CRÉATION D'UNE PUCE BASSE DENSITÉ POUR LA SÉLECTION GÉNOMIQUE EN POULE PONDEUSE

Herry Florian[1,2], Hérault Frédéric[2], Varenne Amandine[1], Burlot Thierry[1], Le Roy Pascale[2] et Allais Sophie[2]

[1]NOVOGEN, Mauguerand 22800 Le Foeil, [2]PEGASE, INRA, Agrocampus Ouest, Domaine de la Prise 35590 Saint-Gilles

florian.herry@inra.fr

RESUMÉ

La sélection génomique utilise des informations sur les polymorphismes de l'ADN, notamment SNP, en complément des mesures de performances, afin de choisir les futurs reproducteurs parmi un ensemble de candidats à la sélection. Depuis 2013, une puce commerciale de génotypages à haute densité (600 000 marqueurs) de l'espèce poule a permis le développement de la sélection génomique dans les filières ponte et chair. Toutefois, les coûts de génotypages avec cette puce restent élevés pour une utilisation en routine sur un grand nombre de candidats à la sélection. Un des enjeux actuels de la sélection génomique est donc le développement de puces de génotypage de basse densité (dites puces LD) à plus faible coût. Il s'agit pour cela de sélectionner un panel de marqueurs SNP permettant une imputation (prédiction) des génotypes manquants sur puce haute densité (puce HD).

Dans cette optique, différentes études ont été menées afin de choisir la stratégie de génotypages basse densité la mieux adaptée à une lignée de poule pondeuse. Différentes densités de génotypages, plusieurs scenarii de populations et deux logiciels ont été testés. Les puces basse densité ont été construites selon deux stratégies : choix des SNP selon un clustering basé sur un seuil de déséquilibre de liaison ou choix de SNP à intervalles réguliers. Trois critères d'efficacité des imputations ont été comparés : taux d'erreur génotypique, taux d'erreur allélique et corrélation entre génotypes imputés et réels. Le taux d'erreur génotypique s'est révélé un critère plus discriminant que les autres. Le logiciel FImpute apparaît plus intéressant que le logiciel Beagle car c'est un meilleur compromis entre efficacité de l'imputation et temps de calcul. L'étude des différentes puces basse densité montre que les imputations sont meilleures lorsque la densité des SNP augmente, lorsque le seuil de DL augmente et lorsque des SNP avec une faible MAF (Fréquence Allélique Mineure) sont pris en compte. Les études de populations montrent que les imputations sont meilleures lorsque la taille de la population de référence augmente. L'influence du degré d'apparentement entre population de référence et candidats ainsi que l'impact sur la précision des évaluations génomiques restent à approfondir.

ABSTRACT

Design of a low density chip for genomic selection in layer chickens

Genomic selection uses information on DNA polymorphisms, in particular SNP, in addition of performance measures, in order to choose breeders of the next generation among a set of selection candidates. Since 2013, a commercial high density genotyping chip (600000 markers) for chicken allowed the implementation of genomic selection in layer and broiler breeding. However, genotyping costs with this chip still remain high for a routine use on a large number of selection candidates. Consequently, one of the current stakes of genomic selection is the development at a lower cost of low density genotyping chips (LD chips). To do so, a set of SNP markers has to be selected to enable an imputation (prediction) of missing genotypes on a high density chip (HD chip).

In this regard, various studies were conducted to choose the low density genotyping strategy the best suited to laying hens line. Different genotyping densities, several population scenarii and two software were tested. Low density genotyping chips were built according to two strategies: a choice of SNP depending on a clustering based on linkage disequilibrium threshold or a choice of SNP at regular intervals. Three criteria of imputation accuracy were compared: genotypic error rate, allelic error rate and correlation. The genotypic error rate appeared to be a criterion more discriminating than the others. FImpute software is preferred to Beagle software because FImpute is a better compromise between imputation accuracy and calculation time. The study of different low density genotyping chips shows that imputations improve with SNP density, when the LD threshold increases and when SNP with low MAF (Minor Allele Frequency) are considered. Population studies show that imputations are better when the size of the reference population increases. The influence of the kinship degree between reference population and target population, and the impact on accuracy of genomic evaluations still remain to be deepened.

