ECOLE NATIONALE d'EQUITATION
Le Cadre Noir-Saumur
les Haras nationaux
Institut français du cheval et de l'équitation

38ème Journée
de la Recherche Equine
Jeudi 1er mars 2012

Premiers résultats de l'évaluation génomique chez les chevaux de concours hippique

Par :

- *A. Ricard[1], A. Legarra[2], J.C. Meriaux[3], S. Danvy[4], G. Guérin[1]*
- *[1] INRA, GABI, UMR1313, 78350 Jouy-En-Josas*
  [2] INRA, UR 631, 31326 Castanet-Tolosan
  [3] LABOGENA, 78350 Jouy-En-Josas
  [4] IFCE, Recherche et Innovation, 61310 Exmes

Résumé

La précision de l'évaluation génomique pour le concours de saut d'obstacle a été testée sur 908 étalons de sport. Les typages ont été réalisés par la puce Illumina, 44441 Single Nucleotide Polymorphism (SNP) ont été retenus. Les performances utilisées sont des pseudo-performances qui résument toutes les performances propres et des apparentés extérieurs à l'échantillon génotypé. Deux évaluations sont réalisées : une à partir des généalogies (évaluation génétique), l'autre à partir typages (évaluation génomique, mèthode GBLUP). Malgré une structure génétique favorable avec des SNP proches qui présentent un assez fort déséquilibre de liaison (r^2=0,24 à 50Kb), les résultats ne donnent qu'un très faible avantage à la génomique comparée à l'indexation classique. Sur l'échantillon de validation, les corrélations entre la pseudo-performance et les évaluations sont 0,39 (génomique) et 0.36 (génétique). Ces corrélations tombent à 0,30 (génomique) et 0,28 (génétique) pour les seuls SF et SE car le groupe des AA, bien différencié par les SNP, explique à lui seul les écarts de performance dus à la race. Aucune application pratique ne peut être à ce jour proposée.

Mots clés : Saut d'obstacle, Evaluation génomique, BLUP, SNP

Summary

Reliability of genomic evaluation for jumping horses was tested on a sample of 908 genotyped stallions (71% Selle français, 17% Selle Etranger, 13% Anglo Arab). Genotyping was performed using Illumina Equine SNO50 BeadChip, 44444 SNP were retained. Pseudo performances combining own and relative performances outside the genotyped sample were constructed. Two evaluations were performed: genomic with genotypes and genetic with genealogy. Cross validation was used. The evaluations on validation sample were obtained without performances. GBLUP was used for genomic evaluations. Inspite of a favorable genetic structure (linkage disequilibrium equal to 0.24 at 50Kb, mean distance between adjacent makers), results showed low advantage to genomic evaluation. On validation sample, correlation between pseudo performance and genomic evaluation was 0.39 and 0.36 for genetic evaluation. For the SF and SE breeds alone, the correlations were 0.30 and 0.28 respectively due to the breed performances differences with AA group. No practical applications are proposed at present. Research is pursued in order to improve the number of couples sire/son with high number of measured progeny.

Key-words: Jumping, Genomic evaluation, BLUP, SNP

Introduction

Les résultats présentés ici sont issus du projet JUMPSNP. Ce projet, financé par l'IFCE, l'INRA et le Fonds Eperon et soutenu par la FNC, l'ANSF et l'ANAA se proposait de tester l'efficacité d'une évaluation génomique chez le cheval de sport. L'évaluation génomique révolutionne actuellement la sélection bovine. L'évaluation est obtenue dès la naissance directement à partir des typages de l'ADN et permet d'avoir une précision très supérieure à celle apportée par la seule connaissance de la généalogie, même si elle n'atteint pas celle du testage sur descendance. La précision obtenue (équivalente à plusieurs années de performances propres ou quelques produits chez nous) a été jugée suffisante pour supprimer le testage sur descendance et choisir directement les taureaux à partir des typages. Nous voulions juger de la précision obtenue pour une évaluation génomique chez les chevaux de sport en typant un échantillon d'étalons rassemblé grâce à la participation de nombreux propriétaires, incluant les ex-Haras Nationaux et des échantillons disponibles à LABOGENA.

