Whole-genome association study for milk protein composition in dairy cattle

G. C. B. Schopen,*1 M. H. P. W. Visker,† P. D. Koks,† E. Mullaart,* J. A. M. van Arendonk,† and H. Bovenhuis†

*CRV, PO Box 454, 6800 AL Arnhem, the Netherlands
†Animal Breeding and Genomics Centre, Wageningen University, PO Box 338, 6700 AH Wageningen, the Netherlands

ABSTRACT

Our objective was to perform a genome-wide association study for content in bovine milk of αS1-casein (αS1-CN), αS2-casein (αS2-CN), β-casein (β-CN), κ-casein (κ-CN), α-lactalbumin (α-LA), β-lactoglobulin (β-LG), casein index, protein percentage, and protein yield using a 50K single nucleotide polymorphism (SNP) chip. In total, 1,713 Dutch Holstein-Friesian cows were genotyped for 50,228 SNP and a 2-step association study was performed. The first step involved a general linear model and the second step used a mixed model accounting for all family relationships. Associations with milk protein content and composition were detected on 20 bovine autosomes. The main genomic regions associated with milk protein composition or protein percentage were found on chromosomes 5, 6, 11, and 14. The number of chromosomal regions showing significant (false discovery rate <0.01) effects ranged from 3 for β-CN and 3 for β-LG to 12 for αS2-CN. A genomic region on Bos taurus autosome (BTA) 6 was significantly associated with all 6 major milk proteins, and a genomic region on BTA 11 was significantly associated with the 4 caseins and β-LG. In addition, regions were detected that only showed a significant effect on one of the milk protein fractions: regions on BTA 13 and 22 with effects on αS1-CN; regions on BTA 1, 9, 10, 17, 19, and 28 with effects on αS2-CN; a region on BTA 6 with an effect on β-CN; regions on BTA 13 and 21 with effects on κ-CN; regions on BTA 1, 5, 9, 16, 17, and 26 with effects on α-LA; and a region on BTA 24 with an effect on β-LG. The proportion of genetic variance explained by the SNP showing the strongest association in each of these genomic regions ranged from <1% for αS1-CN on BTA 22 to almost 100% for casein index on BTA 11. Variation associated with regions on BTA 6, 11, and 14 could in large part but not completely be explained by known protein variants of β-CN (BTA 6), κ-CN (BTA 6), and β-LG (BTA 11) or DGAT1 variants (BTA 14). Our results indicate 3 regions with major effects on milk protein composition, in addition to several regions with smaller effects involved in the regulation of milk protein composition.

Key words: genome-wide association, casein, α-lactalbumin, β-lactoglobulin

INTRODUCTION

Because of their nutritional value, dairy products make a significant contribution to human diets. For the dairy industry, not only milk protein yield but also milk protein composition is important; for example, milk with higher casein content will result in greater cheese yield (e.g., Emmons et al., 1990; Wedholm et al., 2006). In the Netherlands at present, payment of dairy farmers is based on protein yield. However, in practice it might be possible to produce more cheese from milk with the same amount of protein but a higher relative amount of casein.

Several studies have examined the effects of milk protein variants on the content of αS1-CN, αS2-CN, β-CN, κ-CN, α-LA, and β-LG (e.g., Ng-Kwai-Hang et al., 1987; Bobe et al., 1999; Heck et al., 2009). Heck et al. (2009) showed that variants of the β-CN and κ-CN genes, which are both located on BTA 6, and variants of the β-LG gene, which is located on BTA 11, were associated with milk protein composition. For example, β-LG protein variant B is associated with a higher casein content (e.g., Lundén et al., 1997; Heck et al., 2009) and, in addition, has been demonstrated to be associated with higher cheese yields (e.g., Van den Berg et al., 1992).

In addition to associations between known genetic variants of milk proteins and milk protein composition, the results of a genome-wide linkage study that we previously reported revealed significant QTL for milk protein content and composition on BTA 1, 3, 5, 9, 10, 14, 15, and 17 (Schopen et al., 2009b). However, confidence intervals of these QTL were, in general, large, which complicates identifying candidate genes. Confidence intervals can be reduced by moving to association studies and high-density SNP genotyping, which has recently become feasible. Moreover, new chromosomal regions affecting milk protein composition might be detected due to the higher power offered by an association study using high-density SNP genotypes. The objective of
This study, therefore, was to perform a genome-wide association study for milk protein composition (αs1-CN, αs2-CN, β-CN, κ-CN, α-LA, and β-LG), casein index, protein percentage, and protein yield using 50K SNP genotypes in Dutch Holstein-Friesian cows.

MATERIALS AND METHODS

Phenotypes

As part of the Dutch Milk Genomics Initiative, test-day morning milk samples, one from each cow, were collected from 1,912 first-lactation Holstein-Friesian cows from 398 commercial herds throughout the Netherlands. Details about the animals used in this study are available in Schopen et al. (2009a).

