Identification of QTL and candidate mutations affecting major milk proteins in three French dairy cattle breeds

***M.P. Sanchez*[*], A. Govignon-Gion[*], M. Ferrand[†], M. Gele[†], D. Pourchet[‡], M.N. Rossignol[§], S. Fritz[*#], M. Boussaha[*], A. Capitan[*#], D. Rocha[*], G. Miranda[*], P. Martin[*], M. Brochard[†], D. Boichard[*]**

[*]INRA, UMR1313 GABI, 78350 Jouy en Josas, France ; [†]IDELE, 75012 Paris, France ; [‡]ECEL Doubs - Territoire de Belfort, 25640 Roulans, France ; [§]Labogena, 78350 Jouy en Josas, France ; [#]UNCEIA, 75012 Paris, France

ABSTRACT: A whole genome scan was performed to detect QTL for milk αs1, αs2, β, κ-casein, α-lactalbumin and β-lactoglobulin predicted from mid infrared (MIR) spectrometry. About 8,000 cows of Holstein, Montbéliarde and Normande breeds were genotyped mainly with the 50K SNP beadchip. A linkage disequilibrium and linkage analysis with 6 SNP-haplotypes revealed 61 to 75 QTL in each breed. Three of these QTL, common to all breeds (BTA6-87Mb, BTA11-103Mb and BTA20-58Mb), were highly significant and explained 16-51% of the genetic variance. Other very significant QTL were also observed on BTA2, 6, 14, 20 and 29. Concordance analyses, performed between QTL statuses and sequence-derived polymorphisms from 13 bulls, revealed already known causal polymorphisms in ***LGB*** (BTA11) and ***GHR*** (BTA20) genes and also novel mutations on BTA 2, 6, 20 and 29. These preliminary results are encouraging for further studies to identify causal mutations.

Keywords:
Dairy cattle
Milk proteins
QTL

Introduction

Milk protein composition plays an essential role in techno-functional milk properties (milk coagulation time, cheese yield, heat stability…). Whereas measuring protein composition with reference methods is very expensive, mid-infrared (MIR) spectrometry provides an alternative opportunity for large scale analyses. Once MIR spectra are recorded, milk protein composition can be estimated at a large scale and can be included in the breeding goal. Combined with high throughput genotyping technologies, it makes it possible to identify genomic regions responsible for genetic variation (QTL) of individual protein contents and to apply genomic selection.

One objective of the *PhénoFinLait* project was to dissect the genetic architecture of individual milk protein composition. MIR prediction equations were derived for the 6 major bovine milk proteins (αs1, αs2, β, κ-casein, α-lactalbumin, and β-lactoglobulin) from 500 reference samples analyzed with Reverse Phase HPLC. and were applied in the three French Montbéliarde (MO), Normande (NO) and Holstein (HO) dairy cattle breeds. A whole genome scan was carried out on these data to identify QTL affecting individual protein composition in MO, NO and HO cows. Finally, candidate mutations in the most significant QTL regions were searched using whole genome sequences from the 1000 bull genomes project (Daetwyler et al., 2014).

Materials and Methods

Data. Around 900,000 milk MIR spectra were collected from 160,253 MO, NO and HO cows from 2009 to 2012 and used to predict the milk concentration of whole protein and of the six main milk proteins. The accuracy (R^2) of the prediction equations ranged from 59% for α-lactalbumin to 92% for β-casein. Individual proteins were then expressed in proportion of milk or total proteins. Thirteen traits were analyzed: protein content (PC) and the 6 proteins expressed in percentage of milk and of proteins. Data from cows with at least three test-day records per lactation during the first three lactations were pre-corrected for non genetic effects using a mixed model. Herd x test-day, stage of lactation, year x month of calving, laboratory x spectrometer x test-month effects were included as fixed effects while animal genetic and permanent environment effects were assumed random. Pre-corrected data for fixed effects and the random permanent environment effect were then averaged per cow. All analyses were performed within breed.

