Caprine CSN1S1 haplotype effect on gene expression and milk composition measured by Fourier transform infrared spectroscopy

I. Berget,*†¹ H. Martens,†‡ A. Kohler,‡ S. K. Sjurseth,† N. K. Afseth,‡ B. Narum,‡ T. Adnøy,*† and S. Lien*†

*Department of Animal and Aquacultural Sciences, and

†CIGENE, Norwegian University of Life Sciences, Box 5003, 1430 Ås, Norway

‡Centre for Biospectroscopy and Data Modelling, Nofima Mat AS, Osloveien 1, 1430 Ås, Norway

ABSTRACT

The Norwegian dairy goat population has a high frequency of a CSN1S1 (α_{S1} -casein) haplotype with negative effects on protein and fat content. It is characterized by a single point deletion in exon 12 of CSN1S1, leading to a truncated protein and hence a low content of α_{S1} -casein in the milk. This haplotype together with another haplotype with a deletion in exon 9 are called “weak” haplotypes. “Strong” haplotypes, on the other hand, have positive effects on important milk production traits. We show that expression of CSN1S1 in the mammary gland of lactating goats is significantly lower in animals with 2 weak haplotypes. Moreover, the effects of defective alleles were not detected in animals having 1 strong and 1 weak haplotype. Expression levels of other genes in the casein cluster were not affected by the CSN1S1 haplotypes investigated. Milk samples from goats with 2 weak haplotypes could be distinguished from the other milk samples using Fourier transform infrared (FTIR) spectroscopy and partial least squares discriminant analysis (PLS-DA). The PLS-DA components were related to spectra of pure caseins and whey proteins, hence FTIR has a potential for identifying milk samples with low α_{S1} -casein content and different protein composition. The results indicate that FTIR-based measurements can be incorporated into breeding plans, or for selection of milk samples with high casein content, which in turn may improve cheese-making properties of the milk.

Key words: casein, goat, Fourier transform infrared spectroscopy, gene expression

INTRODUCTION

Caseins are one of the major milk protein groups. The caprine casein genes are highly polymorphic, with CSN1S1 (which codes for α_{S1} -casein) being the most complex, having at least 18 variants with strong, medium, and weak effects on protein content, as well as

null alleles with no synthesis of the protein (Leroux et al., 1990; Martin et al., 2002; Neveu et al., 2002; Sacchi et al., 2005; Caroli et al., 2006). Most genetic variants differ only by amino acid substitutions, but some variants arise from exon skipping (Brignon et al., 1990; Leroux et al., 1992).

Hayes et al. (2006) categorized alleles within casein loci into haplotypes and investigated associations between haplotypes and production traits in Norwegian goats. The most common haplotype of the CSN1S1 gene in the Norwegian goat population is characterized by a single point deletion in exon 12, which leads to a truncated protein (Vegarud et al., 1989), and hence a low content of α_{S1} -casein in the milk. According to Hayes et al. (2006) this haplotype, as well as another haplotype with a deletion in exon 9 of the same gene, has negative effects on important milk production traits. In this work, the effects of the major CSN1S1 haplotypes on the transcription levels (mRNA) are investigated using MassARRAY (Sequenom Inc., San Diego, CA; Ding and Cantor, 2003). We also quantify allele-specific expression and document exon skipping for the haplotype containing the exon 12 deletion.

The high frequency (0.72) of the exon 12 deletion is a problem for the dairy industry because the clotting ability of milk depends on the casein composition and the α_{S1} -casein content of the milk in particular (Vegarud et al., 1989). Thus, fast and inexpensive methods for identifying milk samples with low α_{S1} -casein content could be beneficial in at least 2 aspects: 1) to select milk samples to be used for white cheese production, and 2) to use as input to breeding programs to genetically improve Norwegian goat milk quality in the long term.

Fourier transform infrared (FTIR) spectrometers are used in Norway for daily quality control (Foss Electric Milkoscan FT 6000, Foss Electric, Hillerød, Denmark). Several studies have shown that FTIR can be used to study casein content (Hewavitharana and van Brakel, 1997; Luginbühl, 2002; Sorensen et al., 2003) and fatty acid profiles in cow milk (Soyeurt et al. 2006). Milk samples from animals with defective CSN1S1 alleles not only have less protein, but also differ in milk protein composition. In this work, we investigated the ability

Received October 26, 2009.

Accepted May 3, 2010.

¹Corresponding author: ingunn.berget@umb.no

Table 1. Characterization of CSN1S1 (α_{S1} -casein) haplotypes of goats used in the study¹

SNP_ID	Position	Haplotype				
		1	3	4	5	7
CSN1S1prom_264	Promoter	A	A	A	G	A
<i>CSN1S1prom_866²</i>	<i>Promoter</i>	<i>C</i>	<i>C</i>	<i>C</i>	<i>C</i>	<i>T</i>
CSN1S1prom_884	Promoter	C	T	C	T	T
CSN1S1prom_888	Promoter	G	A	G	A	A
CSN1S1prom_1105	Promoter	G	A	G	A	A
CSN1S1prom_1169	Promoter	G	A	G	A	A
CSN1S1prom_1379	Promoter	C	T	C	T	T
CSN1S1prom_1470	Promoter	G	A	G	A	A
CSN1S1Ex4_6075	Exon 4	T	C	T	C	C
CSN1S1Ex4_6091	Exon 4	C	G	C	G	G
CSN1S1Ex9_9889	Exon 9	C	D	C	C	C
CSN1S1in9_9918	Intron 9	A	G	A	G	G
CSN1S1Ex10_10673	Exon 10	C	G	C	G	G
CSN1S1_E12-F3-0	Exon 12	D	A	A	G	G
CSN1S1ex17_16860	Exon 17	C	C	C	T	T

¹The polymorphisms within exon 12 of CSN1S1 were labeled 1 (D, deletion), 3 (G), and 6 (A) in Hayes et al. (2006).