INTRODUCTION

Les années 2000 ont été marquées par l'utilisation massive des SNP pour le génotypage suite à la publication des génomes de référence pour de nombreuses espèces d'élevage. Au sein d'une population, les SNP correspondent à des changements d'une seule base (A, T, G ou C), à un locus donné, très fréquents et apparaissant de façon régulière le long de l'ADN. Depuis 2013, une puce commerciale de génotypages à haute densité (HD) (600 000 SNP) de l'espèce poule (Kranis et al., 2013) a permis le développement de la sélection génomique dans les filières ponte et chair. Avec la connaissance des génotypages et performances d'une population de référence, il est possible d'estimer la valeur génomique d'un individu dont on ne connaît que le génotype. L'objectif principal est de choisir parmi l'ensemble des candidats à la sélection de la génération *n* les meilleurs individus reproducteurs pour un ou plusieurs caractères qui permettront de produire les individus de la génération *n*+1.

Toutefois, les coûts de génotypages avec une puce HD restent élevés pour une utilisation en routine sur un grand nombre de candidats à la sélection. Un des enjeux de la sélection génomique est donc de développer des puces à SNP basse densité (LD) à plus faible coût. À partir des génotypages des candidats obtenus avec une puce LD, la technique de l'imputation permet de remonter aux génotypages HD des candidats à la sélection. L'imputation consiste à prédire les génotypages HD des candidats à la sélection à partir de leur génotypages LD et des génotypages HD de la population parentale. Cette technique s'appuie sur les règles de la transmission mendélienne et sur le déséquilibre de liaison (DL).

De nombreux travaux ont été menés jusqu'à aujourd'hui, aussi bien en filière bovine, porcine, ovine et avicole. Ces travaux ont été menés sur différents logiciels dont FImpute (Sargolzaei et al., 2014) et Beagle (Browning and Browning, 2016) et ont étudié plusieurs facteurs pouvant influencer la qualité des imputations. Ces facteurs doivent être pris en compte pour créer une puce LD et pour obtenir de bons résultats d'imputation. La densité de SNP sur les puces LD (Dassonneville et al., 2012), le seuil de déséquilibre de liaison utilisé pour la construction des puces LD (Hozé et al., 2013), l'intégration de SNP avec une faible MAF (Hayes et al., 2012 ; Heidaritabar et al., 2015), la taille de la population de référence ou le degré d'apparentement entre population de référence et population candidate (Hozé et al., 2013 ; Heidaritabar et al., 2015) sont des facteurs identifiés dans la littérature comme influençant la qualité des imputations et ayant un impact sur la précision des évaluations génomiques (Dassonneville et al., 2011 ; Heidaritabar et al., 2014 ; Wolc et al., 2014). Toutefois, les particularités du génome aviaire notamment en matière de structure du DL n'ont pas été totalement explorées. L'objectif de cette étude est donc de déterminer quelle est la stratégie de construction de puce basse densité la mieux adaptée au monde avicole.

1. MATÉRIELS ET MÉTHODES

1.1 Population d'étude

La population d'étude est constituée de trois générations de coqs et de poules pondeuses issues d'une lignée pure commerciale créée et sélectionnée par la société Novogen du Groupe Grimaud (Le Foeil, 22). La première génération (G0) est constituée de 437 coqs dont 134 ont été mis en reproduction pour produire la deuxième génération (G1), constituée de 565 coqs dont 125 sont les pères de la troisième génération (G2). Cette dernière génération est constituée de 132 coqs et 635 poules pondeuses.

Des prises de sang ont été réalisées au niveau de la veine brachiale des animaux et l'ADN a été extrait et hybridé sur la puce de génotypage 600K Affymetrix® Axiom® HD par le laboratoire Ark-Genomics (Édimbourg, Royaume-Uni). Au total, l'ensemble des animaux a été génotypé pour 580 961 SNP. Les génotypages ont ensuite été filtrés par un contrôle qualité permettant de retenir 282 928 SNP informatifs pour les analyses, répartis sur les macro-chromosomes (1 à 5), les chromosomes intermédiaires (6 à 10), les micro-chromosomes (11 à 33) et le chromosome sexuel Z. Ces SNP sont désignés sous l'appellation 300K par la suite.