1. Matériel et méthodes

1.1. Echantillon

Le choix de la population à génotyper a été motivé par la disponibilité de l'échantillon d'ADN et la bonne précision des évaluations génétiques disponibles. Grace au volontariat d'éleveurs propriétaires, à l'IFCE et à LABOGENA, nous avons pu réaliser 908 génotypages.

Les chevaux sont pour 71% des Selle français (SF), 17% des étalons étrangers (SE, principalement des KWPN, Holsteiner, Belgian Warmblood, Hannovrien, Oldenbourg, Cheval de Sport Belge), 13% d'Anglo-Arabes (AA). La plupart sont des étalons, seuls 17 mères ont été génotypées. Ils sont nés principalement entre 1989 et 2005, uniformément distribués sur cette période, ce qui fait environ 46 chevaux par année de naissance et seuls 14% sont nés avant, jusqu'en 1974. Parmi ces 908 chevaux, 556 sont déjà pères de chevaux ayant des performances en compétition. Ils représentent 78% des étalons SF en activité en 2009, 45% des étalons de sport étrangers et 65% des étalons AA, soit un bon échantillonnage de la population actuelle. A l'intérieur de l'échantillon, 127 étalons génotypés sont parents ou grand parents de chevaux génotypés et ainsi 37% des étalons ont leur père génotypés. La taille moyenne des familles d'étalons génotypés est 2,7. Il y a 82 familles avec 3 demi frères et plus et 17 familles avec 10 demi frères et plus (jusqu'à 39).

Les pedigrees des 908 chevaux ont été remontés jusqu'en 1945 (ce qui correspond à l'information utilisée dans les évaluations génétiques officielles). Cela fait un fichier de 6 562 chevaux.

1.2. Genotypage

Le génotypage a été effectué avec la puce Equine Illumina SNP50 à LABOGENA. Cette puce comprend 54 602 SNP répartis sur tout le génome. Les marqueurs de qualité insuffisante ont été supprimés selon les normes suivantes : genotypes identifiés sur moins de 80% de l'échantillon (CALL_FREQ<80%), fréquence de l'allèle mineur (MAF) <5% (sur l'ensemble de l'échantillon), déviation de l'équilibre de Hardy Weinberg (*p value* du test$<10^{-8}$). Le chromosome X n'a pas été inclus dans l'analyse. Finalement 44 444 SNP ont été retenus. La distance moyenne entre SNPs adjacents est de 50 256 paires de bases (soit environ 0,05cM) et près de la moitié des SNPs adjacents sont proches de moins de 25 000 paires de bases. La distribution de la fréquence de l'allèle mineur est uniformément répartie entre 5% et 50%.

1.3. Modèle

Le modèle le plus classique utilisé pour calculer les évaluations génomiques est le « GBLUP » ou BLUP génomique (Meuwissen *Et Al.* 2001; VanRaden 2008). Ce modèle explique la performance par la somme des effets des SNP :

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

Avec **y** le vecteur des performances, **b** le vecteurs des effets fixes (effets d'environnements) avec **X** la matrice d'incidence, **a** le vecteur des effets des SNP (44 444 effets), et **Z** la matrice d'incidence des SNP. Chaque SNP est bi-allèlique, un des deux allèles est défini comme référence et la matrice **Z** est remplie de 0, 1 ou 2 en fonction du nombre d'allèles de référence portés par le cheval à ce SNP. Les effets des SNP sont des effets aléatoires dont la matrice de variance covariance est diagonale :

$$V(\mathbf{a}) = \mathbf{I}\sigma_a^2$$

Il a été montré que ce modèle était équivalent au modèle suivant :

$$\mathbf{y} = \mathbf{Xb} + \mathbf{u} + \mathbf{e}$$

Avec $\mathbf{u} = \mathbf{Za}$, la valeur génomique du cheval égale à la somme des effets des SNP qu'il possède. Dans ce cas on a :

$$V(\mathbf{u}) = \mathbf{G}\sigma_u^2,$$

avec **G** la matrice de parenté génomique calculée à partir des SNP telle que $\mathbf{G} = \mathbf{ZZ}'/k$ avec $k = 2\sum_{i=1}^{44444} p_i q_i$ avec p_i la fréquence de l'allèle de référence du SNP *i*. C'est le modèle que nous avons utilisé avec le logiciel blupf90 (AGUILAR *et al.* 2010).

