

Identification de mutations candidates affectant la composition du lait en acides gras et protéines par analyse d'association sur la séquence du génome dans les races Holstein, Montbéliarde et Normande.

GOVIGNON-GION A. (1)(2), BOICHARD D. (1)(2), CROISEAU P. (1)(2), BARBAT-LETERRIER A. (1)(2), HOZE C. (1) (3), FRITZ S. (1) (3), BOUSSAHA M. (1)(2), BROCHARD M. (4), SANCHEZ M.P. (1)(2)

(1) INRA, UMR1313 GABI, 78350 Jouy-en-Josas, France

(2) AgroParisTech, UMR1313 GABI, 78352 Jouy-en-Josas, France

(3) Alice, 75012 Paris, France

(4) Idele, 149 rue de Bercy, 75012 Paris, France

RESUME - Cette étude réalisée dans le cadre du programme PhénoFinlait vise à identifier les gènes et les mutations responsables de la variabilité génétique de la composition du lait en acides gras (AG) et protéines dans les trois races bovines laitières Holstein, Montbéliarde et Normande. Pour 23 AG et 6 protéines, le phénotype analysé est la moyenne des mesures par moyen infra rouge à chaque contrôle, corrigées pour les effets de milieu. Les séquences du génome complet de 8746 vaches génotypées avec la puce Illumina 50k ont été reconstituées avec le logiciel FImpute, en utilisant comme référence la séquence de 1147 taureaux du projet « 1000 génomes bovins ». Au total, 28 millions de polymorphismes ont été prédits. L'analyse GWAS a été conduite intra-race avec le logiciel GCTA, le modèle incluant un effet marqueur et un effet polygénique. Cette approche a permis de caractériser 30 à 40 régions par caractère sur tout le génome. On retrouve la majorité des régions déjà détectées avec la puce 50k, mais les intervalles de localisation sont plus précis. Cette première analyse a permis de définir des régions de l'ordre de 500 kb à 3 Mb, analysées par BayesC. Cette méthode multi-marqueurs réduit considérablement l'effet du déséquilibre de liaison à grande distance et permet de proposer des mutations causales candidates. La comparaison entre races et l'annotation des variants aident au choix entre ces propositions. Ce travail est illustré avec 10 exemples sur les chromosomes 5, 14, 17, 19 et 27 pour les AG et 1, 2, 6, 11, et 20 pour les protéines. Cette étude montre l'intérêt du séquençage de génomes entiers pour identifier les mutations responsables du déterminisme génétique des caractères.

Identification of candidate mutations affecting milk composition in fatty acids and proteins through genome wide association study at the sequence level in Holstein, Montbéliarde, and Normande breeds.

GOVIGNON-GION A. (1)(2), BOICHARD D. (1)(2), CROISEAU P. (1)(2), BARBAT-LETERRIER A. (1)(2), HOZE C. (1) (3), FRITZ S. (1) (3), BOUSSAHA M. (1)(2), BROCHARD M. (4), SANCHEZ M.P. (1)(2)

(1) INRA, UMR1313 GABI, 78350 Jouy-en-Josas, France

SUMMARY – This study, part of the PhénoFinlait project, aims at identifying the genes and mutations responsible for the genetic variability of milk fatty acids (FA) and proteins in Holstein, Montbéliarde and Normande dairy cattle breeds. For 23 FA and 6 proteins, the analyzed phenotype was the mean of mid infra-red test-day measurements, adjusted for environmental effects. Whole genome sequences (WGS) of 8746 cows genotyped with the 50k Illumina chip were imputed with FImpute software, using 1147 sequenced bulls of the “1000 bull genomes” project as a reference. Over 28 million polymorphisms were predicted. GWAS analysis was carried out within breed with GCTA software and the model included a marker effect and a polygenic effect. Overall, 30 to 40 QTL regions were detected per trait over the genome. Many regions already detected with the 50k chip were confirmed but location confidence intervals were smaller. From this first analysis, several 500kb – 3Mb long regions were defined and analyzed by BayesC. This multi-marker method strongly reduces the effect of long distance linkage disequilibrium and pinpoints a limited number of candidate mutations. Breed comparison and functional annotation of variants also provide information. This work is illustrated by 10 examples on chromosomes 5, 14, 17, 19 and 27 for FA and 1, 2, 6, 11, and 20 for proteins. This study shows the value of WGS to identify candidate causal variants underlying the genetic variation of complex traits.