Genotyping. DNA was extracted from blood samples of about 8,000 *PhénoFinLait* cows genotyped with the Illumina Bovine beadchips: 50K (~7,500 cows) or 7K and then imputed (~500 cows). A quality control was applied on SNP mapped to the 29 autosomes (UMD3.1 assembly) to remove SNP with a MAF (Minor Allele Frequency) lower than 5%, a call rate lower than 95% and with genotype frequencies departing from the Hardy-Weinberg equilibrium ($P<10^{-4}$). After filtering, imputation and phasing were carried out as described in Boichard et al. (2012). In total, 36,912, 37,363 and 39,683 SNP for 2,773 MO, 2,673 NO and 2,208 HO cows, respectively, were used for QTL detection analyses.

QTL detection analyses. A linkage disequilibrium and linkage analysis (LDLA) approach was applied to detect QTL (Meuwissen and Goddard, 2001). In order to maximize linkage disequilibrium between SNP and QTL, each SNP effect was estimated by considering haplotypes of six consecutive SNP in a model including also a mean and a polygenic effect. For each haplotype, the hypothesis of one QTL (H_1) was compared to the hypothesis of no QTL (H_0). The test statistic was computed as the likelihood ratio test (LRT), $LRT=-2\ln(L_0-L_1)$, where L_0 and L_1 were the maximized likelihoods under H_0 and H_1, respectively. QTL effects were estimated as the proportion of the genetic variance explained by the QTL: $\sigma^2_g/(\sigma^2_g+\sigma^2_u)$ with σ^2_g, the genetic variance due to the QTL and σ^2_u, the polygenic variance. Assuming that the theoretical

distribution of the LRT was a 2-df χ^2 distribution and that the whole analysis was equivalent to 65,000 independent tests (5,000 independent tests per trait and 13 traits), less than three false positives were expected with a LRT greater than 20 in each breed. In order to limit the number of false positive QTL, only results with LRT $\geq$ 20 were thus retained in this study.

Candidate mutation analyses. In most significant QTL regions, candidate mutations were searched by combining QTL statuses and whole genome sequence-derived polymorphisms of bulls. This was done in three steps for each QTL region.

1) Haplotypes of 15 SNP were defined considering 7 SNP at the left and 7 SNP at the right of the most significant SNP in the QTL region.

2) For each bull having more than 50 daughters, QTL statuses were determined by testing the haplotype effect transmitted by the sire to its daughters. Bulls were considered as heterozygous at the QTL when the p-value was lower than 0.05, homozygous if p-value > 0.10 and unknown if 0.05 < p-value < 0.10.

3) For bulls presenting both sequence and QTL status, polymorphisms in 6Mb regions (±3Mb around the most probable location of the QTL) that were concordant to QTL statuses (with one discordance tolerated to take into account sequence errors) were retained as candidate mutations. Only mutations expected to have an impact on the amino acid sequence were first considered.

Results and Discussion

QTL detection analyses. Numerous QTL were detected for milk protein composition in each of the studied population. In total, 62, 75 and 61 SNP x trait combinations were found significant with a LRT $\geq$ 20 in MO, NO and HO breeds, respectively. QTL were located on all the 29 *Bos Taurus* (BTA) autosomes, except BTA 8, 9, 12, 13 and 18 (figure 1). Nine of these QTL regions had very highly significant effects (LRT $\geq$ 50) on protein contents expressed in % of proteins (2, 6a, 6b and 29), in % of milk (20a) or in both units (6c, 11, 14 and 20b). Five of them, common to several breeds, were located within a 1-Mb interval and presented similar effects (same traits were generally affected with the same magnitude) for the different breeds suggesting a common genetic determinism (table 1).

The three most significant regions (6c, 11 and 20b) were evidenced in the MO, NO and HO breeds. The BTA11 QTL had particularly huge values of LRT (from 1,563 to 1,866 depending on the breed) and effects (from 41 to 51% of the genetic variance of the β-lactoglobulin content expressed in % of proteins, depending on the breed). The BTA6c and BTA20b QTL mainly affected the κ-casein and α-lactalbumin contents, respectively (from 16 to 32% of the genetic variance, according to the region and the breed). Other regions exclusively affected casein contents, except the BTA14 QTL region which had effects on both caseins and α-lactalbumin contents. These results are consistent with the QTL results obtained by Schopen et al. (2011) on protein contents estimated by capillary zone electrophoresis. Moreover, the number of QTL detected and the magnitude of their effects are in agreement with the high heritability (h^2) estimates obtained in the *PhénoFinLait* populations ($0.27 < h^2 < 0.86$; Brochard et al., 2013). Protein composition from MIR spectra seems therefore accurate enough for genetic investigations.

