78% pour les chevaux de CSO, et de 83% pour les Pur-Sang Arabes et croisés Arabes.

Tableau 8.1 : Caractéristiques des populations de chevaux utilisées (SF : Selle Français, AA : Anglo-Arabe, SE : chevaux de Sport Etrangers)

		Effectifs			
Discipline	**Race des chevaux génotypés**	**Chevaux génotypés**	**Pédigrée complet**	**Pédigrée limité à 5 générations**	**Pédigrée limité à 3 générations**
CSO	69% SF, 13% AA, 18% SE	1 010	9 802	4 936	2 117
Endurance	Pur-Sang Arabes et croisés Arabes	779	10 412	4 388	1 772
Trot	Trotteur Français	682	6 707	6 050	5 792

Génotypes

Pour les chevaux de CSO et les Trotteurs Français, les génotypages ont été réalisés avec la puce Illumina 50K. Pour les chevaux Pur-Sang Arabes et croisés Arabes, les génotypes sont issus de la puce Illumina 74K. Les SNPs ont été soumis à des tests de qualité : respect de l'équilibre d'Hardy-Weinberg, taux de typage d'au moins 80%, fréquence de l'allèle minimum supérieure à 5%. Avec ces critères, le nombre de SNPs retenus était de 44 424 pour les chevaux de CSO, de 41 711 pour les Trotteurs Français, et de 56 200 pour les Pur-Sang Arabes et croisés Arabes.

8.2.2. Observation du déséquilibre de liaison dans les différentes populations de chevaux

Afin d'observer l'évolution du DL en fonction de l'augmentation de la distance entre SNPs, la moyenne des DL entre 2 SNPs regroupés en fonction de leur distance a été calculée. Pour cela, les marqueurs ont été regroupés par tranche de 0.005cM : un groupe avec tous les SNPs séparés de 0 à 0.005cM, un groupe avec les SNPs séparés de 0.005 à 0.010cM, un groupe avec les SNPs séparés de 0.010 à 0.015cM, etc. Le nombre de SNPs présents dans les classes de distance diminue avec l'augmentation de la classe : il existe pour presque tous les SNPs un SNP à moins de 0.005cM, mais ce n'est plus le cas pour les grandes distances qui ne concernent que les SNPs aux extrémités des chromosomes. Comme nous ne disposions des haplotypes que pour les chevaux de CSO, nous avons choisi de calculer le DL en calculant les corrélations entre les génotypes aux SNPs.

La population des chevaux de CSO génotypés comportant un sous-groupe de chevaux Anglo-Arabes, l'étendue du DL a été calculée dans l'ensemble de la population des chevaux de sport, mais aussi séparément pour les Anglo-Arabes d'une part et pour les Selle Français et les chevaux de sport étrangers d'autre part.

8.2.3. Calcul du nombre de segments indépendants dans le génome

Deux approches ont été utilisées pour calculer le nombre de segments indépendants M_e dans le génome.

Goddard *et al*. (2011) écrivent la valeur génétique g de la façon suivante : $g = \boldsymbol{Wu}$, avec w_{ik} le génotype du i^{ème} animal au k^{ème} marqueur, et $\boldsymbol{u}$ le vecteur contenant les effets des marqueurs. Ils calculent la variance des éléments non-diagonaux de $\boldsymbol{WW'}$ (l'élément ij de $\boldsymbol{WW'}$ étant la covariance des valeurs génétiques des individus i et j, et l'indice l correspondant aux QTL qui sont au nombre de Q):

$$V(w_i' w_j) = E(\sum w_{ik} w_{jk})(\sum w_{ik} w_{jk})$$

$$V(w_i' w_j) = E(\sum\sum(w_{ik}w_{jk})(w_{il}w_{jl}))$$

$$V(w_i' w_j) = \sum\sum E(w_{ik}w_{jk})(w_{il}w_{jl})$$

$$V(w_i' w_j) = \sum\sum E(w_{ik}w_{il})E(w_{jk}w_{jl})$$

$$V(w_i' w_j) = Cov(w_k, w_l)^2$$

$$V(w_i' w_j) = \sum\sum r_{kl}^2 2p_k(1-p_k)2p_l(1-p_l)$$

$$V(w_i' w_j) \cong \sum\sum 2p_k(1-p_k)2p_l(1-p_l)\sum\sum r_{kl}^2 / [Q(1-Q)]$$

$$V(w_i' w_j) = \sigma_g^4 \overline{r^2}$$

avec σ_g^2 la variance génétique additive et $\overline{r^2}$ la moyenne des déséquilibres de liaison entre les SNPs. Ils en concluent que $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{W}'/\sigma_g^2$ est une matrice d'apparentement dont les éléments non-diagonaux ont pour variance $\overline{r^2}$. De ce résultat, ils déduisent deux façons de calculer M_e. Tout d'abord, en considérant que les QTL sont M loci indépendants, alors $r_{kk}^2 = 1$ et $r_{kl}^2 = 1$:

$$M_e = \frac{1}{\overline{r^2}}$$

La deuxième méthode consiste à utiliser la variance des éléments non-diagonaux de $\boldsymbol{G}$ puisque Goddard *et al*. (2011) ont montré qu'ils ont pour variance $\overline{r^2}$. Cependant cette méthode de calcul a été obtenue en supposant les individus non-apparentés. Dans le cas où les animaux sont apparentés, Wientjes *et al*. (2013) proposent d'utiliser $\boldsymbol{D} = \boldsymbol{G} - \boldsymbol{A}$, la matrice $\boldsymbol{D}$ ayant été définie par Goddard *et al.* (2011) comme les déviations observées entre le pédigrée et l'apparentement réalisé. Finalement la seconde méthode utilisée pour calculer M_e est donc :

$$M_e = \frac{1}{Var(\boldsymbol{G} - \boldsymbol{A})}$$

Pour cette méthode basée sur les différences entre $\boldsymbol{G}$ et $\boldsymbol{A}$, pour chacune des populations différentes profondeurs de pédigrée ont été utilisées pour la matrice $\boldsymbol{A}$: le pédigrée a été utilisé dans son intégralité, ou bien limité à 5 ou 3 générations remontées.

8.3. Résultats

8.3.1. Etendue du DL dans les différentes populations de chevaux

La Figure 8.1 présente l'étendue du DL en fonction de la distance entre les marqueurs. On peut remarquer que le DL décroit assez rapidement. Pour les distances proches de 1cM, le DL dans l'ensemble de la population des chevaux de CSO est semblable au DL des Trotteurs Français. En revanche si on distingue les Selles Français et les chevaux de sport étrangers des Anglo-Arabes, le DL est un peu plus élevé chez les Anglo-Arabes et un peu plus faible chez les Selles Français et les chevaux de sport étrangers. Le DL le plus faible à 1cM est observé chez les chevaux Pur-Sang Arabes et croisés Arabes.

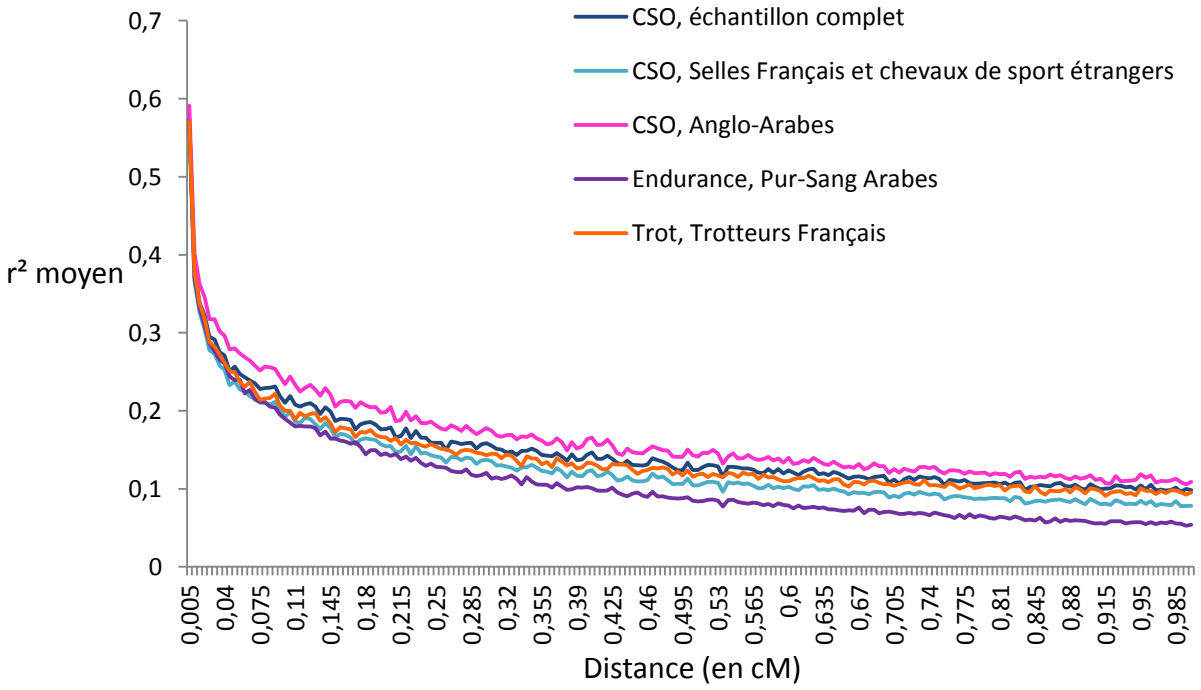
Quand la distance en cM diminue, le DL augmente un peu plus chez les chevaux Anglo-Arabes que chez les Selles Français et chevaux de sport étrangers. Finalement pour des distances très faibles le

DL est le même chez les chevaux Selles français et chevaux de sport étrangers, chez le Trotteur Français, et chez les Pur-Sang Arabes et croisés Arabes, le DL pour ces derniers ayant augmenté plus rapidement que celui des Trotteurs Français et Selles Français et chevaux de sport étrangers. Le DL chez les Anglo-Arabes est un peu plus élevé.

On peut remarquer que quand la distance entre marqueurs diminue le DL des Pur-Sang Arabes, Trotteurs Français, et Selles Français et chevaux de sport étrangers semble converger, alors que dans le même temps l'évolution du DL chez les Anglo-Arabes et les Pur-sang Arabes et croisés Arabes évoluent en parallèle jusqu'à environ 0.20cM.

