Detection of genetic variation affecting milk coagulation properties in Danish Holstein dairy cattle by analyses of pooled whole-genome sequences from phenotypically extreme samples (pool-seq)


*Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark; and †Department of Food Science, Aarhus University, Tjele, Denmark

ABSTRACT: Rennet-induced milk coagulation is an important trait for cheese production. Recent studies have reported an alarming frequency of cows producing poorly coagulating milk unsuitable for cheese production. Several genetic factors are known to affect milk coagulation, including variation in the major milk proteins; however, recent association studies indicate genetic effects from other genomic regions as well. The aim of this study was to detect genetic variation affecting milk coagulation properties, measured as curd-firming rate (CFR) and milk pH. This was achieved by examining allele frequency differences between pooled whole-genome sequences of phenotypically extreme samples (pool-seq). Curd-firming rate and raw milk pH were measured for 415 Danish Holstein cows, and each animal was sequenced at low coverage. Pools were created containing whole genome sequence reads from samples with “extreme” values (high or low) for both phenotypic traits. A total of 6,992,186 and 5,295,501 SNP were assessed in relation to CFR and milk pH, respectively. Allele frequency differences were calculated between pools and 32 significantly different SNP were detected, 1 for milk pH and 31 for CFR, of which 19 are located on chromosome 6. A total of 9 significant SNP, which were selected based on the possible function of proximal candidate genes, were genotyped in the entire sample set (n = 415) to test for an association. The most significant SNP was located proximal to CSN3, explaining 33% of the phenotypic variance. CSN3, coding for κ-casein, is the most studied in relation to milk coagulation due to its position on the surface of the casein micelles and the direct involvement in milk coagulation. Three additional SNP located on chromosome 6 showed significant associations explaining 7, 3.6, and 1.3% of the phenotypic variance of CFR. The significant SNP on chromosome 6 were shown to be in linkage disequilibrium with the SNP peaking proximal to CSN3; however, after accounting for the genotype of the peak SNP within this QTL, significant effects (P-value < 0.1) could still be detected for 2 of the SNP accounting for 2 and 1% of the phenotypic variance. These 2 interesting SNP were located within introns or proximal to the candidate genes—solute carrier family 4 (sodium bicarbonate cotransporter), member 4 (SLC4A4) and LIM and calponin homology domains 1 (LIMCH1), respectively—making them interesting targets for further analysis.

Key words: curd-firming rate, dairy cattle, genetic variation, sequencing, whole-genome sequencing of pools of individuals (pool-seq)


1This project was conducted as part of the “Danish–Swedish Milk Genomics initiative,” supported by the Danish Council for Strategic Research, the Danish Cattle Federation, and Arla Foods as well as the “Coagene” project supported by the Danish National Advanced Technology Foundation, Viking Genetics, and Arla Foods. A special thanks to Mette Jeppesen and Hanne Jørgensen for technical assistance.

2Equal contributions.

3Corresponding author: Christian.Bendixen@mbg.au.dk
Received September 18, 2015.
Accepted January 29, 2016.
### INTRODUCTION

Milk coagulation is central to cheese production and is strongly associated with cheese yield. Rennet-induced coagulation is a 2-step process comprising an enzymatic phase where proteolysis of κ-casein leads to second phase aggregation of casein micelles and gel formation (Horne and Banks, 2004; Poulsen et al., 2013). The occurrence of poorly coagulating (PC) and noncoagulating (NC) milk (Ikonen et al., 2004; Poulsen et al., 2013) is highly disadvantageous for the dairy industry, and indeed, a negative correlation with cheese yield has been established (Wedholm et al., 2006). Furthermore, studies have shown that the mixing of NC milk with well-coagulating milk proportionally reduces the overall coagulation properties (Frederiksen et al., 2011).

It is well established that impaired coagulation properties (PC and NC milk) is associated with lower casein content, lower content of total protein, lower total calcium content, higher milk pH, and larger casein micelle size (Glantz et al., 2010; Hallén et al., 2010; Jensen et al., 2012). Furthermore, the degree of phosphorylation and glycosylation of αS1-casein and κ-casein, respectively, affects milk coagulation properties (MCP; Frederiksen et al., 2011; Jensen et al., 2012). Moreover, MCP are greatly influenced by the different genetic variants of the major milk proteins, particularly CSN2 (casein beta) coding for β-casein, CSN3 (casein kappa) coding for κ-casein, and PAEP (progestagen-associated endometrial protein) coding for β-lactoglobulin (Comin et al., 2008; Bonfatti et al., 2010; Poulsen et al., 2013). However, only a part of the variation in milk coagulation is explained by the variation in the major milk proteins, and other factors are expected to contribute to the trait.

A previous study indicated a role of a sialyltransferase thought to be involved in glycosylation of κ-casein (Tyrisèva et al., 2008). However, even though several studies have identified genomic regions or proposed candidate genes and gene ontologies associated with MCP (Schopen et al., 2011; Gambra et al., 2013; Gregersen et al., 2015), there is limited information about the causative genes responsible for reduced coagulation ability.