1.2 Simulation des puces basse-densité

À partir de la puce 300K, sept puces ont été simulées en « effaçant » une partie des génotypages, le but étant d'imputer les génotypages manquants. Deux méthodologies intra-chromosomes ont été utilisées pour construire les puces :

- La méthode « équidistante », consistant à sélectionner des SNP à intervalles réguliers. 3 puces ont été créées : la puce 20Kequi (18 196 SNP), la puce 10Kequi (9 352 SNP) et la puce 3Kequi (3 337 SNP).
- La méthode « DL », consistant à sélectionner des SNP en se basant sur le DL entre SNP. La sélection des SNP se fait sous R avec le package hclust et la méthode « Ward.D » à partir de la matrice des R^2 obtenue avec Plink. Cette méthode permet d'obtenir des clusters de SNP en très fort DL les uns avec les autres en maximisant la variance inter-cluster et en minimisant la variance intra-cluster. 4 puces ont été créées en fonction du seuil de DL étudié : la puce DL 0.5 (9 820 SNP), la puce DL 0.2 (5 224 SNP), la puce DL 0.1 (3 988 SNP) et la puce DL 0.05 (3 357 SNP).

Enfin, une puce supplémentaire a été créée en ajoutant sur la puce DL 0.5 294 SNP marqueurs de QTL affectant des caractères de production et de qualité des œufs. Il s'agit de la puce QTL (10 114 SNP).

1.3 Scénarii populations

Quatre scenarii de taille et de relations de parenté différentes ont été étudiés. Les générations constituant les populations de référence et candidate sont détaillées dans le Tableau 1.

Tableau 1. Détail des populations de référence et candidate en fonction des scenarii

Scénario	Référence	Candidats
Utopige	G0	G1
Population totale	G0 - G1	G2
G1→ G2	G1	G2
Saut de génération	G0	G2

Utopige correspond au nom d'un projet ANR à l'origine de la mise en place d'une sélection génomique pour l'entreprise et qui a utilisé G0 et G1 comme population d'étude.

1.4 Stratégies d'imputation étudiées

6 stratégies d'imputations différentes ont été étudiées. À partir des puces, nous avons étudié l'effet :

- de la densité des SNP sur les puces LD,
- du seuil de DL utilisé pour construire les puces,
- de l'intégration de SNP marqueurs de QTL à effets forts,
- de l'intérêt de choisir les SNP sur la notion de distance entre SNP ou bien sur le seuil de DL.

Enfin à partir des scenarii population, nous avons étudié l'effet :

- de la taille de la population de référence,
- du degré d'apparentement entre population de référence et population candidate.

1.5 Logiciels étudiés

Les performances des logiciels FImpute V2.2 (Sargolzaei et al., 2014) et Beagle V4.1 (Browning and Browning, 2016) en termes de temps de calcul et d'efficacité d'imputation ont été comparées sur le scénario Utopige.

1.6 Mesure d'efficacité de l'imputation

Quatre critères sont pris en compte pour mesurer l'efficacité de l'imputation (Dassonneville et al., 2012) :

- Taux d'erreur génotypique : différences par SNP entre les génotypages HD et les génotypages imputés. Si les génotypes diffèrent d'un ou deux allèles, une erreur complète est comptabilisée.
- Taux d'erreur allélique : si les génotypes ne diffèrent que d'un allèle, une demi-erreur est alors comptabilisée.
- Corrélations : Calcul par SNP des corrélations de Pearson entre génotypages HD et génotypages imputés.
- Impact sur les évaluations génomiques de trois caractères aux déterminisme génétiques différents (intensité de ponte, couleur de coquille, poids d'œufs) (Romé et al., 2015) : calcul des corrélations de Spearman sur les 150 candidats ayant la plus grande GEBV (Genomic Estimated Breeding Value) à partir des génotypages HD afin d'évaluer leur éventuel reclassement selon leur GEBV calculée à partir des génotypages imputés.

2. RÉSULTATS ET DISCUSSION

2.1 Comparaison des mesures d'efficacité de l'imputation

Les mesures d'efficacité de l'imputation ont été comparées sur le scénario Utopige pour les puces DL 0.5, QTL et 10Kequi, avec FImpute. On constate que le taux d'erreur génotypique pour la puce 10Kequi est de 3.08%, taux supérieur à celui de la puce DL 0.5 (2.36%), et celui de la puce QTL (2.31%). Des résultats similaires sont obtenus en utilisant le taux d'erreur allélique ou les corrélations. Le taux d'erreur génotypique reste cependant plus intéressant car plus discriminant. De plus, on peut aussi considérer que si l'imputation est mauvaise au niveau d'un des deux allèles, cela pourra avoir des conséquences importantes si elle concerne un SNP à effet fort. Enfin, il est intéressant de noter que le taux d'erreur génotypique est 1.98 fois supérieur au taux d'erreur allélique ce qui indique que la majorité des erreurs d'imputation est une erreur au niveau d'un seul allèle.