INTRODUCTION

Les triglycérides, eux-mêmes constitués d'acides gras (AG), et les protéines sont les constituants majeurs du lait en termes de valeur ajoutée, de valeur nutritionnelle et de propriétés techno-fonctionnelles. Alors que leur mesure avec les méthodes de référence est très coûteuse, la spectrométrie moyen infra-rouge (MIR) est une méthode alternative utilisable à grande échelle pour mesurer la composition en AG et protéines et ainsi développer des applications diverses, dont la prise en compte en sélection. En les combinant avec les technologies de génotypage à haut débit, il est possible d'identifier les QTL (ou régions

génomiques responsables de la variabilité génétique) de ces composantes individuelles et de mettre en œuvre une sélection génomique.

Un objectif du programme PhénoFinLait est de disséquer l'architecture génétique de la composition en AG et protéines du lait (Brochard et al, 2014). Les équations MIR construites pour 23 acides gras et les 6 protéines majeures ont été appliquées sur trois populations bovines Montbéliarde, Normande et Holstein. Lors d'études antérieures, un « genome scan » a été réalisé sur un échantillon de vaches génotypées avec la puce 50k pour identifier les QTL d'AG

(Govignon-Gion et al, 2012) et de protéines (Sanchez et al, 2013). Cette étude reprend les mêmes données.

En parallèle, le consortium international "1000 génomes bovins" (Daetwyler et al, 2014) a produit une base de 1147 séquences de génomes complets dont ceux des ancêtres majeurs de près de 30 races bovines. L'objectif de cette étude est de combiner toutes les informations disponibles (phénotype, génotype et séquence) pour identifier les mutations génétiques causales sous-jacentes aux principaux QTL de composition fine du lait. Dans la suite du texte, le terme « variant » sera utilisé pour décrire l'ensemble des polymorphismes de la séquence et le terme « mutation » sera réservé aux polymorphismes causaux candidats.

1. MATERIEL ET METHODES

1.1. Phénotypes et génotypes SNP

Les phénotypes d'acides gras et de protéines, ainsi que les génotypes SNP, ont été obtenus comme décrit par Govignon-Gion et al (2012) et Sanchez et al (2013). Au total, 23 AG et 6 protéines ont été analysés. Pour chaque caractère, les prédictions MIR à chaque contrôle ont été corrigées pour les effets du troupeau-date de contrôle, du mois de mise bas et du stade de lactation estimés par un modèle BLUP. Le phénotype final, basé sur un minimum de 3 contrôles, est la moyenne par vache de ses performances ajustées. Parmi les vaches phénotypées, 8746 ont été génotypées avec la puce Bovine SNP50 Beadchip d'Illumina.

1.2. Données de séquence

Les données de séquence de génome complet de 1147 taureaux de 30 races ont été utilisées, correspondant au Run 4 du projet « 1000 génomes bovins ». Les filtres utilisés ont retenu 28 millions de variants considérés comme fiables. Ces données ont été utilisées comme référence pour imputer la séquence complète de 8746 vaches génotypées, à l'aide du logiciel FImpute (Saerger et al, 2014). Pour une meilleure précision, l'imputation a été réalisée en deux étapes, de 50k à 631k en utilisant les nombreux typages haute densité (HD) disponibles dans les trois races, puis vers la séquence.

1.3. Analyses d'association

Deux analyses successives ont été conduites. Dans un premier temps, chaque caractère a fait l'objet d'une analyse d'association (GWAS) intra race sur l'ensemble du génome, en testant individuellement l'effet des 28 millions de variants, dans un modèle prenant également en compte la valeur polygénique des animaux pour éviter tout biais de structuration. La parenté entre animaux a été estimée à partir de 631 000 variants de la puce HD. Les analyses ont été réalisées avec le logiciel GCTA (Yang et al, 2011). Un variant a été déclaré significatif si la probabilité sous H0 était inférieure à 10^{-6} , soit $\log_{10}(1/p) > 6$. Une région QTL a été définie comme l'intervalle, centré sur le variant le plus significatif dans un intervalle de 2 Mb maximum et incluant tous les variants avec une valeur de $\log_{10}(1/p)$ au moins égale aux 2/3 de la valeur au pic.