Le DL sur de courtes distances est représentatif de l'histoire ancienne des populations, tandis que le DL sur de longues distances renseigne sur l'histoire récente des populations. Ainsi, le DL le plus faible est observé chez les Pur-Sang Arabes à 1cM, ce qui signifierait que ces chevaux sont ceux chez qui il y a le plus de diversité. Cependant sur de courtes distances le DL des Pur-Sang Arabes est similaire à celui des Trotteurs Français, une race dont le stud-book est très fermé. L'évolution parallèle du DL chez les Anglo-Arabes et chez les Pur-Sang Arabes pourrait témoigner d'une sélection similaire, l'Anglo-Arabe étant issu du croisement de Pur-Sang Arabes et de Pur-Sang Anglais, et le Pur-Sang Arabe étant élevé depuis plus longtemps que l'Anglo-Arabe. Le DL plus élevé chez les Anglo-Arabes que chez les Pur-Sang Arabes est en accord avec les résultats de McCue *et al*. (2012) qui obtiennent un DL plus élevé chez les Pur-Sang Anglais que chez les Pur-Sang Arabes. L'augmentation plus rapide du DL sur de courtes distances pour les Anglo-Arabes que pour les Pur-Sang Arabes peut témoigner d'une sélection plus importante sur le Pur-Sang Anglais observée par McCue *et al*. (2012).

Figure 8.1: Déséquilibre de liaison (r^2) moyen en fonction de la distance génomique exprimée en cM pour différentes populations de chevaux, pour des distances comprises entre 0 et 1cM



8.3.2. Nombre de segments indépendants dans les populations

Les différentes valeurs de M_e obtenues à partir de l'inverse du DL moyen sont présentées dans le Tableau 8.2. Le plus faible M_e est obtenu pour les Anglo-Arabes, ce qui est cohérent avec le déséquilibre de liaison observé dans la figure Figure 8.1. Le nombre de segments indépendant estimé pour les Trotteurs Français est quasiment deux fois supérieur à celui des Anglo-Arabes. Les valeurs obtenues pour les Selles Français et les chevaux de sport étrangers, ainsi que pour l'échantillon complet de CSO sont intermédiaires entre le nombre de segments indépendants chez les Anglo-Arabes et les Trotteurs Français.

Tableau 8.2 : Nombres de segments indépendants (M_e) estimés à partir de moyennes de déséquilibre de liaison (DL) dans différentes populations de chevaux (SF : Selle Français, SE : Sport Etranger, AA : Anglo-Arabe)

Discipline	Races	DL moyen	M_e
CSO	SF, SE et AA	3.78^{E}-02	359
	SF et SE	3.33^{E}-02	456
	Anglo-Arabe	1.22^{E}-02	226
Endurance	Pur-Sang Arabe et croisés Arabes	3.09^{E}-02	552
Trot	Trotteur Français	4.02^{E}-02	396

Le Tableau 8.3 présente les valeurs de M_e obtenues à partir de $Var(\boldsymbol{G}-\boldsymbol{A})$, pour différentes profondeur du pédigrée. Pour l'échantillon complet des chevaux de CSO, les Selles Français et chevaux de sport étrangers, et les Trotteurs Français, la limitation de la profondeur du pédigrée augmente le nombre de segments indépendants estimés de quelques dizaines à un peu plus d'une centaine (variation de 10 à 15%). En revanche, pour les Anglo-Arabes et les Pur-Sang Arabes ou croisés Arabes, le fait de tronquer le pédigrée diminue le M_e estimé de 29% pour les Anglo-Arabes, et de 37% pour les Pur-Sang-Arabes et croisés Arabes.

Quand le pédigrée est complet, le nombre le plus important de segments indépendants est estimé chez les Pur-Sang Arabes et croisés Arabes. L'échantillon complet des chevaux de CSO et les Selles Français et chevaux de sport étrangers obtiennent les M_e les plus faibles, et les Anglo-Arabes et les Trotteurs Français ont des valeurs de M_e intermédiaires. Quand la profondeur du pédigrée est limitée à 3 générations, ce classement change du fait de l'évolution différente du M_e en fonction de la profondeur du pédigrée : les Trotteurs Français ont le M_e le plus important, les Anglo-Arabes et les Pur-Sang Arabes et croisés Arabes ont les M_e les plus faibles, et l'échantillon complet des chevaux de CSO et les Selles Français et Selle étrangers obtiennent des M_e intermédiaires.

Tableau 8.3 : Nombres de segments indépendants (M_e) estimés à partir de la variance des éléments de G-A dans différentes populations (SF : Selle Français, SE : Sport Etranger, AA : Anglo-Arabe)

		M_e		
Discipline	**Races**	**Pédigrée complet**	**Limité à 5 générations**	**Limité à 3 générations**
CSO	SF, SE et AA	897	918	1 000
	SF et SE	982	1 011	1 126
	Anglo-Arabe	1 240	1 218	884
Endurance	Pur-Sang Arabe et croisés Arabes	1 467	1 291	923
Trot	Trotteur Français	1 236	1 268	1 358

8.4. Discussion sur l'estimation de M_e

8.4.1. Différence d'échelle des valeurs obtenues

Les valeurs de M_e calculées à partir du DL ou bien à partir de $Var(\boldsymbol{G}-\boldsymbol{A})$ sont très différentes. Quand le DL est utilisé le M_e estimé est de quelques centaines, alors que quand on utilise $Var(\boldsymbol{G}-\boldsymbol{A})$ M_e est plus proche de 1 000. Plusieurs formules déterministes ont été développées pour calculer M_e à partir de la longueur du génome (et éventuellement du nombre de chromosomes et de leur longueur moyenne) et de la taille efficace de la population N_e (Stam 1980, Goddard 2009, Hayes *et al*. 2009c, Goddard *et al*. 2011, Meuwissen *et al*. 2013). L'article de Leroy *et al.* (2013) porte sur le calcul de N_e dans différentes espèces basé sur les données du pédigrée, et pour les chevaux (mais aussi pour les autres espèces étudiées) des valeurs très différentes de N_e sont obtenues (d'une centaine à quasiment 2000) suivant la méthode et le type de données utilisées (nombres de reproducteurs mâles et femelles, variance du nombre de descendants, consanguinité…). Le nombre de segments indépendants M_e dépendant de la taille efficace de la population N_e, comme la valeur de N_e estimée est très dépendante de la méthode utilisée, on pouvait s'attendre à obtenir des résultats très variables également pour M_e. Par ailleurs, les formules déterministes proposant de calculer M_e à partir de N_e donnent elles aussi des valeurs de M_e très différentes pour une même valeur de N_e. Cependant les différences de résultats observées entre les deux méthodes, moyennes de DL ou variance des éléments de $\boldsymbol{G}-\boldsymbol{A}$, restent surprenantes dans la mesure où d'après Goddard *et al.* (2011) ces méthodes sont censées être équivalentes.

8.4.2. Des valeurs relatives différentes également

Le calcul de M_e basé sur des moyennes de DL ou bien sur la variance des éléments non-diagonaux de $\boldsymbol{G}-\boldsymbol{A}$ donne donc des valeurs différentes dans l'absolu. Si on compare les résultats non pas en terme de valeurs, mais en classant les populations en fonction de leur nombre de segments indépendants, les résultats diffèrent également. En effet, quand M_e est calculé à partir de moyennes de DL, ce sont les Pur-Sang Arabe et croisés Arabes qui ont le plus grand M_e, suivis par les Selles Français et les chevaux de sport étrangers et les Trotteurs Français. Les Anglo-Arabes ont le M_e le plus faible. Quand M_e est calculé à partir de la variance des éléments de $\boldsymbol{G}-\boldsymbol{A}$ en utilisant le pédigrée complet, on trouve également que les Pur-sang Arabes et croisés Arabes ont le M_e le plus élevé, suivis par les Trotteurs Français. En revanche les Anglo-Arabes ont un M_e intermédiaire, et le M_e le plus faible est obtenu par les Selles Français et chevaux de sport étrangers.

On a vu dans les résultats que limiter la profondeur du pédigrée à 3 générations modifie les résultats comparés à ceux obtenus avec les pédigrées complets. Si on compare les résultats obtenus en limitant le pédigrée à 3 générations à ceux obtenus avec des moyennes de DL, on trouve avec les 2 méthodes un M_e faible pour les Anglo-Arabes et intermédiaire pour les Selles Français et chevaux de sport étrangers. En revanche, les Pur-Sang Arabes et croisés Arabes qui avaient un M_e basé sur les moyennes de DL élevé obtiennent un M_e faible quand on utilise la variance des éléments de $\boldsymbol{G}-\boldsymbol{A}$. De même, les Trotteurs Français qui obtenaient un M_e intermédiaire avec les moyennes de DL ont un M_e élevé quand on utilise la variance des éléments de $\boldsymbol{G}-\boldsymbol{A}$.

8.4.3. Cohérence entre les 2 méthodes : la singularité des Anglo-Arabes et des Pur-Sang Arabes et croisés Arabes

Un point de cohérence entre les méthodes est peut-être la différence de résultats observés pour les Anglo-Arabes et pour les chevaux Pur-Sang Arabes et croisés Arabes. En effet, les courbes d'évolution

du DL en fonction de la distance entre les marqueurs sont parallèles pour ces 2 groupes de chevaux. Quand M_e est calculé à partir de la variance des éléments de $\boldsymbol{G} - \boldsymbol{A}$, la modification de la profondeur du pédigrée a le même effet sur ces 2 groupes de chevaux également. Cependant cette cohérence ne se retrouve pas dans les valeurs de M_e, car celles-ci sont très proches quand la variance des éléments de $\boldsymbol{G} - \boldsymbol{A}$ est utilisée, alors qu'avec les moyennes de DL il y a quasiment un facteur 2 entre le M_e des Anglo-Arabes et le M_e des Pur-Sang et croisés Anglo-Arabes.

8.4.5. Comparaison des résultats obtenus en se limitant aux populations utilisées pour tester la sélection génomique

Si on se limite aux populations qui ont été utilisées pour tester la sélection génomique, c'est-à-dire qu'on ne distingue pas les Anglo-Arabes de l'échantillon des chevaux de CSO, les résultats obtenus avec les 2 méthodes sont cohérents quand le pédigrée complet est utilisé. On trouve alors que ce sont les Pur-Sang Arabes et croisés Arabes qui ont le M_e le plus élevé, suivis par les Trotteurs Français, et les chevaux de sport français et étrangers obtiennent le M_e le plus faible, qu'on le calcule à partir de moyenne de DL ou de la variance des éléments de $\boldsymbol{G} - \boldsymbol{A}$. C'est le classement des M_e obtenus pour ces 3 populations que nous utiliserons pour comparer les résultats obtenus dans les différents tests de la sélection génomique, en tenant compte des autres paramètres influant sur la précision.