2.2 Comparaison des logiciels

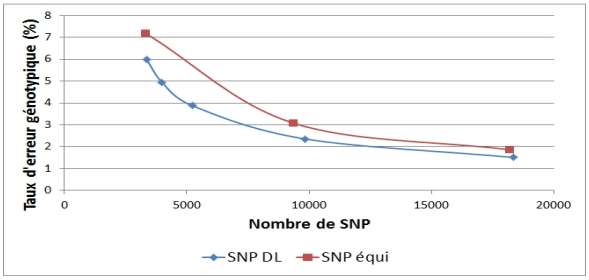
Une comparaison du taux d'erreur génotypique obtenu avec les deux logiciels testés est réalisée sur le scénario Utopige pour les puces DL 0.5, QTL et 10Kequi. On constate que les résultats sont meilleurs sur les trois puces avec Beagle avec des taux d'erreurs génotypiques de 1.98%, 1.55% et 1.54% pour les puces 10Kequi, DL 0.5 et QTL, respectivement. Le classement des puces reste inchangé.

Toutefois, le temps mis par Beagle pour réaliser l'imputation sur la puce QTL est supérieur à 24h contre seulement moins de 4min pour FImpute, ce qui explique l'utilisation de FImpute par la suite. Ces résultats sont bien en accord avec ceux indiqués par Sargolzaei et al. (2014). Ainsi, une utilisation en routine de Beagle sur l'ensemble des candidats à la sélection sera compliquée compte tenu du temps de calcul et du gain en terme d'efficacité qui ne justifient pas une telle contrainte.

2.3 Influence de la densité de marqueurs

La Figure 1 illustre sur le scénario Utopige l'évolution du taux d'erreur génotypique en fonction du nombre de SNP présents sur la puce LD. On constate, pour les deux méthodologies utilisées une diminution du taux d'erreur avec une augmentation du nombre de SNP sur les puces LD. Avec des SNP choisis selon le DL, pour 3 357 SNP et 9 820 SNP, les taux d'erreurs sont

Figure 1. Évolution du taux d'erreur génotypique en fonction du nombre de SNP pour des puces basées sur le seuil de DL ou bien sur la distance entre SNP

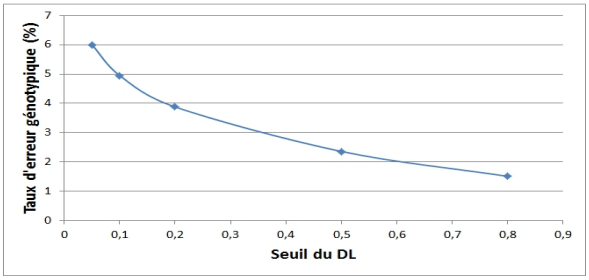


respectivement de 6.00% et 2.36%. De même avec des SNP équidistants, pour 3 337 SNP et 9 352 SNP, les taux d'erreurs sont respectivement de 7.18% et 3.18%. Avec un plus grand nombre de SNP sur puces LD, le nombre de génotypages présents permettant d'identifier les haplotypes de référence correspondants augmente, ce qui diminue la probabilité d'identifier par hasard des haplotypes en commun entre les populations de référence et candidate (Dassonneville et al., 2012).

2.4 Influence du seuil de déséquilibre de liaison

La Figure 2 montre que, pour les puces DL 0.05, DL 0.2 et DL 0.5, les taux d'erreurs génotypiques sont respectivement de 6.00%, 3.88% et 2.36%. Le taux d'erreur diminue avec une augmentation du seuil de DL. Or avec une augmentation du seuil de DL, le nombre de SNP sur les puces augmente. Pour les mêmes raisons que précédemment, le taux d'erreur diminue avec l'augmentation du nombre de SNP sur puces LD. De plus, en augmentant le seuil de DL, un SNP fortement associé aux autres SNP du cluster dans lequel il se trouve est choisi comme représentant du cluster. C'est un « bon » SNP pour l'imputation.