Protein percentage was determined by infrared spectroscopy. To calculate test-day morning milk protein yield, we multiplied protein percentage by test-day morning milk yield. Milk yields were missing for 141 cows, leaving 1,771 cows with records on protein yield. Milk protein composition was evaluated using capillary zone electrophoresis, as described by Heck et al. (2008). Using capillary zone electrophoresis, we quantified αs1-CN, αs2-CN, β-CN, κ-CN, α-LA, and β-LG. All 6 major milk proteins were expressed as weight-proportion of the total protein fraction (wt/wt%). Furthermore, the casein index was calculated as follows:

\[
\text{Casein index} = \frac{\sum \text{ casein}}{\sum \text{ casein} + \sum \text{ whey}} \times 100, \tag{1}
\]

where \( \sum \text{ casein} \) was defined as the sum of the percentages of αs1-CN, αs2-CN, β-CN, and κ-CN, and \( \sum \text{ whey} \) was calculated by summing the percentages of α-LA and β-LG. Table 1 gives the mean, phenotypic standard deviation, and intraherd heritability for milk protein composition, casein index, protein percentage, and protein yield. The intraherd heritability was calculated by dividing the additive genetic variance by the sum of the additive genetic variance and the error variance. More detailed information about the estimation of the heritability and the proportion of variance explained by herd is described by Schopen et al. (2009a).

Genotypes

We isolated DNA from blood samples of the cows. For this study, DNA was available from cows of 5 large paternal half-sib families (consisting of 214, 187, 175, 174, and 97 cows) and from cows of 53 small paternal half-sib families (consisting of 9–29 cows). A 50K SNP chip was designed by CRV (cooperative cattle improvement organization, Arnhem, the Netherlands), obtained from Illumina (San Diego, CA), and used to genotype all animals with the Infinium assay (Illumina). Charlier et al. (2008) provide more information about this custom 50K SNP chip. Sequences of all SNP mentioned in the paper are given in the supplementary information (Supplemental Table S1; available online at http://www.journalofdairyscience.org/). Genotyping resulted in 50,856 technically successful SNP. Of all SNP, 48,910 had a genotyping rate >99%, 49,965 SNP had a genotyping rate >95%, and 50,249 SNP had a genotyping rate >90%. The average minor allele frequency was 0.24. The SNP were mapped using the bovine genome assembly BTAU4.0 (Li et al., 2009). Of the 50,856 SNP, 778 could not be mapped to one of the 29 bovine autosomes or the X chromosome; these unmapped SNP were assigned to chromosome Null. Of the 50,856 SNP, 233 were excluded from the association study because they were monomorphic and 395 were excluded because they had a genotyping rate <80%. More details can be found in the supplementary information (Supplemental Table S2; available online at http://www.journalofdairyscience.org/). The final set of genotypes used in the association study consisted of 50,228 SNP.

In total, 1,868 animals were genotyped; however, not all genotyped animals had phenotypes, and not all phenotyped animals had genotypes. The data set that was used in the association study consisted of 1,713 animals with both phenotypes and genotypes, of which 130 animals had no record for protein yield.

The polymorphisms responsible for the known protein variants of κ-CN (A, B, and E) and β-LG (A and...
B), as well as the DGAT1 K232A polymorphism were included in the 50K SNP chip (Supplemental Table S3; available online at http://www.journalofdairyscience.org/). Some of the polymorphisms responsible for the known protein variants of β-CN (A₁, A₂, A₃, B, and I) were included in the 50K SNP chip (Supplemental Table S3) and all polymorphisms responsible for the known protein variants of β-CN were genotyped using a SNaPshot assay as described by Visker et al. (2011).

**Whole-Genome Association**

The whole-genome association study was performed using a 2-step procedure. In the first step, a single SNP analysis was performed using the SNPassoc package (González et al., 2007) in R (R Development Core Team, 2005) using the following general linear model:

\[
y_{ij} = \text{sire}_i + \text{SNP}_j + e_{ij},
\]

where \( y_{ij} \) was the phenotype adjusted for systematic environmental effects: days in lactation, age at first calving, season of calving, and herd; \( \text{sire}_i \) was the fixed effect of sire \( i \); \( \text{SNP}_j \) was the fixed effect of the \( j \)th class of the SNP; and \( e_{ij} \) was the random residual effect \( e_{ij} \sim N(0, \sigma^2_e) \). Phenotypes used in the general linear model were preadjusted for systematic environmental effects that were estimated based on all 1,912 cows with phenotypes, as described by Schopen et al. (2009a). Sire effect was included in the general linear model to account for a sire family effect. The SNP effects were modeled as class variables.

Additional genetic relationships that might exist among individuals, besides the ones accounted for by the sire effect, were not considered in the general linear model. Not accounting for those genetic relationships might lead to an inflated test statistic (Kennedy et al., 1992). Therefore, we accounted for all genetic relationships among individuals in the second step of the whole genome association study. In this step, we reanalyzed all SNP in genomic regions containing SNP that were significantly associated with one of the traits (false discovery rate, FDR <0.01) as determined by the qvalue package (Storey and Tibshirani, 2003) in R. The FDR was calculated based on the \( P \)-values obtained from the general linear model for all significant SNP in the second step of the analysis. Each single SNP was simultaneously adjusted for systematic environmental effects and for all genetic relationships among individuals in ASReml (Gilmour et al., 2002) by using the following animal model:

\[
y_{ijklmn} = \mu + b_i \times \text{lactst}_{ijklmn} + b_2 \times e^{-0.05 \times \text{lactst}_{ijklmn}} + b_3 \times \text{ca}_{ijklmn} + b_4 \times \text{ca}_{ijklmn}^2 + \text{season}_i \times \text{scode}_j + \text{SNP}_k \times \text{animal}_l + \text{herd}_m + e_{ijklmn},
\]

