J. Dairy Sci. 96:5333–5343
http://dx.doi.org/10.3168/jds.2012-6324

Genomic architecture of bovine κ-casein and β-lactoglobulin

R. Gambra,* F. Peñagaricano,† J. Kropp,† K. Khateeb,† K. A. Weigel,‡ J. Lucey,§ and H. Khatib†[1]
*Department of Animal Science, Facultad de Agronomía e Ingeniería Forestal, Pontificia Universidad Católica de Chile, Santiago, RM 7820436, Chile
†Department of Animal Sciences,
‡Department of Dairy Science, and
§Department of Food Science, University of Wisconsin-Madison, Madison 53706

ABSTRACT

The objective of this study was to characterize the genetic architecture underlying the absolute concentrations of 2 important milk proteins, κ-casein (κ-CN) and β-lactoglobulin (β-LG), in a backcross population of (Holstein × Jersey) × Holstein cattle. A genome-wide association analysis was performed using a selective DNA pooling strategy and the Illumina BovineHD BeadChip assay [777,000 (777K) SNP markers; Illumina Inc., San Diego, CA]. After correction for multiple testing, 25 single nucleotide polymorphisms were found to be associated with κ-CN and 36 single nucleotide polymorphisms were associated with β-LG. A pathway association analysis revealed 15 Gene Ontology (GO) terms associated with the κ-CN trait and 28 GO terms associated with β-LG. In addition, several GO terms were associated with both milk proteins. Further analysis revealed that κ-CN and β-LG production is regulated by both kinase and phosphatase activity, including mechanisms regulating the extracellular matrix. These results are in concordance with the complex multihormonal process controlling the expression of milk proteins and interactions between mammary epithelial cells and extracellular matrix components. Although κ-CN and β-LG milk proteins are expressed by single genes, the results from this study showed that many loci are involved in the regulation of the concentration of these 2 proteins.
Key words: κ-casein, β-lactoglobulin, backcross, quantitative trait pathway

INTRODUCTION

Milk proteins have a crucial role in contributing to the nutritional qualities and properties of milk. The total milk protein composition strongly depends on the expression and secretion of individual proteins. Polymorphisms within the major milk protein genes have been identified and characterized within and among breeds of the bovine species (Caroli et al., 2009) and have been reported to be associated with milk composition traits. For example, polymorphisms in the β-LG gene have been previously associated with higher β-LG protein concentration (Ng-Kwai-Hang et al., 1987; Lum et al., 1997; Folch et al., 1999; Kuss et al., 2003; Heck et al., 2009; Huang et al., 2012). Also, concentrations of κ-CN variant B have been associated with cheese production due to reduction of rennet clotting time and increased whey expulsion (Lucey and Kelly, 1994). Despite the fact that the traits relating to milk protein composition are highly heritable, their genetic architecture remains to be elucidated.

Even though each milk protein is expressed by one gene, recent studies have shown that other loci might contribute to the final concentration of the protein. Schopen et al. (2011) reported significant associations between major milk proteins and specific SNP across the genome using a custom design of the Illumina (50,000-marker) 50K BeadChip (Illumina Inc., San Diego, CA). The authors reported that those SNP explained 100% of the additive genetic variation for β-CN and 81.6% of the β-LG variation, whereas the additive genetic variance explained for α-CN, κ-CN, and α-LA ranged from 25 to 35% (Schopen et al., 2011). However, it is possible that the mutations responsible for the variation observed in one population could have segregated differently in a different population. For example, a substantial difference exists in milk protein content between Holsteins and Jerseys; therefore, the introduction of Jersey alleles to the mapping population is expected to generate genetic variation, which in turn facilitates the mapping of DNA variants responsible for milk protein composition traits (Huang et al., 2012).

To date, no genome-wide association study (**GWAS**) has been reported for Holstein × Jersey crossbred populations aimed at the identification of genes affecting milk composition traits. Therefore, the objective of this study was to gain a better understanding of the genomic architecture of the absolute concentra-

Received October 29, 2012.
Accepted April 20, 2013.
[1]Corresponding author: hkhatib@wisc.edu

tions of κ-CN and β-LG in milk in this population. To achieve this, a GWAS was performed using the bovine high-density (BovineHD) platform (Illumina Inc.) followed by pathway association analysis. Genome-wide association study analyses have been very successful in identifying loci related to traits using a single nucleotide marker-based association test that examines the relationship between each SNP marker and the trait of interest. However, because of the need to correct for multiple testing in these analyses, genes that may truly be associated with the trait will not be detected, as they do not reach the significance threshold. One alternative method is a pathway association analysis, which examines if a group of related genes has more significant genes than expected by chance (Wang et al., 2010). Erbe et al. (2012) showed that the accuracy of prediction depends on the architecture of the trait and the model used. Thus, the pathway analysis can add strength to the genomic data and allows for a better understanding of cellular processes, as intricate networks of functionally related genes can unravel the biological basis of the association findings. Understanding the genomic architecture underlying a trait will allow for better selection and further genetic improvement.

MATERIALS AND METHODS

Collection of Milk Samples and HPLC Phenotyping

Milk samples were collected from 164 lactating cows belonging to a crossbred population of Holstein × Jersey described in detail by Maltecca et al. (2009). All animals were housed at the University of Wisconsin-Madison Arlington Agriculture Experimental Station under the same conditions and milked twice daily. The concentrations of κ-CN and β-LG proteins were measured by reverse-phase HPLC as described by Huang et al. (2012).

Selective DNA Pooling and Genotyping

A selective DNA pooling procedure (Darvasi and Soller, 1994) was adopted to construct 8 DNA pools, with 4 pools for each trait under study (i.e., κ-CN and β-LG protein concentration). Pools were constructed using the phenotype of each animal after correction for lactation, DIM, and date of milk sample collection and processing. For each trait, 30% of the animals with the highest and 30% with the lowest concentrations were selected for creating high and low pools, respectively, for selective DNA pooling. A total of 50 animals were selected for each trait and randomly assigned to 2 subpools, thereby generating 2 replicates among the high and low pools. Supplemental Table S1 (available online at http://www.journalofdairyscience.org/) summarizes the pool compositions and measured phenotypes in this study. Genomic DNA from each selected milk sample was extracted as described by Huang et al. (2010) and concentration was measured using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE). Equal amounts of DNA (50 ng) from each individual were used to construct the pools. Genotyping of the pools was performed by GeneSeek (Lincoln, NE) using the BovineHD BeadChip (Illumina Inc.).