Figure 2. Évolution du taux d'erreur génotypique en fonction du seuil de DL



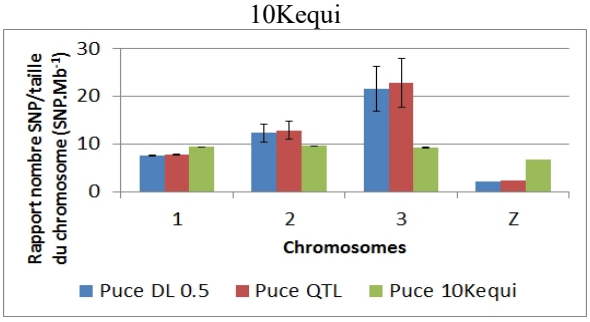
2.5 Influence des marqueurs avec faible MAF

Avec la comparaison des mesures d'efficacité de l'imputation, il a été observé que la puce QTL présentait un taux d'erreur génotypique plus faible que la puce DL 0.5 (2.31% contre 2.36%). Cette amélioration s'explique par les 294 SNP supplémentaires, marqueurs de QTL à effets forts, inclus sur la puce LD. Cela s'explique aussi par la faible MAF des 294 SNP supplémentaires qui seront alors moins bien imputés car ne sont pas forcément retrouvés dans la population de référence, ce qui aboutit à une mauvaise imputation. Les prendre en compte améliore la qualité des imputations.

2.6 Choix de la méthodologie de sélection des SNP

Il ressort des études précédentes que la méthodologie DL, tout en tenant compte des marqueurs avec une faible MAF, serait, à densité de SNP équivalente, la plus adaptée pour obtenir de bons résultats d'imputation.
En effet, avec la méthodologie équidistante, le nombre de SNP retenus sur la puce LD est proportionnel à la taille du chromosome, ce qui n'est pas le cas avec la méthodologie DL (Figure 3). Cela s'explique par une persistance du DL différente entre les macro-chromosomes et les micro-chromosomes (Robert et al., 2015). La persistance du DL diminue des macro-chromosomes aux micro-chromosomes. Il faut donc, en proportion, un plus grand nombre de SNP sur les micro-chromosomes pour couvrir tout le chromosome. En rajoutant en plus des SNP avec une faible MAF, on améliore les résultats pour chaque type de chromosome. La puce QTL construite avec la méthodologie DL et tenant compte de marqueurs avec une faible MAF serait donc la mieux adaptée pour obtenir de bonnes imputations.

Figure 3. Évolution du rapport Nombre de SNP/Taille du chromosome en fonction du type de chromosome sur le scénario Utopige pour les puces DL 0.5, QTL et 10Kequi



1 : Macro-chromosomes (1 à 5) ; 2 : Chromosomes intermédiaires (6 à 10) ; 3 : Micro-chromosomes (11 à 33) ; Z : chromosome sexuel Z

2.7 Effet de la taille de la population de référence

En comparant le scénario « Utopige » avec le scénario « Population totale », nous avons pu étudier l'effet d'une augmentation de taille de la population de référence sur l'imputation des puces DL 0.5, QTL et 10Kequi. On constate ainsi, pour les trois puces, une amélioration du taux d'erreur génotypique qui passe de 3.08% à 2.33% pour la puce 10Kequi, de 2.36% à 2.03% pour la puce DL 0.5 et de 2.31 à 2.00% pour la puce QTL. Le classement des puces reste inchangé. En effet, une augmentation de la taille de la population de référence augmente la taille de la librairie d'haplotypes de référence. La probabilité

d'identifier par hasard, pour un candidat, un mauvais haplotype dans la librairie de référence diminue.

2.8 Effet des relations de parenté entre les populations de référence et candidates

Sur le scénario « Saut de génération », on constate que le taux d'erreur génotypique augmente par rapport aux deux scenarii précédents pour la puce 10Kequi en passant à 3.33%. Plus les individus sont proches en terme de parenté, plus ils ont en commun des fragments d'haplotypes de grandes tailles. En conséquence, la probabilité d'identifier par hasard un mauvais fragment d'haplotype pour un candidat diminue, ce qui permet d'obtenir de bonnes imputations. À l'inverse, en diminuant les relations de parenté (en sautant une génération par exemple), on va donc dégrader le taux d'erreur génotypique (Dassonneville et al., 2011 ; Hayes et al., 2011 ; Hozé et al., 2013). Or, pour les puces DL 0.5 et QTL, avec un saut de génération, les taux d'erreurs diminuent (1.78% et 1.74% respectivement). Ce résultat surprenant peut s'expliquer par une différence de comportement entre les puces équidistantes construites avec une méthodologie mendélienne et les puces basées sur le DL construites avec une méthodologie populationnelle.