Nous le comparerons au modèle des évaluations génétiques classique qui est le même mais avec

$$V(\mathbf{u}) = \mathbf{A}\sigma_u^2,$$

où **A** est la matrice de parenté calculée à partir de la connaissance des généalogies.

1.4. Choix de la performance

Dans la plupart des applications actuelles, inspirées des modèles utilisés en sélection bovine laitière, la performance utilisée dans le modèle décrit précédemment n'est pas directement la performance de l'animal génotypé, pour la bonne raison qu'il s'agit de taureaux laitiers. En fait, on utilise une pseudo-performance qui est dans leur cas les déviations des productions des filles (DYD=Daughter Yield Deviation). Ce DYD est la moyenne des productions laitières des filles corrigées pour les effets fixes (estimés dans le système d'indexation classique) et pour la valeur génétique des mères (estimées elles aussi dans le système d'indexation classique). Mais, pour les chevaux de sport, les performances des produits ne sont pas toujours la principale source d'information pour le calcul des valeurs génétiques. Les performances propres, et les performances d'apparentés autres (comme les demi-frères) peuvent jouer un rôle important, d'autant que les productions sont moins nombreuses que chez les bovins. Nous ne pouvions cependant utiliser uniquement la performance propre car il fallait résumer l'ensemble des informations apportées aussi par tous les apparentés. Ces pseudo-performances devaient contenir l'information de tout ce qui était extérieur à l'échantillon génotypé (par exemple les performances des produits performeurs non génotypés) mais ne pas introduire l'information apportée par la parenté entre individus génotypés (par exemple les performances d'un petit enfant issus d'un produit génotypé d'un cheval génotypé). Nous avons donc développé une méthode originale qui permet d'obtenir ces pseudo-performances à partir du fichier des indices génétiques, de leur précision et des généalogies. L'élaboration de ces pseudo-performances permet de comparer deux modèles, l'un génomique (avec la matrice de parenté basée sur les typages) et l'autre génétique (avec la matrice de parenté basée sur la connaissance des généalogies) sur un strict pied d'égalité. Avec chaque pseudo-performance un poids est calculé qui donne la quantité d'information introduite dans la pseudo-performance.

Les évaluations génétiques utilisées pour calculer les pseudo-performances sont celles de l'évaluation officielle pour le concours de Saut d'obstacle (BSO), basées sur 2 critères le classement dans chaque épreuve et une somme de points cumulée dans l'année en fonction de la place et de la difficulté de l'épreuve. La moyenne des BSO de l'échantillon est 10,9 avec un écart type de 10,2. La moyenne des coefficients de détermination (CD) est 0,70. 94% des CD sont supérieurs à 0,50, ce qui correspond à la précision obtenue après une carrière de performeur, mais seulement 18% des CD sont supérieurs à 0,90.

1.5. Méthode de validation des évaluations proposées – Echantillon de validation

Grace aux pseudo-performances et aux poids attachés à chacune d'elles, nous pouvons calculer des valeurs génétiques, en utilisant les généalogies pour calculer les covariances entre individus génotypés, ou des valeurs génomiques, en utilisant les SNP pour calculer les covariances entre individus génotypés. Pour comparer finement les deux évaluations, qui sont toujours très fortement corrélées, la communauté scientifique a mis au point un processus d'apprentissage et de validation. Les évaluations sont calculées à partir d'un sous-échantillon dit d'apprentissage à partir des pseudo-performances et de la généalogie ou des typages. Puis les évaluations de l'ensemble des chevaux restant, dit échantillon de validation, sont

calculées sans utiliser leurs pseudo-performances uniquement à partir des relations créées grâce à la généalogie ou aux typages avec la population d'apprentissage. On calcule enfin la corrélation entre les pseudo-performances de cette population et les deux évaluations proposées dans la population de validation pour estimer dans quelle proportion les évaluations prédisent les pseudo-performances.