Association Analysis of DNA Pooling

The allele A frequency (f_alleleA) for each replicate was estimated on the basis of the raw data as follows: f_alleleA = Xraw/(Xraw + Yraw), where Xraw and Yraw are the normalized intensities of the 2 channels red and green used to genotype SNP on the Illumina Inc. platform. The association analysis was performed based on the following statistic (Tc), which combines experimental and sampling errors. This procedure was described in detail by Abraham et al. (2008) and is reviewed briefly here:

$$Tc = \frac{\left(\overline{f}_H - \overline{f}_L\right)^2}{v_H + v_L + \varepsilon_H^2 + \varepsilon_L^2},$$

This statistic combines the following: (a) a chi-squared T_1 for testing differences between 2 proportions (i.e., allele frequencies between high and low pools) accounting for the sampling variance:

$$T_1 = \frac{\left(\overline{f}_H - \overline{f}_L\right)^2}{v_H + v_L},$$

where

$$\overline{f}_k = \frac{n_{k1} \cdot f_{k1} + n_{k2} \cdot f_{k2}}{(n_{k1} + n_{k2})}$$

is the mean of the allele frequency $\left(\overline{f}_k\right)$ in the kth pool (where k = H or L), calculated using the 2 replicates; n_{k1} and n_{k2} are the number of individuals in each of the 2 replicates; and

$$v_k = \frac{\overline{f}_k \cdot \left(1 - \overline{f}_k\right)}{2 \cdot (n_{k1} + n_{k2})}$$

is the binomial sampling variance (v_k); (b) a Z-statistic for testing the difference in mean allele frequencies between high and low pools:

$$Z = \frac{\left(\overline{f}_H - \overline{f}_L\right)}{\sqrt{\varepsilon_H^2 + \varepsilon_L^2}}$$

where $\varepsilon_k^2 = 0.5 \cdot \sum_{i=1}^{2}\left(f_{ki} - \overline{f}_k\right)^2$ is the square of the standard error due to experimental error in the kth pool.

Selective DNA pooling analysis can lead to biased results due to hidden population stratification. Therefore, the genomic control method proposed by Devlin and Roeder (1999) was applied. In this analysis, 20% of the markers (n = 155,592), evenly distributed across the whole genome, were used to calculate the inflation factor λ as follows:

$$\lambda = \text{median}\ (Tc_1,\ Tc_2,\ \ldots,\ Tc_n)/0.456.$$

Under the null hypothesis $\left(\text{i.e., } H_0 : \overline{f}_H - \overline{f}_L = 0\right)$, Tc/λ follows (approximately) a chi-squared distribution with 1 degree of freedom (Devlin and Roeder, 1999). Hence, upper-tail one-sided *P*-values were determined by comparing each Tc/λ value to a χ_1^2 distribution. Finally, to account for multiple hypothesis testing and to control the false discovery rate, the procedure proposed by Benjamini and Hochberg (1995) was adopted, and the q-value (defined as the minimum false discovery rate at which the test may be called significant) for each test was obtained.

Individual Genotyping and Statistical Analysis

The most significant SNP for κ-CN (5 SNP) and β-LG (2 SNP) found in GWAS were validated by individual genotyping. All the SNP selected for κ-CN genotyping were located on chromosome 6, upstream of the κ-CN gene. In addition, the 2 SNP selected for β-LG were located on chromosome 11, downstream of the β-LG gene, and chromosome 20. The individual DNA samples used to construct the pools were genotyped by PCR restriction fragment length polymorphism (PCR-RFLP). Sequences of the SNP regions were obtained from the National Center for Biotechnology Information database (http://www.ncbi.nlm.nih.gov). Primers were designed using the Primer3Plus program (Untergasser et al., 2007). Primer sequences, PCR conditions, and restriction enzymes used in RFLP are provided in Supplemental Table S2 (available online at http://www.journalofdairyscience.org/).

Associations between each SNP and the concentrations of κ-CN or β-LG were analyzed using the following mixed liner model:

$$y_{ijklmn} = \beta_0 + \text{Lactation}_j + \beta_1 DIM_k + \beta_2 (DIM_k)^2 + \beta_3 SNP_{il} + \text{Date}_m + \text{Sire}_n + e_{ijklmn},$$

where y_{ijklmn} is the phenotype of the ith animal; the constant β_0 represents a general mean; Lactation_j represents the fixed effect of the lactation class (j = 1, 2, or 3); β_1 and β_2 are the linear and quadratic coefficients of DIM, respectively; SNP_{il} is the number of copies of one allele of the SNP (corresponding to 0, 1, or 2 copies) carried by the ith animal; β_3 is the regression coefficient for the SNP considered (known as the allele substitution effect); Date_m is a random effect for the date of sample collection and analysis, which accounted for seasonal and batch effect on HPLC; Sire_n represents the random effect of the sire of the ith animal; and e_{ijklmn} is the random residual for each observation. Association between SNP and phenotypes was tested using a likelihood ratio test by comparing the aforementioned model to a reduced model without the SNP effect.

Linkage disequilibrium (**LD**) was estimated between significant SNP for κ-CN and 12 SNP within the κ-CN gene (*CSN3*) previously genotyped by Huang et al. (2012). Also, LD was estimated between SNP BovineHD2100006549, located outside the β-LG gene, and 8 SNP within the β-LG gene (progestagen-associated endometrial protein, *PAEP*) as well as 3 SNP upstream of *PAEP*. Linkage disequilibrium [measured as the squared correlation coefficient (r^2)] was calculated using the genetics package in the R program (R Development Core Team, 2009).

Pathway Analysis

To improve our interpretations of the GWAS results, a gene set-enrichment analysis was performed to identify quantitative trait+ pathways related to κ-CN and β-LG concentrations in milk. The methodology of the pathway analysis was described in detail by Peñagaricano et al. (2012) and is reviewed briefly here.