2.9 Impact sur les évaluations génomiques

Les sélectionneurs travaillant sur un classement des individus les uns par rapport aux autres, les corrélations de Spearman (rang) ont été calculées sur les 150 candidats ayant les meilleurs résultats d'évaluations à partir des génotypages HD afin d'étudier leur éventuel reclassement selon leurs résultats obtenus à partir des génotypages imputés.
Il est noté, pour les trois caractères étudiés (intensité de ponte, couleur de la coquille et poids d'œufs), une diminution significative des corrélations avec une diminution du nombre de SNP sur puces LD quand on compare la puce 10Kequi (9 352 SNP) avec la 3Kequi (3 337 SNP). De même, une baisse significative des résultats est observée entre la puce DL 0.5 et DL 0.05. Les corrélations entre GEBV diminuent lorsque le nombre de SNP sur puces LD et le déséquilibre de liaison diminuent, et donc lorsque les imputations sont moins bonnes.
En revanche, à densité de SNP équivalente, les résultats sont meilleurs significativement pour la puce 10Kequi comparée à la puce DL 0.5 et QTL pour la couleur de coquille (0.9695 contre 0.9460 et 0.9499) et le poids d'œufs (0.9855 contre 0.9741 et 0.9755). De même pour la puce 3Kequi comparée avec la puce DL 0.05, les résultats sont significativement meilleurs pour la couleur de coquille avec la puce 3Kequi. Et cela malgré des taux d'erreurs plus élevés sur les puces équidistantes comparées aux puces DL.
Trois hypothèses peuvent être avancées et seront à étudier par la suite. La première est que certaines erreurs d'imputations pourraient avoir plus d'importance que d'autres. Se tromper sur un SNP à effet fort peut avoir plus de conséquences que de se tromper un SNP à effet plus faible sur les caractères étudiés. La deuxième est que des erreurs techniques pourraient se produire en laboratoire lors du génotypage HD et subsister malgré le contrôle qualité. La dernière est que les évaluations génomiques sont réalisées en BLUP Single Step (Legarra et al, 2009), méthode qui favoriserait alors les puces équidistantes.

CONCLUSION

Cette étude a permis de poser les bases d'une problématique qui sera à approfondir par la suite : faut-il choisir une puce LD construite sur le seuil du DL ou bien sur la notion de distance entre SNP ? En effet, des taux d'erreurs plus faibles sont obtenus avec les puces DL plutôt que les puces équidistantes. Toutefois l'objectif de la sélection génétique est de choisir les individus ayant le meilleur potentiel génétique pour les caractères étudiés. Nous avons pu constater que les résultats des évaluations génomiques étaient meilleurs, à densité de SNP équivalente, pour les puces équidistantes. Conclure sur l'intérêt d'utiliser une puce basée sur le seuil de DL ou bien sur la notion de distance entre SNP est donc actuellement compliqué.

BIBLIOGRAPHIE

1. Browning B.L. et Browning S.R., 2016. *Am. J. Hum. Genet.* Vol. 98, n° 1, pp. 116-126.
2. Dassonneville R., Brøndum R.F., Druet T. et al., 2011. *J. Dairy Sci.* Vol. 94, n° 7, pp. 3679-3686.
3. Dassonneville R., Fritz S., Ducrocq V. et al., 2012. *J. Dairy Sci.* Vol. 95, n° 7, pp. 4136-4140.
4. Hayes B.J., Bowman P.J., Daetwyler H.D. et al., 2012. *Anim. Genet.* Vol. 43, n° 1, pp. 72-80.
5. Heidaritabar M., Calus M.P.L., Vereijken A. et al., 2014. *10th W.C.G.A.L.P.* pp. 3.
6. Heidaritabar M., Calus M.P.L., Vereijken A. et al., 2015. *BMC Genet.* Vol. 16, n° 1, pp. 101-114.
7. Hozé C., Fouilloux M.N., Venot E. et al., 2013. *Genet Sel Evol.* Vol. 45, n° 1, pp. 33-43.
8. Kranis A., Gheyas A.A., Boschiero C. et al., 2013. *BMC Genom.* Vol. 14, n° 1, pp. 59-71.
9. Legarra A., Aguilar I. et Misztal I., 2009. *J. Dairy Sci.* Vol. 92, n° 9, pp. 4656-4663.
10. Romé H., Varenne A., Hérault F. et al., 2015. *Genet. Selec. Evol.* Vol. 47, n° 1, pp. 83-93.
11. Sargolzaei M., Chesnais J.P. et Schenkel F.S., 2014. *BMC Genom.*. Vol. 15, n° 1, pp. 478.
12. Wolc A., Kranis A., Arango J. et al., 2014. *10th W.C.G.A.L.P.* pp. 6.
13. Robert R., Hérault F., Romé H. et al., 2015. *9th E.S.P.G.* pp. 43.