The efforts in relation to MCP have until now concentrated on identification of genomic regions affecting the traits analyzed through genomewide association studies (GWAS) using SNP chip data. A number of candidate genes have been suggested, but there is still limited knowledge on the causal variants. Genomewide association studies based on SNP chips rely on the selected SNP on the chips. The SNP on the bovine high density HD chip are selected based on a minimum allele frequency of 5% in specific breeds as well as additional criteria based on the position in the genome (Illumina, 2015). Access to low-frequen-

### MATERIALS AND METHODS

#### Ethical Statement

All procedures were approved by the National Guidelines for Animal Experimentation and the Danish Animal Experimental Ethics Committee, and all sampling was restricted to routine on-farm procedures that did not cause any inconvenience or stress to the animals, and hence, no specific permission was required. All sampling was approved by the owners of the 20 farms who gave permission to conduct the studies on site.

#### Milk and Tissue Samples

Morning milk and ear clips were sampled from 456 privately owned Danish Holstein (DH) cows according to standard farming procedures and as a part of the Danish–Swedish Milk Genomics Initiative (https://djextranet.agrsci.dk/sites/milkgenomics/public/pages/front.aspx, accessed February 17 2016). Between 19 and 25 samples were collected from each of 20 herds visited in the period October to December of 2009. The cows were selected to be as unrelated as possible on the different farms; however, sires could father cows from different farms. Most cows (174) did not have sires in common with anyone, but 7 sires fathered more than 10
cows (10 to 17 cows). Records were kept of days in milk and parity. All sampling procedures were performed as described in Poulsen et al. (2013). All animals were in mid lactation. Somatic cell count (SCC) for individual milk samples was determined by flow cytometry using a Fossomatic 5000 (Foss Analytical A/S, Hillerød, Denmark) at the Eurofins Laboratory, Holstebro, Denmark. Samples with high SCC (SCC > 500,000 cells/mL; 37 DH) were excluded. Rennet-induced coagulation was determined on individual milk samples in duplicate using a ReoRox4 rheometer (MediRox AB, Nyköping, Sweden) as described in Poulsen et al. (2013). Milk pH was measured in skimmed milk samples, prior to coagulation, with a PHM 220 pH meter (RadioMeter, Copenhagen, Denmark).

Genomic DNA was purified from ear tissue by salt precipitation. Cells were lysed by adding 500 μL lysis buffer (1 M Tris pH 8.0, 0.5 M EDTA pH 8.0, 2 M NaCl, and 10% SDS) and 30 μL proteinase K (10 mg/mL) to 25 to 50 mg of ground tissue. After storage overnight at 55°C, 300 μL 6 M NaCl was added and the sample was centrifuged for 15 min at 3,500 × g at room temperature. Seven hundred microliters of supernatant were transferred to a new tube and DNA was precipitated by adding 500 μL isopropanol and centrifuged for 10 min at 3,500 × g at room temperature. Precipitated DNA was washed with 0.5 mL 70% ethanol followed by centrifugation for 10 min at 3,500 × g at room temperature and discarding the supernatant before resuspension in 250 μL Tris-EDTA (1 M Tris pH 8.0 and 0.5 M EDTA pH 8.0).

**Sequence Library Preparation, Sequencing, and Read Mapping**

After exclusion of samples with high SCC and samples with a low amount of DNA or lack of trait measures, each of 415 remaining DH cows were sequenced on a HiSeq 2000 (Illumina, San Diego, CA) at low genome-wide coverage (approximately 0.6x depth of coverage). Paired-end sequencing libraries were constructed for the HiSeq platform using the NEBNext DNA library Prep Master Mix Set for Illumina (New England BioLabs, Ipswich, UK) for sequence library preparation according to the manufacturer’s guidelines. The libraries were amplified by PCR using the PCR index primers (PCR primer index 1–12, part numbers 1005714–1005725; Illumina) to generate library-specific barcodes according to the manufacturer’s guidelines. Amplified libraries were multiplexed into 36 pools each containing 12 indexed samples. Each pool was sequenced in a single lane on the Illumina HiSeq 2000. Reads were base called and demultiplexed using the manufacturer’s standard procedures. Resulting paired-end reads were individually mapped against the *Bos taurus* reference genome Btau_UMD3.1 (Zimin et al., 2009) using the Burrows–Wheeler alignment tool BWA version 0.5.9-r16 (Li and Durbin, 2009). Multisample SNP calling was performed using UnifiedGenotyper from the Genome Analysis Toolkit (DePristo et al., 2011). NGS-SNP (Grant et al., 2011) was used as described in Das et al. (2015) to obtain a comprehensive functional annotation of the called SNP. NGS-SNP used Ensembl (Flicek et al., 2011), dbSNP (Sayers et al., 2011), Entrez Gene (Sayers et al., 2011), and Uniprot (The Uniprot Consortium, 2011) as source databases and *Homo sapiens* for the model annotations to obtain the annotation of the genetic variants.

A more detailed description of library construction, sequencing, read mapping, and SNP annotation can be found in Supplementary File S1 (see the online version of the article at http://journalofanimalscience.org).