where \( y_{ijklmn} \) was the dependent variable; \( \mu \) was the overall mean; \( b \) is a constant; \( \text{lactst}_{ijklmn} \) was the covariate describing the effect of days in lactation; \( \text{ca}_{ijklmn} \) was the covariate describing the effect of age at first calving; \( \text{season} \) was the fixed effect with 3 classes for calving season (June–August 2004, September–November 2004, and December 2004–February 2005); \( \text{scode}_j \) was the fixed effect accounting for possible differences in genetic level between proven bull daughters and young bull daughters; \( \text{SNP}_k \) was the fixed effect of the SNP, modeled as a class variable; \( \text{animal}_l \) was the random additive genetic effect of animal \( l \); \( \text{herd}_m \) was the random herd effect; and \( e_{ijklmn} \) was the random residual effect. The variance-covariance structure of the additive genetic effects was \( \text{Var}(\text{animal}) = \mathbf{A} \sigma^2_a \), where \( \mathbf{A} \) was a matrix of additive genetic relationships among individuals and \( \sigma^2_a \) was the additive genetic variance. Heritabilities were re-estimated in these analyses.

After running all analyses using the animal model, the most significant SNP for each trait in each genomic region was examined for 3 criteria. This examination was performed to check the sensitivity of the association for genotype classes with few, possibly extreme, observations. If the SNP was discarded based on one of the criteria, the next most significant SNP in that region was examined for the same 3 criteria. Regarding the first criterion, when the number of animals for 1 of the genotype classes was <10, the association analysis was repeated without that genotype class. The SNP was discarded if the association was no longer significant. Second, when a phenotype in one of the genotype classes was more than 3 phenotypic standard deviations from the mean of that genotype class, the association analysis was repeated without that phenotype. The SNP was discarded if the association was no longer significant. Third, when a SNP had a minor allele frequency of <2%, the SNP was discarded.

**Significance Thresholds and SNP Variance**

Significance thresholds were obtained by calculating the FDR using the qvalue package (Storey and Tibshirani, 2003) in R. The FDR was calculated based on the \( P \)-values obtained from the general linear model for all
 Associations Using the Mixed Model

After running all analyses using the mixed model, the most significant SNP in each genomic region was examined for 3 quality criteria, which resulted in elimination of 4 genomic regions (2, 7, 12, and 23). Regions 7, 12, and 23 each consisted of 1 single significant SNP, whereas region 2 consisted of 5 SNP, of which 3 were significant. The SNP were eliminated because the significance could be ascribed to a genotype class of fewer than 10 animals (regions 12 and 23), because the significance could be ascribed to an extreme phenotype in one of the genotype classes (regions 2 and 12), or because the minor allele frequency of the SNP was <2% (regions 2 and 7). Consequently, 31 genomic regions distributed over 20 chromosomes showed significant associations with one or more of the studied traits (Table 2). Genomic regions were identified based on the bovine genome assembly BTAU4.0. Mapping the most significant SNP to the UMD3.0 physical map (ftp://ftp.cbcb.umd.edu/pub/data/Bos_taurus/) showed some differences in positions, but the genomic regions showing significant association were the same (Supplemental Table S6; available online at http://www.journalofdairyscience.org/). For each of the 6 milk proteins investigated in this study, we found more than one chromosomal region significantly associated with the trait, ranging from 3 regions for β-CN and β-LG to 12 regions for αS2-casein (Table 2). For 5 regions (regions 6_2, 10_1, 11, 14, and 29), multiple traits showed significant association (Table 2). On BTA 6, region 6_2 was significantly associated with all 6 major milk proteins, whereas a chromosomal region on BTA 11 was significantly associated with all major milk proteins except α-LA. Four regions (regions 6_2, 10_1, 14, and 29) were significantly associated with protein percentage and at least 1 of the 6 major milk proteins. The other 26 regions were uniquely associated with one of the traits studied: αS1-CN (region 13_1, 22_1, and 22_2), αS2-CN (regions 1_2, 9_2, 10_2, 17_1, 19, 28_1, and 28_2), β-CN (region 6_1), κ-CN (regions 13_2 and 21), α-LA (regions 1_1, 5_1, 5_2, 9_1, 16, 17_2, and 26), β-LG (region 24), and protein percentage (region 5_3, 15, 20, 25_1, and 25_2) (Table 2). The genomic regions showing highly significant associations [−log10(P-value) >10] on milk protein composition or protein percentage were found on BTA 5, 6, 11, and 14 (Table 2).

On BTA 5 in region 5_1, SNP ULGR_SNP_U63109_1966 was significantly associated with α-LA [−log10(P-value) = 21.45]. The estimated proportion of genetic variance for α-LA explained by this SNP was 9.9% (Table 2). The SNP ULGR_SNP_U63109_1966 is known from previous studies as “+15 relative to the transcription start point” (Bleck and Bremel, 1993). No significant associations of this SNP with any of the other milk proteins investigated in this study were detected.

On BTA 6 in region 6_2, several SNP were significantly associated with the 6 major milk proteins, casein index, and protein percentage. The significance level of the associations ranged from −log10(P-value) = 6.31 for β-LG to −log10(P-value) = 100.69 for β-CN. The estimated proportion of genetic variance explained by these SNP ranged from 2.0% for β-LG to 95.7% for β-CN (Table 2). The SNP ULGR_BTC-060550 is known from previous studies as “+15 relative to the transcription start point” (Bleck and Bremel, 1993). No significant associations of this SNP with any of the other milk proteins investigated in this study were detected.