Candidate mutation analyses. In each of the most significant regions where QTL were detected in several breeds, concordances were searched between QTL statuses and sequences of bulls. The two QTL regions 6a and 20a found only in the HO breed were additionally investigated because polymorphisms associated to the protein composition were previously described in these two regions. Among all the bulls of the *PhénoFinLait* project, 7 MO, 2 NO and 4 HO had both sequences and a sufficient number of daughters with phenotypes and genotypes (from 53 to 226) to properly estimate their QTL statuses. Assuming identical causal polymorphisms between breeds, the 13 bulls were jointly analyzed. Depending on the region, QTL statuses were evaluated for 9 to 11 bulls (table 1). For the 2 to 4 remaining bulls, the QTL statuses could not be determined because bulls were homozygous for the tested haplotype or because the p-value associated to the haplotype effect was intermediate (0.05 < p-value < 0.10). Among all the bulls with a QTL status, from 1 (14 and 20a) to 6 (6b and 11) were found heterozygous to the QTL.

No concordance was found for two of the seven QTL regions analyzed: BTA6c and BTA14. On BTA6c, genes coding for caseins are tightly linked in a 250-kb cluster (***CSN1S1***, ***CSN1S2***, ***CSN2*** and ***CSN3***). Five known polymorphisms in these genes were additionally tested in the 13 bulls and none was concordant with the QTL statuses. It was probably due to the presence of different haplotypes formed by several polymorphisms in strong linkage disequilibrium in this region (Martin et al., 2002). On BTA14, the known ***DGAT1*** polymorphism (K232A; Grisart et al., 2001) could not be detected and it was probably due to the low sequence quality in the BTA14 centromeric region for the only heterozygous bull (HO).

However, with the approach used in our study, we succeeded in isolating known mutations in two QTL regions and it was particularly notable that only these mutations were found as concordant. Two mutations described by Ganai et al. (2009) as the causal genetic polymorphisms of β-lactoglobulin protein variants A and B were found in the ***LGB*** gene on BTA11 (103,303,475 and 103,304,757pb). Another known mutation previously found to affect PC was identified in the ***GHR*** gene on BTA20a at the position 31,909,478 (F279Y; Blott et al., 2003).

In addition, we have detected 3, 3, 4, 3 and 1 new concordant mutations in BTA 2, 6a, 6b, 20b and 29 regions, respectively. They were located in genes coding for known proteins except the 4 mutations in the 6b region that were located in a novel protein coding gene. The Y581S mutation in the ***ABCG2*** gene (Cohen-Zinder et al., 2003) located in the vicinity of the 6a QTL region was also tested and it could not explain the effects observed in our study because all bulls were homozygous *A/A* for this mutation.

Conclusion

The identification of numerous QTL with large effects on protein composition predicted from MIR spectra is very promising to select these traits from routinely collected MIR spectra. The *PhénoFinLait* population, consisting of 2,208 to 2,773 phenotyped and genotyped cows depending on the breed, represents a first reference population for genomic selection. In the near future, the MIR spectra combined with large scale cow genotyping will make rapidly grow the reference populations. If a clear breeding objective was defined, genomic selection could be easily implemented for these traits. Moreover, the present paper shows the efficiency of concordance analyses to highlight candidate mutations provided that only one causal polymorphism is present and that the sequence of the region is well characterized. In more complex regions with several causative mutations, association studies applied to cow sequences imputed from the 50K genotypes (via the 800K genotypes) would be preferable.

Literature Cited

Blott S., Kim J.J., Moisio S. et al. (2003). Genetics, 163, 253-266.

Boichard D., Guillaume F., Baur A. et al. (2012). Anim. Prod. Sci., 52, 115-120.

Brochard M., Sanchez M.P., Govignon-Gion et al. (2013). Renc. Rech. Ruminants, 20, 158.

Cohen-Zinder M., Seroussi E., Larkin D.M. et al. (2003). Genome Res., 15, 936-944.