8.5. Discussion sur les précisions obtenues en fonction des différents paramètres

Le Tableau 8.4 rassemble les précisions des évaluations génomiques obtenues dans le cadre de cette thèse, et les différents paramètres pouvant affecter la précision de la sélection génomique. En plus de la précision observée, qui est dans notre cas la corrélation entre les valeurs génétiques estimées de chevaux dont on a supprimé les performances et leurs performances réalisées et corrigées pour les effets fixes, nous avons utilisé comme critère de comparaison la précision attendue. La précision attendue est calculée de la façon suivante :

$$r_{attendue} = \frac{cov(\hat{g},y)}{\sigma_{\hat{g}}\sigma_y} = h\sqrt{CD},$$

avec $\hat{g}$ les valeurs génétiques estimées, y les performances, $\sigma_{\hat{g}}$ et σ_y les écart-types correspondants, h la racine de l'héritabilité. CD correspond au CD moyen calculé à partir de la variance d'erreur de l'estimation des valeurs génétiques σ_r^2, obtenue par l'inversion des équations du modèle mixte : $CD = 1 - \sigma_r^2/\sigma_{\hat{g}}^2$.

8.5.1. Comparaison des résultats intra-échantillons

Pour l'échantillon des chevaux de CSO, la précision attendue a été calculée pour 3 sous-groupes : un par année de naissance des chevaux utilisés comme candidats pour tester la sélection génomique. Ces 3 groupes se différencient par une taille de la population de référence croissante (877 chevaux dans la population de référence pour les candidats nés en 2003, 921 pour ceux nés en 2004 et 977 pour ceux nés en 2005). La taille de la population de référence semble avoir un effet sur la précision attendue, car que celle-ci augmente également avec les années de naissance des chevaux : 0.33 pour ceux nés en 2003, 0.35 pour ceux nés en 2004, et 0.37 pour ceux nés en 2005. Pour les 3 groupes de candidats, la précision de la sélection génomique a été vérifiée en calculant la corrélation entre les valeurs génétiques estimées des candidats et le logarithme de leurs gains annuels reçus au cours de leur carrière corrigé pour les effets fixes. La précision variait en fonction de l'année de compétition,

mais l'augmentation globale de la précision observée avec l'augmentation de la taille de la population de référence est vérifiée : pour les chevaux nés en 2003 elle était comprise entre 0.12 et 0.54, entre 0.20 et 0.56 pour les chevaux nés en 2004, et entre 0.40 et 0.58 pour les chevaux nés en 2005.

Pour les chevaux d'endurance, l'estimation des valeurs génétiques avec une évaluation génomique a été testée pour trois caractères de performance : la vitesse, la distance et le code d'état en fin de course. Pour ces 3 caractères la taille de la population de référence, le nombre de marqueurs et le nombre de segments indépendants étaient les mêmes, en revanche l'héritabilité était de 0.22 pour la vitesse et de 0.10 pour les 2 autres caractères. Cette différence s'observe dans les précisions attendues, qui étaient respectivement de 0.18, 0.13 et 0.13, mais pas dans les précisions observées correspondantes qui étaient de 0.29, 0.43 et 0.68.

Pour les courses au trot, les valeurs génétiques étaient estimées pour 5 caractères dans un même échantillon, avec le même nombre de marqueurs, la même population de référence et le même nombre de segments indépendants. Les caractères étudiés avaient des héritabilités différentes, autour de 0.30 pour les logarithmes de gains, et de 0.56 pour la qualification. La précision attendue était de 0.37 pour la qualification, de 0.36 pour le gain à 2 ans, de 0.27 pour le gain à 5 ans, et de 0.23 et 0.22 pour les gains à 3 ans et à 4 ans. Le fait que le gain à 2 ans ait une précision attendue proche de celle de la qualification malgré une héritabilité plus faible peut être dû au modèle. En effet le SNP *BIEC2-620109*, lié au gène majeur *DMRT3*, a été utilisé en effet fixe dans le modèle. Or ce gène a un fort effet sur les gains précoces. Cette remarque est également valable mais dans une moindre mesure pour les gains tardifs. Les précisions observées sont plus élevées que les précisions attendues pour les gains, alors que la précision observée est plus faible que la précision attendue pour la qualification. Les précisions observées élevées peuvent être dues au fait que la population est sélectionnée.

8.5.2. Comparaison des résultats obtenus dans les différentes populations

On peut remarquer que malgré une héritabilité de 0.22, 56 200SNPs, et 552 individus dans la population de référence, la précision attendue pour la vitesse en endurance est de 0.18, alors que pour un caractères d'héritabilité proche, le gain en courses au trot à 4 ans (h^2=0.25), la précision attendue est un peu plus élevée (0.22) malgré un nombre de marqueurs plus faible (41 711 SNPs) et une population de référence plus petite (424 chevaux). La précision attendue un peu plus élevée malgré des paramètres apparemment plus défavorables pourrait être due au nombre de segments indépendants à estimer, plus faible chez le Trotteur Français (396) que chez les Pur-Sang Arabes et croisés Arabes (552).

Si on compare les résultats obtenus pour le gain en courses au trot à 3 ans (h^2=0.32) aux résultats obtenus pour le gain annuel en CSO (h^2=0.28), la précision attendue est meilleure pour le gain en CSO que pour le gain en courses au trot à 3 ans (autour de 0.35 contre 0.23). Ce résultat pourrait être principalement dû à la population de référence plus importante pour les chevaux de CSO (environ 900 chevaux, contre un peu plus de 400 pour les Trotteurs), car les nombres de SNPs utilisés et les nombres de segments indépendants à estimer sont du même ordre de grandeur dans les 2 cas.

En revanche pour le gain à 2 ans qui a la même héritabilité que le gain annuel en CSO, la précision attendue est aussi élevée que pour le gain annuel en CSO, alors que comme pour les gains en courses au trot à 3 ans la taille de la population de référence est quasiment moitié plus faible que pour le

CSO. Là encore les 2 caractères ont une héritabilité similaire, et les valeurs génétiques sont estimées avec quasiment le même nombre de SNPs, avec un nombre de segments indépendants proches également. Il est possible que le désavantage pour le gain à 2 ans en courses au trot de la population de référence plus petite soit compensé par la prise en compte du gène à effet majeur sur le gain précoce, ce qui permettrait d'atteindre une précision attendue aussi élevée qu'en CSO.

L'estimation du nombre de segments indépendants dans les différentes populations de chevaux dans lesquelles la sélection génomique a été testée était nécessaire pour discuter les précisions attendues et obtenues, car ce paramètres a un fort effet sur la précision. Avec deux méthodes supposées équivalentes, l'une utilisant des moyennes de déséquilibre de liaison entre SNPs, l'autre la variance des différences entre les coefficients non-diagonaux des matrices d'apparentement génomique et classique, nous avons obtenu des valeurs de M_e très différentes. Ces valeurs était cependant cohérentes si on classait les populations de chevaux sur lesquelles la sélection génomique a été testée en fonction de leur nombre de segments indépendants : avec les 2 méthodes, le M_e le plus faible était obtenu par les chevaux de CSO, le M_e le plus élevé par les chevaux d'endurance, et les trotteurs avaient un M_e intermédiaire. Néanmoins la variabilité des valeurs obtenues en fonction de la méthode utilisée est problématique dans la mesure où nous conseillons d'estimer ce paramètre au même titre que l'héritabilité afin d'en tenir compte dans la mise en place de l'évaluation génomique.

Les précisions attendues et obtenues pour l'évaluation génomique ont été comparées en fonction des paramètres connus pour leur effet sur la précision : nombre de segments indépendants, taille de la population de référence, nombre de marqueurs, héritabilité. Les comparaisons ont été réalisées au sein de chaque discipline, et entre disciplines quand les valeurs des paramètres le permettaient. Sur les comparaisons réalisées, les précisions attendues semblent globalement cohérentes avec les valeurs prises par les paramètres.

Tableau 8.4 : Paramètres et précision de la sélection génomique observée et attendue dans différentes populations de chevaux pour l'aptitude à la performance en CSO, en courses d'endurance et en courses au trot.

Discipline	Méthode	Type de données	Performance	Nombre de SNPs	Nombre de segments indépendants	Taille de la population de référence	héritabilité	Précision observée	Précision attendue	Biais
CSO	Single-step	Performances propres	Log(gain 2007)	44 424	359	877 (candidats nés en 2003)	0.28	0.12	0.33	0.16
			Log(gain 2008)					0.32		0.58
			Log(gain 2009)					0.42		0.95
			Log(gain 2010)					0.54		1.45
			Log(gain 2011)					0.38		1.23
			Log(gain 2012)					0.48		1.14
			Log(gain 2008)			921 (candidats nés en 2004)		0.20	0.35	0.39
			Log(gain 2009)					0.56		1.34
			Log(gain 2010)					0.50		1.08
			Log(gain 2011)					0.39		1.08
			Log(gain 2012)					0.40		1.05
			Log(gain 2009)			977 (candidats nés en 2005)		0.58	0.37	1.54
			Log(gain 2010)					0.54		1.38
			Log(gain 2011)					0.43		1.19
			Log(gain 2012)					0.40		1.27
Endurance	GBLUP	Indices de performance sur la carrière multi-caractères corrigés pour les effets fixes	Vitesse	56 200	552	552	0.22	0.29	0.18	0.96
			Distance				0.10	0.43	0.13	0.85
			Code d'état				0.10	0.38	0.13	0.84
Trot	GBLUP multi-caractère, SNP lié au gène majeur en effet fixe	Moyennes de performances corrigées pour les effets fixes	Log(gain 2 ans)	41 711	396	424	0.28	0.44	0.36	1.44
			Log(gain 3 ans)				0.32	0.27	0.23	1.16
			Log(gain 4 ans)				0.25	0.26	0.22	1.06
			Log(gain 5 ans)				0.26	0.45	0.27	1.25
			Qualification				0.56	0.25	0.37	0.78

