Variants on *Bos Taurus Autosome 12* are genetically differentiated between Swedish Red cows with non-coagulating and well-coagulating milk

S. I. Duchemin¹, V. R. Gregersen², M. Glantz³, M. Paulsson³, F. A. L. Pinto⁴, W. F. Fikse⁴ & D-J de Koning¹

¹Swedish University of Agricultural Sciences, Animal Breeding and Genetics, PO Box 7023, SE-750 07, Uppsala, Sweden  
sandrine.duchemin@slu.se (Corresponding Author)  
²Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark  
³Department of Food Technology, Engineering and Nutrition, Lund University, Lund, Sweden  
⁴Växa Sverige, Uppsala, Sweden

Summary

Milk not coagulating within a defined time-span after rennet addition is known as non-coagulating (NC) milk. The high prevalence of NC milk found in Swedish Red (SR) cows is undesirable for the cheese processing industry. We used pooled sequences of SR cows with extreme phenotypes to identify new genomic regions associated with NC milk. Cows were divided in two groups: one group of NC milk and one group of well-coagulating (WC) milk. For each group, DNA of the cows were pooled and sequenced. Subsequently, we estimated the allele-frequency differences between the two groups of cows, tested these differences with the Fisher’s exact test, and calculated pairwise FST between the two groups of cows. For chromosome (BTA) 12, 11 significant variants were considered as genetically differentiated between the two pools. Furthermore, 2 out of these 11 results are: rs515854473, which is annotated as a deletion, and rs516248211 which is annotated as an SNP preceded by an insertion. BTA12 has not previously been associated with NC milk.

Keywords: milk coagulation properties, pooled sequences, FST, allele-frequency differences

Introduction

Milk not coagulating within a defined time-span after rennet addition is known as non-coagulating (NC) milk. Gustavsson et al. (2014) estimated a prevalence of 18% for NC milk in the Swedish Red (SR) breed. In addition, their study showed that NC milk is heritable (h²=0.45) and NC milk is unfavourably correlated with protein content (rg=0.38). Therefore, selecting SR breeding cows for higher protein content may increase the prevalence of NC milk. This increase is not desirable, especially for the cheese processing industry.

One of the challenges in studying NC milk is the identification of mutations that segregate at low minor allele frequency (MAF). Duchemin et al. (2016) identified a quantitative trait locus (QTL) associated with NC milk on *Bos taurus autosome* (BTA) 18, when performing a region-wise association study with imputed sequences. In their study, the most significant variant associated with this QTL segregated at low MAF. While it is interesting to find important low MAF variants, sample size, costs of sequencing and imputation remain as limiting factors. To overcome these limitations, Schlötterer at al. (2014) suggested sequencing pools of individuals over sequencing individuals separately, and Emond et al. (2012) suggested analysing individuals at both ends of the phenotype distribution (those with extreme phenotypes). The aim of this study was to use pooled sequences of SR cows with extreme phenotypes to identify new genomic regions associated with NC milk. For this purpose, cows were divided in two groups: NC milk and well-coagulating (WC) milk. For each group, DNA of the cows were pooled and sequenced. Subsequently, we applied principles of population
To estimate the allele-frequency differences, test these differences with the Fisher’s exact test, and calculate pairwise FST between the two groups of cows.

Material and methods

Animals, phenotypes and genotypes

Morning milk was sampled from approximately 400 SR cows belonging to 21 herds in the south of Sweden. Milk samples were collected during the indoor season in two distinct periods: April through May 2010, and September 2010 through April 2011. Directly after collection, fresh milk samples were cooled, defatted by centrifugation and preserved as described in Hallén et al. (2007). Rennet-induced coagulation of individual skim milk samples was determined using a ReoRox4 rheometer (MediRox AB, Nyköping, Sweden) as described by Poulsen et al. (2013). The addition of the chymosin represented time zero. The curd-firming rate of some samples after 60 min of rennet was equal to zero, therefore these samples were defined as NC milk. Subsequently, two pools with extreme phenotypes were selected as follows: a) pool 1 consisted of 30 SR cows with NC milk, and b) pool 2 consisted of 30 SR cows with WC milk with an average curd-firming rate equal to 13.5 Pa/min. Regarding pool 1, 16 cows are unrelated, and 14 cows are half-sibs. Among these half-sibs, 2 cows were mothered by 1 dam, and 12 cows were fathered by 3 sires. Regarding pool 2, 20 cows are unrelated, and 10 cows are half-sibs fathered by 5 sires. In the two pools, cows were multiparous (1-3 lactations) and were in different stage of lactation (5 - 54 weeks in milk).

For each pool of SR cows, equal amount of DNA extracted from blood samples were mixed and sequenced resulting in pooled sequences for pool 1 and 2. Paired-end sequencing libraries were prepared following manufacturer’s guidelines, were amplified by PCR and sequenced using the HiSeq2000 (Illumina, San Diego, CA). After sequencing, the read-to-variant workflow followed the GATK Best Practices guidelines put forward by the Broad Institute (e.g., DePristo et al, 2011). After using Trimmomatic to trim our pair-ended data, BWA was used to align the reads to the bovine genome assembly UMD3.1 (Zimin et al., 2009). The aligned reads were sorted, and duplicates marked, using the Picard tools. The resulting BAM files were used for further analyses. The average depth of coverage of the pooled sequences for pool 1 was 43.0x and for pool 2 was 47.6x.