Dans un second temps, une analyse multi-marqueurs a été conduite sur des régions ciblées définies à partir des résultats du GWAS. Pour éviter tout biais de structuration familiale, le modèle prenait aussi en compte la valeur polygénique des individus. La méthode BayesC, implémentée dans le logiciel GS3 (Legarra et al, 2014), réalise, à chaque itération, une sélection des variants réellement utiles en éliminant la majorité de ceux qui, même à grande distance, ressortent significatifs dans l'analyse GWAS du fait de leur déséquilibre de liaison avec au moins une mutation causale sans avoir de rôle par eux-mêmes. Le critère de sélection utilisé est la probabilité d'inclusion du variant dans le modèle sur l'ensemble des itérations. Cependant, à l'échelle de la séquence, de nombreux variants sont en déséquilibre quasi complet entre eux, de sorte que l'analyse multi-marqueurs ne peut pas les distinguer et répartit leur probabilité d'inclusion

globale entre eux. Pour éviter cet écueil, les petites régions présentant les signaux les plus forts sont détectées en sommant les probabilités d'inclusion sur des fenêtres de 5000 paires de bases et les mutations candidates sont ensuite préférentiellement recherchées dans ces petites régions sélectionnées.

Les mutations candidates retenues après ces 2 étapes sont encore assez nombreuses et des informations complémentaires sont donc utilisées : - l'annotation fonctionnelle des variants (c'est-à-dire leur prédiction d'effet biologique quand elle est possible) réalisée dans le cadre du projet « 1000 génomes bovins » (Grant et al, 2011). - on suppose qu'une mutation causale ségrégeant dans plusieurs races a des effets similaires.

Les résultats acquis sont très nombreux et il n'est pas possible ici de les présenter tous. Nous ferons donc un focus sur quelques exemples.

2. RESULTATS

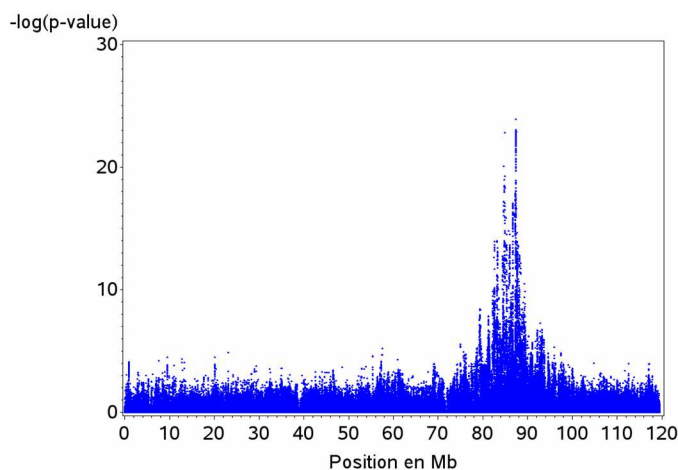
2.1. Analyse d'association mono marqueur

Le tableau 1 fournit un décompte des QTL détectés dans chaque race pour un groupe de caractères. Les QTL détectés dans les analyses 50k antérieures et dans la bibliographie (Stoop et al, 2009 ; Schenning et al, 2009 ; Shopen et al, 2011 ; Bouwman et al, 2011) sont confirmés. On retrouve ainsi les QTL des chromosomes 5, 14 (*DGAT1*), 17, 19, 26 (*SCD*), 27 (*AGPAT6*) affectant les AG et ceux des chromosomes 6 (caséines), 11 (*PAEP*), 14 (*DGAT1*), 20 (*GHR*) affectant les protéines. La résolution est augmentée de sorte que dans certains cas, on peut décomposer un QTL détecté précédemment en plusieurs. La figure 1 illustre un exemple de profil, pour la teneur en caséine κ et le chromosome 6 en race Montbéliarde.

Tableau 1. Nombre de QTL détectés par GWAS

| | Caractère | Montbéliarde | Normande | Holstein |
|-----------|------------------------|--------------|----------|----------|
| AG | C4:0 | 24 | 30 | 28 |
| | C12:0 | 24 | 18 | 31 |
| | Total Saturés | 23 | 11 | 26 |
| | Total C18:1 | 22 | 15 | 20 |
| | Oméga3 | 25 | 21 | 44 |
| Protéines | Caséine α s1 | 69 | 23 | 87 |
| | Caséine α s2 | 22 | 19 | 13 |
| | Caséine β | 18 | 18 | 35 |
| | Caséine κ | 24 | 26 | 26 |
| | α Lactalbumine | 27 | 19 | 19 |
| | β Lactoglobuline | 20 | 21 | 22 |

Figure 1. Race Montbéliarde, chromosome 6 : statistique de test (« Manhattan Plot ») pour la teneur en caséine κ .