Dans la mesure où nous avons développé une méthode qui calcule les pseudo-performances en faisant strictement abstraction des relations intra population génotypée, théoriquement, n'importe quel échantillon d'apprentissage ou de validation pourrait être utilisé sans risque d'introduire de biais par la connaissance de performances successives dans le temps. Mais en pratique, il a été montré que l'efficacité de la sélection génomique dépendait principalement 1° de la quantité d'information introduite dans les pseudo performances, donc en clair du CD des étalons de la population d'apprentissage et de validation, 2° des relations de parenté proches entre population d'apprentissage et de validation, donc en clair de la présence de couple père/produit entre les deux populations. Nous avons donc choisi comme population de validation les étalons qui remplissent les conditions suivantes :

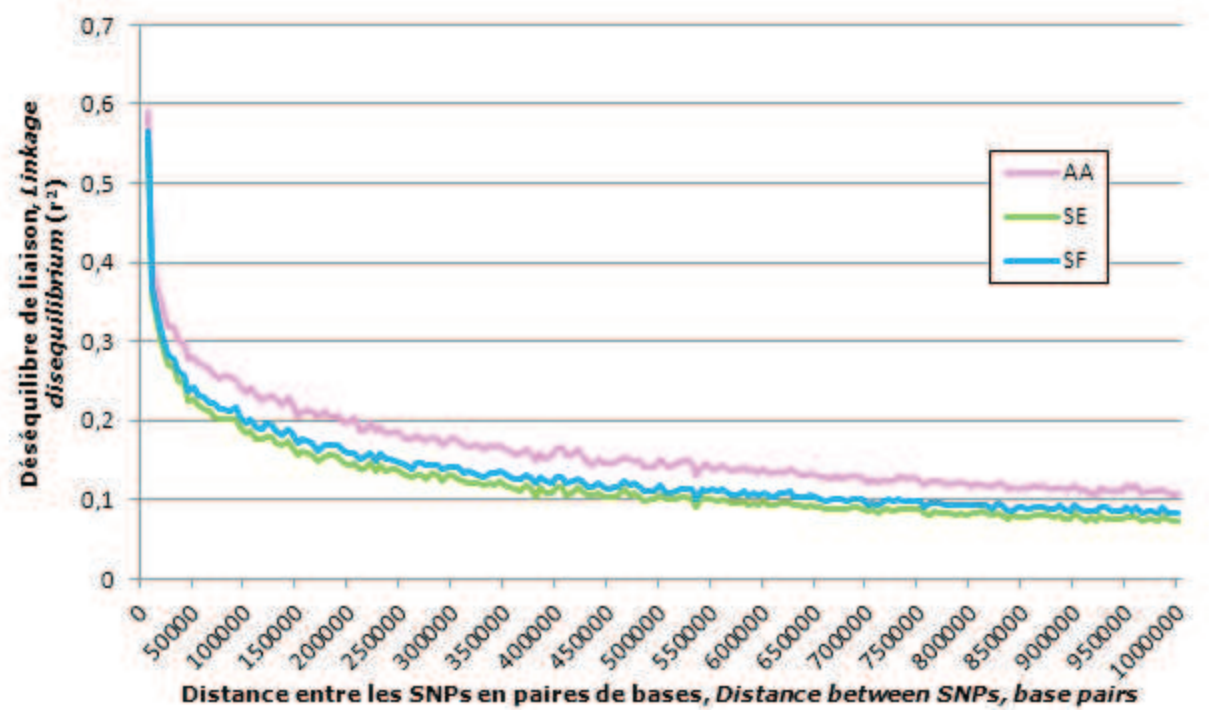
- avoir leur père génotypé présent dans la population d'apprentissage
- avoir un nombre de pseudo-performances important (performances propres + performances des produits et autres apparentés en dehors de l'échantillon génotypé, ce qui équivaut à un « poids » supérieur à 3)
- être issu de familles de 3 demi-frères génotypés au moins.