**Selection of Animals to Pool**

Because low-coverage sequencing (approximately 0.6x depth of coverage) is not sufficient for precise detection of genotypes in relation to single individuals, it is not possible to conduct a traditional GWAS based on the data generated. Instead, a pool-seq approach, where differences in allele frequencies are used as an indication of association, was performed by combining sequence reads of individuals showing extreme phenotypes as described below. Samples were pooled according to 2 different traits: CFR, calculated from consecutive points of the linear part of the gelation profile (Poulsen et al., 2013), and pH of the raw skimmed milk prior to the coagulation analysis. The distribution of CFR and pH is shown in Fig. 1 and 2, respectively. Two pools from phenotypically extreme samples were generated for each of the 2 traits. For CFR, the pools were created based on the average of the duplicate measurements of CFR. Based on this distribution, the different pools were selected with the aim of obtaining pools of similar sizes containing the extreme low and high values for each trait. Eight samples had a CFR of 0 and were classified as NC. For CFR, 19 samples (11 PC and 8 NC samples) were selected for pool 1 (low CFR) and 23 well-coagulating samples were selected for pool 2 (high CFR). The distribution of pH is relatively narrow, and 18 samples were selected for pool 1 (low milk pH) and 19 samples for pool 2 (high milk pH).

A principal components analysis using genomically estimated kinship as previously described in Buitenhuis et al. (2013) showed no sign of population stratification (Supplementary Fig. S4; see the online version of the article at http://journalofanimalscience.org). However, in order to assure low or absence of
stratification between pools, within-traits family relations were examined. Family relations were investigated at the sire to great-grand sire level for all pools, and it was found that the level of relations between the pools of extreme CFR or milk pH was approximately the same. It was observed that if sires fathered more than 1 cow (daughter to great-granddaughter) in 1 pool, this sire would also have daughters or granddaughters or great-granddaughters represented in the pool of the other extreme. Descriptive statistics of the pools are shown in Table 1. Sequence data of the 4 pools can be accessed via the European Nucleotide Archive (study accession number PRJEB8408; http://www.ebi.ac.uk/ena, accessed February 17, 2016).

Measurement of Allele Frequency Differences between Pools

The BAM files for each individual sample in each pool were merged with SAMtools (Li et al., 2009) ‘merge’ and then, for each pool, a mpileup file was generated using SAMtools ‘mpileup.’ The coverage for each pool was calculated as the average of the read depth for each position in the mpileup file. Each mpileup file was subsequently split by chromosome before conversion to a synchronized file format using PoPoolation2 (version 1.201; Kofler et al., 2011) with a minimum base quality of 20 and the fastq-type “Sanger.” In total, the analyses predicted 16,871,942 SNP across all samples (n = 415) whereof 6,992,186 segregated within the pools of CFR (n = 42) and 5,295,501 segregated within the pools of milk pH (n = 37).

To identify possible genomic regions associated with CFR or pH, the difference in allele frequency between groups was calculated for individual SNP by means of the fixation index (Fst) using PoPoolation2 (Kofler et al., 2011). The Fst was calculated from the allele frequencies using the classical approach (Hartl and Clark, 2007) with the following settings: a minimum count of the minor allele of 5 combined for the pools, a minimum coverage for each pool of 8, a maximum coverage of 20, the option ‘suppress noninformative,’ and a window size of 1. A Fisher’s exact test was performed using PoPoolation2 with the same settings as for the calculation of the Fst in regard to minimum count and coverage to estimate the significance of the allelic difference between populations. Quantile-quantile (Q-Q) plots were constructed for the P-values using an R script (http://www.gettinggeneticsdone.com/2010/07/qq-plots-of-p-values-in-r-using-base.html, accessed February 17, 2016). Based on the results of the Q-Q plots, genomewide significance levels were set to $9.5 \times 10^{-7}$ and $7.3 \times 10^{-7}$ for CFR and milk pH, respectively.

Genotyping

As the pool-seq analyses combined with P-values calculated using the Fisher’s exact test is indicative of only potential associations, further analysis were performed to investigate these associations. A subset of 12
SNP was selected for further analysis. The selected SNP were all significant in relation to the Fst P-value. Seven significant SNP located within introns of genes were selected for subsequent analyses based on a predicted function considered relevant for milk coagulation (kinases, phosphatases, Guanosine TriPhosphatases, and transcription factors; involvement in metal binding or transport; and calcium dependent or calcium regulation of pathways). In addition, 2 significant SNP located in the intergenic regions proximal to the gene SLC4A4 were selected. The selection in relation to SLC4A4 was based on the findings of 2 independent studies performed in DH and the Swedish Red breed, respectively, which pinpoint the region around the gene to be relevant for cheese production (Gregersen et al., 2014, 2015). Also, 3 SNP in close proximity to CSN3 were selected based on the known involvement of the protein in milk coagulation. In order to obtain genotyping information separately for each sample, custom TaqMan SNP genotyping assays (Applied Biosystems, Foster City, CA) were designed for each of the selected SNP. All animals were genotyped for the selected SNP with the purpose of performing a traditional association test. The initial preparation was performed as described in Poulsen et al. (2013); however, the analysis was performed using the ViiA7 real-time PCR equipment and ViiA7 software version 1.2.1 (Applied Biosystems). For one of the selected SNP, a TaqMan assay could not be designed and genotyping was performed by PCR followed by Sanger sequencing of the PCR products. Positions of the selected SNP as well as of the constructed assays and PCR primers are listed in Supplementary Table S1 (see the online version of the article at http://journalofanimalscience.org).