Assignment of SNP to Genes. Single nucleotide polymorphisms were assigned to genes, based on the UMD3.1 bovine genome sequence assembly (http://www.bovinegenome.org/cgi-bin/gbrowse/bovine_UMD31/), if they were located within the genomic sequence of an annotated gene or within 20 kb of the 5′ or 3′ ends of the first and last exon, respectively. A 20-kb distance was used to capture proximal regulatory and other functional regions that may lie outside, but close to the gene. If a SNP was located within or proximal to more than 1 gene, all of those genes were included in the subsequent analyses. Finally, a gene was considered to be significantly associated with the concentration of κ-CN or β-LG if that gene contained at least 1 SNP with a nominal *P*-value <0.025.

Assignment of Genes to Pathways. Gene Ontology (**GO**) was used to define functional sets of genes (Ashburner et al., 2000). The GO database assigns bio-

logical descriptors (named GO terms) to genes on the basis of the properties of their encoded products. These terms fall into 3 domains: biological process, molecular function, and cellular component. Genes are assigned to the same GO term if they are more closely related in terms of their biological functions compared with random sets of genes.

Pathway-Based Association Analysis. The association of a given pathway with concentration of κ-CN or β-LG was analyzed using a test of proportions based on the cumulative hypergeometric distribution (Tavazoie et al., 1999). This test was performed to search for an overrepresentation of significantly associated genes among all the genes in the pathway. The *P*-value of observing k significant genes in the pathway was calculated as follows:

$$P\text{-value} = 1 - \sum_{i=0}^{k-1} \frac{\binom{S}{i}\binom{N-S}{m-i}}{\binom{N}{m}},$$

where S is the total number of genes that are significantly associated with the trait, N is the total number of genes that are analyzed in the study, and m is the number of genes in the pathway.

To avoid testing overly narrow or broad functional categories, only GO terms with more than 30 genes that were located between levels 5 and 9 in the GO hierarchy were tested. Only those functional categories with a q-value <0.01 were considered significant. All analyses were performed using the procedure FatiGO (Al-Shahrour et al., 2004), implemented on the platform Babelomics (Medina et al., 2010).

RESULTS

The BovineHD BeadChip was used to search for SNP associated with κ-CN and β-LG concentrations using a selective DNA pooling strategy. Frequencies for each SNP in the DNA pools were calculated from the raw signals of the intensities of the red and green channels. Then, a statistical approach combining the *Tc* statistic (Abraham et al., 2008) with the genomic control approach (Devlin and Roeder, 1999) to control for population stratification was used to search for significant SNP associated with κ-CN or β-LG. For each trait, 20% of the markers were used to estimate the inflation factor λ as described by Devlin and Roeder (1999). The inflation factor was λ = 1.32 and λ = 0.83, for κ-CN and β-LG, respectively. Finally, the *Tc* statistic was adjusted using these factors. Supplemental Figure S1 (available online at http://www.journalofdairyscience.org/) shows the quantile-quantile plots of *P*-values (logarithmic scale) obtained in the association analysis for κ-CN and β-LG concentrations. Inspection of these quantile-quantile plots reveals an excess of smaller *P*-values compared with what was expected under the null hypothesis.

SNP Associated with κ-CN Concentration

A total of 771,054 SNP were used for association analysis. After correction for multiple testing, 25 SNP distributed on several chromosomes were found to be significant for κ-CN concentration (Figure 1). Information on SNP position, surrounding genes, and q-values is available in Supplemental Table S3 (available online at http://www.journalofdairyscience.org/). The association found in the DNA pooling analysis was validated for 4 of the 5 individually genotyped SNP (Table 1). Low LD (r^2 <0.25) was observed between the individually genotyped SNP and SNP in the *κ-CN* gene reported by Huang et al. (2012). Moderate LD was found between SNP BTB-00248845 and BovineHD0600017929 (r^2 = 0.41). The LD values between SNP BovineHD2700004460 and BovineHD060001885, BovineHD0600034748, and BovineHD060001885 were 0.37, 0.49, and 0.41, respectively.

SNP Associated with β-LG Concentration

A total of 775,870 SNP were used in the association analysis for β-LG concentration. Thirty-six SNP located on chromosomes 2, 5, 6, 9, 10, 11, 13, 14, 16, 20, and 21 were found to be significant for β-LG concentration (Figure 2). Specific information pertaining to the significant SNP, such as chromosome position, genes nearby, and q-value, is available in Supplemental Table S4 (available online at http://www.journalofdairyscience.org/). Two SNP that were tested in individual genotyping were found to be significant for β-LG concentration (Table 1). High LD (r^2 > 0.55) was found between SNP BovineHD1100030070 and SNP in the β-LG gene (Figure 3).

Pathway Association Analysis

To better understand the genetic architecture of κ-CN and β-LG, a quantitative trait pathway analysis was performed. Of a total 777K SNP used in this study, 386,304 were located within or 20 kb upstream or downstream from annotated genes. This set of SNP defined a total of 24,369 genes annotated in the UMD3.1 bovine genome sequence assembly, which in turn were evalu-

Table 1. Validation analysis of SNP associated with κ-CN and β-LG in milk

SNP name	Chromosome	Position (UMD3.1)[1]	Nearest gene	Distance[2] (bp)	q-value (pool based)	*P*-value (individual genotyping)
κ-CN						
BTB-00248845	6	61,905,913	*LIMCH1*	Within	1.48E-05	<0.001
BovineHD0600017929	6	64,763,165	*KCDT8*	12,266	7.40E-03	0.018
BovineHD0600018885	6	68,440,001	*TXK*	Within	6.42E-04	0.007
BovineHD0600019260	6	69,661,473	*SPATA18*	−53,102	8.38E-03	0.035
BovineHD0600034748	6	69,907,488	*LOC100138990*	Within	9.61E-04	0.17
β-LG						
BovineHD1100030070	11	103,308,330	*PAEP*	−1,949	5.23E-05	<0.001
BovineHD2100006549	21	22,331,732	*BLM*	Within	1.40E-06	0.005

[1]http://www.bovinegenome.org/cgi-bin/gbrowse/bovine_UMD31/.