The proportion of genetic variance explained by a SNP was calculated from the estimated genotype effects obtained from the animal model and the observed genotype frequencies. The SNP variance was expressed as proportion of the total additive genetic variance.
(Visker et al., 2011), whereas SNP rs43703016 is 1 of the 2 SNP that are causal for protein variants A and B of κ-CN. We investigated whether the known protein variants of β-CN and κ-CN could explain the detected associations for milk protein composition on BTA 6. Therefore, we performed additional analyses in which all SNP in region 6_2 were reanalyzed after adjusting for protein variants. First, the effects of the β-CN or κ-CN protein variants were estimated for each trait using the mixed model. Subsequently, adjusted records were calculated by subtracting the effects of the protein variants. These adjusted records were used in a mixed model analysis to obtain SNP effects after adjusting for the known β-CN or κ-CN protein variants. Adjusting for β-CN protein variants A\(^1\), A\(^2\), A\(^3\), B, and I or κ-CN protein variants A, B, and E changed the SNP showing the most significant association (Table 3). The proportion of genetic variance explained by SNP in region 6_2 decreased after adjusting for the known protein variants. A dramatic decrease was observed for variance in β-CN content explained by region 6_2 after accounting for β-CN variants. After accounting for κ-CN variants, the SNP most significantly associated with κ-CN content had a –log\(^{10}\) (P-value) of 29.27 and still explained 13% of the genetic variance for κ-CN (Table 3).

On BTA 11, significant associations were found with α\(_S\)-CN, α\(_S\)-CN, β-CN, κ-CN, β-LG, and casein index. The most significant association was found for β-LG...
Table 2. Single nucleotide polymorphisms showing the most significant association with the 6 major milk proteins, casein index, or protein percentage in each of 31 genomic regions with their position, name, minor allele frequency (MAF), the –log10(P-value) based on the association analysis using the mixed model, and the proportion of explained genetic variance (VarSNP)