La région des caséines (86 Mb) présente un gros pic qui semble composé de plusieurs pics élémentaires et il est possible qu'il y ait plusieurs QTL proches. A cette échelle de la mégabase, une forte proportion des QTL, et en particulier

les plus forts, apparaissent souvent partagés entre races, ce qui suggère a priori le partage d'un certain nombre de mutations causales.

Les quelques mutations déjà connues pour leur effet sur la composition du lait apparaissent en général avec des statistiques de test élevées, mais pas forcément au maximum de chaque QTL. Bien que l'analyse soit en principe quasi exhaustive (presque tous les variants existants sont analysés), ce constat illustre que le pic ne correspond pas forcément aux mutations causales, pour plusieurs raisons : - la statistique de test dépend de l'effet estimé (en principe maximal pour la mutation causale) mais aussi de la précision d'estimation de cet effet variant et donc de la fréquence allélique. Elle tend donc à favoriser les variants aux fréquences équilibrées ; - il peut y avoir plusieurs mutations causales, de sorte que le variant le plus significatif peut être celui le plus lié à l'ensemble de ces mutations, sans être causal lui-même. Globalement, cette analyse fournit déjà une quantité énorme d'information. Il est important de noter que la plupart de ces études se limitent à la seule l'analyse GWAS.

Cette analyse a aussi permis de cibler les régions à étudier plus en détail dans l'intervalle de confiance de localisation du QTL, incluant donc la ou les mutations causales. Le tableau 2 montre les régions présentées dans cet article.

2.1. Analyse d'association multi-marqueurs

Les régions analysées couvrent entre 500 kb et 3 Mb et contiennent plusieurs milliers ou dizaines de milliers de variants (environ un variant tous les 120 bases). Chaque variant est caractérisé par sa probabilité d'inclusion dans le modèle estimée par sa fréquence de sélection par le processus de Monte Carlo au cours des itérations. Idéalement, si les variants étaient indépendants, la mutation causale serait beaucoup plus souvent sélectionnée que les autres. Conditionnellement à la mutation causale, des variants liés n'apporteraient pas d'information supplémentaire et ne seraient donc pas sélectionnés. Mais avec un variant toutes les 120 bases en moyenne, le niveau d'association entre variants proches est très élevé, de sorte qu'il peut être difficile de distinguer leurs effets apparents. La probabilité d'inclusion globale d'une petite région, souvent forte, va alors se répartir sur un certain nombre de variants de cette région, chacun d'eux ne se voyant attribuer finalement qu'une probabilité faible. Il devient difficile de discerner les mutations causales des variants liés. Dans un premier temps, on a considéré les sommes de probabilité d'inclusion sur de petites régions de 5 kb (40 variants environ).

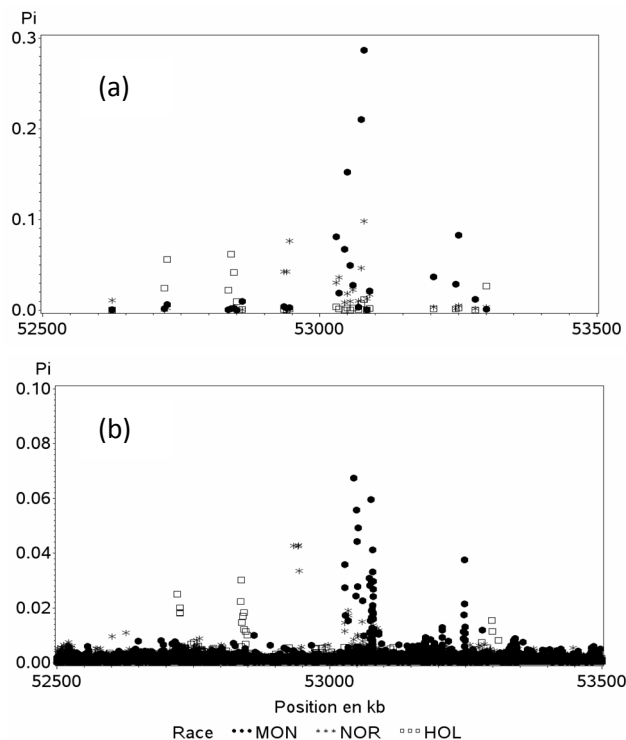
La figure 2, correspondant au caractère C4:0 sur une région du chromosome 17, montre l'intérêt de cette approche qui permet de visualiser les régions cibles, alors que les variants individuels fournissent un signal plus dilué. Les mutations candidates peuvent ensuite être recherchées plus spécifiquement dans ces petites régions de 5 kb.