Discussion générale et perspectives

Du point de vue des aspects théoriques de la mise en place de la sélection génomique, les travaux réalisés au cours de cette thèse ont permis de souligner l'importance d'un paramètre de la précision de l'évaluation : le nombre de segments indépendants dans le génome. Nous avons montré que ce paramètre est celui qui a le plus de poids dans les formules développées ces dernières années pour prédire la précision de l'évaluation génomique, avec la taille de la population de référence. Les écarts entre les précisions observées dans la bibliographie au cours de la méta-analyse et les précisions prédites par les formules seraient principalement causés par ce paramètre. En effet, M_e peut être calculé à partir de la longueur du génome et de la taille efficace de la population, ou bien estimé à partir des génotypages (et éventuellement du pédigrée). Dans la méta-analyse, nous avons montré que la valeur de M_e varie beaucoup en fonction de la formule utilisée pour le calcul, ce qui conduit à surestimer ou à sous-estimer la précision de l'évaluation génomique. Par ailleurs, les formules de calcul de M_e utilisent la taille efficace de la population N_e, paramètre qui peut être estimé en utilisant différentes méthodes basées sur des données variées (consanguinité, sex-ratio dans la population, etc.), et donnant des résultats variés également. Nous avons pu vérifier la difficulté d'estimer M_e sur les populations de chevaux étudiées au cours de la thèse, les formules de calcul de M_e ayant donné des valeurs très différentes. Nous avons également calculé M_e à partir de données réelles, d'une part en utilisant seulement les génotypages (moyennes de DL entre SNPs), d'autre part en utilisant les matrices d'apparentement classique et génomique (variance de la différence des coefficients non-diagonaux de **G** et **A**). Là encore, alors que les méthodes sont supposées équivalentes, les valeurs obtenues variaient suivant la méthode. Il nous semble donc que l'estimation de ce paramètre doit être améliorée, car son importance nécessiterait de le connaitre avant d'envisager la sélection. En effet, plus le nombre de segments indépendants à estimer est important, plus il faudra d'informations pour estimer correctement leurs effets. Le nombre de segments indépendant devrait donc être estimé pour chaque population et pour chaque caractère, au même titre que l'héritabilité. Tant qu'aucune méthode d'estimation sûre ne sera disponible pour M_e, il est envisageable de l'estimer dans une expérience d'évaluation génomique où l'on connait la précision de l'évaluation et les autres paramètres. La valeur obtenue pourra être réutilisée pour la même population.

Un premier essai de sélection génomique avait été réalisé pour les performances en CSO, en réalisant l'évaluation avec un GBLUP et en utilisant comme pseudo-phénotype des indices dérégressés. La précision de l'évaluation génomique était très proche de la précision de l'évaluation classique. Dans la thèse, la méthode de l'évaluation génomique en une étape a été testée. L'intérêt de cette méthode était d'utiliser les performances de tous les chevaux dans l'évaluation, qu'ils soient génotypés ou non, en s'affranchissant des approximations qui étaient nécessaires pour prendre toute l'information en compte dans des valeurs génétiques dérégressées. Là encore, l'amélioration de la précision de l'évaluation est faible, et ne permet pas pour l'instant d'envisager de remplacer l'évaluation classique par une évaluation génomique. Parmi les raisons possibles pour expliquer le faible gain en précision, les deux causes suivantes devraient être investiguées prochainement.

La première explication envisagée est l'impossibilité de prendre correctement en compte un effet race dans le modèle. En effet l'importance de l'effet race ou type de cheval était visible dans l'estimation des paramètres génétiques, mais cet effet n'étant pas héritable, il n'a pas pu être inclus simplement dans le modèle. Une solution pour prendre en compte l'effet de la race serait d'utiliser des groupes de parents inconnus, mais dans notre cas les groupes de parents inconnus ont conduit à

estimer des valeurs génétiques visiblement fausses. Ce problème a été rencontré dans plusieurs espèces, l'hypothèse pour expliquer ces résultats étant que la matrice d'apparentement génomique capte des informations non contenues dans la matrice d'apparentement classique. Une amélioration du single-step consistant à connecter les groupes de parents inconnus a été développée très récemment et est en cours d'implémentation dans le logiciel d'évaluation.

Une autre cause possible pour expliquer les faibles précisions obtenues en CSO est le fait que les chevaux génotypés utilisés soient des étalons, c'est-à-dire d'anciens candidats qui ont été sélectionnés sur leurs bons résultats et sont déjà reproducteurs. L'échantillon contenait surtout des bons chevaux, et une précision meilleure pourrait peut-être être obtenue en incluant dans la population de référence plus de chevaux moins bons performeurs. En effet une population de référence doit être représentative de la population afin d'entrainer correctement le modèle d'évaluation. L'inclusion de chevaux moins sélectionnés dans la population de référence devrait être permise par le projet SoGen. Dans le cadre de ce projet, 2 000 chevaux seront génotypés, en incluant dans l'échantillon de bons performeurs et de mauvais performeurs. Ces génotypages seront également utilisés pour confirmer ou infirmer les résultats de l'analyse d'association réalisée au cours de la thèse, qui avait mis en évidence un gène candidat potentiel sur le chromosome 1 pour l'aptitude à la performance en CSO.

On peut espérer que la prise en compte de l'effet race d'une part et l'utilisation de données plus représentatives de la variabilité dans la population d'autre part permettront d'améliorer la précision de l'évaluation génomique. Dans ce cas, le fait d'obtenir plus précocement des valeurs génétiques aussi précises que celles estimées après plusieurs années de performances devrait permettre de réduire l'intervalle de génération.

Pour les courses d'endurance, la sélection génomique a été testée avec un BLUP génomique sur des moyennes de performances corrigées. Ici, à l'inverse des données de CSO, l'échantillon de chevaux génotypés était principalement composé de performeurs et incluait peu de reproducteurs. Comme pour le CSO, l'évaluation génomique était un peu plus précise que l'évaluation classique, mais l'écart trop faible des résultats entre les deux méthodes ne permet pas non plus d'envisager l'utilisation de la sélection génomique à l'heure actuelle.

Pour les courses au trot, l'évaluation génomique a été testée dans des conditions un peu différentes de celles du CSO et de l'endurance, car un gène ayant un effet majeur sur les performances des trotteurs a été mis en évidence récemment. La population de référence utilisée était plus favorable que celles de CSO et d'endurance, car beaucoup d'individus génotypés avaient leur père génotypé dans la population de référence, et malgré un taux de qualification supérieur à celui de la population l'échantillon comptait des chevaux non-qualifiés. Nos résultats montrent qu'une évaluation génomique multi-caractères prenant en compte un SNP lié au gène majeur permet d'améliorer sensiblement la précision des valeurs génétiques pour les caractères affectés par le gène majeur (capacité du cheval à se qualifier pour participer aux courses, gains précoces et dans une moindre mesure gain tardifs), et diminue le biais des valeurs génétiques estimées pour les caractères peu affectés par le gène (gains en milieu de carrière). Au vu des précisions obtenues, l'utilisation de l'évaluation génomique semble prometteuse. Cependant, des travaux supplémentaires sont nécessaires concernant l'effet du gène majeur. La raison biologique expliquant la supériorité inattendue des hétérozygotes en fin de carrière n'est pas encore identifiée. Le projet GenOtrot a

pour objectifs de vérifier l'effet du gène majeur sur différents critères de performances en courses en génotypant 600 chevaux supplémentaires, de vérifier également les effets d'autres QTL détectés chez le Trotteur Français en plus de *DMRT3*, et d'identifier les causes mécaniques des différences de performances des chevaux de génotypes différents grâce à un phénotypage fin des allures. Si l'effet positif du gène majeur sur les gains tardifs est bien dû à des raisons biologiques, son utilisation dans le cadre d'une évaluation génomique devrait permettre de sélectionner les trotteurs avec différents objectifs : par exemple produire des chevaux faciles à qualifier, et des chevaux plus difficiles à qualifier mais plus performants en fin de carrière. Il sera alors important de réfléchir le schéma de sélection de façon à conserver des hétérozygotes dans la population tout en limitant l'apparition d'homozygotes.

Annexe

FICHIER COMPETITIONS EQUESTRES – CHEVAUX DE SPORT

HISTORIQUE

- Fichier des épreuves détaillées :
  - De 1985 à 1996 : fichier de France GALOP , pour toutes les épreuves de compétitions équestres et épreuves d'élevage.
  - De 1997 à 1998 : fichier France GALOP, pour les épreuves d'élevage.
  - De 1997 à 1998 : fichier de la DNSE pour les compétitions équestres.
  - De 1999 à XXXX : fichier SIRE pour les compétitions équestres et les épreuves d'élevage.

ARCHIVAGE DES FICHIERS

Les fichiers sont archivés sous ADA et sont conservés 10 ans actuellement , les noms varient en fonction du style du fichier et de l'organisme qui nous les a transférés. Les noms des fichiers sont les suivants :

- Fichier performances détaillées :
  - STEEPLE et France GALOP :
    ARCH_STEEPLE_EPR_1985 … 1998
  - DNSE
    ARCH_ATOS_EPR_1997 … 1998
  - SIRE
    ARCH_SIRE_SPORT_EPR_1998 ….xxxx

DESCRIPTION DES ENREGISTREMENTS

Performances détaillées.

- Fichier performances détaillées France GALOP de 1985 à 1991 pour les épreuves d'élevage et les épreuves des compétitions équestres.