<table>
<thead>
<tr>
<th>Region</th>
<th>Trait</th>
<th>Position (bp)</th>
<th>SNP name</th>
<th>MAF (%)</th>
<th>–log10(P-value)</th>
<th>VarSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_1</td>
<td>α-LA</td>
<td>109,682,944</td>
<td>ULGR_BTA-120188</td>
<td>46.8</td>
<td>5.96</td>
<td>2.6</td>
</tr>
<tr>
<td>1_2</td>
<td>αs2-CN</td>
<td>149,189,841</td>
<td>ARS-BFGL-NGS-8140</td>
<td>44.9</td>
<td>6.30</td>
<td>2.3</td>
</tr>
<tr>
<td>5_1</td>
<td>α-LA</td>
<td>34,386,828</td>
<td>ULGR_SNP_U63109_1966</td>
<td>24.9</td>
<td>21.45</td>
<td>9.9</td>
</tr>
<tr>
<td>5_2</td>
<td>α-LA</td>
<td>75,273,391</td>
<td>rs41604573</td>
<td>12.9</td>
<td>3.81</td>
<td>1.8</td>
</tr>
<tr>
<td>5_3</td>
<td>Protein (%)</td>
<td>98,502,764</td>
<td>rs41569048</td>
<td>24.2</td>
<td>4.72</td>
<td>2.0</td>
</tr>
<tr>
<td>6_1</td>
<td>β-CN</td>
<td>44,186,279</td>
<td>ULGR_BTC-054780</td>
<td>20.4</td>
<td>6.10</td>
<td>3.2</td>
</tr>
<tr>
<td>6_2</td>
<td>αs2-CN</td>
<td>88,139,061</td>
<td>ULGR_BTC-043582</td>
<td>19.5</td>
<td>26.87</td>
<td>16.4</td>
</tr>
<tr>
<td>6_2</td>
<td>Protein (%)</td>
<td>83,566,807</td>
<td>ULGR_BTC-053514</td>
<td>24.5</td>
<td>38.09</td>
<td>15.2</td>
</tr>
<tr>
<td>6_2</td>
<td>β-CN</td>
<td>88,330,265</td>
<td>ULGR_BTC-060550</td>
<td>30.4</td>
<td>100.69</td>
<td>95.7</td>
</tr>
<tr>
<td>6_2</td>
<td>κ-CN</td>
<td>88,532,332</td>
<td>rs43703016</td>
<td>30.9</td>
<td>63.88</td>
<td>25.9</td>
</tr>
<tr>
<td>6_2</td>
<td>Protein (%)</td>
<td>88,532,332</td>
<td>rs43703016</td>
<td>30.9</td>
<td>16.36</td>
<td>7.2</td>
</tr>
<tr>
<td>6_2</td>
<td>β-LG</td>
<td>88,532,332</td>
<td>rs43703016</td>
<td>30.9</td>
<td>6.31</td>
<td>2.0</td>
</tr>
<tr>
<td>9_1</td>
<td>Casein index</td>
<td>17,377,901</td>
<td>rs43702921</td>
<td>31.6</td>
<td>4.60</td>
<td>3.1</td>
</tr>
<tr>
<td>9_2</td>
<td>Protein (%)</td>
<td>107,166,278</td>
<td>rs41604573</td>
<td>38.8</td>
<td>4.42</td>
<td>1.8</td>
</tr>
<tr>
<td>10_1</td>
<td>Protein (%)</td>
<td>144,126,768</td>
<td>rs41569048</td>
<td>48.2</td>
<td>4.42</td>
<td>1.8</td>
</tr>
<tr>
<td>10_1</td>
<td>Protein (%)</td>
<td>83,566,807</td>
<td>ULGR_BTC-060550</td>
<td>30.4</td>
<td>100.69</td>
<td>95.7</td>
</tr>
<tr>
<td>10_1</td>
<td>Protein (%)</td>
<td>88,532,332</td>
<td>ULGR_BTC-043582</td>
<td>30.9</td>
<td>16.36</td>
<td>7.2</td>
</tr>
<tr>
<td>11</td>
<td>αS1-CN</td>
<td>107,166,278</td>
<td>rs41604573</td>
<td>42.3</td>
<td>9.00</td>
<td>3.6</td>
</tr>
<tr>
<td>11</td>
<td>Protein (%)</td>
<td>107,166,278</td>
<td>rs41569048</td>
<td>42.3</td>
<td>6.10</td>
<td>3.2</td>
</tr>
<tr>
<td>11</td>
<td>Protein (%)</td>
<td>107,166,278</td>
<td>rs41569048</td>
<td>42.3</td>
<td>9.10</td>
<td>3.6</td>
</tr>
<tr>
<td>13_1</td>
<td>αs2-CN</td>
<td>38,193,375</td>
<td>ULGR_SNP_AJ318490_1c7</td>
<td>40.1</td>
<td>10.19</td>
<td>6.0</td>
</tr>
<tr>
<td>14</td>
<td>αs2-CN</td>
<td>445,086</td>
<td>ULGR_SNP_AJ318490_1c7</td>
<td>40.1</td>
<td>15.23</td>
<td>5.7</td>
</tr>
<tr>
<td>15</td>
<td>Protein (%)</td>
<td>51,890,939</td>
<td>ARS-BFGL-NGS-107234</td>
<td>17.7</td>
<td>8.21</td>
<td>3.8</td>
</tr>
<tr>
<td>16</td>
<td>α-LA</td>
<td>53,592,299</td>
<td>rs41604569</td>
<td>4.7</td>
<td>4.16</td>
<td>2.3</td>
</tr>
<tr>
<td>17_1</td>
<td>αs2-CN</td>
<td>19,358,686</td>
<td>ULGR_AAF03022572_96396</td>
<td>31.7</td>
<td>4.26</td>
<td>1.8</td>
</tr>
<tr>
<td>17_2</td>
<td>α-LA</td>
<td>35,660,215</td>
<td>rs41569048</td>
<td>36.7</td>
<td>2.83</td>
<td>3.0</td>
</tr>
<tr>
<td>19</td>
<td>αs2-CN</td>
<td>36,162,641</td>
<td>rs41604573</td>
<td>24.5</td>
<td>3.73</td>
<td>2.2</td>
</tr>
<tr>
<td>20</td>
<td>Protein (%)</td>
<td>38,325,406</td>
<td>rs41604170</td>
<td>15.9</td>
<td>6.88</td>
<td>2.9</td>
</tr>
<tr>
<td>21</td>
<td>κ-CN</td>
<td>47,056,558</td>
<td>BTB-00821654</td>
<td>15.0</td>
<td>4.44</td>
<td>1.9</td>
</tr>
<tr>
<td>22_1</td>
<td>αs2-CN</td>
<td>40,264,878</td>
<td>rs41604569</td>
<td>2.1</td>
<td>3.97</td>
<td>2.0</td>
</tr>
<tr>
<td>22_2</td>
<td>αs2-CN</td>
<td>52,343,527</td>
<td>ARS-BFGL-NGS-15809</td>
<td>16.8</td>
<td>3.39</td>
<td>0.5</td>
</tr>
<tr>
<td>24</td>
<td>β-LG</td>
<td>35,667,068</td>
<td>rs41604569</td>
<td>46.6</td>
<td>5.01</td>
<td>1.6</td>
</tr>
<tr>
<td>25_1</td>
<td>Protein (%)</td>
<td>39,290,012</td>
<td>ULGR_BTA-60666</td>
<td>19.2</td>
<td>3.72</td>
<td>1.4</td>
</tr>
<tr>
<td>25_2</td>
<td>Protein (%)</td>
<td>34,973,002</td>
<td>ARS-BFGL-NGS-38306</td>
<td>34.2</td>
<td>3.77</td>
<td>1.7</td>
</tr>
<tr>
<td>26</td>
<td>α-LA</td>
<td>33,137,824</td>
<td>ULGR_BTC-053514</td>
<td>24.1</td>
<td>5.08</td>
<td>2.1</td>
</tr>
<tr>
<td>28_1</td>
<td>αs2-CN</td>
<td>18,554,359</td>
<td>rs41604573</td>
<td>44.0</td>
<td>3.67</td>
<td>1.4</td>
</tr>
<tr>
<td>28_2</td>
<td>αs2-CN</td>
<td>43,216,794</td>
<td>rs41588847</td>
<td>12.7</td>
<td>3.79</td>
<td>1.2</td>
</tr>
<tr>
<td>29</td>
<td>αs2-CN</td>
<td>45,241,089</td>
<td>ULGR_SNP_CAPN1-AF248054-7813</td>
<td>9.6</td>
<td>4.51</td>
<td>1.5</td>
</tr>
<tr>
<td>29</td>
<td>κ-CN</td>
<td>45,924,822</td>
<td>rs41586192</td>
<td>29.4</td>
<td>5.87</td>
<td>5.7</td>
</tr>
<tr>
<td>29</td>
<td>Protein (%)</td>
<td>45,488,026</td>
<td>rs41586192</td>
<td>32.6</td>
<td>5.08</td>
<td>2.1</td>
</tr>
</tbody>
</table>

1Regions are identified by the chromosome number and the region on that chromosome: _1 is the first region, _2 is the second region, and _3 is the third region on a chromosome.