2.1.1. Analyses de régions affectant les AG

a) Chromosome 27

Dans cet exemple qui couvre 500 kb et inclut 5031 variants au total, les fenêtres couvrant la région de 36,205 à 36,220 Mb sont les plus informatives. Cette sous-région contient 237 variants. L'élimination des variants de probabilité faible (<1%) dans les trois races permet de réduire le nombre de candidats à 21. Parmi ceux-ci, 12 variants présentent des probabilités plus élevées (mais encore assez faibles, comprises entre 2 et 5%), et la plupart dans au moins deux races. Ces variants très proches les uns des autres sont localisés près du gène *AGPAT6*, dont 8 dans la partie amont avec un rôle régulateur possible, et 4 parmi ces 8 avec les probabilités les plus élevées. Nous supposons donc que ces 4 variants localisés sur 3 kb entre les positions 36 209 319 et 36 212 352 sont candidats pour expliquer le QTL. Ce résultat est en accord avec Daetwyler et al (2014) qui proposent plusieurs variants possibles dans la même petite région.

Figure 2. Probabilités d'inclusion (P_i) dans l'analyse multi-marqueurs du C4:0 sur le chromosome 17. (a) somme des $P_i > 0,01$ par intervalles de 5 kb ; (b) P_i de chaque variant.



b) Chromosome 17

Ce QTL, peu décrit dans la littérature, affecte les AG saturés à chaîne courte, donc le début de la chaîne d'élongation. La région analysée, située entre 52,5 et 55 Mb, contient 21 812 variants. Une seule petite région de 10 kb autour de 53,080 Mb présente une somme de probabilité élevée. Parmi les 70 variants de cette région, 22 sont en déséquilibre de liaison quasi-total et ont la même probabilité. Ils sont tous dans des régions introniques du gène *BRI3BP*.

c) Chromosome 19

Ce QTL, l'un des plus importants sur la matière grasse, est largement décrit dans la littérature mais n'est pas encore caractérisé, même si plusieurs gènes candidats sont suspectés. Il est sans doute composé de plusieurs QTL. La partie analysée en multi-marqueurs couvre la région entre 50 et 53 Mb et contient 27 806 variants. Plusieurs fenêtres distinctes sont mises en évidence : la région autour de 50,990 Mb contient un variant à forte probabilité, situé dans la partie amont du gène *LOC100336469* ; la région autour de 51,150 Mb contient un autre variant à forte probabilité dans la région 3' du gène *LOC781977* ; la région autour de 51,440 Mb contient plusieurs variants liés de probabilité comprise entre 4 et 6% chacun, tous localisés dans la partie aval du gène *RFNG* ; enfin, la région principale entre 51,360 et 51,385 Mb comprend 6 variants à forte probabilité (jusqu'à 43%), identifiés dans les trois races, et tous présents dans le gène *FASN*, la plupart dans la zone amont du gène. *FASN* est un gène impliqué dans le métabolisme des AG.

d) Chromosome 5

La région de 92,5 à 94,5 Mb comprend 24 472 variants. Les signaux sont très forts mais légèrement décalés entre races. Les pics situés à 93,548 et 93,913 Mb en races Normande et Montbéliarde correspondent à des zones intergéniques, tandis que le pic en race Holstein est compris entre 93,944 et 93,956 Mb. Dans cette région du gène *MGST1*, 11 variants très liés ressortent de l'analyse, dont 4 avec les probabilités les plus élevées (5 à 9%) et localisés dans la région régulatrice amont vers 93,954 Mb.