Variable	***Position***	**Longueur**	
Information concours			
Code enregistrement	1-4	4	Alpha EN10
Filler	5-16	12	
Discipline	17	1	Alpha 1=CD,2=CCE,3=CSO
Lieu	18-23	6	Alpha
Date	24-29	6	Numérique jjmmaa
Nom du lieu	34-63	30	Alpha
Centre organisateur	64-87	24	Alpha
Information épreuve			
Code enregistrement	1-4	4	Alpha EN12
Filler	5-17	13	
Lieu	18-23	6	Alpha
Date	24-29	6	Numérique jjmmaa

Discipline	30	1	Alpha 1=CD,2+cce,3=CSO
Epreuve	31-33	3	Alpha N° en CSO , Niveau en CD
Niveau	34	1	Binaire Niveau CSO
Nom du prix	35-64	30	Alpha
Filler	65-90	26	
Code International	91	1	Alpha I=inter J=junior
Information détaillée			
Code enregistrement	1-4	4	Alpha EN15
Filler	5-17	13	
Lieu	18-23	6	Alpha
Date	24-29	6	Numérique jjmmaa
Discipline	30	1	Numérique 1=CD,2=CCE,3=CSO
Épreuve	31-33	3	Alpha
Filler	34-38	5	
N° cheval	39-46	8	Alpha
N° propriétaire	47-52	6	Alpha
N° cavalier	53-59	6	Alpha
Filler	60-102	43	
Cavalier	103-125	23	Alpha
Filler	126-127	2	
Propriétaire	128-150	23	Alpha
Cheval	151-175	25	Alpha
Sexe	176	1	Alpha M,F,H
Robe	177-178	2	Alpha
Age	179-180	2	Numérique
Filler	181-260	80	
Race PERE de MERE	261-262	2	Alpha
Race cheval	263-264	2	Alpha
Filler	265-267	3	
Race MERE	268-269	2	Alpha
Race PERE	270-271	2	Alpha
Filler	272-275	4	
Place	276-278	3	Numérique
Filler	279-287	9	
Gain	288-292	5	Numérique
Filler	293-303	9	
Code place	304	1	Alpha A=1er,B=classé,C=partant
Filler	305-310	6	

- Fichier performances détaillées France GALOP de 1992 pour les épreuves d'élevage et les épreuves des compétitions équestres. Identique aux années précédentes sauf sur les informations détaillées.

Variable	**Position**	**Longueur**	
Information détaillée			
Code enregistrement	1-4	4	Alpha EN15
Filler	5-17	13	
Lieu	18-23	6	Alpha
Date	24-29	6	Numérique jjmmaa
Discipline	30	1	Numérique 1=CD,2=CCE,3=CSO
Épreuve	31-33	3	Alpha
Filler	34-38	5	
N° cheval	39-46	8	Alpha
N° propriétaire	47-52	6	Alpha
N° cavalier	53-59	6	Alpha
Filler	60-102	43	
Cavalier	103-125	23	Alpha
Filler	126-127	2	
Propriétaire	128-150	23	Alpha
Cheval	151-175	25	Alpha
Sexe	176	1	Alpha M,F,H
Robe	177-178	2	Alpha
Age	179-180	2	Numérique
Filler	181-260	80	
Race PERE de MERE	261-262	2	Alpha
Race cheval	263-264	2	Alpha
Filler	265-267	3	
Race MERE	268-269	2	Alpha
Race PERE	270-271	2	Alpha
Filler	272-275	4	
Place	276-278	3	Numérique
Filler	279-287	9	
Gain	288-295	8	Numérique
Primes	296-303	8	Numérique
Code place	304	1	Alpha G=1er,P=classé,N=partant
Filler	305-310	6	

- Fichier performances détaillées France GALOP de 1993 à 1998 pour les épreuves d'élevage et de 1993 à 1996 pour les épreuves des compétitions équestres. Identique aux années précédentes sauf sur les informations détaillées.

Variable	**Position**	**Longueur**	
Information détaillée			
Code enregistrement	1-4	4	Alpha EN15
Filler	5-17	13	
Lieu	18-23	6	Alpha
Date	24-29	6	Numérique jjmmaa

Discipline	30	1	Numérique 1=CD,2=CCE,3=CSO
Épreuve	31-33	3	Alpha
Filler	34-38	5	
N° cheval	39-46	8	Alpha
N° propriétaire	47-52	6	Alpha
N° cavalier	53-59	6	Alpha
Filler	60-102	43	
Cavalier	103-125	23	Alpha
Filler	126-127	2	
Propriétaire	128-150	23	Alpha
Cheval	151-175	25	Alpha
Sexe	176	1	Alpha M,F,H
Robe	177-178	2	Alpha
Age	179-180	2	Numérique
Filler	181-261	81	
Race PERE de MERE	262-263	2	Alpha
Race cheval	264-265	2	Alpha
Filler	266-268	3	
Race MERE	269-270	2	Alpha
Race PERE	271-272	2	Alpha
Filler	273-276	4	
Place	277-279	3	Numérique
Filler	280-288	9	
Gain	289-296	8	Numérique
prime	297-304	8	Numérique
Code place	305	1	Alpha G=1er,P=classé,N=partant
Filler	306-310	5	

- Fichier performances détaillées DNSE de 1997 à 1998 pour les épreuves des compétitions équestres.

Variable	**Position**	**Longueur**	
Information concours			
Code enregistrement	1-4	4	Alpha EN10
N° concours	5-13	9	Numérique AAAADDNNN
Discipline	14	1	Numérique 1=CD,2=CCE,3=CSO
Département	15-17	3	Alpha
Date concours	18-25	8	Numérique JJMMAAAA
Libellé lieu concours	26-50	25	Alpha
Nom société	51-75	25	Alpha
Information épreuve			
Code enregistrement	1-4	4	Alpha EN12
N° concours	5-13	9	Numérique AAAADDNNN
N° épreuve	14-16	3	Alpha
Discipline	17	1	Numérique

Departement	18-20	3	Alpha
Date épreuve	18-28	8	Numérique JJMMAAAA
Classe épreuve	29-34	6	Alpha
Nom prix	35-59	25	Alpha
Indicateur international	60	1	Alpha I=inter
Information détaillée			
Code enregistrement	1-4	4	Alpha EN15
N° concours	5-13	9	Numérique AAAADDNNN
N° épreuve	14-16	3	Alpha
Discipline	17	1	Numérique
Département	18-20	3	Alpha
Date épreuve	21-28	8	Numérique JJMMAAAA
N° SIRE CHEVAL	29-36	8	Alpha
N° cavalier	37-43	7	Numérique
Nom cavalier	44-68	25	Alpha
Nom propriétaire	69-98	30	Alpha
Nom cheval	99-123	25	Alpha
Sexe	124	1	Alpha M,F,H
Libellé race	125-132	8	Alpha
Robe	133-140	8	Alpha
An. Nais cheval	141-144	4	Numérique
Place	145-147	3	Numérique
Gain	148-155	8	Numérique
Prime	156-163	8	Numérique

- Fichier performances détaillées SIRE de 1999 à 2000 pour les épreuves des compétitions équestres et les épreuves d'élevage.

Variable	**Position**	**Longueur**	
Information concours			
Code enregistrement	1-2	2	Numérique 10
N° concours	3-11	9	Numérique AANNNN AAAADDNNN AAAAFXNNN
Discipline	12	1	Numérique
Department	13-15	3	Numérique
Date concours	16-23	8	Numérique JJMMAAAA
Libelle lieu concours	24-49	26	Alpha variable
Information épreuve			
Code enregistrement	1-2	2	Numérique 30
N° concours	3-11	9	Numérique
N° épreuve	12-14	3	Alpha
N° société	15-22	8	Alpha
N° index épreuve	23-29	7	Alpha
Classe épreuve	30-35	6	Alpha
Nom du prix	36-75	40	Alpha
Barème epr. CSO	76-85	10	Alpha

Nombre engages	86-88	3	Numérique
Number forfaits	89-91	3	Numérique
Nombre partants	92-94	3	Numérique
Nombre non-partants	95-97	3	Numérique
Information cheval			
Code enregistrement	1-2	2	Numérique 40
N° concours	3-11	9	Numérique
N° épreuve	12-14	3	Alpha
N° SIRE cheval	15-22	8	Alpha
N°licence cavalier	23-30	8	Numérique
Titre cavalier	31-35	5	Alpha
Prénom cavalier	36-62	27	Alpha
Nom cavalier	63-94	32	Alpha
Place	95-97	3	Numérique
Gain	98-105	8	Numérique (8,2)
Code devise	106	1	Alpha
Note moyenne DR	107-111	5	Numérique (5,2)
Point dressage CCE	112-117	6	Numérique (6,2)
Point fond CCE	118-123	6	Numérique (6,2)
Point obstacle CCE	124-129	6	Numérique (6,2)
Total CCE	130-135	6	Numérique (6,2)
Point 1er phase CSO	136-141	6	Numérique (6,2)
Point 2ieme phase CSO	142-147	6	Numérique (6,2)
Point 3ieme phase CSO	148-153	6	Numérique (6,2)
Temps 1er phase CSO	154-159	6	Numérique (6,2)
Temps 2ième phase CSO	160-165	6	Numérique (6,2)
Temps 3ième phase CSO	166-171	6	Numérique (6,2)
Point Hunter	172-177	6	Numérique (6,2)
Point reprise Attelage	178-183	6	Numérique (6,2)
Point marathon AT	184-189	6	Numérique (6,2)
Point maniabilité AT	190-195	6	Numérique (6,2)
Total Attelage	196-201	6	Numérique (6,2)
Code résultat	202	1	Alpha G=1er,P=classé,N=partant
Point Endurance	203-206	4	Numérique (4,2)
Vitesse Endurance	207-210	4	Numérique (4,2)
Freq. cardiaque Endurance	211-213	3	Numérique
Point Elite Endurance	214-215	2	Numérique (1,0,-1)
Qualif. Jeunes Chevaux	216-217	2	Alpha (EL,EX,TB)
Prime réussite HN	218-225	8	Numérique (8,2)
Prime réussite DNSE	226-233	8	Numérique (8,2)

- Fichier performances détaillées SIRE de 2001 à 2002 pour les épreuves des compétitions équestres et les épreuves d'élevage , les informations sont identiques pour le niveau concours et niveau épreuve.