²The data in the second row (italics) were not included in this study but were included in Hayes et al. (2006), where the haplotypes were defined.

of FTIR to distinguish milk samples from animals with different CSN1S1 haplotypes.

Because the O-H bending signal (around 1,640cm⁻¹) from water in liquid milk measurements masks important protein information in the so-called amide I region of the infrared spectrum, 2 sampling techniques were used for FTIR measurements: 1) the Foss Milkoscan instrumentation used in routine milk analysis of liquid milk samples, and 2) a new technique based on the Bruker HTGS system (Bruker Optik GmbH, Ettlingen, Germany) developed for FTIR of dried thin films. The new technique preserves the information in the amide I region.

The feasibility of separating milk samples from animals with different CSN1S1 haplotypes from FTIR spectra of liquid and dried milk samples was investigated by using partial least squares discriminant analysis (PLS-DA; Wold, 1982). Reference spectra of pure α_S -casein, β -casein, κ -casein, α -lactalbumin, and β -lactoglobulin were projected onto the PLS-DA model to interpret the PLS-DA components in terms of milk proteins. This reveals the major differences in the composition of milk from animals with defective CSN1S1 alleles in contrast to alleles with high expression levels of CSN1S1.

MATERIALS AND METHODS

Sampling of Animals and Characterization of Haplotypes Included in Study

This study focused on differences between weak and strong CSN1S1 haplotypes at the gene expression level

and in milk composition as measured by FTIR. In particular, we focused on haplotypes 1, 3, 4, and 5 from Hayes et al. (2006). Haplotypes 1, 4, and 5 represent 3 polymorphisms genotyped in the same assay in Hayes et al. (2006). The D (deletion) allele is the most frequent and is characteristic for haplotype 1, whereas the A and G alleles are present in haplotypes 4 and 5, respectively. Haplotype 3 has a deletion in exon 9 and the A variant in exon 12. Haplotypes 1 and 4 diverge only at the exon 12 polymorphism, whereas haplotype 5 is more similar to haplotype 3 although it has the normal variant (C) where haplotype 3 has a deletion in exon 9. See Table 1 for a full description of the haplotypes.

Forty-five animals were selected from 2 farms (herds A and B) in a larger sample (627 goats distributed on 5 farms). Because some haplotypes are very rare, not all feasible combinations were found, and some haplotypes had fewer than 3 occurrences. To compensate for other sources of variation, a larger number of samples was selected for the overrepresented genotypes. The number of samples selected for each of the genotypes is shown in Table 2. Although the focus in this work was on the mentioned haplotypes, 2 animals heterozygous for haplotype 7 were also included. According to Hayes et al. (2006), this haplotype has intermediate effects on protein and fat content, and it was grouped together with the weak haplotypes in this study.

Throughout this paper, subscripts S and W are used to indicate strong and weak haplotypes, respectively. Genotypes are designated (1_W,1_W) for animals with 2 copies of haplotype 1_W, (1_W,3_W) for animals heterozygous for haplotype 1_W and 3_W, and so on.

Table 2. Number of animals for each haplotype combination

Haplotype combination ¹	Observations, n
(1 _w ,1 _w)	17
(1 _w ,3 _w)	4
(1 _w ,4 _s)	5
(1 _w ,5 _s)	3
(1 _w ,7 _w)	1
(3 _w ,4 _s)	4
(3 _w ,5 _s)	1
(4 _s ,4 _s)	4
(4 _s ,5 _s)	3
(5 _s ,5 _s)	2
(3 _w ,5 _s)	1

¹_w = weak haplotype; _s = strong haplotype; (1_w,1_w) indicates an animal with 2 copies of haplotype 1_w, and (1_w,3_w) indicates an animal heterozygous for haplotype 1_w and 3_w. Haplotypes 1, 3, 4, and 5 are from Hayes et al. (2006).

Milk Sampling, Biopsy Procedure, and RNA Preparation

Milk samples were collected during the morning milking in herd A and during the evening milking in herd B. For each individual, midstream milk samples were collected in a 50-mL centrifuge tube and snap frozen in liquid nitrogen before being analyzed. The goats were moderately sedated with an intramuscular injection of 0.2 mL of xylazine (Rompun 20 mg/mL, Bayer AG, Leverkusen, Germany). The hair on a 10- × 10-cm area of skin caudal on the left gland was shaved and the skin was sanitized with 70% ethanol. The biopsy needle (Quick-Core Biopsy Needle, 14 gauge, 6 cm; William Cook Europe, Bjaeverskov, Denmark) was prepared by locking it into the ready position with the specimen notch (20 mm) completely covered by the cutting cannula. The left gland was held firmly and the needle was inserted through the skin slightly proximal to the midline. The needle tip was positioned between the skin of the mammary gland and the mammary gland tissue, the stylet was advanced to expose the specimen notch, and the cutting cannula was fired by pressing the plunger, all according to the manufacturer's instructions. It is possible to capture approximately 35 mm³ of tissue in the specimen notch of the needle. Because mammary gland tissue is elastic, and therefore not easily sampled, the notch was usually not entirely filled. Two biopsies were therefore collected from each animal to ensure the collection of an adequate amount of tissue. Collected mammary gland samples were submerged in RNAlater RNA Stabilizing Reagent (Qiagen Ltd., West Sussex, UK) immediately after harvesting. The RNeasy Lipid Tissue Mini Kit (Qiagen Ltd.) was used for isolation of total RNA from the tissue. The concentration of RNA was measured using a Nanodrop ND-1000 spectrophotometer (Saveen Werner, Malmö, Sweden). Degradation

of RNA was measured on random samples using a RNA 6000 Nano LabChip Kit on a 2100 BioAnalyser (both from Agilent Technologies, Santa Clara, CA).