2This value is expressed as proportion of the additive genetic variance.

3The SNP is located within the promoter region of the α-LA gene and is known from previous studies as “+15” (Bleck and Bremel, 1993).

4The SNP is located in intron 7 of the β-CN gene and is in full linkage disequilibrium with the SNP in exon 7 (rs43703011) that is causal for protein variants A1 and A2 of β-CN (Visker et al., 2011).

5The SNP is 1 of the 2 SNP that are causal for protein variants A and B of κ-CN.

6The SNP is located in the promoter region of the β-LG gene and is in full linkage disequilibrium with the SNP in exon 3 (ULGR_SNP_X14710_3984) and the SNP in exon 4 that are causal for protein variants A and B of β-LG (Ganai et al., 2009).

7The SNP is in full linkage disequilibrium with SNP ULGR_SNP_AJ318490_1b. Both SNP gave similar results and are the 2 SNP responsible for the DGAT1 K232A polymorphism.
The estimated proportion of genetic variance explained by this SNP ranged from 3.6% for αS2-CN to 99.8% for casein index (Table 2). The SNP rs41255679 is known to be in linkage disequilibrium with protein variants A and B of β-LG (Ganai et al., 2009). Adjusting for β-LG protein variants A and B decreased the proportion of genetic variance explained by this region on BTA 11, most considerably for β-LG content and casein index (Table 3).

On BTA 14, significant associations were detected with αS1-CN, αS2-CN, and protein percentage. The most significant association was found for protein percentage \([-\log_{10}(P\text{-value}) = 45.23]\). The estimated proportion of genetic variance explained by this SNP ranged from 5.7% for αS2-CN to 17.5% for protein percentage (Table 2). The SNP ULGR_SNP_AJ318490_1c is 1 of the 2 SNP responsible for the DGAT1 K232A polymorphism. Adjusting for the DGAT1 K232A polymorphism reduced the proportion of genetic variance explained by this region to 1.5% for αS1-CN, 0.9% for αS2-CN, and 1% for protein percentage (Table 3).

Among the unmapped SNP, SNP were significantly associated with αS1-CN, αS2-CN, κ-CN, α-LA, β-LG, casein index, and protein percentage (Supplemental Table S4; available online at http://www.journalofdairyscience.org/). A BLAST analysis was performed for these SNP, using the UMD3.0 physical map instead of the BTAU4.0 genome assembly. All 9 significant unmapped SNP could be mapped to positions on BTA 5, 6, 11, and 14, which were located in regions already defined in this study (regions 5_1, 6_2, 11, and 14 in Table 2) as significantly associated with milk protein composition or protein percentage (Supplemental Table S4).

### Genetic Variance Explained

For each trait and region, the most significant SNP were identified and these SNP were analyzed simultaneously (multiple SNP analysis) using the mixed model. In total, between 25 and 35% of the additive genetic variation for αS1-CN, αS2-CN, κ-CN, and α-LA was explained by the regions identified in this study. All additive genetic variation for β-CN was explained by SNP in 3 genomic regions, and most of the genetic variation for β-LG (81.6%) was explained by the 3 re-
regions identified in this study (Table 4). The estimated proportion of genetic variance explained in the multiple SNP analysis was, in most cases, similar to the sum of variances estimated based on the single SNP analyses (Table 4). For αS2-CN, simultaneous analysis of the most significant SNP resulted in a lower proportion of explained genetic variance. This result is in agreement with the reported significant interaction between QTL affecting αS2-CN on BTA 6 and 17 (Schopen et al., 2009b).

### DISCUSSION

This study reports on genome-wide association in dairy cattle for the 6 major milk proteins (αS1-CN, αS2-CN, β-CN, κ-CN, α-LA, and β-LG), casein index, protein percentage, and protein yield. This work is, to our knowledge, the first to describe results of a genome-wide association study for bovine milk protein composition.

Using a mixed model and accounting for existing family relationships is the most appropriate analysis for an association study. However, this analysis is time consuming, especially if multiple traits are being analyzed. Therefore, whole-genome association analysis was performed using a 2-step procedure. In the first step, a general linear model accounting for sire family relationships only was used, and in the second step, a mixed model accounting for all family relationships was used. The \( P \)-values from the linear model were in general slightly lower than those from the mixed model. This is in agreement with Kennedy et al. (1992) and Yu et al. (2006), who both showed that not accounting for family relationships would result in an inflated test statistic. For SNP analyzed using the mixed model, the correlation between \( P \)-values from the general linear model and the mixed model ranged from 0.85 for α-LA to 0.92 for κ-CN. This is in agreement with Aulchenko et al. (2007), who showed that \( P \)-values from a general linear model are highly correlated with \( P \)-values from a mixed model.

Single nucleotide polymorphisms significantly associated with milk protein composition or protein percentage were found on 20 bovine autosomes (BTA 1, 5, 6, 9, 10, 11, 13, 14, 15, 16, 17, 19, 20, 21, 22, 24, 25, 26, 28, and 29). The results of this association study are in good agreement with our previously reported linkage study performed on a subset of the same population (Schopen et al., 2009b). All genome-wide significant QTL from the linkage study were confirmed in the present study except for a QTL affecting β-CN on BTA 3. Several suggestive QTL from the linkage study were not confirmed, whereas associations on BTA 20, 21, and 22 were not significant in the linkage study. The identification of these new regions on BTA 20, 21, and 22 illustrates the increased power of the present study compared with our previous study (Schopen et al., 2009b). The increased power can be ascribed to the difference between an association analysis and a linkage analysis and, in addition, to the fact that around 1,000 extra animals were included in this study compared with our linkage study.