[2]Positive numbers represent distances downstream (bp) and negative numbers are distance upstream (bp) of the SNP.

ated in the pathway analysis. Of these genes, a subset of 3,382 and 3,244 genes showed significant associations with κ-CN and β-LG concentrations, respectively.

For GO analysis, we tested GO categories with more than 30 genes and located between levels 5 and 9 in the GO hierarchy. A total of 755 GO terms met these requirements and hence were tested by a hypergeometric test for enrichment of significant genes associated with each trait. Fifteen GO terms showed significant overrepresentation of genes statistically associated with concentration of κ-CN, in which 10 GO terms belonged to the molecular function domain and 5 GO terms classified into the cellular component domain. For β-LG concentration, 28 GO terms had overrepresentation of genes significantly associated with this trait. Out of those, 13 correspond to GO terms that belonged to the molecular function domain, 3 GO terms were classified into the cellular component domain, and 12 GO terms belonged to the biological process domain. The number of genes in each functional category (out of 24,369 genes), the expected number of significant genes under the null hypothesis (i.e., with no significant overrepresentation), and the observed number of significant genes per category (out of the 3,382) are shown in Tables 2 and 3, for κ-CN and β-LG, respectively.

DISCUSSION

In this study, a selective DNA pooling strategy was used to search for SNP associated with milk composition traits using 775,870 SNP for β-LG and 771,054 SNP for κ-CN. A total of 36 SNP and 25 SNP were found to be associated with β-LG and κ-CN, respectively. Interestingly, individual genotyping of significant SNP showed that 6 of 7 examined SNP confirmed the results obtained with the DNA pooling, which dem-

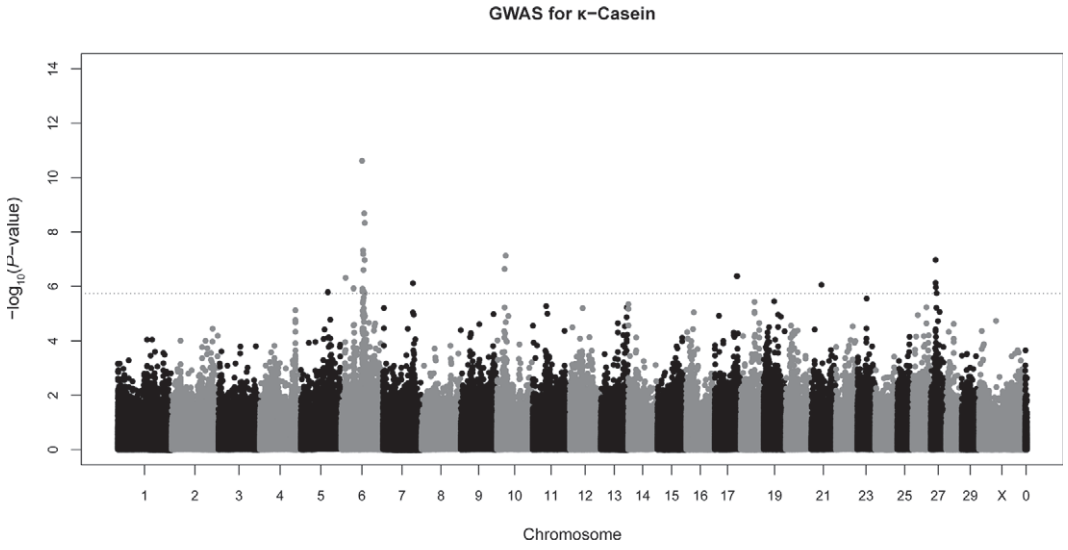

Figure 1. Manhattan plot of *P*-values for the genome-wide association study (GWAS) for κ-CN. The dotted horizontal line indicates a false discovery rate (FDR) of 5%.

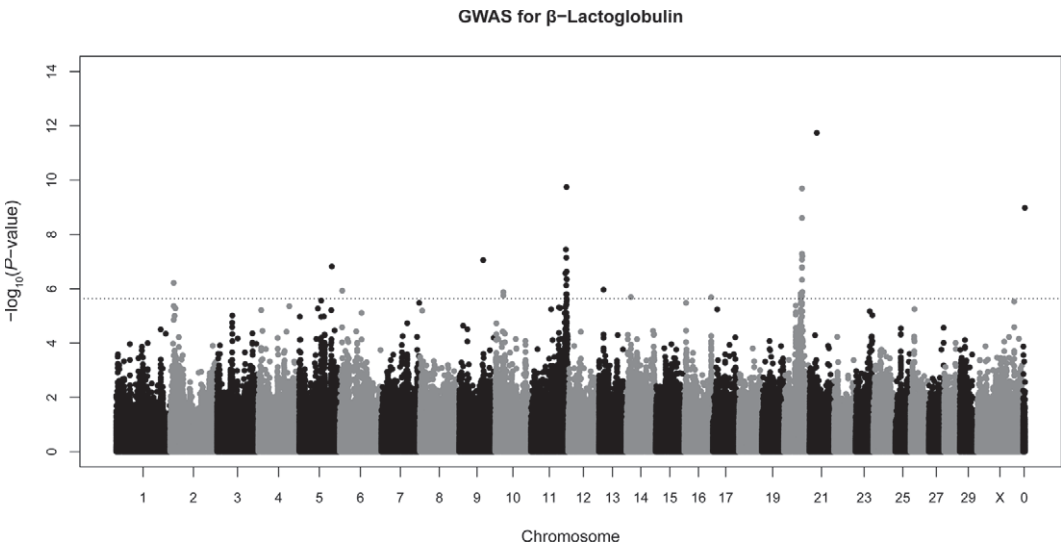

Figure 2. Manhattan plot of *P*-values for the genome-wide association study (GWAS) for β-LG. The dotted horizontal line indicates a false discovery rate (FDR) of 5%.

onstrates the effectiveness of this approach. The GO analysis revealed 28 and 15 terms with significant overrepresentation of genes associated with β-LG and κ-CN, respectively.