Gene and Allele Expression Analysis with MassARRAY

Quantitative gene and allele expression were assessed by real-time competitive PCR (**rcPCR**) in combination with high-throughput matrix-assisted laser desorption/ionization time-of-flight mass spectroscopy (**MALDI-TOF MS**) as described by the manufacturer (Sequenom Inc., San Diego, CA; Ding and Cantor, 2003). Seven assays within CSN1S1, CSN1S2, and CSN3 were designed for allele expression, and 1 assay (CSN2) was designed for gene expression only (Table 3). Note that for all genes except CSN2, expression was measured at several positions within each gene (Table 3). For all assays, we used 11 concentrations of the competitor ranging between -14.0 and -9.3 on the log scale, as well as a positive control (no competitor). In addition, 1 assay was designed explicitly to quantify skipping of CSN1S1 exon 12. The primer for this assay was located within exon 11 and extended into exon 12 or 13 for cDNA where exon 12 was present or skipped, respectively. Two replicates of the rcPCR were run for all samples and assays. Scores for the relative amount of cDNA to competitor for each allele at the different concentrations were estimated by the MassARRAY system and exported for further analyses in MATLAB (Mathworks Inc., Natick, MA).

FTIR Measurements

Milk Samples. Spectra of liquid milk samples were obtained from the Foss Milkoscan instrumentation used in routine milk analysis. Dry milk samples were measured using a new technique developed for measurements of dried thin films by using a High Throughput Screening eXTension (HTS-XT) unit, which was coupled to a Tensor 27 spectrometer (both from Bruker Optik GmbH). Samples of goat's milk (3.5 µL) were transferred onto an infrared (IR)-light-transparent, silicon, 96-well microliter plate. Subsequently, the samples were dried under moderate vacuum (0.9 bar) to generate an even, thin film suitable for IR measurements. Spectra were recorded in transmission mode in the spectral region of 4,000 and 500 cm⁻¹ with a resolution of 4 cm⁻¹ and an aperture of 6.0 mm. Background spectra of the silicon substrate were collected before each sample measurement to account for variation in water vapor and CO₂. For each spectrum, 40 scans were averaged. The samples were kept on ice during the preparation. The milk sample from 1 animal was missing; hence,

measurements were made on milk samples from 44 animals.

Reference Spectra. All reference spectra were obtained using an Equinox 55 FT-IR spectrometer (Bruker Optik GmbH) equipped with a Pike MIRacle single reflection ATR unit incorporating a horizontal diamond crystal plate (Pike Technologies Inc., Madison, WI). A deuterated triglycine sulfate (DTGS) detector was employed for detection, and spectra were recorded in the region between 4000 cm^{-1} and 370 cm^{-1} with a spectral resolution of 4 cm^{-1} and an aperture of 5.0 mm. For each spectrum, 64 scans were averaged. Powders of β -CN (Sigma-Aldrich Inc., St. Louis, MO), κ -CN (Sigma-Aldrich Inc.), α_S -CN ($\alpha_{S1} + \alpha_{S2}$, purified in our laboratory), β -LG (Davisco Foods International, Inc., Le Sueur, MN), and α -LA (Arla Foods, Aarhus, Denmark) were fastened onto the attenuate total reflectance (ATR) crystal with a high-pressure clamp. All protein powders were obtained from bovine milk. The purities of the different proteins were $\geq 98\%$ (β -CN), 70% (κ -CN), 80% (α_S -CN), 95% (β -LG), and 60% (α -LA). For all samples, 3 replicate measurements were obtained and averaged. Between 4 and 6 samples were measured for each protein.

Preprocessing of FTIR Spectra. To remove technical variation in the spectral measurements, the newly developed method for removing inter-replicate variation (Kohler et al., 2009) was applied on the spectra of dried milk samples and the reference (protein) spectra. The averages of the replicated spectra after pre-processing were used for further analyses. Extended multiplicative signal correction (Martens and Stark, 1991; Martens et al., 2003) was applied for the Foss spectra of the liquid milk samples, as no replicate measurements were available.

Data Analysis

Gene Expression Analyses. Let s_{ijk} be the relative amount of cDNA to competitor for allele k at concentration i and PCR reaction j ($k = 1, \dots, \text{number of alleles}$, $i = 1, \dots, 12$, $j = 1, 2$) as estimated by the MassARRAY system. For a given gene, the expression level x is estimated by fitting the summed scores for all alleles $\left(S_{ij} = \sum_{k=1}^{\text{alleles}} s_{ijk} \right)$, and c_{ij} , the log of the competitor concentration to the model in Equation [1], where the parameter H determines the steepness of the sigmoid:

$$S_{ij} = \frac{1}{1 + 10^{H(x - c_{ij})}} \quad (i = 1, 2, \dots, 12 \quad j = 1, 2). \quad [1]$$

Table 3. Assays for gene expression¹

Gene ²	Position	Alleles
CSN1S1	Exon 4	C, T
CSN1S1	Exon 9	D, C
CSN1S1	Exon 12	D, A, G
CSN2	Splice	
CSN1S2	Exon 15	C, T
CSN1S2	Exon 16	A, T
CSN3	Exon 3	A, G
CSN3	Exon 4	C, T

¹The CSN2 assay was not designed for allele-specific expression and was used for normalization. See supplementary data file 1 (<http://www.journalofdairyscience.org/current>) for details on primers and competitors. One additional assay was designed for measuring skipping of CSN1S1 exon 12.