### BTA 5

Region 5_1 on BTA 5 starts at 13.5 Mbp and ends at 42.9 Mbp, and shows a strong effect on α-LA. The SNP most significantly associated with α-LA is located at 34.4 Mbp within the promoter region of the α-LA gene, and is known from previous studies as “+15” (Bleck and Bremel, 1993). α-Lactalbumin is 1 of the 2 proteins involved in lactose synthesis, and a high concentration

<table>
<thead>
<tr>
<th>Trait</th>
<th>Regions2</th>
<th>Sum3</th>
<th>Multiple SNP analysis4</th>
</tr>
</thead>
<tbody>
<tr>
<td>αS1-CN</td>
<td>6_2, 11, 13_1, 14, 22_1, 22_2</td>
<td>32.4</td>
<td>31.6</td>
</tr>
<tr>
<td>αS2-CN</td>
<td>1_2, 6_2, 9_2, 10_1, 10_2, 11, 14, 17_1, 19, 28_1, 28_2, 29</td>
<td>39.8</td>
<td>30.2</td>
</tr>
<tr>
<td>β-CN</td>
<td>6_1, 6_2, 11</td>
<td>110.5</td>
<td>104.9</td>
</tr>
<tr>
<td>κ-CN</td>
<td>6_2, 11, 13_2, 21, 29</td>
<td>37.0</td>
<td>35.2</td>
</tr>
<tr>
<td>α-LA</td>
<td>1_1, 5_1, 5_2, 6_2, 9_1, 16, 17_2, 26</td>
<td>30.7</td>
<td>26.5</td>
</tr>
<tr>
<td>β-LG</td>
<td>6_2, 11, 24</td>
<td>85.6</td>
<td>81.6</td>
</tr>
<tr>
<td>Casein index</td>
<td>6_2, 11</td>
<td>104.5</td>
<td>101.7</td>
</tr>
<tr>
<td>Protein (%)</td>
<td>5_3, 6_2, 10_1, 14, 15, 20, 25_1, 25_2, 29</td>
<td>39.0</td>
<td>40.0</td>
</tr>
</tbody>
</table>

1 This value is expressed as proportion of the additive genetic variance.
2 Regions are identified by the chromosome number and the region on that chromosome: _1 is the first region, _2 is the second region, and _3 is the third region on a chromosome.
3 This is the sum of the genetic variances explained by the most significant SNP in each region from the single SNP analysis.
4 This value results from the multiple SNP analysis, in which SNP are analyzed simultaneously.
of α-LA is necessary to ensure maximum synthesis of lactose (e.g., Fitzgerald et al., 1970; Menzies et al., 2009). Therefore, an additional analysis was performed for association between SNP in region 5_1 and lactose percentage. Results showed that several SNP in this region were associated ($P < 0.01$) with lactose percentage. This confirms the results of Lundén and Lindersson (1998) who reported an effect of the “+15” SNP on lactose percentage. Our study indicates that this region on BTA 5 affects both α-LA and lactose percentage, and that the effects of each of the “+15” SNP alleles are in the same direction for both α-LA and lactose percentage.

**BTA 6**

Region 6_2 on BTA 6 starts at 61.1 Mbp and ends at 97.7 Mbp, and SNP located within this region were significantly associated with protein composition and protein percentage. The SNP ULRG_BTC-060550 was most significantly associated with β-CN. This SNP is located at 88.3 Mbp in intron 7 of the β-CN gene, and is known to be in linkage disequilibrium with β-CN protein variants A and A (Visker et al., 2011). Effects of β-CN protein variants A, A, A, B, and I in this population have been reported by Visker et al. (2011). The SNP rs43703016 was most significantly associated with κ-CN, β-LG, casein index, and protein percentage. This SNP is located at 88.5 Mbp in exon 4 of the κ-CN gene, and is 1 of the 2 SNPs that are causal for κ-CN protein variants A and B. Effects of κ-CN protein variants A, B, and I in this population have been reported by Heck et al. (2009).

Adjusting the data for β-CN protein variants A, A, B, A, B, and I or κ-CN protein variants A, B, and E resulted in a considerable decrease in the variation associated with region 6_2. However, our study shows that not all variation associated with this region can be attributed to known protein variants of β-CN or κ-CN.

Another significant association for β-CN on BTA 6 was found at 44.2 Mbp (region 6_1). The osteopontin (OPN) gene and the ABCG2 gene are located around 38 Mbp and have previously been associated with milk protein percentage (Cohen-Zinder et al., 2005; Schnabel et al., 2005). This suggests that OPN or ABCG2 is a candidate gene for the association at 44.2 Mbp. An additional analysis was performed in which those SNP most significantly associated with β-CN in regions 6_1 and 6_2 were simultaneously included in the mixed model. This analysis strongly reduced the significance of the SNP in region 6_1 ($P$-value from 7.94E−07 to 0.07). The significance of the SNP in region 6_2 was also reduced but remained highly significant ($P$-value from 2.04E−101 to 4.89E−81). This suggests that the SNP in regions 6_1 and 6_2 are in linkage disequilibrium, and that the region surrounding the casein genes (6_2) is most likely responsible for the observed associations.