β-LG

Of the 36 significant SNP associated with β-LG, 1 SNP (BovineHD1100030070) was in high LD with the SNP responsible for the variants A/B of the β-LG protein. Variant A of β-LG has been associated with high milk protein concentration, especially with high concentration of β-LG, whereas variant B was associated with less total milk protein, but higher casein concentration (Ng-Kwai-Hang et al., 1987; Lundén et al., 1997; Heck et al., 2009; Huang et al., 2012). Single nucleotide polymorphism BovineHD1100030070 was also in high LD with SNP rs41255679 ($r^2 = 0.55$) and SNP rs41255681 ($r^2 = 0.65$). Both SNP are also in high LD with the SNP responsible for the A/B variants in this population ($r^2 > 0.88$), which is in agreement with the results of Lum et al. (1997). Single nucleotide polymorphism rs41255679 was reported to be associated with β-LG concentration (Kuss et al., 2003; Schopen et al., 2011; Huang et al.,

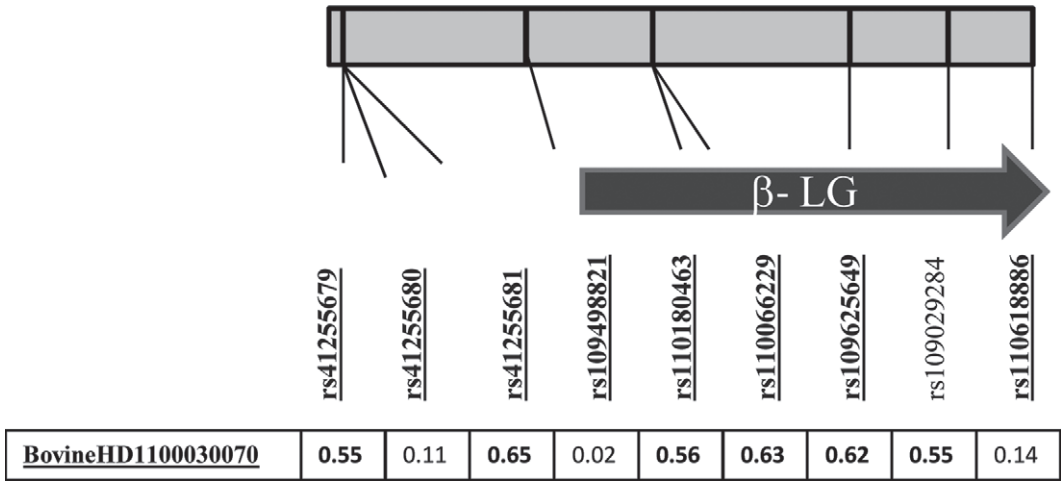

Figure 3. Linkage disequilibrium (correlation coefficient squared, r^2) between SNP BovineHD1100030070 and SNP within the β-LG gene. The top horizontal rectangle shows the relative location of each SNP on a bovine chromosome. Each vertical bar represents a SNP. Bold and underlined SNP were significantly association with β-LG concentration in Huang et al. (2012). Numbers in the boxes represent the calculated linkage disequilibrium. The β-LG gene is shown by the horizontal arrow, with the direction of the arrowhead indicating the transcriptional strand.

Table 2. Gene Ontology (GO; Ashburner et al., 2000) molecular function terms and GO cellular component terms significantly overrepresented with genes statistically associated with concentration of κ-CN

GO identification	Term	No. of genes in the GO term	Expected no. of significant genes	Actual no. of significant genes	q-value
GO molecular function					
0004672	Protein kinase activity	574	80	117	0.0007
0004674	Protein serine/threonine kinase activity	490	68	102	0.0007
0004713	Protein tyrosine kinase activity	491	68	101	0.0008
0004725	Protein tyrosine phosphatase activity	91	13	25	0.0089
0005089	Rho guanyl-nucleotide exchange factor activity	72	10	21	0.0094
0005509	Calcium ion binding	646	90	123	0.0028
0008081	Phosphoric diester hydrolase activity	63	9	21	0.0017
0016301	Kinase activity	753	104	147	0.0007
0016773	Phosphotransferase activity, alcohol group as acceptor	702	97	137	0.0007
0042578	Phosphoric ester hydrolase activity	277	38	64	0.0007
GO cellular component					
0005578	Proteinaceous extracellular matrix	196	27	46	0.0048
0005856	Cytoskeleton	751	104	141	0.0033
0015629	Actin cytoskeleton	217	30	49	0.0060
0044420	Extracellular matrix part	71	10	21	0.0067
0044430	Cytoskeletal part	475	66	96	0.0033

Table 3. Gene Ontology (GO; Ashburner et al., 2000) molecular function terms, GO cellular component terms, and GO biological process terms significantly overrepresented with genes statistically associated with concentration of β-LG

GO identification	Term	No. of genes in the GO term	Expected no. of significant genes	Actual no. of significant genes	q-value
GO molecular function					
0004672	Protein kinase activity	574	76	110	0.0018
0004674	Protein serine/threonine kinase activity	490	65	95	0.0030
0004713	Protein tyrosine kinase activity	491	65	100	0.0006
0005083	Small guanosine triphosphatase (GTPase) regulator activity	207	28	51	0.0006
0005085	Guanyl-nucleotide exchange factor activity	130	17	35	0.0015
0005096	GTPase activator activity	128	17	32	0.0061
0005099	Ras GTPase activator activity	59	8	19	0.0041
0005509	Calcium ion binding	646	86	115	0.0086
0016301	Kinase activity	753	100	132	0.0075
0016462	Pyrophosphatase activity	744	99	132	0.0061
0016773	Phosphotransferase activity, alcohol group as acceptor	702	93	125	0.0061
0016818	Hydrolase activity, acting on acid anhydrides	747	99	132	0.0061
0017111	Nucleoside-triphosphatase activity	727	97	127	0.0092
GO cellular component					
0005578	Proteinaceous extracellular matrix	196	26	51	<0.0001
0019898	Extrinsic to membrane	61	8	20	0.0018
0044459	Plasma membrane part	823	110	164	<0.0001
GO biological process					
0006468	Protein amino acid phosphorylation	718	96	137	0.0005
0006897	Endocytosis	108	14	32	0.0005
0007264	Small GTPase-mediated signal transduction	444	59	94	0.0002
0007265	Ras protein signal transduction	205	27	55	<0.0001
0007266	Rho protein signal transduction	102	14	27	0.0099
0009966	Regulation of signal transduction	607	81	125	<0.0001
0016192	Vesicle-mediated transport	328	44	73	0.0005
0016310	Phosphorylation	924	123	169	0.0005
0043087	Regulation of GTPase activity	88	12	26	0.0021
0046578	Regulation of Ras protein signal transduction	174	23	48	<0.0001
0048468	Cell development	613	82	114	0.0044
0051056	Regulation of small GTPase-mediated signal transduction	217	29	60	<0.0001