### Table 1. Descriptive statistics of the selected pools and all samples for curd-firming rate and pH

<table>
<thead>
<tr>
<th></th>
<th>Pool1a</th>
<th>Pool2b</th>
<th>All samplesc</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Curd-firming rate</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.326</td>
<td>Mean</td>
<td>19.846</td>
</tr>
<tr>
<td>Median</td>
<td>1.65</td>
<td>Median</td>
<td>19.2</td>
</tr>
<tr>
<td>SD</td>
<td>1.206</td>
<td>SD</td>
<td>3.1164</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>Minimum</td>
<td>16.1</td>
</tr>
<tr>
<td>Maximum</td>
<td>2.85</td>
<td>Maximum</td>
<td>26.55</td>
</tr>
<tr>
<td>Count</td>
<td>19</td>
<td>Count</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td></td>
<td>418</td>
</tr>
<tr>
<td><strong>pH</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>6.548</td>
<td>Mean</td>
<td>6.851</td>
</tr>
<tr>
<td>Median</td>
<td>6.56</td>
<td>Median</td>
<td>6.84</td>
</tr>
<tr>
<td>SD</td>
<td>0.023</td>
<td>SD</td>
<td>0.044</td>
</tr>
<tr>
<td>Minimum</td>
<td>6.5</td>
<td>Minimum</td>
<td>6.81</td>
</tr>
<tr>
<td>Maximum</td>
<td>6.57</td>
<td>Maximum</td>
<td>6.98</td>
</tr>
<tr>
<td>Count</td>
<td>18</td>
<td>Count</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td></td>
<td>419</td>
</tr>
</tbody>
</table>

a–c: Significant differences (t-test) within traits between pools are indicated by different letters (P < 0.05).

### Population-Based Association Study for Selected SNP

A number of SNP showed an appropriate significance level in relation to the Fst P-values as well as a promising position within the genome in relation to candidate genes as described above. These SNP were genotyped in the entire sample population using the TaqMan assays and tested for an additive allele effect. Calculation of the genomic relationship matrix (G) was performed using 494,984 SNP as previously described (Buitenhuis et al., 2013) and the linear animal model described below was applied for the association mapping:

\[
Y_{ijk} = \mu + \text{Herd}_i + \text{Parity}_j + b \times \text{SNP} + \text{Animal}_k + e_{ijk},
\]

in which \(Y_{ijk}\) is the phenotype of individual \(k\) in herd \(i\) and lactation \(j\); \(\mu\) is the fixed mean effect; \text{Herd}_i is a fixed effect \((i = 1, 2, \ldots, 20)\); \text{Parity}_j is a fixed effect \((j = 1, 2, 3)\); \(b\) is the allele substitution effect, which is a count in individual \(k\) of 1 of the 2 alleles (with arbitrary labeling); and Animal is the random additive genetic effect based on \(G\) of animal \(k\) (Yang et al., 2010). The effect of the SNP was tested by a Wald test with a null hypothesis of \(b = 0\). The analyses were performed using REML in the R interface of DMU (Madsen et al., 2007). The amount of phenotypic variance explained was estimated using the output from the ANOVA statistics. The total variance was computed as the sum of the variances for the model (residual, animal, and SNP), where the variance of the SNP allele effect \((\text{Var}_{\text{allele1}})\) was estimated according to

\[
\text{Var}_{\text{allele1}} = 2 \times \text{freq}_{\text{allele1}} \times (1 - \text{freq}_{\text{allele1}}) \times (\text{effect}_{\text{allele1}})^2,
\]

in which \text{freq}_{\text{allele1}} and \text{effect}_{\text{allele1}} are the frequency and the estimated additive genetic effect, respectively, of allele 1. The variance of the SNP allele effect was then divided by the total variance and multiplied by 100 to get the value in percentage. Box plots were created for SNP showing a significant association to CFR using R (version 3.0.2; http://www.r-project.org; accessed February 17, 2016), and statistical significance between genotypes was determined by ANOVA followed by a t-test for 2 samples assuming unequal variance.