2. Résultats

2.1. Structure de la population

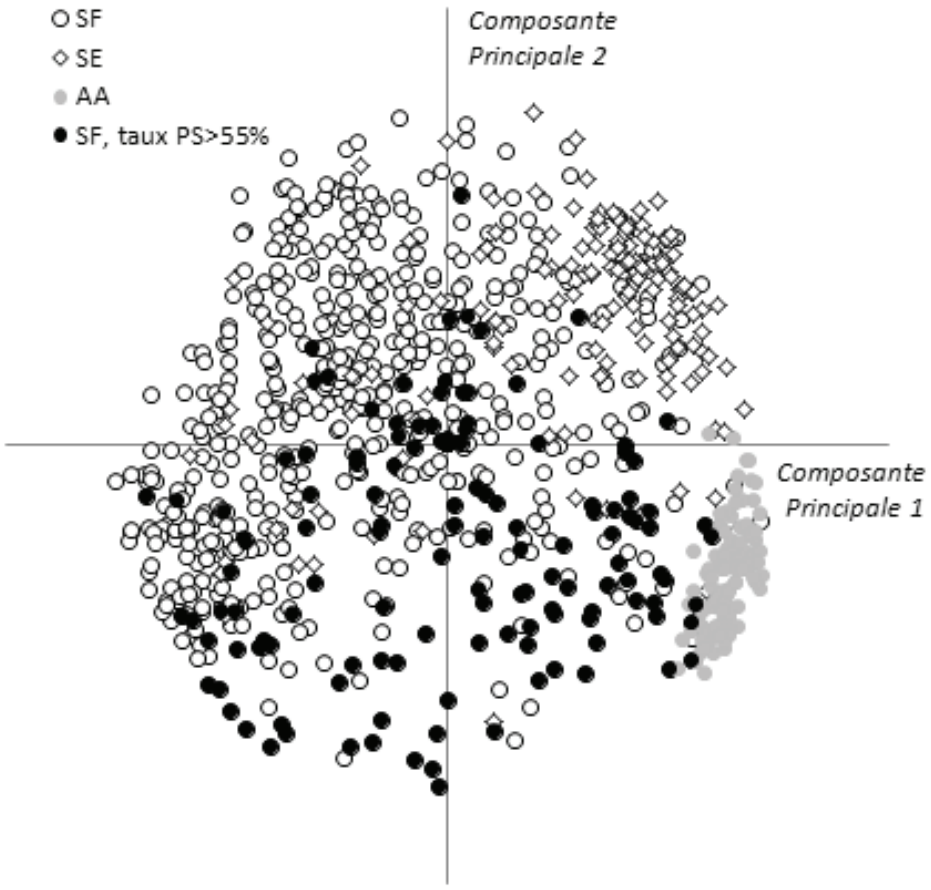
La figure I représente la moyenne des déséquilibres de liaisons entre paires de SNP distants de moins d'une mégabase (1Mb=10^6 paires de bases ou pb), observée sur notre population, mesurés par le carré de la corrélation (r^2), et regroupés par pas de 5Kb paires de bases. Le déséquilibre de liaison moyen observé à une distance de 50Kb (qui correspond à la distance moyenne entre SNPs adjacents) est de 0,280 pour les AA, 0,228 pour les SE et 0,243 pour les SF. Il y a 216 SNPs monomorphiques chez les AA, et seulement 1 chez les SF et les SE. La corrélation entre le déséquilibre de liaison mesuré entre SNPs adjacents dans les différentes races (44 001 paires) est 0,956 entre le SF et le SE, 0,918 entre le SF et l'AA et 0,907 entre le SE et l'AA.