Variable	**Position**	**Longueur**	
Information cheval			
Code enregistrement	1-2	2	Numérique 40
N° concours	3-11	9	Numérique
N° épreuve	12-14	3	Alpha
N° SIRE cheval	15-22	8	Alpha
N°licence cavalier	23-30	8	Numérique
Titre cavalier	31-35	5	Alpha
Prénom cavalier	36-62	27	Alpha
Nom cavalier	63-94	32	Alpha
Place	95-97	3	Numérique
Gain	98-105	8	Numérique (8,2)
Code devise	106	1	Alpha
Note moyenne DR	107-111	5	Numérique (5,2)
Point dressage CCE	112-117	6	Numérique (6,2)
Point fond CCE	118-123	6	Numérique (6,2)
Point obstacle CCE	124-129	6	Numérique (6,2)
Total CCE	130-135	6	Numérique (6,2)
Point 1er phase CSO	136-141	6	Numérique (6,2)
Point 2ieme phase CSO	142-147	6	Numérique (6,2)
Point 3ieme phase CSO	148-153	6	Numérique (6,2)
Temps 1er phase CSO	154-159	6	Numérique (6,2)
Temps 2ième phase CSO	160-165	6	Numérique (6,2)
Temps 3ième phase CSO	166-171	6	Numérique (6,2)
Point Hunter	172-177	6	Numérique (6,2)
Point reprise Attelage	178-183	6	Numérique (6,2)
Point marathon AT	184-189	6	Numérique (6,2)
Point maniabilité AT	190-195	6	Numérique (6,2)
Total Attelage	196-201	6	Numérique (6,2)
Code résultat	202	1	Alpha G=1er,P=classé,N=partant
Point Endurance	203-206	4	Numérique (4,2)
Vitesse Endurance	207-210	4	Numérique (4,2)
Freq. cardiaque Endurance	211-213	3	Numérique
Point Elite Endurance	214-215	2	Numérique (1,0,-1)
Qualif. Jeunes Chevaux	216-217	2	Alpha (EL,EX,TB)
Prime réussite HN	218-225	8	Numérique (8,2)
Prime réussite DNSE	226-233	8	Numérique (8,2)
Point réussite HN	234-239	6	Numérique (6,2)
Point réussite DNSE	240-245	6	Numérique (6,2)

- Fichier performances détaillées SIRE de 2003 à 2009 pour les épreuves des compétitions équestres et les épreuves d'élevage , les informations sont identiques pour le niveau concours et niveau épreuve, pour les informations sur le cheval tout est identique sauf une nouvelle variable est ajoutée en fin d'enregistrement.

Variable	Position	Longueur	
Information cheval			
……	…-245		
Gain engageur	246-253	8	Numérique (8,2)

- Fichier performances détaillées SIRE de 2010 pour les épreuves des compétitions équestres et les épreuves d'élevage.

Variable	Position	Longueur	
Information concours			
Code enregistrement	1-2	2	Numérique 10
N° concours	3-11	9	Numérique AANNNN AAAADDNNN AAAAFXNNN
filler	12	1	
Department	13-15	3	Numérique
Date concours	16-23	8	Numérique JJMMAAAA
Libelle lieu concours	24-73	50	Alpha variable
Information épreuve			
Code enregistrement	1-2	2	Numérique 30
N° concours	3-11	9	Numérique
N° épreuve	12-14	3	Alpha
Discipline	15-16	2	Alpha DR,CC,SO
N° société	17-24	8	Alpha
N° index épreuve	25-31	7	Alpha
Classe épreuve	32-37	6	Alpha
Nom du prix	38-292	255	Alpha
Barème epr. CSO	293-302	10	Alpha
Hauteur	303-305	3	Numérique
Distance	306-308	3	Numérique
Nombre engages	309-311	3	Numérique
Number forfaits	312-314	3	Numérique
Nombre partants	315-317	3	Numérique
Nombre non-partants	318-320	3	Numérique
Dotation initiale	321-329	9	Numérique (9,2)
Nombre de prix	330-332	3	Numérique
Montant distribute	333-341	9	Numérique (9,2)
Information cheval			
Code enregistrement	1-2	2	Numérique 40
N° concours	3-11	9	Numérique
N° épreuve	12-14	3	Alpha
N° SIRE cheval	15-22	8	Alpha
N°licence cavalier	23-30	8	Numérique
Titre cavalier	31-35	5	Alpha
Prénom cavalier	36-62	27	Alpha

Nom cavalier	63-94	32	Alpha
Catégorie licence	95-134	40	alpha
Place	135-137	3	Numérique
Gain	138-145	8	Numérique (8,2)
Gain cavalier en euros	146-153	8	Numérique (8,2)
Gain engageur	154-161	8	Numérique (8,2)
Point réussite HN	162-169	8	Numérique (8,2)
Point réussite DNSE	170-177	8	Numérique (8,2)
Surprime SHF	178-185	8	Numérique (8,2)
Prime SHF en euros	186-193	8	Numérique (8,2)
Caractère prime	194	1	Alpha 1=1er prime, 2=2ime prime, E=prime espoir
Etat résultat	195-234	40	Alpha
Cause elimination	235-274	40	Alpha
Indice résultat	275-314	40	Alpha
Mention champ. SHF	315-354	40	Alpha
Point épreuve jeune Poney	355-357	3	Numérique
Nombre de point	358-366	9	Numérique (9,3)
Temps 1er phase CSO	367-372	6	Numérique (6,2)
Point 2ieme phase CSO	373-378	6	Numérique (6,2)
Temps 3ieme phase CSO	379-384	6	Numérique (6,2)
Note présentation CSO	385-390	6	Numérique (6,3)
Pourcentage final DR	391-396	6	Numerique (6,3)
Point juge C DR	397-399	3	Numérique
Point juge H DR	400-402	3	Numérique
Point juge M DR	403-405	3	Numérique
Point juge B DR	406-408	3	Numérique
Point juge E DR	409-411	3	Numérique
Note ensemble DR	412-414	3	Numérique
Point test dressage CC/AT	415-419	5	Numérique (5,2)
Point fond CC	420-425	6	Numérique (6,2)
Point temps dépassé fond CC	426-430	5	Numérique (5,2)
Point temps dépassé obstacle CC	431-435	5	Numérique (5,2)
Point obstacle CC	436-441	6	Numérique (6,2)
Point marathon AT	442-447	6	Numérique (6,2)
Point maniabilité AT	448-453	6	Numérique (6,2)
Indice test dressage AT	454	1	Alpha
Indice test marathon AT	455	1	Alpha
Indice test maniabilité AT	456	1	Alpha
Vitesse moyenne RE	457-461	5	Numérique (5,3)
Distance RE	462-467	6	Numérique
Fréquence cardiaque	468-470	3	Numérique

RE			
Temps récup. RE	471-475	5	Numérique (5,2)
Temps réel RE	476-483	8	HH :MM :SS
Point de performance WE	484-489	6	Numérique (6,2)
Point apparence WE	490-495	6	Numérique (6,2)

- Fichier performances détaillées SIRE de 2011 à …. pour les épreuves des compétitions équestres et les épreuves d'élevage, pas de changement pour les informations concours et épreuves .

Variable	**Position**	**Longueur**	
Information cheval			
Code enregistrement	1-2	2	Numérique 40
N° concours	3-11	9	Numérique
N° épreuve	12-14	3	Alpha
N° SIRE cheval	15-22	8	Alpha
N°licence cavalier	23-30	8	Numérique
Titre cavalier	31-35	5	Alpha
Prénom cavalier	36-62	27	Alpha
Nom cavalier	63-94	32	Alpha
Catégorie licence	95-134	40	alpha
Place	135-137	3	Numérique
Gain	138-145	8	Numérique (8,2)
Gain cavalier en euros	146-153	8	Numérique (8,2)
Gain engageur	154-161	8	Numérique (8,2)
Point réussite HN	162-169	8	Numérique (8,2)
Point réussite DNSE	170-177	8	Numérique (8,2)
Surprime SHF	178-185	8	Numérique (8,2)
Prime SHF en euros	186-193	8	Numérique (8,2)
Caractère prime	194	1	Alpha 1=1er prime, 2=2ime prime, E=prime espoir
Etat résultat	195-234	40	Alpha
Cause elimination	235-274	40	Alpha
Indice résultat	275-314	40	Alpha
Mention champ. SHF	315-354	40	Alpha
Point épreuve jeune Poney	355-357	3	Numérique
Nombre de point	358-366	9	Numérique (9,3)
Temps 1er phase CSO	367-373	7	Numérique (7,2)
Point 2ieme phase CSO	374-379	6	Numérique (6,2)
Temps 3ieme phase CSO	380-386	7	Numérique (7,2)
Note présentation CSO	387-392	6	Numérique (6,3)
Pourcentage final DR	393-398	6	Numerique (6,3)
Point juge C DR	399-401	3	Numérique
Point juge H DR	402-404	3	Numérique
Point juge M DR	405-407	3	Numérique
Point juge B DR	408-410	3	Numérique

Point juge E DR	411-413	3	Numérique
Note ensemble DR	414-416	3	Numérique
Point test dressage CC/AT	417-421	5	Numérique (5,2)
Point fond CC	422-427	6	Numérique (6,2)
Point temps dépassé fond CC	428-432	5	Numérique (5,2)
Point temps dépassé obstacle CC	433-437	5	Numérique (5,2)
Point obstacle CC	438-443	6	Numérique (6,2)
Point marathon AT	444-449	6	Numérique (6,2)
Point maniabilité AT	450-455	6	Numérique (6,2)
Indice test dressage AT	456	1	Alpha
Indice test marathon AT	457	1	Alpha
Indice test maniabilité AT	458	1	Alpha
Vitesse moyenne RE	459-463	5	Numérique (5,3)
Distance RE	464-469	6	Numérique
Fréquence cardiaque RE	470-472	3	Numérique
Temps récup. RE	473-477	5	Numérique (5,2)
Temps réel RE	478-485	8	HH :MM :SS
Point de performance WE	486-491	6	Numérique (6,2)
Point apparence WE	492-497	6	Numérique (6,2)

Liste des figures

Liste des tableaux

Liste des travaux

Brard, S., Ricard, A., 2015. Genome-wide association study for jumping performances in French sport horses. Anim. Genet. 46, 78–81. doi:10.1111/age.12245

Brard, S., Ricard, A., 2015. Is the use of formulae a reliable way to predict the accuracy of genomic selection? J. Anim. Breed. Genet. 132, 207–217. doi:10.1111/jbg.12123

Brard, S., Ricard, A., 2015. Should we use the SNP linked to DMRT3 in genomic evaluation of French Trotter? Accepté par J. Anim. Sci, doi: 10.2527/jas2015-9224

Bibliographie

ACA, 2014a. L'Arabe, en ligne sur <http://www.haras-nationaux.fr/information/accueil-equipaedia/races-dequides/chevaux-de-sang/arabe.html>. [28 avril 2015].

ACA, 2014b. Règlement du stud-book français du Cheval Arabe, en ligne sur <http://www.acafrance.org/FR/l_aca/studbook.asp>. [28 avril 2015].

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93, 743–752. doi:10.3168/jds.2009-2730

ANAA, 2014a. L'Anglo-Arabe – Le standard de la race, en ligne sur <http://www.anaa.fr/fr/anglo-arabe/>. [28 avril 2015].