### RESULTS

#### Genetic Variation between Pools

Sequence reads from individual samples were merged into 4 pools including extreme samples for CFR and milk pH. The number of mapped sequence reads
was 683,515,752 in pool 1 and 856,340,275 in pool 2 for CFR, corresponding to 82.8 and 85.8%, respectively, of the sequence reads being mapped to the *Bos taurus* reference genome. For milk pH, the number of sequence reads was 644,741,118 in pool 1, corresponding to 83.8% of the sequence reads, and 660,732,261 in pool 2, corresponding to 85.6% being mapped to the reference genome. The average read depth for pools 1 and 2 related to CFR was 10.7 and 13.7, whereas the average read depth related to pH was 10.2 and 10.6, respectively, only considering mappings of high quality. This corresponds to the coverage per sample of 0.6x depth of coverage in all pools. The allele frequency for each position in the reference was calculated for each pool separately, and subsequently, the *F*~st~ values were generated between pools within the different traits taking into account coverage and total minimum allele frequency as previously described. In total, the data contained 6,992,186 SNP related to CFR and 5,295,501 SNP related to pH. To identify statistically significant SNP, a Fisher’s exact test was performed and cutoff values to correct for multiple testing were identified by calculation of Q-Q plots. The measurement of allelic differences between populations for the individual traits resulted in 6 highly significant SNP for CFR and 1 significant SNP for pH (Table 2). However, additional regions were also suggested by the analysis. Significant SNP located in introns or upstream of potential candidate genes are listed in Table 3. A complete list of significant SNP is available in Supplementary Table S2 (see the online version of the article at http://journalofanimalscience.org).

### Table 2. Highly significant SNP identified by the pool contrast fixation index (*F*~st~) values

<table>
<thead>
<tr>
<th>BTA</th>
<th>SNP ID<del>1</del></th>
<th>Alleles</th>
<th>Position, bp</th>
<th>Coverage</th>
<th><em>F</em><del>st</del> value</th>
<th><em>P</em>-value</th>
<th>Annotation</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curd-firming rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>rs381890252</td>
<td>A/G</td>
<td>62,133,423</td>
<td>22<del>2</del></td>
<td>0.560</td>
<td>9.47 × 10^-8</td>
<td>Intron</td>
<td>LIMCH1<del>3</del></td>
</tr>
<tr>
<td>5</td>
<td>rs384508820</td>
<td>A/C</td>
<td>83,705,799</td>
<td>12</td>
<td>0.779</td>
<td>8.31 × 10^-8</td>
<td>Intergenic</td>
<td>UGT2A3<del>4</del></td>
</tr>
<tr>
<td>6</td>
<td>rs381126510</td>
<td>C/T</td>
<td>86,780,858</td>
<td>17</td>
<td>0.700</td>
<td>4.51 × 10^-8</td>
<td>Intergenic</td>
<td>UGT2A3<del>4</del></td>
</tr>
<tr>
<td>6</td>
<td>rs137665213</td>
<td>A/G</td>
<td>86,844,984</td>
<td>13</td>
<td>1.000</td>
<td>1.47 × 10^-8</td>
<td>Intergenic</td>
<td>UGT2A3<del>4</del></td>
</tr>
<tr>
<td>6</td>
<td>rs385267572</td>
<td>C/T</td>
<td>88,582,180</td>
<td>14</td>
<td>0.647</td>
<td>8.50 × 10^-8</td>
<td>Intergenic</td>
<td>SLC4A4<del>3,4</del></td>
</tr>
<tr>
<td>18</td>
<td>rs41892680</td>
<td>C/G</td>
<td>56,322,108</td>
<td>12</td>
<td>1.000</td>
<td>5.75 × 10^-8</td>
<td>Intron</td>
<td>CCDC155<del>3</del></td>
</tr>
<tr>
<td>pH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>rs455511793</td>
<td>A/C/G/T</td>
<td>28,578,472</td>
<td>20</td>
<td>0.468</td>
<td>7.30 × 10^-8</td>
<td>Intron</td>
<td>RNF122<del>3,5</del></td>
</tr>
</tbody>
</table>

~1~ID = Identification number. 
~2~Manually selected based on mapping quality, even though a coverage of 22 was above the threshold of coverage. 
~3~Selected SNP validated by PCR. 
~4~Close proximity to candidate gene (<1 Mb). 
~5~Unable to genotype variant.
~6~Not detected in our samples or in the 1000 Bulls Genome data set ([Daetwyler et al., 2014]).

DISCUSSION

In this study, we have used whole-genome NGS to detect SNP with a significant allele frequency difference between pools of extreme samples based on the traits CFR and pH of milk in Holstein cows. Low CFR has previously been correlated with poor MCP and lower cheese yield, and samples with CFR = 0 are described as NC ([Wedholm et al., 2006; Poulsen et al., 2013]). Furthermore, a correlation between pH of unprocessed milk and coagulation properties has been detected ([Ikonen et al., 2004]). Due to the limited number of samples in this project, the size of the pools are relatively small; however, increasing the...
size of each pool would reduce the phenotypic difference between the pools and potentially decrease the genetic difference between pools as well as reduce the power of discovering genetic variation related to CFR and milk pH. Especially for milk pH, where the range of the phenotype is quite narrow, this could pose a potential problem. The values reported are, however, close to the normal range, considered to be between 6.6 and 6.8, with the tails extending outside the range.