e) Chromosome 14

Cette région comprend le gène *DGAT1* dont la mutation K232A a un effet majeur sur le taux butyreux (Grisart et al, 2001) et des effets pléiotropiques sur de nombreux autres

caractères. Cette région est cependant suspectée de porter d'autres mutations causales, car des QTL sont détectés dans des populations fixées pour la mutation K232A. La région analysée couvre environ 2 Mb et contient 27 308 variants. Les résultats sont complexes, puisqu'on met en évidence au moins 3 QTL. Le premier se situe vers 1,62 Mb, en amont de *DGAT1*, et contient un variant avec forte probabilité, localisé en 3' du gène *GPT* (impliqué dans le catabolisme des acides aminés). Ce variant présente une probabilité d'inclusion de 1 en race Montbéliarde ! Le second, comme attendu, est dans le gène *DGAT1* et dans son voisinage (gènes *HSF1*, *BOP1*, *MROH1*). Par contre, la mutation K232A n'apparaît pas comme la plus probable, fortement concurrencée par plusieurs de ses voisines qui présentent parfois de très fortes probabilités. Enfin, la troisième région est en aval, vers 2,705 Mb, avec plusieurs variants dans le gène *CYP11B1*. D'une façon générale, cette région du début du chromosome 14 semble particulièrement riche et le nombre de mutations causales est sans doute élevé.

2.1.2. Analyses de régions affectant les protéines

a) Chromosome 6

La région étudiée couvre le cluster des gènes des caséines, très significatif dans les trois races. Elle comprend 22 554 variants au total. Plusieurs pics sont visibles dans cette région complexe. La région la plus significative correspond à une petite fenêtre de 20kb avec un fort effet sur le taux de caséine κ , entre 87,390 et 87,410 Mb. Dans cette petite région, les 10 variants les plus probables sont dans la région régulatrice en aval du gène *CSN3*.

b) Chromosome 11

La région étudiée (102,3-104,3 Mb) comprend le gène *PAEP* (ou *LGB*) codant pour la β lactoglobuline et contient 23 143 variants. L'effet sur les protéines mais aussi les AG est considérable et il est trouvé dans les trois races. Le caractère le plus affecté est le taux de β lactoglobuline lui-même. La littérature rapporte un effet de deux mutations non synonymes totalement liées dans la partie codante du gène *PAEP*. Nos résultats montrent une situation moins simple. Le pic principal est situé dans le gène *PAEP* dans une région de 30 kb entre les positions 103,285 et 103,315 Mb. 19 variants présentent des probabilités très élevées. Tous sont dans les régions amont et aval du gène, sauf la mutation connue en 103 303 475, en région codante et responsable d'un changement d'acide aminé. Un deuxième pic est observé dans le gène *GPSM1*, 500kb en aval de *PAEP*, avec 3 mutations candidates introniques.

c) Chromosome 20

Le chromosome 20 est bien connu pour affecter le taux protéique, avec une mutation dans le gène *GHR* décrite par Blott et al (2003). Toutefois, le pic le plus significatif, trouvé dans les trois races, est dans une région très différente, entre 57,4 et 59,4 Mb, citée par Kemper et al (2013). Dans notre étude de cette région de 23 769 variants, le pic le plus marqué affecte l' α -lactalbumine dans une petite région de 30 kb, située entre 58,410 et 58,440 Mb. Dans cette région, tous

les variants sont dans des introns du gène *ANKH* qui code pour un régulateur du transport des phosphates sur-exprimé dans le tissu mammaire. On retient 10 variants présentant des probabilités élevées, comprises entre 10 et 85%.

d) Chromosomes 1 et 2

D'après l'analyse GWAS, un QTL situé sur le chromosome 1 autour de 144 Mb affecte le taux de caséine κ dans les trois races ; et un QTL du chromosome 2 vers 131 Mb affecte le taux de caséine α_2 . Deux petites régions de 7 kb sont définies à partir des analyses multi-marqueurs dans les intervalles 144,395-144,405 et 131,810-131,835 Mb. Sur le chromosome 1, 30 variants sont en complet déséquilibre et ont la même probabilité. Ils sont tous situés en région intronique du gène *SLC37A1*. Sur le chromosome 2, un seul variant ressort comme candidat (avec une probabilité modérée), en position intronique dans le gène *ALPL*. De par leurs fonctions respectives dans la synthèse du lait, ces 2 gènes sont de bons candidats.