2012). Interestingly, SNP rs41255679 is located in the binding region of the activator protein 2 (AP-2; Lum et al., 1997). In addition, SNP BovineHD110030069, located in the β-LG gene, showed significant association with β-LG concentration. This SNP was found to be in LD with SNP responsible for the variant A/B of the β-LG protein (Ganai et al., 2009).

Nine genes had significant SNP associated with β-LG, including the 2 nonsynonymous SNP BovineHD1100030073 and BovineHD2100006549. Single nucleotide polymorphism BovineHD1100030073 is located in exon 5 of the glycosyltransferase 6 domain containing 1 (*GLT6D1*) gene on chromosome 11 and causes an amino acid change from serine to arginine. Glycosyltransferase 6 domain containing 1 is located in vicinity of the *PAEP* gene, a condition conserved in mammals, implying co-regulation of expression in mammary epithelial cells (Lemay et al., 2009). The BovineHD2100006549 SNP is positioned on chromosome 21 and located in the first exon of the Bloom syndrome (*BLM*) gene and causes an amino acid change of serine for glycine. The BLM protein belongs to the racQ family of DNA helicases that are responsible for the maintenance of structure and integrity of DNA (Ellis et al., 1995). Further research is needed to better understand the specific mechanism by which *GLT6D1* and *BLM* affect the β-LG concentration.

Even though β-LG is expressed by a single gene, results of this study suggest that β-LG concentration is regulated by many genomic regions distributed in 12 chromosomes, which is consistent with the results of Schopen et al. (2011). Those authors reported significant associations of SNP on chromosomes 6, 11, and 24 with β-LG concentrations in a Dutch Holstein-Friesian population. However, the SNP located on chromosome 24 reported by Schopen et al. (2011) was not significant in our study, probably due to population and sample size differences. The SNP located on chromosomes 6 and 11 were not present in the BovineHD panel used in the current study.

κ-CN

A total of 25 SNP were found to be significantly associated with κ-CN concentration across the genome, in which 14 SNP were located on chromosome 6 between positions 14 and 69.9 Mb. However, no significant LD was found between the individually genotyped SNP in this region and those within the κ-CN (*CSN3*) gene. These results suggest that at least 2 different causative mutations could be responsible for the κ-CN concentration: one in the casein cluster and the other in the 14- to 69.9-Mb region. Interestingly, QTL associated with protein percentage and protein yield have been reported in the 37 and 70 Mb region (Spelman et al., 1996; Nadesalingam et al., 2001; Chen et al., 2006).

Schopen et al. (2011) reported significant associations of SNP on chromosomes 6, 11, 13, 21, and 29 with κ-CN concentration in the Dutch Holstein-Friesian population. Single nucleotide polymorphisms on chromosomes 6, 11, and 29 were not present in the BovineHD panel used in our study, and the SNP on chromosomes 13 and 29 were not significant in our population. This discrepancy could be due to the small sample size used in our study compared with that of Schopen et al. (2011). The analysis of a small number of animals could limit the discovery of SNP that contribute to a small portion of the genetic variance.

Pathway Association Analysis

Correction for multiple testing is necessary to minimize false positive associations in genome-wide association analysis. Consequently, the identified SNP represent only a small portion of the genetic variation of the investigated trait. Therefore, a pathway association analysis approach was applied using all significant SNP before correction for multiple testing. Although SNP with small effects could not reach statistical significance, they could explain part of the genetic variation of the trait (Kemper and Goddard, 2012). Interestingly, the pathway analysis revealed GO enrichment domains that were significantly associated with both milk proteins (kinase activity, protein kinase activity, protein serine/threonine kinase activity, protein tyrosine kinase activity, calcium ion binding, and proteinaceous extracellular matrix), whereas others were exclusive for each protein.

The kinase pathway is an almost universal mechanism used by cells to control intra- and extracellular activity through a cascade of signals that results in a nonstable phosphorylation of proteins (Kyriakis and Avruch, 1996; Cohen, 2000). The functions of these terms have not been reported to affect milk composition traits in cattle. However, Bayat-Sarmadi et al. (1995) reported that Ser/Thr protein kinase mediates the induction of casein expression by prolactin in rabbits, which suggests that these proteins may play roles in the κ-CN synthesis in cattle. It is likely that the same signaling cascade is involved in β-LG expression, which is induced by prolactin (Feuermann et al., 2004). Also, caseins and whey proteins share the same secretory package of vesicular transport (Keenan et al., 1979; Devinoy et al., 1995). The vesicular transport process has been shown to be regulated by several protein kinases including the cyclic AMP-dependent protein kinase (Muñiz et al., 1996).

The association of the calcium ion-binding GO term with κ-CN and β-LG was consistent with previous re-

sults reported by Lemay et al. (2009), where several milk proteins across different mammalian species were used to search for conserved pathways. Furthermore, calcium metabolism has been reported as a potential regulator of the kinase signaling cascade (Cohen, 2000) and the secretion process of milk (Brooks and Landt, 1984). Brooks and Landt (1984) identified protein kinase in the rat mammary acini that is calcium-dependent and may be important in the phosphorylation of κ-CN. Before secretion by the mammary epithelial cells, caseins are phosphorylated by protein kinases and formed into micelles (Brooks and Landt, 1984). Another domain significantly enriched with genes affecting the concentration of both proteins includes pathways pertaining to the extracellular matrix such as proteinaceous extracellular matrix. This pathway is in concordance with the synergistic relationship between the lactogenic hormones and the extracellular matrix observed by Aggeler et al. (1988) in primary mouse epithelial cells. The relationship between extracellular components and protein expression is enhanced by specific GO terms found for β-LG and κ-CN.