**Correction for Multiple Testing**

A Fisher’s exact test taking into account the sequence coverage within the given position was calculated in order to estimate the P-values for comparison of allele frequencies of the identified SNP. The Q-Q plot of the P-values in relation to CFR (Fig. 3A) indicated that the data was distributed as expected, with only few SNP being significant. From the figure, it can be seen that a shift in distribution occurs at $9.5 \times 10^{-7}$, and hence, this level was selected as a threshold for genomewide significance. The Q-Q plot of the P-values in relation to milk pH showed a different picture with only 1 data point above the diagonal ($P = 7.3 \times 10^{-8}$), indicating either population stratification or too low a sample size for this trait given the narrow phenotypic distribution of milk pH (Fig. 3B).

**Coverage of the Pools**

The average genome coverage for the entire study was identified to be 10x per pool, indicating that the observed coverage for some regions was found to be low (<10x depth of coverage for the pool). This will most likely result in falsely assigning some of the SNP as nonsignificant. A pool-seq study with pool sizes of 20 diploid individuals has shown that a coverage of 11x depth of coverage, corresponding to 0.55x depth of coverage per sample, is sufficient to obtain a high accuracy of allele frequency estimates ($R^2 = 0.95$; Table 3).

**Table 3.** Significant SNP associated to curd-firming rate identified by the pool contrast fixation index ($F_{st}$) values selected based on location within introns or in close range of functionally interesting genes

<table>
<thead>
<tr>
<th>BTA</th>
<th>SNP ID$^1$</th>
<th>Allele</th>
<th>Position</th>
<th>Coverage</th>
<th>$F_{st}$ value</th>
<th>$P$-value</th>
<th>Annotation</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs133117587</td>
<td>A/G</td>
<td>120,048,763</td>
<td>11</td>
<td>1.000</td>
<td>7.40 × 10$^{-7}$</td>
<td>Intron</td>
<td>HLTF$^2$</td>
</tr>
<tr>
<td>5</td>
<td>rs110899624</td>
<td>A/C</td>
<td>65,249,933</td>
<td>12</td>
<td>0.848</td>
<td>9.20 × 10$^{-7}$</td>
<td>Intron</td>
<td>ANO4</td>
</tr>
<tr>
<td>6</td>
<td>rs208009847</td>
<td>A/G</td>
<td>87,420,318</td>
<td>9</td>
<td>1.000</td>
<td>4.89 × 10$^{-7}$</td>
<td>Intergenic</td>
<td>CSN3$^3$</td>
</tr>
<tr>
<td>6</td>
<td>rs109846405</td>
<td>C/T</td>
<td>88,610,440</td>
<td>12</td>
<td>0.899</td>
<td>1.50 × 10$^{-7}$</td>
<td>Intergenic</td>
<td>SLC4A4$^2,3$</td>
</tr>
<tr>
<td>6</td>
<td>rs132814786</td>
<td>A/G</td>
<td>90,391,609</td>
<td>16</td>
<td>0.687</td>
<td>2.42 × 10$^{-7}$</td>
<td>Intron</td>
<td>RASSF6$^2,4$</td>
</tr>
<tr>
<td>17</td>
<td>rs447225243</td>
<td>C/T</td>
<td>72,185,024</td>
<td>17</td>
<td>0.614</td>
<td>9.32 × 10$^{-7}$</td>
<td>Intron</td>
<td>LIMK2</td>
</tr>
</tbody>
</table>

1ID = Identification number.
2Selected SNP validated by PCR.
3Close proximity to candidate gene (<1 Mb).
4Unable to genotype variant.
Rellstab et al., 2013). As this resembles our pool size and coverage, our data set should be sufficient to identify differences in allele frequency between pools. A potential risk is that the observed allele frequencies might be biased due to unequal representation of the individual samples in the pool, but because each sample was individually sequenced, manual inspections showed that this seems not to be the case. Previous studies revealed that the accuracy of estimating the allele frequency in pools is reliable; however, it was also evident that precision decreases with decreasing coverage (Ingman and Gyllensten, 2009; Rellstab et al., 2013). We investigated the performance of our pool-seq approach by comparing allele frequency data from sequencing with data from genotyping (Supplementary Table S3; see the online version of the article at http://journalofanimalscience.org). The data was split into pools and allele frequencies were calculated. The table shows that the larger pool (pool 2) is more accurate in the prediction compared with the smaller pool (pool 1) where larger discrepancies are observed. The $r^2$ for pool 1 is 0.75 and 0.99 for pool 2. The lower correlation in pool 1 indicates that the allele frequency estimates obtained from our pool-seq analysis might have been more accurate by increasing the coverage per sample (currently approximately 0.6x depth of coverage). Nevertheless, it seems safe to assume that the difference between pools still is a reliable estimate due to the random distribution of sequence reads. These findings, however, stress the need to perform additional tests in relation to the identified SNP to verify potential associations.