Figure I : Déséquilibre de liaison entre toutes les paires de SNPs distantes de moins de 1Mb dans les races Anglo-Arabe (AA), Selle Français (SF) et Selle Etranger (SE) par pas de 5Kb.
Figure I: Linkage disequilibrium between SNPs pairs distant less than 1Mb by 5Kb step in Anglo Arab (AA), Selle Français (SF) and foreign sport horses (SE).



La figure II représente les 908 chevaux en fonction des deux premières composantes principales de l'analyse en composantes principales de la matrice de parenté génomique. Le groupe des AA est bien différencié alors que les groupes formés par le SF et les SE sont plus mélangés. Il faut noter que les appellations de races sont celles enregistrées par le SIRE mais que les règles d'attribution en fonction des origines varient dans le temps. Quoi qu'il en soit, la catégorisation par race ne donne pas des groupes strictement séparés. En différenciant les SF en fonction de leur taux de Pur Sang (PS), il y a un léger gradient qui distingue les SF proches des SE et éloignés des PS du cas inverse.

Figure II. Graphique des 908 chevaux en fonction des deux premières composantes principales de l'analyse en composantes principales de la matrice de parenté génomique (AA=Anglo-Arabe, SF=Selle Français, SE=Selle Etranger, PS=Pur Sang)
Figure II: Plot of the 908 genotyped horses with the two first principal components of the analysis of genomic matrix. (AA=Anglo-Arab, SF=Selle Français, SE=Foreign Sport horses, PS=Thoroughbred)



2.2. Validation

Les corrélations entre la pseudo-performance et les évaluations génétiques réalisées sur la population de validation sont dans le tableau 1

Tableau 1 : Corrélation entre les pseudo-performances et les évaluations génétique et génomiques sur l'échantillon de validation
Table 1: Correlation between pseudo-phenotypes and genetic and genomic evaluations in the validation sample.

	Toutes races *(all breeds)*	SF+SE
Effectif *(Number of horses)*	103	84
Génétique(généalogies) *Genetic (genealogy)*	0,36	0,28
Génomique (typage) *Genomic (genotyping)*	0,39	0,30

3. Discussion

3.1. Structure génétique

La mesure du déséquilibre de liaison (DL) dans les populations de chevaux de sport françaises nous place en assez bonne situation pour détecter des gènes ou effectuer de la sélection génomique par rapport aux autres espèces domestiques. (McKay *et al.* 2007) ont étudié le DL de 8 races bovines et trouvent des valeurs moyennes de r^2 de 0,50 à 5Kb et 0,22 à 100Kb, donc très légèrement supérieures à nos valeurs (respectivement pour le SF 0,47 et 0,20). Alors que les ovins (Kemper *et al.* 2011) sont en situation plus défavorable avec un r^2 entre 0,12 et 0,19 à 50Kb contre 0,24 à la même distance pour le SF. L'existence d'un DL important à courte distance nous permet d'espérer que les gènes d'intérêt seront associés à des SNP de la puce proches physiquement. En revanche, le DL demeure élevé à d'assez grandes distances, nous risquons donc d'être peu précis pour la localisation des gènes.

3.2. Structure de la population

Comme le montre l'analyse de la matrice génomique, les Anglo Arabes se distinguent du reste des étalons génotypés. C'est un point négatif pour l'utilisation des données de façon conjointe AA et SF pour estimer les valeurs génétiques et c'est ce que nous retrouvons dans l'analyse des résultats de l'évaluation génomique : l'effet le plus notable en faveur de l'évaluation génomique est qu'on estime la différence de niveau entre les AA et les SF, qui existe effectivement pour le CSO. En revanche, nous avons obtenu de bonnes corrélations pour les DL observés dans les différentes races (supérieures à 0,91), ce qui tempère la remarque précédente et devrait permettre d'utiliser efficacement les résultats dans les différentes espèces (les SNP étant liés de la même manière) à condition que les mêmes gènes soient les facteurs limitants dans ces races.