**BTA 11**

The region on BTA 11 associated with milk protein composition starts at 84.3 Mbp and ends at 110.2 Mbp. The SNP rs41255679 was most significantly associated with αs-CN, αs-CN, β-CN, κ-CN, β-LG, and casein index. This SNP is located at 107.2 Mbp in the promoter region of the β-LG gene, and is known to be in linkage disequilibrium with β-LG protein variants A and B (Ganai et al., 2009). Effects of β-LG protein variants A and B in this population have been reported by Heck et al. (2009). Adjusting the data for β-LG protein variants A and B drastically reduced the variation for β-LG content and casein index associated with this region. However, a significant proportion of genetic variance remains associated with this region on BTA 11 after adjusting for β-LG protein variants. This supports findings by Ganai et al. (2009) who reported that besides β-LG protein variants A and B, other SNP in this region explain a significant part of the variation in β-LG content.

**BTA 14**

The region on BTA 14 associated with αs-CN, αs-CN, and protein percentage starts at 0.0 Mbp and ends at 13.6 Mbp. The SNP showing the strongest association was located at 0.4 Mbp and is 1 of the 2 SNP responsible for the DGAT1 K232A polymorphism, which has previously been associated with protein percentage (Grisart et al., 2002). The results suggest that DGAT1 not only affects protein percentage, but remarkably also αs-CN and αs-CN content. The DGAT1 K232A allele was associated with a higher content of αs-CN, a lower content of αs-CN, and a lower protein percentage (0.26%). Adjusting the data for the DGAT1 K232A polymorphism resulted in a considerable decrease in the variation associated with this region; however, a small fraction of the variation remains associated with this region after accounting for the DGAT1 K232A polymorphism.

**Yield of Individual Proteins**

In this study, we used relative percentages instead of absolute yields of individual proteins. For the dairy industry and selective breeding, protein percentage and relative protein composition are complementary sources of information. A potential disadvantage of using relative concentrations is that correlations among
protein fractions are introduced. However, the problem of auto-correlations is limited when a reasonable number of fractions, such as the 6 protein fractions in this study, are being considered.

From a biological point of view, absolute yields as well as relative concentrations of individual proteins are of interest. To investigate if using relative or absolute values results in different associations, we repeated the association analysis for BTA 6 using absolute yields for the 6 major milk proteins. For αS2-CN, β-CN, and κ-CN, the SNP showing the strongest association was identical in both analyses. For αS1-CN, α-LA, and β-LG, the same regions were identified; however, the most significant SNP was different. These results suggest that both relative percentages and absolute yields of individual proteins are associated with the same genomic regions.

No significant associations for test-day morning milk protein yield were detected. These findings suggest that several genes exist that have large effects on milk protein composition and protein percentage, whereas no genes have very large effects on protein yield. However, it should be realized that protein yield for only 1 morning test-day milk sample was analyzed; analyzing the cumulative protein yield of a full lactation is expected to be more powerful.

**Regulation of the 6 Major Milk Proteins**

This study identified genomic regions on BTA 5, 6, 11, and 14 with large effects on milk protein composition or protein percentage. In some cases, these genomic regions explained more than 80% of the genetic variation in the milk proteins (e.g., BTA 6 for β-CN and BTA 11 for β-LG). These variances are likely overestimates of the true variance explained by these regions, especially for SNPs that explain a smaller proportion of the variance (Beavis, 1994).

The largest effects for all 6 milk proteins were found in the region where the gene coding for the corresponding milk protein is located: for αS1-CN, αS2-CN, β-CN, and κ-CN on BTA 6; for α-LA on BTA 5; and for β-LG on BTA 11. In addition to these 3 regions with large effects on one or more of the 6 major milk proteins, we detected 21 regions that were uniquely associated with one specific major milk protein. Furthermore, the number of chromosomal regions affecting a specific milk protein is quite different, varying from 3 regions for β-CN and β-LG to 12 regions for αS2-casein. This indicates that milk protein composition is regulated by a few genes with large effects on all milk proteins and several genes with smaller effects on specific milk proteins.

**CONCLUSIONS**

In total, 31 genomic regions on 20 bovine autosomes showed significant association with milk protein composition or protein percentage. Four genomic regions, on BTA 5, 6, 11, and 14, showed large effects on milk protein composition or protein percentage. The region on BTA 6 was significantly associated with all 6 major milk proteins, whereas the region on BTA 11 was significantly associated with all major milk proteins except α-LA. Additionally, 27 regions showed smaller effects on individual proteins or protein percentage. The associations of BTA 6, 11, and 14 could, in large part, be attributed to known protein variants of β-CN, κ-CN, or β-LG, or the DGAT1 K232A polymorphism. However, these polymorphisms could not explain all the variation associated with chromosomal regions on BTA 6, 11, and 14.

**ACKNOWLEDGMENTS**

This study is part of the Dutch Milk Genomics Initiative, funded by Wageningen University, NZO (Dutch Dairy Association), CRV (Co-operative Cattle Improvement Organization, Arnhem, the Netherlands), and the Dutch technology foundation, STW (Utrecht, the Netherlands). The authors thank the owners of the herds for their help in collecting the data and Sylvia Kinders (Animal Breeding and Genomics Centre, Wageningen, the Netherlands) for her assistance in generating the genotypes.

**REFERENCES**


major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 15:936–944.