In more traditional pool-seq studies, DNA of extreme phenotype samples are first pooled and then the sequencing is performed. In this study, however, each sample was separately sequenced and then pooled, including equal amounts of randomly selected reads for each individual sample. In this way, we were sure that each cow gave equal contributions to the pool and we have the opportunity to perform analyses for many different traits. However, the additional cost of making individual libraries as well as individual coverage greatly restricts the use of this particular experimental design. Furthermore, because the study examines only differences between allele frequencies, it makes no difference if sequences are derived from pooled DNA or from pooled sequence data and the test statistics applied are the same.

### Significant Genomic Regions

A total of 31 and 1 SNP were identified with significantly different allele frequencies between the pools of extreme CFR and pH in milk values, respectively (Supplementary Table S2; see the online version of the article at http://journalofanimalscience.org). The relatively small proportion of significant SNP compared with the total number of SNP is, in part, related to the stringent selection criteria with respect to false positives. The coverage throughout the data set also has an effect. A threshold for the minimum allele count of 5 considering the total coverage in our data of, on average, 20x depth of coverage in the combined pools of a trait results in not all variants being detected. Alleles of high frequency in one pool can be missed if absent in the other pool as can low-frequency variants within these regions. In regions where the coverage is higher, we are, however, able to detect variants of low frequency. Furthermore, in relation to milk pH, the results indicate that only small effects are at work and the data set is not able to identify SNP having an effect on the trait due to the relatively low average coverage. In addition, the distribution of the phenotype, with a quite narrow range, also affects

<table>
<thead>
<tr>
<th>Closest gene</th>
<th>SNP ID$^1$</th>
<th>Allele$^2$</th>
<th>No.</th>
<th>Effect</th>
<th>Allele frequency</th>
<th>Percent variance</th>
<th>$P$-value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANO4</td>
<td>rs110899624</td>
<td>A/C</td>
<td>337</td>
<td>0.648</td>
<td>0.45</td>
<td>1.2</td>
<td>$6.87 \times 10^{-2}$</td>
<td></td>
</tr>
<tr>
<td>CCDC155</td>
<td>rs41892680</td>
<td>C/G</td>
<td>363</td>
<td>-0.000003</td>
<td>0.26</td>
<td>33</td>
<td>$9.99 \times 10^{-1}$</td>
<td></td>
</tr>
<tr>
<td>CSN3</td>
<td>rs200809847</td>
<td>A/G</td>
<td>343</td>
<td>3.46</td>
<td>0.44</td>
<td>33</td>
<td>$1.58 \times 10^{-25}$</td>
<td>*</td>
</tr>
<tr>
<td>LIMCH1</td>
<td>rs381890252</td>
<td>A/G</td>
<td>350</td>
<td>-1.149</td>
<td>0.38</td>
<td>33</td>
<td>$2.65 \times 10^{-3}$</td>
<td>**</td>
</tr>
<tr>
<td>LIMK2</td>
<td>rs447225243</td>
<td>C/T</td>
<td>341</td>
<td>-0.0000003</td>
<td>0.32</td>
<td>33</td>
<td>$9.99 \times 10^{-1}$</td>
<td></td>
</tr>
<tr>
<td>SLC4A4</td>
<td>rs385267572</td>
<td>C/T</td>
<td>342</td>
<td>-2.980</td>
<td>0.02</td>
<td>7</td>
<td>$2.54 \times 10^{-25}$</td>
<td>*</td>
</tr>
<tr>
<td>SLC4A4</td>
<td>rs109846405</td>
<td>C/T</td>
<td>336</td>
<td>-2.199</td>
<td>0.47</td>
<td>7</td>
<td>$2.54 \times 10^{-10}$</td>
<td>*</td>
</tr>
</tbody>
</table>

$^1$ID = identification number.

$^2$Minor allele in bold.

*1% genomewide significance after adjusting for multiple testing (cutoff utilising the entire data set); **5% significance level, adjusting for number of tests performed ($n = 7$).
the possibility to detect the genetic contribution, especially because the heritability for milk pH, on average, has been found to be only moderate ($h^2 = 0.21$; Bittante et al., 2012). This indicates that a larger sample size is probably needed to detect the associations and that this trait is affected, to a large extent, by environmental factors, for example, feeding or infection. The resulting Q-Q plot (Fig. 3B) indicates stratification. That we detect associations from the data set for CFR is related to the high heritability ($h^2 = 0.75$ in DH) that previously was found for this trait (Poulsen et al., 2015).

**Identification of Possible Candidate Genes**

Several studies have detected a major QTL on BTA6 associated with milk protein composition located around the casein cluster (Schopen et al., 2009, 2011; Huang et al., 2012; Gambra et al., 2013) as well as associations to milk coagulation traits (Gregersen et al., 2014; Glantz et al., 2015). In our study, we have detected a major locus comprising SNP in close proximity to CSN3 (Fig. 4). Several studies have previously linked CSN3 to coagulation properties (Comin et al., 2008; Bonfatti et al., 2010; Poulsen et al., 2013), and hence, it was not surprising that we saw a very large effect explaining 33% of the CFR phenotype variance by a single SNP (Table 3).