²CSN1S1 = α_{S1} -casein; CSN2 = β -casein; CSN1S2 = α_{S2} -casein; CSN3 = κ -casein.

The relative expression of allele k was determined as the average of the ratio s_{ik}/S_{ik} across all points below the estimated concentration.

Even though the MassARRAY system measures gene expression relative to a competitor, the estimated expression levels are prone to experimental variation at several steps. To correct for such errors, the gene expression data were normalized to the measured expression of the assay in the CSN2 gene (CSN2-splice, Table 3). In addition, a fourth root transformation was applied to reduce heteroscedasticity.

One-way ANOVA was applied on normalized and transformed group means for the 8 different haplotype combinations shown in Table 2 to test differential expression between the genotypes for all groups with more than 1 observation. The Tukey-Kramer criterion was applied for multiple comparisons. To test the haplotype-specific effects on gene expression, the model in Equation [2] was fitted to the data for each gene:

$$y_i = \mu + \beta_1 H_{1i} + \beta_2 H_{3i} + \beta_3 H_{5i} + \beta_4 H_{7i} + \beta_5 (H_1 \times H_3)_i + \beta_6 (H_1 \times H_5)_i + \beta_7 H_{1i}^2 + \beta_8 H_{5i}^2 + e_i, \quad [2]$$

where, for a given gene, y_i is the measured expression for animal i , and $H_{ki} = 1$ if animal i has haplotype k ($i = 1, 2, \dots, 45$ and $k = 1, 3, 4, 5, 7$) and 0 otherwise; $(H_k \times H_l)_i$ is the interaction between haplotype k and l for individual i . The error is assumed to be normally distributed with variance σ^2 . With this parameterization, all effects measure deviations from the genotype ($4_S, 4_S$). The model can also be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{y} = \{y_1 \dots y_{45}\}$, $\boldsymbol{\beta} = \{\beta_1 \dots \beta_8\}$, $\mathbf{e} = \{e_1 \dots e_{45}\}$, and $\mathbf{X}$ is a matrix with 8 columns containing the dummy variables in Equation [2].

Classification of Animals from FTIR Spectra.

One goal of the FTIR analysis was to investigate if spectral measurement of milk samples could distinguish milk samples from animals with different haplotype combinations. Based on the expression analysis, the milk samples were divided into 2 groups. One group contained genotypes with low CSN1S1 expression. This group consisted of milk samples from animals with 2 weak haplotypes; that is, animals with genotypes (1_w,1_w), (1_w,3_w), and (1_w,7_w). This class will be labeled (w,w). The other class, labeled (s,x), contained the rest of the samples; that is, those having at least one strong haplotype.

The problem of identifying the two groups from the FTIR data can be formulated as a classification problem where the goal is to predict class labels (w,w) and (s,x) from the (preprocessed and centered) spectral measurements of the milk samples. Several methods for doing classification with multivariate data exist (Næs et al., 2002); we applied PLS-DA (Wold, 1982). In PLS-DA, classification is achieved by predicting the class labels (coded by dummy variables) from a partial least squares regression (PLSR) model. The predicted class labels are then obtained by using a cut-off, typically 0.5, for the predicted y-values.

Interval PLS-DA. Different regions of the FTIR spectra are associated with signals from different vibrational modes. To identify and illustrate those parts of the spectra that provided the best information about the casein content, we applied the interval-PLS-DA (i-PLS-DA) approach described in Norgaard et al. (2000). For both sets of spectral measurements (wet and dry milk samples measured by Foss and Bruker methods, respectively), the full spectrum was divided into intervals of approximately 100 cm⁻¹, corresponding to regions with known assignments such as the amide I, II, and III bands. The PLS-DA was then run for all intervals as well as the full spectrum. The cross-validated error rate (systematic cross validation with 7 segments) was used to select the optimal number of components as described by Norgaard et al. (2000). Based on these results, a larger region for fitting a final PLS-DA model was selected manually. The regression coefficients of the final model were tested by permuting the class label vector **y** 10,000 times to obtain approximate *P*-values for each wavelength.

To relate the PLS components to milk proteins, scores for the measured protein spectra were estimated as $\hat{\mathbf{T}}_{\text{ref}} = \mathbf{K}\hat{\mathbf{P}}^t$. Here, $\mathbf{K} = [\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_m]$ is the matrix with the $m = 5$ measured reference spectra, and $\mathbf{P}$ is the X-loadings from the final PLS-DA model.

All analyses were conducted using MATLAB (version 7.7 release 2008b; Mathworks Inc., Natick, MA).

RESULTS AND DISCUSSION

Gene Expression

Forty-five goats with extreme CSN1S1 haplotypes, selected from a larger population of 627 animals, were analyzed for gene and allele expression analyses using rcPCR (Ding and Cantor, 2003). The haplotypes under investigation were those identified in Hayes et al. (2006). Animals to be included in the study had different combinations of strong (4_s and 5_s) and weak (1_w and 3_w) haplotypes. The weak haplotypes were characterized by different single-point deletions in exon 9 or 12 of CSN1S1. In addition, 2 of the goats had haplotype 7_w, which have intermediate effects (Hayes et al., 2006).