CONCLUSION

Cette étude montre l'apport considérable des données de séquence pour approcher les mutations causales impliquées dans le déterminisme de caractères complexes. En général, la mutation causale n'est pas identifiée avec certitude du fait de la forte association entre variants proches, mais un petit nombre de variants liés peut être proposé. Ces variants sont généralement situés dans des gènes, mais pas dans les séquences codantes. Bien que les QTL soient fréquemment colocalisés entre races, cette étude confirme que ce sont les mêmes gènes qui sont concernés, mais suggère que les mutations causales peuvent être différentes d'une race à l'autre.

Les auteurs remercient le Consortium PhenoFinLait et les partenaires du projet « 1000 génomes bovins ». Les séquences ont été obtenues dans le cadre de projets financés par l'ANR (Cartoseq) et Apisgene.

- Blott S., et al. 2003. *Genetics*, 163, 253-266
 Boichard D., et al. 2014. *INRA Productions Animales*, 27, 283-298.
 Bouwman A.C., et al. 2011. *BMC Genetics* 12, 43
 Bouwman A.C., et al. 2014. *J Dairy Sci*, 97, 1139-1149.
 Brochard M., et al. 2013. *Renc. Rech. Rum.*, 16, 423.
 Daetwyler H.D., et al. 2014. *Nature Genetics*, 46, 858-867.
 Ganai N.A., et al. 2009. *Anim. Genet.*, 40, 127-133.
 Govignon-Gion A. et al. 2012. *Renc. Rech. Rum.*, 19, 65-72.
 Grant J.R., et al. 2011. *Bioinformatics* 27:2300-2301.
 Grisart B., et al. 2001. *Genome Research* 12:222-231.
 Kemper K.E., et al. 2015. *Genet Sel Evol*, 47, 29
 Legarra A., et al. 2014.. http://snp.toulouse.inra.fr/~alegarra/manualgs3_last.pdf
 Raven L.A., et al. 2014. *BMC Genomics* 15:62.
 Sanchez M.P., et al. 2013. *Renc. Rech. Ruminants*, 20, 149.
 Sargolzaei M., et al. 2014. *BMC Genomics*, 15:478
 Schopen G.C.B., et al. 2011. *J Dairy Sci* 94, 3148-3158
 Yang J., et al. 2011. *Am J Hum Genet.* 88: 76-8

Tableau 2. Bilan des analyses ciblées réalisées

| Chromosome | Région (Mb) | Caractère | Intervalle de confiance (kb) | Mutations candidates | Gènes | Annotations | Référence |
|------------|-------------|-------------------------|------------------------------|----------------------|-----------------------------|--------------|-----------------------|
| 5 | 92,5-94,5 | SAT | 93 940-93 955 | 4 | <i>MGST1</i> | Amont | Raven et al. 2014 |
| 14 | 1,3-3,8 | SAT, POLY | 1620-1625 | 1 | <i>GPT</i> | 3'UTR | - |
| | | SAT | 1790-1870 | 4 | <i>DGAT1</i> , <i>MROH1</i> | Diverses | Grisart et al. 2001 |
| | | POLY | 2700-2720 | 4 | <i>CYP11B1</i> | Amont / aval | - |
| 17 | 52,5-55,0 | C4:0 | 53 075-53 085 | 22 | <i>BRI3BP</i> | Introns | - |
| 19 | 50,0-53,0 | C12:0 | 51 360-51 385 | 6 | <i>FASN</i> | Amont | Bouwman et al. 2014 |
| 27 | 36,0-36,5 | C16:0 | 36 205-36 220 | 4 | <i>AGPAT6</i> | Amont | Daetwyler et al. 2014 |
| 6 | 86,3-88,4 | caséine κ | 87 390-87 410 | 10 | <i>CSN3</i> | Diverses | - |
| 11 | 102,3-104,3 | β -lactoglobuline | 103 285-103 315 | 19 | <i>PAEP</i> | Diverses | Ganai et al. 2009 |
| | | | 103 860-103 870 | 3 | <i>GPSM1</i> | Introns | - |
| 20 | 57,4-59,5 | α -lactalbumine | 58 410-58 440 | 10 | <i>ANKH</i> | Introns | Kemper et al. 2015 |
| 1 | 143,4-145,4 | caséine κ | 144 395-144 405 | 30 | <i>SLC37A1</i> | Introns | Kemper et al. 2015 |
| 2 | 130,8-132,9 | caséine α_2 | 131 810-131 835 | 1 | <i>ALPL</i> | Introns | - |