A SNP on chromosome 6 at position 88,610,440 bp located in an intergenic region close to the gene SLC4A4 was shown to have a genomewide significant association to CFR, explaining 7% of the phenotypic variance. As illustrated in the box plot (Fig. 5A), the distribution of the phenotypes in the different genotype groups has an additive genetic effect on the phenotype. This SNP could, therefore, explain some of the variation associated with noncoagulation in DH, as the homozygous group 2 (reference allele) contains no NC samples. The SLC4A4 gene is located in close proximity to the major locus on BTA6 (spanning about 1 Mb) containing the casein gene cluster, which also previously was identified to affect MCP (Gregersen et al., 2014). The association could potentially be caused by linkage disequilibrium (LD) with this cluster. To examine this further, a new association test adjusting for the genotype effect of a SNP (BTA6: 87,420,138) situated upstream of CSN3 detected within the major QTL was performed. The results of this association test indicated that the major part of the effect observed for the SLC4A4 SNP was due to linkage with the SNP in the casein cluster. The revised association test revealed that our SNP located in close proximity to SLC4A4 still explained 2% ($P$-value $= 0.07$) of the genetic variation. Another SNP related to SLC4A4 (BTA6:88,582,180) was also adjusted for the effect of the CSN3 SNP and was not found to be significant after the association test was performed. SLC4A4 codes for an electronegic sodium bicarbonate cotransporter (NBC1), which belongs to a family of integral membrane proteins mediating the transport of sodium and bicarbonate across epithelial cells and contributes to the regulation of intracellular pH (Abuladze et al., 2000). In a previous study, SLC4A4 was suggested as a possible candidate gene associated with clinical mastitis in Norwegian Red cattle (Sodeland et al., 2011). In this study, we found no significant difference in SCC between the pools, indicating an additional function of the gene.
Furthermore, we detected a SNP in intron 6 of the gene *LIMCH1* being significantly associated with CFR and explaining 3.6% of the phenotypic difference. The phenotypic distribution is depicted in Fig. 5B, illustrating that homozygous group 2 (reference allele) is not associated with the NC property of the trait. Again, an association test was performed incorporating the genotype of the *CSN3* SNP in the model. This revealed that the association detected in this position is mainly due to LD with SNP in the major QTL; however, our SNP within *LIMCH1* still explains 1% of the variance at the 10% significance level. In a previous study, Gambra et al. (2013) suggests that a causative mutation for κ-casein concentration in milk was not different in the transgenic mice compared with wild-type mice (Le et al., 2012). Interestingly, the gene *ANO4* (anotamin 4) on chromosome 5, which belongs to the family of Ca\(^{2+}\)-activated Cl\(^{-}\) channels (Tian et al., 2012), was also downregulated in transgenic mice overexpressing *miR-30b* (Le et al., 2012). In our study, the allele frequencies of a SNP in intron 7 of *ANO4* were found to be significantly different between the 2 pools; however, no significance was obtained by the association test. The findings of our study together with the study of Gregersen et al. (2014) identify 3 separate effects on CFR in a 26.5-Mb region on BTA6. This could indicate an advantage during evolution of codirectional selection, because the probability of markers being inherited together is increased when they are in close proximity.

We were not able to identify any SNP for milk pH. First of all, the only SNP that was identified by the pool-seq study could not be genotyped, indicating a false prediction of a variant caused either by sequencing errors or by alignment problems due to a repetitive region. Second of all, the Q-Q plot indicated some kind of stratification of the data. This could easily be caused by environmental factors as previously mentioned. Many of the samples used for the study were shown to be outside the normal range of milk pH where external factors such as bacteria in the milk making the milk acidic or mastitis making the milk more alkaline might affect the tails of the distribution (Atasever et al. 2010). We analyzed the tails of the trait within this experiment, and therefore, we might be analyzing the impact of the external factors instead of the intended trait. For this reason, milk pH might not be a trait suited for a pool-seq analysis and, in this particular case, a normal GWAS would have been preferred.

As predicted and shown by others (Yang et al., 2015), we were able to identify markers located within a QTL that was previously identified by GWAS, which confirms pool-seq as an alternative to GWAS. Further analysis is, however, needed in order to assess all SNP identified by the pool-seq analyses, and whether it might include new loci and low-frequency variants is uncertain but such analyses become very costly. The approach using

**Figure 5.** Box plots representing the distribution of curd-firming rate (CFR) in relation to genotypes of 3 SNP associated with the candidate genes *SLC4A4*, chromosome 6, 88610440 (A) and *LIMCH1*, chromosome 6, 62133423 (B). *a* indicates significant difference between genotypes.
low-sequence coverage of individual cows in the present study might result in missed SNP in high LD or SNP in coding regions of the suggested candidate genes; further work is required to identify and verify potentially causal genetic variants affecting the coagulation ability of milk.

**Conclusion**

Using a pool-seq strategy aimed at detecting SNP with significant allele frequency differences between pools of extreme samples related to CFR and milk pH revealed 31 and 1 significant SNP, respectively. Three SNP significantly associated with CFR were located proximal to the potential candidate genes CSN3, SLC4A4, and LIMCH1.

**LITERATURE CITED**