For β-LG concentration, GO terms belonging to the biological function domain were Ras protein signal transduction, Rho protein signal transduction, regulation of signal transduction related to the regulation of guanosine triphosphatase (**GTPase**) activity, regulation of Ras protein signal transduction, and regulation of small GTPase-mediated signal transduction. All terms are related to each other as part of the Ras superfamily, in which GTPases function as guanosine diphosphate/guanosine triphosphate (GDP/GTP)-regulated molecular switches (Vetter and Wittinghofer, 2001). The Rho proteins serve as key regulators of extracellular-stimulus-mediated signaling networks that regulate actin organization, cell cycle progression, and gene expression (Etienne-Manneville and Hall, 2002). Two GO enriched cellular component domains specific for κ-CN concentration were cytoskeleton and actin cytoskeleton. The actin cytoskeleton mediates a variety of essential biological functions in all eukaryotic cells. In addition to providing a structural framework in which cell shape and polarity are defined, its dynamic properties provide the driving force for cells to move and to divide (Hall, 1998). It has been noticed that primary mammary cells change their shape when they are actively secreting milk proteins (Aggeler et al., 1988). Overall, the cytoskeleton assembly and organization is controlled by Rho, Rac, and Cdc42 proteins that belong to the Rho protein family (Hall, 1998; Quilliam et al., 2002). The Rho guanyl-nucleotide exchange factor activity (Rho-GEF) is the most common link between cell-surface ligand/receptor binding and Rho, Rac, and Cdc42 GTPases (Quilliam et al., 2002). Interestingly, the Rho-GEF term was found to be associated with κ-CN concentration in the current study (Table 2), which supports our findings on the association between cytoskeleton and κ-CN.

The pathway association results are in concordance with the complex multi-hormonal process controlling the expression of caseins and interaction between the mammary epithelial cells and the extracellular matrix (Rosen et al., 1999). Several pathways that were associated with κ-CN concentration in the current study have been previously described for their association with caseins within milk in both the mouse and rabbit. However, these results warrant further studies to obtain a full understanding of the involvement of the significant pathways in the modulation of the caseins in bovine milk. Importantly, the characterization of pathways involved in κ-CN concentration based on single significant SNP found in the GWAS analysis testifies to the efficacy of the quantitative trait pathway analysis used in this study.

CONCLUSIONS

The genome-wide association analysis used in this study revealed the complexity of the genomic architecture of milk composition traits. Single nucleotide polymorphisms significant for κ-CN and β-LG were distributed in several chromosomes. Subsequent pathway analysis identified several pathways contributing to κ-CN and β-LG concentrations. These pathways have been previously described in mammary gland epithelial cells in rodents, which support the notion that the genetic architecture of milk composition traits is conserved across mammalian species. Furthermore, the quantitative trait pathways identified in this study can be used in breeding programs to improve milk and cheese production.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their helpful comments and suggestions. This study was supported by a US Department of Agriculture Hatch act formula fund 142-PRJ52YY from the University of Wisconsin, Madison.

REFERENCES

Abraham, R., V. Moskvina, R. Sims, P. Hollingworth, A. Morgan, L. Georgieva, K. Dowzell, S. Cichon, A. M. Hillmer, M. C. O'Donovan, J. Williams, M. J. Owen, and G. Kirov. 2008. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. BMC Med. Genomics 1:44.

Aggeler, J., C. S. Park, and M. J. Bissell. 1988. Regulation of milk protein and basement membrane gene expression: The influence of the extracellular matrix. J. Dairy Sci. 71:2830–2842.

Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo. 2004. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20:578–580.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppiq, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene Ontology: Tool for the unification of biology. Nat. Genet. 25:25–29.

Bayat-Sarmadi, M., C. Puissant, and L. M. Houdebine. 1995. The effects of various kinase and phosphatase inhibitors on the transmission of the prolactin and extracellular matrix signals to rabbit alpha S1-casein and transferrin genes. Int. J. Biochem. Cell Biol. 27:707–718.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc., B 57:289–300.

Brooks, C. L., and M. Landt. 1984. Calcium-ion and calmodulin-dependent kappa-casein kinase in rat mammary acini. Biochem. J. 224:195–200.

Caroli, A. M., S. Chessa, and G. J. Erhardt. 2009. Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. J. Dairy Sci. 92:5335–5352.

Chen, H. Y., Q. Zhang, C. C. Ying, C. K. Wang, W. J. Gong, and G. Mei. 2006. Detection of quantitative trait loci affecting milk production traits on bovine chromosome 6 in a Chinese Holstein population by the daughter design. J. Dairy Sci. 89:782–790.

Cohen, P. 2000. The regulation of protein function by multisite phosphorylation—A 25 year update. Trends Biochem. Sci. 25:596–601.

Darvasi, A., and M. Soller. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics 138:1365–1373.

Devinoy, E., M. G. Stinnakre, F. Lavialle, D. Thépot, and M. Ollivier-Bousquet. 1995. Intracellular routing and release of caseins and growth hormone produced into milk from transgenic mice. Exp. Cell Res. 221:272–280.

Devlin, B., and K. Roeder. 1999. Genomic control for association studies. Biometrics 55:997–1004.

Ellis, N. A., J. Groden, T.-Z. Ye, J. Straughen, D. J. Lennon, S. Ciocci, M. Proytcheva, and J. German. 1995. The Bloom's syndrome gene product is homologous to RecQ helicases. Cell 83:655–666.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114–4129.

Etienne-Manneville, S., and A. Hall. 2002. Rho GTPases in cell biology. Nature 420:629–635.

Feuermann, Y., S. J. Mabjeesh, and A. Shamay. 2004. Leptin affects prolactin action on milk protein and fat synthesis in the bovine mammary gland. J. Dairy Sci. 87:2941–2946.