ANAA, 2014b. Règlement relatif au stud-book du cheval Anglo-Arabe et au registre du Demi-Sang Anglo-Arabe, en ligne sur <http://www.anaa.fr/fr/anglo-arabe/stud-book.html>. [28 avril 2015].

Asoro, F.G., Newell, M.A., Beavis, W.D., Scott, M.P., Jannink, J.-L., 2011. Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. The Plant Genome Journal 4, 132. doi:10.3835/plantgenome2011.02.0007

Bacanu, S.-A., Devlin, B., Roeder, K., 2002. Association studies for quantitative traits in structured populations. Genet. Epidemiol. 22, 78–93. doi:10.1002/gepi.1045

Banks, R.G., van der Werf, J.H.J., 2009. Economic evaluation of whole genome selection using meat sheep as a case study. http://www.aaabg.org/livestocklibrary/2009/banks430.pdf.

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. doi:10.1186/1297-9686-44-3

Berry, D. P., J. F. Kearney, and B. L. Harris. 2009. Genomic selection in Ireland. In Proc. Interbull International Workshop, Uppsala, Sweden. Bulletin No. 39, 2009. ISSN 1011–6079. Interbull, Uppsala, Sweden.

Berry, D.P., McClure, M.C., Mullen, M.P., 2014. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. J. Anim. Breed. Genet. 131, 165–172. doi:10.1111/jbg.12067

Bijma, P., 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. J. Anim. Breed. Genet. 129, 345–358. doi:10.1111/j.1439-0388.2012.00991.x

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. doi:10.1186/1471-2156-12-80

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. doi:10.1534/genetics.107.080838

Calus, M.P.L., 2010. Genomic breeding value prediction: methods and procedures. Animal 4, 157–164. doi:10.1017/S1751731109991352

Christensen, O.F., Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol 42, 2. doi:10.1186/1297-9686-42-2

Christensen, O.F., 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genetics Selection Evolution 44, 37. doi:10.1186/1297-9686-44-37

Clark, S.A., Hickey, J.M., Daetwyler, H.D., Werf, J.H. van der, 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution 44, 4. doi:10.1186/1297-9686-44-4

Cleveland, M.A., Hickey, J.M., Forni, S., 2012. A Common Dataset for Genomic Analysis of Livestock Populations. G3 (Bethesda) 2, 429–435. doi:10.1534/g3.111.001453

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. doi:10.1371/journal.pone.0003395

Daetwyler, H.D., Hickey, J.M., Henshall, J.M., Dominik, S., Gredler, B., van der Werf, J.H.J., Hayes, B.J., 2010. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Anim. Prod. Sci. 50, 1004–1010.

Daetwyler, H.D., Kemper, K.E., van der Werf, J.H.J., Hayes, B.J., 2012. Components of the Accuracy of Genomic Prediction in a Multi-Breed Sheep Population. J. Anim. Sci. doi:10.2527/jas.2011-4457

Daetwyler, H.D., Calus, M.P.L., Pong-Wong, R., de Los Campos, G., Hickey, J.M., 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193, 347–365. doi:10.1534/genetics.112.147983

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M.-N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P.J., Coote, D., Chamberlain, A.J., Anderson, C., VanTassell, C.P., Hulsegge, I., Goddard, M.E., Guldbrandtsen, B., Lund, M.S., Veerkamp, R.F., Boichard, D.A., Fries, R., Hayes, B.J., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46, 858–865. doi:10.1038/ng.3034

Dassonneville, R., Brøndum, R.F., Druet, T., Fritz, S., Guillaume, F., Guldbrandtsen, B., Lund, M.S., Ducrocq, V., Su, G., 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J. Dairy Sci. 94, 3679–3686. doi:10.3168/jds.2011-4299

Devlin, B., Roeder, K., Wasserman, L., 2001. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60, 155–166. doi:10.1006/tpbi.2001.1542

Dubois, C., Ricard, A., 2007. Efficiency of past selection of the French Sport Horse: Selle Français breed and suggestions for the future. Livestock Science, Special section: Non-Ruminant Nutrition Symposium 112, 161–171. doi:10.1016/j.livsci.2007.02.008

Dubois, C., Manfredi, E., Ricard, A., 2008. Optimization of breeding schemes for sport horses. Livest. Sci. 118, 99-112. doi:10.1016/j.livsci.2008.01.005

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

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. doi:10.1371/journal.pone.0081046

FFE, 2014a. Règlement des compétitions- Dispositions spécifiques au Concours de Saut d'Obstacles, en ligne sur <http://www.ffe.com/Disciplines/General/CSO/Reglement>. [9 avril 2015].

FFE, 2014b. Règlement des compétitions- Dispositions spécifiques Endurance, en ligne sur <http://www.ffe.com/Disciplines/General/Endurance/Reglement>. [28 avril 2015].

FFE, 2015a. Engagés et épreuves, en ligne sur <https://ssl.ffecompet.com/ffecompet/index3.php?pagev3=ENG.Statistiques.PFO_ENG_StatistiquesEngagements2>. [28 avril 2015].

FFE, 2015b. Règlement des compétitions- Dispositions spécifiques au Concours Complet d'Equitation, en ligne sur <http://www.ffe.com/Disciplines/General/CCE/Reglement>. [28 avril 2015].

Gao, H., Christensen, O.F., Madsen, P., Nielsen, U.S., Zhang, Y., Lund, M.S., Su, G., 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genet Sel Evol 44, 8. doi:10.1186/1297-9686-44-8

Gilmour, A.R., Gogel, B.J., Cullis, B.R., and Thompson, R. 2006 ASReml User Guide Release 2.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK

Goddard, M. E., B. J. Hayes, H. McPartlan and A. J. Chamberlain, 2006. Can the same genetic markers be used in multiple breeds? Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil, August 13–18, 2006

Goddard, M.E., Hayes, B.J., 2007. Genomic selection. J. Anim. Breed. Genet. 124, 323–330. doi:10.1111/j.1439-0388.2007.00702.x

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

Goddard, M. E., Hayes, B. j., Meuwissen, T. H. E., 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of Animal Breeding and Genetics 128, 409–421. doi:10.1111/j.1439-0388.2011.00964.x

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M., Snell, R., 2002. Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Res. 12, 222–231. doi:10.1101/gr.224202

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. doi:10.1534/genetics.107.081190

Habier, D., Fernando, R.L., Dekkers, J.C.M., 2009. Genomic selection using low-density marker panels. Genetics 182, 343–353. doi:10.1534/genetics.108.100289

Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12, 186. doi:10.1186/1471-2105-12-186

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. doi:10.1534/genetics.113.152207

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

Hayes, B.J., Bowman, P.J., Chamberlain, A.C., Verbyla, K., Goddard, M.E., 2009b. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41, 51. doi:10.1186/1297-9686-41-51

Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009c. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res (Camb) 91, 47–60. doi:10.1017/S0016672308009981

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, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet 6, e1001139. doi:10.1371/journal.pgen.1001139

Hayes, B.J., Bowman, P.J., Daetwyler, H.D., Kijas, J.W., van der Werf, J.H.J., 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43, 72–80. doi:10.1111/j.1365-2052.2011.02208.x

Henderson, C.R., 1973. Sire evaluation and genetic trends. Journal of Animal Science 1973, 10–41. doi:/1973.1973Symposium10x

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theoret. Appl. Genetics 38, 226–231. doi:10.1007/BF01245622

Ibáñẽz-Escriche, N., Fernando, R.L., Toosi, A., Dekkers, J.C., 2009. Genomic selection of purebreds for crossbred performance. Genetics Selection Evolution 41, 12. doi:10.1186/1297-9686-41-12

IFCE, 2015. Elevage et production, en ligne sur <http://statscheval.haras-nationaux.fr/core/zone_menus.php?zone=229&r=1318>. [28 avril 2015].

IFCE-OESC, 2014. Courses-Organisation des courses en France, en ligne sur <http://statscheval.haras-nationaux.fr/core/tabbord.php?zone=229&r=1326>. [28 avril 2015].

IFCE-OESC, 2015a. Le cheptel équin français, en ligne sur <http://www.haras-nationaux.fr/information/accueil-equipaedia/filiere-economie/chiffres-cles-sur-les-entreprises-ressources-et-territoire/le-cheptel-equin-francais.html>. [4 mai 2015].

IFCE-OESC, 2015b. Les chiffres sur les activités équestres, en ligne sur <http://www.haras-nationaux.fr/information/accueil-equipaedia/filiere-economie/chiffres-cles-sur-les-activites-equines/les-chiffres-sur-les-activites-equestres.html>. [4 mai 2015].

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., Sorrells, M.E., 2014. Training set optimization under population structure in genomic selection. Theor Appl Genet 128, 145–158. doi:10.1007/s00122-014-2418-4

Jäderkvist, K., Andersson, L.S., Johansson, A.M., Árnason, T., Mikko, S., Eriksson, S., Andersson, L., Lindgren, G., 2014. The DMRT3 “Gait keeper” mutation affects performance of Nordic and Standardbred trotters. J. Anim. Sci. 92, 4279–4286. doi:10.2527/jas.2014-7803

Jannink, J.-L., 2010. Dynamics of long-term genomic selection. Genetics Selection Evolution 42, 35. doi:10.1186/1297-9686-42-35

Jensen, J., Su, G., Madsen, P., 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genet. 13, 44. doi:10.1186/1471-2156-13-44

Koenen, E. P. C., Aldridge, L. I., 2002. Testing and genetic evaluation of sport horses in an international perspective. Proc. 7th WCGALP, Montpellier, August 2002

Legarra, A., Robert-Granié, C., Manfredi, E., Elsen, J.-M., 2008. Performance of Genomic Selection in Mice. Genetics 180, 611–618. doi:10.1534/genetics.108.088575

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

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I., Misztal, I., 2015. Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. Genetics genetics.115.177014. doi:10.1534/genetics.115.177014

Leroy, G., Mary-Huard, T., Verrier, E., Danvy, S., Charvolin, E., Danchin-Burge, C., 2013. Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. Genetics Selection Evolution 45, 1. doi:10.1186/1297-9686-45-1

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

Liu, T., Qu, H., Luo, C., Shu, D., Wang, J., Lund, M.S., Su, G., 2014. Accuracy of genomic prediction for growth and carcass traits in Chinese triple-yellow chickens. BMC Genetics 15, 110. doi:10.1186/s12863-014-0110-y

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

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. doi:10.1534/genetics.109.107391

Mccue, M. E., Bannasch, D. L. , Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., Distl, O., Guérin, G., Hasegawa, T., Hille, E. W., Leeb, T., Lindgren, G., Penedo, M. C. T., RØed, K. H., Ryder, O. A., Swinburne, J. E., Tozaki, T., Valberg, S. J., Vaudin, M., Lindblad-Toh, K., Wade, C. M., Mickelson, J. R., 2012. A high density SNP array for the domestic horse and extant perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. Plos Genet 8(1), e1002451. doi:0.1371/journal.pgen.1002451

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

Meuwissen, T.H.E., 2009. Accuracy of breeding values of “unrelated” individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41, 35. doi:10.1186/1297-9686-41-35

Meuwissen, T., Hayes, B., Goddard, M., 2013. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci 1, 221–237. doi:10.1146/annurev-animal-031412-103705

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet and D. H. Lee. 2002. BLUPF90 and related programs (BGF90). In: Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France, XXVIII:1–2.