The expression of the CSN1S1 gene was measured using 3 positions, each in different exons (4, 9, and 12; Table 3). The average correlations between the assays across all samples were 0.90 (exon 4 and 9), 0.76 (exon 4 and 12), and 0.73 (exon 9 and 12). The variation between the 3 assays can be explained by experimental factors such as primer efficiency and varying noise levels in the mass spectra, as well as molecular mechanisms such as alternative splicing of the mRNA. Also, the assay for exon 12 was more complicated than for the others because there were 3 possible alleles.

Differential expression of the CSN genes between haplotype combinations with more than one observation (see Table 3) was tested using one-way ANOVA. The group means were significantly different for all assays within the CSN1S1 locus ($P < 0.001$), but not in any of the other loci within the casein gene cluster. The group means and the corresponding standard deviations for each of the CSN1S1 assays and the means of these assays are presented in Figure 1. Because the mean is the more robust estimate of the expression, we will refer to the mean of the 3 CSN1S1 assays when the expression of CSN1S1 is discussed throughout this paper.

Animals with genotypes (1_w,1_w) and (1_w,3_w) had significantly less expression of CSN1S1 than animals with other genotypes. Interestingly, animals with 1 weak and 1 strong haplotype usually had similar expression levels as those with 2 strong haplotypes. The single animal with genotype (1_w,7_w) had expression similar to that of the low group, whereas the single animal with genotype (5_s,7_w) had expression similar to that of the high expression group. Therefore, animals with haplotype 7_w were grouped together with the weak haplotypes.

Table 4 shows the ANOVA table for the means of the CSN1S1 assays, according to the model indicated in Equation [2]. The only significant effects ($P < 0.05$) were the interaction terms $H_1 \times H_3$ and H_1^2 ; main effects were not significant. This result implies that 2

Figure 1. Group means for expression levels for the 3 CSN1S1 (α_{s1} -casein) loci. $\times$ = exon 4, $+$ = exon 9, $\diamond$ = exon 12, and $\circ$ = average of all 3 loci. For the average, the standard deviation is marked for all groups with >1 observation. Expression levels are normalized and transformed. The estimated log (base 10) concentration of the mRNA transcripts varied between -11.2 and -13.4 . The difference between the mean of (1,1) and (4,4) corresponds to a fold change of ~ 16 .

weak alleles are needed to obtain a significant deviation in expression from the wild type or normal genotype ($4_S, 4_S$). The P -value for H_1^2 was much smaller than that for $H_1 \times H_3$, but as the number of observations for the first group was much larger, the P -values are not directly comparable.

These results indicate that strong haplotypes can, to a certain degree, compensate for underexpression of defective haplotypes, because the expression of CSN1S1 in animals with 1 strong and 1 weak haplotype was closer to that of animals with 2 strong haplotypes than to that of animals with 2 weak haplotypes.

The goat casein genes are known to display multiple mRNA forms because of alternative splicing (reviewed

by Martin et al., 2002). Initial sequencing of the mRNA from goats homozygous for the single point deletion in exon 12 of CSN1S1 provided evidence for partial skipping of this exon. The exon skipping results in an mRNA protein lacking 14 amino acids encoded by exon 12. When the single point deletion in exon 12 is present, transcription of this exon introduces a shift in the reading frame, resulting in a truncated protein.

Thus, an assay was designed to specifically quantify the extent of exon skipping. The extent of exon skipping varied between 11 and 67% for the ($1_W, 1_W$) animals. This result demonstrates that the CSN1S1 gene produces at least 2 mRNA variants of α_{s1} -casein in animals homozygous for the exon 12 deletion. Moreover, the expression of CSN1S1 was weakly but positively correlated with the degree of exon 12 skipping.

Allele Expression

Figure 2 shows the relative expression of the different alleles in CSN1S1 exon 12 assay for animals with different SNP alleles at this locus. Recall that D is characteristic for haplotype 1_W , whereas A is present in haplotypes 3_W and 4_S , and G is present in haplotype 5_S and 7_W . The D allele from haplotype 1_W is only expressed for the combination ($1_W, 3_W$) and the A in 3_W is not expressed in genotype ($3_W, 5_S$). Hence, from Figure 2, it is clear that the weak haplotypes are only apparent relative to other weak haplotypes, whereas for genotypes with one strong (4_S or 5_S) and one weak haplotype (1_W or 3_W), only the strong allele is detectable. Together with results on the mRNA levels above, this indicates that dominance can be a result of differential expression between alleles.

Although Figure 2 clearly shows the differences in relative expression between weak and strong alleles, it also reveals an artifact related to the technology, because the G allele seems to be expressed in 2 animals

Table 4. Analysis of variance table for test of haplotypes

Source ¹	Sum of squares	df	Mean square	<i>F</i>	Prob > <i>F</i>
H_1	0.0165	1	0.0165	2.25	0.1429
H_3	0.01783	1	0.01783	2.43	0.1282
H_5	0.0019	1	0.0019	0.26	0.6141
H_7	0.00001	1	0.00001	0	0.9741
$H_1 \times H_3$	0.05435	1	0.05435	7.41	0.0102*
$H_1 \times H_5$	0.02357	1	0.02357	3.21	0.082
H_1^2	0.27313	1	0.27313	37.23	0***
H_5^2	0.00181	1	0.00181	0.25	0.6229
Error	0.24941	34	0.00734	—	—
Total	1.24081	42	—	—	—

¹ H_k = effect of haplotype k ($k = 1, 3, 4, 5, 7$); $H_k \times H_l$ = interaction effect of haplotype k and l ($k = 1, l = 3, 5$) (heterozygous); H_l^2 = squared effect of haplotype k (homozygous).