Daetwyler H.D., Capitan A., Pausch H. et al. (2014). Nature Genetics, in press.

Ganai N.A., Bovenhuis H., Van Arendonk, J.A.M. et al. (2009). Anim. Genet., 40, 127-133.

Grisart B., Coppieters W., Farnir F. et al. (2001). Genome Res., 12, 222-231.

Martin P., Szymanowska M., Zwierzchowski L. et al. (2002). Reprod. Nutr. Dev., 42, 433-459.

Meuwissen T.H.E. and Goddard M.E. (2001). Genet. Sel. Evol., 33, 605-634.

Schopen G.C.B., Visker M.H.P.W, Koks P.D. et al. (2011). J. Dairy Sci., 94, 3148-3158

Figure 1. Genomic locations of detected QTL

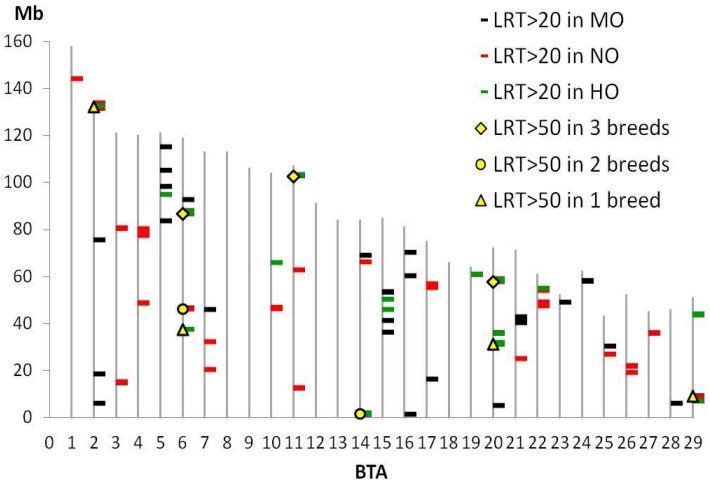

Table 1. Characteristics of the most significant QTL regions (LRT > 50) and results of concordance analyses

BTA	*Breed*	*QTL detection analyses*										*Concordance analyses*			
		Most sign. SNP			**Traits affected by the QTL region [2]**							**Bulls [3]**		**N pol [4]**	**Genes**
		Location (Mb)	**LRT**	**Effect [1]**	**PC**	**α-lact**	**β-lact**	**Cas αs1**	**Cas αs2**	**Cas β**	**Cas κ**	**HET**	**HOM**		
2	NO	132.6	55	7.3					**P**			3	9	3	3 genes (known proteins)
6a	HO	37.7	102	30.7				**P**				5	6	3	3 genes (known proteins)
6b	MO	46.7	86	6.6				**P**				6	5	4	1 gene (novel protein)
6b	NO	46.6	86	10.0				**P**							
6c	MO	87.0	310	16.9	M	M	P	M	MP	MP	**MP**	5	5	0	-
6c	NO	86.9	230	16.0	M	M		M	MP	MP	**MP**				
6c	HO	87.4	281	20.3							**MP**				
11	MO	103.3	1866	46.3			**MP**	P		P	P	6	3	2	1 gene (***LGB***)
11	NO	103.3	1752	41			**MP**	P			P				
11	HO	103.3	1563	51.3			**MP**			P	P				
14	NO	1.8	117	11.0	M	**M**		M	M	M	MP	1	9	-	-
14	HO	1.7	185	8.8	M	MP		**M**	M	M	M				
20a	HO	31.6	51	7.9							**M**	1	8	1	1 gene (***GHR***)
20b	MO	58.2	404	31.8		**MP**						3	7	3	3 genes (known proteins)
20b	NO	59.0	364	25.0		**MP**									
20b	HO	58.3	392	25.7		**MP**									
29	MO	9.3	66	4.9				**P**				2	10	1	1 gene (known protein)

[1]QTL effects are expressed in % of the genetic variance of the trait; [2]QTL has an effect on the trait expressed in % of protein (P) or in % of milk (M); in bold, the trait affected by the most significant QTL; [3]Number of bulls from the 1000 bull genome project homozygous (HOM) or heterozygous (HET) for the QTL; [4]Number of concordant polymorphisms with an impact on the amino acid sequence