Misztal, I., Legarra, A., Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92, 4648–4655. doi:10.3168/jds.2009-2064

Misztal, I., Vitezica, Z. g., Legarra, A., Aguilar, I., Swan, A. a., 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130, 252–258. doi:10.1111/jbg.12025

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. doi:10.1186/1297-9686-42-37

Muir, W.M., 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124, 342–355. doi:10.1111/j.1439-0388.2007.00700.x

Promerová, M., Andersson, L.S., Juras, R., Penedo, M.C.T., Reissmann, M., Tozaki, T., Bellone, R., Dunner, S., Hořín, P., Imsland, F., Imsland, P., Mikko, S., Modrý, D., Roed, K.H., Schwochow, D., Vega-Pla, J.L., Mehrabani-Yeganeh, H., Yousefi-Mashouf, N., G. Cothran, E., Lindgren, G., Andersson, L., 2014. Worldwide frequency distribution of the "Gait keeper" mutation in the DMRT3 gene. Anim Genet 45, 274–282. doi:10.1111/age.12120

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. doi:10.3168/jds.2009-2840

Pszczola, M., Strabel, T., Mulder, H.A., Calus, M.P.L., 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95, 389–400. doi:10.3168/jds.2011-4338

REFErences, 2011a. L'évolution des usages du cheval depuis le 19ème siècle, en ligne sur <http://www.haras-nationaux.fr/information/accueil-equipaedia/utilisations/differentes-utilisations/levolution-des-usages-du-cheval.html>. [9 avril 2015].

REFErences, 2011b. Les utilisations du cheval en France, en ligne sur <http://www.haras-nationaux.fr/information/accueil-equipaedia/utilisations/differentes-utilisations/les-utilisations-du-cheval-en-france.html>. [9 avril 2015].

Resende, M.F.R., Muñoz, P., Resende, M.D.V., Garrick, D.J., Fernando, R.L., Davis, J.M., Jokela, E.J., Martin, T.A., Peter, G.F., Kirst, M., 2012. Accuracy of Genomic Selection Methods in a Standard Dataset of Loblolly Pine (Pinus taeda L.). Genetics genetics.111.137026. doi:10.1534/genetics.111.137026

Ricard, A., 2008. 'Les différents indices actuellement publiés' in L'amélioration génétique des équidés, édition Les Haras Nationaux, Imprimerie du Corrézien, Naves, pp. 171-207.

Ricard, A., Danvy, S., Blouin, C., et Tavernier, L., 2010. Les indices des chevaux de sport revisités. 36ème Journée de la Recherche Equine, Paris, France, présenté le 4 mars 2010.

Ricard, A., Danvy, S., Legarra, A., 2013. Computation of deregressed proofs for genomic selection when own phenotypes exist with an application in French show-jumping horses. J. Anim. Sci. 91, 1076–1085. doi:10.2527/jas.2012-5256

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V.M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C.-C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., Moreau, L., 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics 192, 715–728. doi:10.1534/genetics.112.141473

Roos, A.P.W. de, Hayes, B.J., Goddard, M.E., 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183, 1545–1553. doi:10.1534/genetics.109.104935

Schaeffer, L.R., 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123, 218–223. doi:10.1111/j.1439-0388.2006.00595.x

Schennink, A., Stoop, W.M., Visker, M.H.P.W., Heck, J.M.L., Bovenhuis, H., van der Poel, J.J., van Valenberg, H.J.F., van Arendonk, J. a. M., 2007. DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. Anim. Genet. 38, 467–473. doi:10.1111/j.1365-2052.2007.01635.x

SECF, 2011, Règlement du stud-book du Trotteur Français, en ligne sur <http://www.cheval-francais.eu/fr/le-trot-de-a-a-z/le-trotteur-francais/elevage-du-trotteur-francais.html>. [4 mai 2015].

SECF, 2013, Bilan d'activité 2012, en ligne sur <http://www.cheval-francais.eu/pdfCom/bilan2012.pdf>. [10 juin 2015].

SECF, 2014, Bilan d'activité 2013, en ligne sur <http://www.cheval-francais.eu/pdfCom/BILAN_2013.pdf>. [10 juin 2015].

SECF, 2015, Bilan d'activité 2014, en ligne sur < http://www.cheval-francais.eu/pdfCom/BILAN_2014.pdf>. [10 juin 2015].

SECF, 2015. Règlement de la Société d'Encouragement à l'élevage du Cheval Français formant le code des courses au trot. Bulletin de la SECF : 13bis, en ligne sur <http://www.letrot.com/publi.php?type=BIS>. [28 avril 2015].

Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Ødegard, J., Meuwissen, T.H., 2009. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. Genet Sel Evol 41, 53. doi:10.1186/1297-9686-41-53

Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Meuwissen, T.H., 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet Sel Evol 41, 29. doi:10.1186/1297-9686-41-29

Sonesson, A.K., Woolliams, J.A., Meuwissen, T.H.E., 2012. Genomic selection requires genomic control of inbreeding. Genet. Sel. Evol. 44, 27. doi:10.1186/1297-9686-44-27

Stam, P., 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genetics Research 35, 131–155. doi:10.1017/S0016672300014002

Stud-book Selle Français, 2012. La race Selle Français, en ligne sur <http://www.sellefrancais.fr/la-race-selle-francais-82-rubrique.html>. [9 avril 2015].

Stud-book Selle Français, 2015. Règlement du stud-book Selle Français 2015, en ligne sur <http://www.sellefrancais.fr/reglement-du-stud-book-88-rubrique.html>. [9 avril 2015].

Su, G., Guldbrandtsen, B., Gregersen, V.R., Lund, M.S., 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J. Dairy Sci. 93, 1175–1183. doi:10.3168/jds.2009-2192

Teyssèdre, S., Dupuis, M.C., Guérin, G., Schibler, L., Denoix, J.M., Elsen, J.M., Ricard, A., 2012. Genome-wide association studies for osteochondrosis in French Trotter horses. J. Anim. Sci. 90, 45–53. doi:10.2527/jas.2011-4031

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073. doi:10.1038/nature09534

Thomasen, J.R., Sørensen, A.C., Su, G., Madsen, P., Lund, M.S., Guldbrandtsen, B., 2013. The admixed population structure in Danish Jersey dairy cattle challenges accurate genomic predictions. J. Anim. Sci. 91, 3105–3112. doi:10.2527/jas.2012-5490

Tibshirani, R., 1996, Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society 58:1, 267-288.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423. doi:10.3168/jds.2007-0980

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S., 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92, 16–24. doi:10.3168/jds.2008-1514

VanRaden, P.M., Null, D.J., Sargolzaei, M., Wiggans, G.R., Tooker, M.E., Cole, J.B., Sonstegard, T.S., Connor, E.E., Winters, M., van Kaam, J.B.C.H.M., Valentini, A., Van Doormaal, B.J., Faust, M.A., Doak, G.A., 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96, 668–678. doi:10.3168/jds.2012-5702

Ventura, R.V., Lu, D., Schenkel, F.S., Wang, Z., Li, C., Miller, S.P., 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J. Anim. Sci. 92, 1433–1444. doi:10.2527/jas.2013-6638

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. doi:10.1017/S0016672309990243

Verbyla, K.L., Bowman, P.J., Hayes, B.J., Goddard, M.E., 2010. Sensitivity of genomic selection to using different prior distributions. BMC Proc 4, S5. doi:10.1186/1753-6561-4-S1-S5

Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blöcker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N., Penedo, M.C.T., Raison, J.M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D.F., Biagi, T., Binns, M.M., Chowdhary, B.P., Coleman, S.J., Valle, G.D., Fryc, S., Guérin, G., Hasegawa, T., Hill, E.W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J.R., Murray, J., Nergadze, S.G., Onofrio, R., Pedroni, S., Piras, M.F., Raudsepp, T., Rocchi, M., Røed, K.H., Ryder, O.A., Searle, S., Skow, L., Swinburne, J.E., Syvänen, A.C., Tozaki, T., Valberg, S.J., Vaudin, M., White, J.R., Zody, M.C., Lander, E.S., Lindblad-Toh, K., 2009. Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. Science 326, 865–867. doi:10.1126/science.1178158

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

Weber, K.L., Thallman, R.M., Keele, J.W., Snelling, W.M., Bennett, G.L., Smith, T.P.L., McDaneld, T.G., Allan, M.F., Van Eenennaam, A.L., Kuehn, L.A., 2012. Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes. J. Anim. Sci. 90, 4177–4190. doi:10.2527/jas.2011-4586

Weigel, K.A., de los Campos, G., González-Recio, O., Naya, H., Wu, X.L., Long, N., Rosa, G.J.M., Gianola, D., 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92, 5248–5257. doi:10.3168/jds.2009-2092

Wientjes, Y.C.J., Veerkamp, R.F., Calus, M.P.L., 2013. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics 193, 621–631. doi:10.1534/genetics.112.146290

Zhang, X.-S., Hill, W.G., 2005. Predictions of patterns of response to artificial selection in lines derived from natural populations. Genetics 169, 411–425. doi:10.1534/genetics.104.032573

Zhong, S., Dekkers, J.C.M., Fernando, R.L., Jannink, J.-L., 2009. Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics 182, 355–364. doi:10.1534/genetics.108.09827