* P -value < 0.05; *** P -value = 0.

with genotype (1_W , 3_W), which do not have this variant. The reason for this is that there are systematic errors in the MALDI-TOF mass spectra, leading to small peaks throughout the spectra. Occasionally these peaks overlap with the signal from the cDNA or competitor. In our data, a small peak in the vicinity of the signal from the G allele was observed for all samples. Hence, the score for the G allele is biased, leading to 2 anomalous observations in Figure 2, and a seemingly higher expression of G than A in animals with the genotype (4_S , 5_S). This could be corrected numerically by estimating and subtracting a baseline for each allele. We have chosen not to do this because it could introduce other artifacts. Moreover, this problem was only observed for the exon 12 assay, which contained 3 alleles. Despite this obvious error, the results show a clear pattern between the strong and weak alleles.

Use of FTIR Spectroscopy to Identify Weak Genotypes

Measurements by FTIR on dry (Bruker) and liquid (Foss) milk samples were used to investigate the potential of FTIR to distinguish milk samples from animals with extreme haplotypes. More specifically, the animals were divided into 2 groups corresponding to the expression level of CSN1S1 (Figure 1). The group with low expression was labeled (w,w) and consisted of animals with 2 weak haplotypes. This group could be expected to have decreased α_{S1} -casein content in milk. In the

other group, with high expression of CSN1S1, the animals had at least 1 strong haplotype (4_S or 5_S); hence, this class was labeled (s,x).

To identify the most important regions of the spectra, i-PLS-DA was applied on Bruker and Foss spectra. The results are presented in Figures 3 and 4 for Bruker and Foss results, respectively. The figures show the error rate for each interval, as well as for the global models. The average spectra are visualized in the same plots.

The best classification from the Bruker FTIR data was obtained in the amide II region (interval 17, $1,600$ – $1,500$ cm^{-1}); the model in this region gave 3 out of 44 errors (5 components), and for the global model, 6 errors were made (6 components). For the Foss data, the best classification was obtained in the amide III region (interval 20, $1,300$ to $1,200$ cm^{-1}); when using 5 components, there were 6 out of 45 mistakes, and for the global model 14 errors were made (12 components). As expected, using spectra obtained from dried milk samples resulted in a better classification than using spectra from wet milk samples. The dried milk spectra have, in general, better quality and higher resolution than the wet milk spectra. This result may be due to the differences in the sampling technique and to instrumental differences. The amide I region, which is the most specific region for protein measurements, is not available for the wet milk measurements. It is, however, surprising that the 2 instruments differed so much in their performance in the amide III band.

Interpretation of PLS Model

In the following discussion, we will focus on the Bruker spectra, because they gave the best classification and can be expected to contain the most information about the protein composition of the milk samples. Although the best classification was obtained in the amide II region ($1,600$ to $1,500$ cm^{-1}), Figure 3 clearly shows that a larger region within the protein region had good predictive ability. The final model was therefore fitted to the region $1,700$ to $1,400$ cm^{-1} (intervals 16 to 18). Based on the cross validation results, 4 components were used in the final model. During cross validation, only 1 observation was misclassified. The significant wavelengths ($P < 0.05$, 10,000 permutations) were within $1,634$ to $1,630$, $1,616$ to $1,610$, $1,520$ to $1,514$, $1,506$ to $1,504$, and at $1,416$ cm^{-1} . Of these, $1,630$, $1,616$, $1,516$, and $1,416$ cm^{-1} corresponded to true minima in the second derivative of the milk spectra and the true protein spectra (Figure 5), indicating that differences in the milk spectra are related to qualitative differences in protein composition.

Figure 6 shows the score plot for the first 2 PLS-DA components. Superimposed on this, the predicted

Figure 2. Relative expression of D, A, and G in CSN1S1 (α_{S1} -casein) exon 12 for animals with different alleles at this position. The haplotype combinations for each individual are indicated in the boxes above. The D allele is present in haplotype 1_W only, whereas the A allele occurs in haplotype 3_W and 4_S , and the G allele in haplotype 5_S and 7_W . D, A, and G represent different polymorphisms in exon 12; D = deletion; w and s denote weak and strong haplotypes, respectively.

Figure 3. Cross-validated error rate for interval-partial least squares discriminant analysis using Bruker spectra after extended multiplicative signal replicate correction with 3 components. The error rates are shown for 5 components for the local models and 6 components for the global model. The bars represent the error rate for each interval and the dotted line the error rate for the global model. The solid curve represents the mean milk spectrum. The gray box marks the Fourier transform infrared region with no information. In this region, the error rate is randomly distributed around 0.5. The error rate for interval 17 ($1600\text{--}1500\text{ cm}^{-1}$) is 0.07, whereas for the global model it is 0.14.

scores for various purified cow milk protein fractions are included; these were obtained from their preprocessed FTIR spectra and then fitted to the same PLS-DA model. The caseins have high positive scores along the second component and are located together with the

(s,x) group, which has high expression of CSN1S1. The 2 whey proteins, on the other hand, have high negative scores on this component and are located together with the (w,w) group, which has low expression of CSN1S1. Hence, the second component appears to be related to

Figure 4. Cross-validated error rate for interval-partial least squares discriminant analysis using Foss spectra after extended multiplicative signal correction. The error rates are shown for 5 components for the local models and 12 components for the global model. The bars represent the error rate for each interval and the dotted line the error rate for the global model. The solid curve represents the mean milk spectrum. The gray box marks regions not used in calibration due to saturation by water absorption. The error rate for interval 20 ($1300\text{--}1200\text{ cm}^{-1}$) is 0.12, whereas for the global model it is 0.31.