3.3. Efficacité de l'évaluation génomique

Comparé aux résultats obtenus chez les bovins laitiers, notre gain de précision apportée à la naissance par la génomique par rapport à ce que donnent déjà des indices classiques est assez dérisoire. En valeur absolue, les chiffres sont difficiles à comparer entre espèce car très dépendants de la quantité d'information contenue dans les pseudo-performances chez nous et les DYD chez eux (corrélation entre 0,31 et 0,59 selon les caractères chez la Holstein en France, (Legarra *et al.* 2011)) mais la comparaison génétique/génomique suffit à considérer les avantages actuels de la génomique comme limités. Plusieurs facteurs peuvent expliquer ce manque de résultat. Notre effectif est très réduit (une centaine d'individus pour la population de validation et 800 dans la population d'apprentissage mais dont beaucoup sont peu reliés) comparativement aux bovins (près de 16000 taureaux aujourd'hui dans le programme EuroGenomics (Lund *et al.* 2010)). Nous n'avons pas de nombreux couple père/fils ayant chacun un CD très élevé dans notre échantillon génotypé, alors que chez les bovins laitiers, tous les taureaux de la population de validation sont dans ce cas. L'hétérogénéité et l'utilisation de plus de 3 races différentes, comme nous l'avons vu, fait perdre de l'efficacité intra race en se concentrant sur les différences entre race. En tout état de cause, nous ne pouvons proposer aujourd'hui une application pratique à ces premiers résultats.

3.4. Alors que faire ?

D'une part le temps va nous apporter des informations supplémentaires : les jeunes étalons des couples pères/produits génotypés vont voir leur descendance augmenter et nous gagnerons naturellement en précision. Chez les bovins laitiers, de multiples méthodes statistiques ont été testées mais ont toutes données des résultats très comparables. C'est peut être plutôt dans la définition de l'échantillon de validation et d'apprentissage, soit la population de référence qu'il faut chercher des améliorations notables. En effet, même si les effectifs sont moindres, nous possédons la même exhaustivité dans la description des génomes en présence que les espèces à fort effectif puisque nous avons génotypé une grande partie de l'ensemble des reproducteurs mâles en activité, ce qui laisse entrevoir une solution possible. C'est un axe de recherche qui va être développé conjointement avec nos collègues méthodologistes et nos collègues de la sélection ovine qui rencontrent les mêmes problèmes dans le cadre du métaprogramme de l'INRA SelGen. L'information existe au sein de notre matrice génomique, reste à l'extraire de façon originale, ce qui ne passe pas forcément par les mêmes chemins que ceux développés dans les espèces où la quantité d'information ne nécessite pas d'autres voies.

Remerciements

Les auteurs remercient les propriétaires des chevaux qui ont participé à l'analyse, la FNC, l'ANSF et l'ACA qui ont soutenu le projet, LABOGENA pour les analyses, et l'IFCE, l'INRA et le Fond Eperon qui ont financé les recherches.

Références

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta *et al.*, 2010 Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* 93: 743-752.

Kemper, K. E., D. L. Emery, S. C. Bishop, H. Oddy, B. J. Hayes *et al.*, 2011 The distribution of SNP marker effects for faecal worm egg count in sheep, and the feasibility of using these markers to predict genetic merit for resistance to worm infections. *Genetics Research* 93: 203-219.

Legarra, A., C. Robert-Granie, P. Croiseau, F. Guillaume and S. Fritz, 2011 Improved Lasso for genomic selection. *Genetics Research* 93: 77-87.

Lund, M. S., A. P. W. de Roos, A. G. de Vries, T. Druet, V. Ducrocq *et al.*, 2010 Improving genomic prediction by eurogenomics collaboration, pp. in *9th World Conference on Genetics Applied to Livestock Production*, Leipzig, Germany.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts *et al.*, 2007 Whole genome linkage disequilibrium maps in cattle. *Bmc Genetics* 8.

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

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91: 4414-4423.