Folch, J. M., P. Dovc, and J. F. Medrano. 1999. Differential expression of bovine β-lactoglobulin A and B promoter variants in transiently transfected HC11 cells. J. Dairy Res. 66:537–544.

Ganai, N. A., H. Bovenhuis, J. A. van Arendonk, and M. H. Visker. 2009. Novel polymorphisms in the bovine β-lactoglobulin gene and their effects on β-lactoglobulin protein concentration in milk. Anim. Genet. 40:127–133.

Hall, A. 1998. Rho GTPases and the actin cytoskeleton. Science 279:509–514.

Heck, J. M., A. Schennink, H. J. van Valenberg, H. Bovenhuis, M. H. Visker, J. A. van Arendonk, and A. C. van Hooijdonk. 2009. Effects of milk protein variants on the protein composition of bovine milk. J. Dairy Sci. 92:1192–1202.

Huang, W., B. Kirkpatrick, G. Rosa, and H. Khatib. 2010. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. Anim. Genet. 41:570–578.

Huang, W., F. Peñagaricano, K. Ahmad, J. Lucey, K. Weigel, and H. Khatib. 2012. Association between milk protein gene variants and protein composition traits in dairy cattle. J. Dairy Sci. 95:440–449.

Keenan, T. W., M. Sasaki, W. N. Eigel, D. J. Morré, W. W. Franke, I. M. Zulak, and A. A. Bushway. 1979. Characterization of a secretory vesicle-rich fraction from lactating bovine mammary gland. Exp. Cell Res. 124:47–61.

Kemper, K. E., and M. E. Goddard. 2012. Understanding and predicting complex traits: Knowledge from cattle. Hum. Mol. Genet. 21:R45–R51.

Kuss, A. W., J. Gogol, and H. Geldermann. 2003. Associations of a polymorphic AP-2 binding site in the 5′-flanking region of the bovine β-lactoglobulin gene with milk proteins. J. Dairy Sci. 86:2213–2218.

Kyriakis, J. M., and J. Avruch. 1996. Protein kinase cascades activated by stress and inflammatory cytokines. Bioessays 18:567–577.

Lemay, D. G., D. J. Lynn, W. F. Martin, M. C. Neville, T. M. Casey, G. Rincon, E. V. Kriventseva, W. C. Barris, A. S. Hinrichs, A. J. Molenaar, K. S. Pollard, N. J. Maqbool, K. Singh, R. Murney, E. M. Zdobnov, R. L. Tellam, J. F. Medrano, J. B. German, and M. Rijnkels. 2009. The bovine lactation genome: Insights into the evolution of mammalian milk. Genome Biol. 10:R43.

Lucey, J. A., and J. Kelly. 1994. Cheese yield. J. Soc. Dairy Technol. 47:1–14.

Lum, L. S., P. Dovč, and J. F. Medrano. 1997. Polymorphisms of bovine β-lactoglobulin promoter and differences in the binding affinity of activator protein-2 transcription factor. J. Dairy Sci. 80:1389–1397.

Lundén, A., M. Nilsson, and L. Janson. 1997. Marked effect of β-lactoglobulin polymorphism on the ratio of casein to total protein in milk. J. Dairy Sci. 80:2996–3005.

Maltecca, C., K. A. Weigel, H. Khatib, M. Cowan, and A. Bagnato. 2009. Whole-genome scan for quantitative trait loci associated with birth weight, gestation length and passive immune transfer in a Holstein × Jersey crossbred population. Anim. Genet. 40:27–34.

Medina, I., J. Carbonell, L. Pulido, S. C. Madeira, S. Goetz, A. Conesa, J. Tárraga, A. Pascual-Montano, R. Nogales-Cadenas, J. Santoyo, F. García, M. Marbà, D. Montaner, and J. Dopazo. 2010. Babelomics: An integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 38:W210–W213.

Muñiz, M., M. Alonso, J. Hidalgo, and A. Velasco. 1996. A regulatory role for cAMP-dependent protein kinase in protein traffic along the exocytic route. J. Biol. Chem. 271:30935–30941.

Nadesalingam, J., Y. Plante, and J. P. Gibson. 2001. Detection of QTL for milk production on chromosome 1 and 6 of Holstein cattle. Mamm. Genome 12:27–31.

Ng-Kwai-Hang, K. F., J. F. Hayes, J. E. Moxley, and H. G. Monardes. 1987. Variation in milk protein concentrations associated with genetic polymorphism and environmental factors. J. Dairy Sci. 70:563–570.

Peñagaricano, F., K. A. Weigel, G. J. M. Rosa, and H. Khatib. 2012. Inferring quantitative trait pathways associated with bull fertility from a genome-wide association study. Front. Genet. 3:307.

Quilliam, L. A., J. F. Rebhun, and A. F. Castro. 2002. A growing family of guanine nucleotide exchange factors is responsible for activation of Ras-family GTPases. Prog. Nucleic Acid Res. Mol. Biol. 71:391–444.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosen, J. M., S. L. Wyszomierski, and D. Hadsell. 1999. Regulation of milk protein gene expression. Annu. Rev. Nutr. 19:407–436.

Schopen, G. C. B., M. H. P. W. Visker, P. D. Koks, E. Mullaart, J. A. M. van Arendonk, and H. Bovenhuis. 2011. Whole-genome association study for milk protein composition in dairy cattle. J. Dairy Sci. 94:3148–3158.

Spelman, R. J., W. Coppieters, L. Karim, J. A. van Arendonk, and H. Bovenhuis. 1996. Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. Genetics 144:1799–1807.

Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. 1999. Systematic determination of genetic network architecture. Nat. Genet. 22:281–285.

Untergasser, A., H. Nijveen, X. Rao, T. Bisseling, R. Geurts, and J. Leunissen. 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 35:W71–W74.

Vetter, I. R., and A. Wittinghofer. 2001. The guanine nucleotide-binding switch in three dimensions. Science 294:1299–1304.

Wang, K., M. Li, and H. Hakonarson. 2010. Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet. 11:843–854.