Figure 5. Regression coefficients for the final partial least squares discriminant analysis model after 4 components, together with mean casein and whey spectra. Significant wavelengths (permutation P -values < 0.05 with 10,000 permutations) corresponding to true minima in the milk spectra are marked by arrows. In addition, wavelengths between 1,506 and 1,504 cm^{-1} were significant. All the vectors were normalized to length one before plotting.

differences in milk protein composition for animals with high or low expression of CSN1S1, and indicates that goats with low expression of CSN1S1 have a relatively higher amount of whey proteins in their milk. This shows that the FTIR spectra contain information not only on protein content, but also on protein composition, which is related to protein quality.

In Figure 7, the correlation loadings; that is, the correlation between the original variables (including dummy variables for the class labels) and the PLS-DA components, are shown for the first 2 components. Only the significant wavelengths are marked in the plot. The most interesting feature in Figure 7 is that the minimum at 1,630 cm^{-1} is highly correlated with the second component, which is related to the differences in casein-to-whey composition of the milk samples. Moreover, the minimum at 1,416 cm^{-1} is positively correlated with the (w,w) class; that is, the class with low expression of CSN1S1. The minimum occurring around 1,516 to 1,520 cm^{-1} is highly correlated with the first component. This region of the FTIR spectrum is in the amide II region, reflecting a secondary protein structure of unidentified nature. As can be seen in the score plot (Figure 6), this first component is strongly influenced by FTIR spectra of a few extreme animals. The second component is strongly correlated with 1,630 to 1,634 cm^{-1} , which probably reflects a better known secondary protein structure, namely an aggregated β -sheet. The variation along this component is spanned by many animals. The third component (not shown) is

highly correlated to the peak around 1,616 cm^{-1} , which is located in the middle of the plot in Figure 7; this component is not easy to interpret, and is dominated by only a few observations.

Animals included in this study were selected with the objective of maximizing the contrast between the weak and strong haplotypes. Consequently, 2 levels of CSN1S1 expression were observed, and classification was applied to relate the spectral measurements of milk samples to the mRNA levels for individual goats. In the population as a whole, a more continuous distribution should be expected, and a regression approach could be more useful for relating mRNA levels and FTIR spectra. We have shown that FTIR has potential for classifying milk samples from animals with high versus low expression of CSN1S1, which indicates that FTIR has the ability to separate milk samples with low and high α_{S1} content and different protein compositions, as shown in Figure 6. Although this study was based on selected specific CSN1S1 haplotypes present in the Norwegian dairy goat population, FTIR may also be applied to separate other CSN1S1 variants with strong and weak production of α_{S1} -casein.

Implications

Both types of FTIR could distinguish, to a large degree, the samples with high versus low expression of CSN1S1. In our data, the best classification was obtained with the Bruker system because these spectra were more informative about the protein structure. An

Figure 6. Score plot for the first 2 components with predicted scores for the pure proteins. + = animals with 2 weak haplotypes referred to as (w,w) in text; O = animals with 1 strong and 1 weak haplotype, referred to as (s,x) in the text; ◇ = predicted scores for the pure proteins. The labels have been moved manually to improve readability.

Figure 7. Correlation loading plot for the 2 first components; that is, the correlation between the original variables and the scores from the partial least squares model. Only the significant wavelengths ($P < 0.05$) are marked in the plot (+); the class variables are marked by circles (O). The inner and outer circles mark 50 and 100% explained variance, respectively; hence, variables between the circles are more than 50% explained by the shown components. Moreover, variables close to each other are positively correlated, whereas variables in opposite directions are negatively correlated.

obvious advantage of the Foss spectra, however, is that they are already used for routine quality control of milk. The routine quality control of milk samples considers protein percentage, whereas FTIR provides additional information about protein structure and composition. This implies that the use of FTIR in quality control of goat milk can be extended. Furthermore, spectral measurements are fast and inexpensive, and because the protein composition is directly related to the genotype, our results may pave the way for including protein composition in breeding goals.

Perfect classification cannot be obtained by FTIR. It is, however, much faster and inexpensive than genotyping, so some misclassification could be acceptable. In PLS-DA the class labels are dummy variables predicted by a PLS regression model. In our analyses, we used a cut-off limit of 0.5. By adjusting the cut-off limit, the classification could be adjusted to avoid certain types of misclassification errors. One could also use FTIR as a first screening, and then genotype animals where the predicted class label is uncertain.

Finally, it must be noted that FTIR separates animals that have no strong alleles from animals with 1 or 2 strong alleles; it is difficult to distinguish between animals with 1 or 2 strong alleles. Hence, eliminating weak alleles from the Norwegian goat population by this approach would be difficult. However, because the frequency of haplotype 1 is so high (0.72), applying

selection pressure against animals homozygous for this allele would reduce the frequency of this haplotype to some extent. This process would at least decrease the number of animals that are homozygous for weak alleles.

CONCLUSIONS

Lactating goats having defective alleles in both CSN1S1 haplotypes have very low expression of the gene in the mammary gland. When one of these weak haplotypes is combined with a strong haplotype, the expression level of CSN1S1 is not significantly different from expression levels in animals with 2 strong haplotypes. Moreover, expression of weak alleles was not detectable in animals being heterozygous for weak and strong haplotypes. Expression of other CSN genes was not affected by the haplotypes. The differential expression has clear effects on milk composition, because milk samples from animals with low and high expression could be separated using FTIR. The differences in the FTIR spectra are related to both protein content and protein composition.

ACKNOWLEDGMENTS

Ole Taugbøl (Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences) is acknowledged for the development of the biopsy procedure. MassARRAY gene expression analysis was run by Kristil K. Sundsaasen (CIGENE, Norwegian University of Life Sciences). Arne Roseth (CIGENE, Norwegian University of Life Sciences) assisted when taking biopsies. Kjetil Jørgensen (Tine) provided the Foss measurements of the milk samples. Financial support from the Agricultural Food Research Foundation and the grant 185135/I10 (Quality goat milk for cheese production) financed by the Research Council of Norway is also greatly acknowledged.

REFERENCES

- Brignon, G., M. F. Mahe, B. Ribadeaudumas, J. C. Mercier, and F. Grosclaude. 1990. Two of the three genetic-variants of goat alpha-s1-casein which are synthesized at a reduced level have an internal deletion possibly due to altered RNA splicing. *Eur. J. Biochem.* 193:237–241.
- Caroli, A., F. Chiatti, S. Chessa, D. Rignanese, P. Bolla, and G. Pagnacco. 2006. Focusing on the goat casein complex. *J. Dairy Sci.* 89:3178–3187.
- Ding, C., and C. R. Cantor. 2003. A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *Proc. Natl. Acad. Sci. USA* 100:3059–3064. doi:10.1073/pnas.0630494100
- Hayes, B., N. Hagesaether, T. Ådnøy, G. Pellerud, P. R. Berg, and S. Lien. 2006. Effects on production traits of haplotypes among casein genes in Norwegian goats and evidence for a site of preferential recombination. *Genetics* 174:455–464.

- Hewavitharana, A. K., and B. van Brakel. 1997. Fourier transform infrared spectrometric method for the rapid determination of casein in raw milk. *Analyst (Lond.)* 122:701–704.
- Kohler, A., U. Böcker, J. Warringer, A. Blomberg, S. W. Omholt, E. Stark, and H. Martens. 2009. Reducing inter-replicate variation in FTIR spectroscopy by extended multiplicative signal correction (EMSC). *Appl. Spectrosc.* 63:296–305.
- Leroux, C., P. Martin, M. F. Mahe, H. Leveziel, and J. C. Mercier. 1990. Restriction fragment length polymorphism identification of goat alpha s1-casein alleles: A potential tool in selection of individuals carrying alleles associated with a high level protein synthesis. *Anim. Genet.* 21:341–351.
- Leroux, C., N. Mazure, and P. Martin. 1992. Mutations away from splice site recognition sequences might cis-modulate alternative splicing of goat alpha-s1-casein transcripts—Structural organization of the relevant gene. *J. Biol. Chem.* 267:6147–6157.
- Luginbühl, W. 2002. Evaluation of designed calibration samples for casein calibration in Fourier transform infrared analysis of milk. *Lebensm. Wiss. Technol.* 35:554–558.
- Martens, H., J. P. Nielsen, and S. B. Engelsen. 2003. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Anal. Chem.* 75:394–404.
- Martens, H., and E. Stark. 1991. Extended multiplicative signal correction and spectral interference subtraction—New preprocessing methods for near-infrared spectroscopy. *J. Pharm. Biomed. Anal.* 9:625–635.
- Martin, P., M. Szymanowska, L. Zwierzchowski, and C. Leroux. 2002. The impact of genetic polymorphisms on the protein composition of ruminant milks. *Reprod. Nutr. Dev.* 42:433–459.
- Næs, T., T. Isakson, T. Fearn, and T. Davies. 2002. *A User-Friendly Guide to Multivariate Calibration and Classification*. 1st ed. NIR Publications, Chichester, UK.
- Neveu, C., A. Riaublanc, G. Miranda, J. F. Chich, and P. Martin. 2002. Is the apocrine milk secretion process observed in the goat species rooted in the perturbation of the intracellular transport mechanism induced by defective alleles at the alpha(s1)-CN locus? *Reprod. Nutr. Dev.* 42:163–172.
- Norgaard, L., A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen. 2000. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* 54:413–419.
- Sacchi, P., S. Chessa, E. Budelli, P. Bolla, G. Ceriotti, D. Soglia, R. Rasero, E. Cauvin, and A. Caroli. 2005. Casein haplotype structure in five Italian goat breeds. *J. Dairy Sci.* 88:1561–1568.
- Sørensen, L. K., M. Lund, and B. Juul. 2003. Accuracy of Fourier transform infrared spectrometry in determination of casein in dairy cows' milk. *J. Dairy Res.* 70:445–452.
- Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. *J. Dairy Sci.* 89:3690–3695.
- Vegarud, G. E., T. S. Molland, M. J. Brovold, T. G. Devold, and P. Al-eström. 1989. Rapid separation of genetic-variants of caseins and whey proteins using urea-modified gels and fast electrophoresis. *Milchwissenschaft* 44:689–691.
- Wold, S. 1982. Food research and data analysis. Pages IX, 535 s in *Proc. International Union of Food Science and Technology Symposium*, Oslo, Norway. Applied Science, London, UK.