Statistical analyses of pooled sequences

We used PoPoolation2 version 1.201 (Kofler et al., 2011) to calculate allele frequency between the two pools, pairwise FST and its implemented Fisher’s exact test to determine the significance of these allele frequency differences. A variant from a pool was genetically differentiated from another at FST > 0.5, and significant differences in allele frequency at $-\log_{10}(P\text{-values}) > 4.5$. We used Integrative Genome Viewer (IGV) version 2.3.98 (Thorvaldsdóttir et al., 2013) to visualize pairwise comparisons between pools 1 and 2. The annotation of the significant results was done using the Ensembl database. QQ-plots were generated in R, and genomic inflation ($\lambda$) was estimated and corrected for in R using custom made script.

Results

Genome-wide results considering FST > 0.5 and $-\log_{10}(P\text{-values}) > 4.5$ showed a total of 58 significant variants distributed on 24 chromosomes. Here we focus on the results of BTA12, for which 11 significant variants were genetically differentiated between Pools 1 and 2 (Table 1; Figure 2). The coverage of these 11 variants ranged from 13 to 27; the pairwise FST ranged from 0.53 to 0.78 and the significance of the difference in allele frequency ranged from 4.53
\[ \text{-Log}_{10}(P\text{-values}) \leq 5.65 \] (Table 1). Three out of these 11 variants are located around 37.7 Mbp and 3 out of 11 variants are located between 70 and 73 Mbp. Furthermore, QQ-plots after correcting for the genomic inflation (Figure 1), suggest that the few variants deviating from the \(x=y\) line are possible true significant results.

Among the 11 interesting results, we chose variants rs515854473 (Figure 3) and rs516248211 (Figure 4) to illustrate the genomic differences between cows of pools 1 and 2. Variant rs515854473 is located at 56.76 Mbp on BTA12 (Table 1; Figure 3), and is an intergenic variant characterized in the Ensembl database as a 1 bp deletion. This deletion is in close proximity to two SNP upstream from a deletion and two SNP downstream to an insertion/deletion. Variant rs515854473 is heterozygous in pool 1, with 6 sequences with C-alleles and 20 sequences with G-alleles, and is homozygous in pool 2, with 24 sequences with C-alleles. Variant rs516248211 is located at 84.63 Mbp on BTA12 (Table 1; Figure 4), and is an intergenic variant characterized in the Ensembl database as an SNP proceeded by an insertion. Variant rs516248211 is heterozygous in pool 1, with 24 sequences with A-alleles, 1 sequence with C-allele preceded by 1 insertion, and is heterozygous in pool 2, with 6 sequences with A-alleles, and 26 sequences with C-alleles preceded by 21 insertions.

**Discussion**

Selection of SR cows for higher protein content may be resulting in selecting undesirable variants in parts of the genome. The genetic diversity of these genomic regions can be compared between DNA pools of extreme phenotypes. On BTA12, our results indicate that many variants are genetically differentiated (FST \(\geq 0.5\)) between SR cows with NC and WC milk. There is no consensus on the appropriate threshold necessary to indicate that one population is genetically differentiated from another. For this reason, we chose 0.50 as an appropriate measure, because it is comparable with FST measurements used in human studies.

Inflation of estimated P-values suggests a systematic bias in our analyses. One explanation for this bias may be the genetic relationships between the SR cows in each of the pools. In pool 1, 14 out of the 30 SR cows were half-sibs, and in pool 2, 10 out of the 30 SR cows were half-sibs. After correcting for genomic inflation, the QQ-plot (Figure 1) supports the possibility of identification of distinct genomic regions between cows with NC and WC milk.

The results of BTA12 seem to be located in regions that are between insertion and deletions, on top of deletions, and in copy number variations (CNV) (Table 1). The region located between 70 and 76 mega base-pairs (Figure 1) shows promising differences between cows producing NC milk and WC milk. Moreover, when assessing the BAM files in IGV, variant rs515854473 (Figure 3) in pool 1 showed 6 sequences with C-Alleles and 20 sequences with G-Alleles whereas in pool 2, 24 sequences were homozygote with C-alleles. Interestingly, 14 out the 20 sequences with G-alleles seemed to be followed by an insertion (Figure 3). On the other hand, variant rs516248211 (Figure 4) in pool 1 only showed one sequence with C-allele preceded by one insertion whereas in pool 2, 21 out of the 26 sequences with C-alleles were preceded by insertions. Based on these findings, it is possible that cows producing NC milk do have different genomic regions from cows producing WC milk. Furthermore, it is possible that undesirable parts of the genome, such as insertions or deletions might be under strong artificial selection in SR cows.

Despite these interesting results, caution should be considered when interpreting biological significance of insertions and deletions and CNV associated with NC and WC milk. Because of the limitations of our study (e.g., differences in coverage between pools 1 and 2 and small samples size), we cannot exclude that these might not affect milk coagulation. However, other studies with comparable sample sizes (e.g., Bertelsen et al., 2016) have been successful in identifying genetically distinct genomic regions. Nevertheless, the detection of 6 out of 11 variants associated with NC milk in two smaller region on BTA12 deserves further investigation.
List of References


Table 1. Description of variants showing position, coverage, FST values, significance of difference in allele frequency, and annotation details from Ensembl database between the Swedish Red cows with non-coagulating and well-coagulating milk.

<table>
<thead>
<tr>
<th>Chr:BP1</th>
<th>Coverage</th>
<th>FST Value</th>
<th>-Log10(P-values)</th>
<th>Identifiers</th>
<th>Variant description</th>
<th>Annotation details from Ensembl database</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:3752165</td>
<td>23</td>
<td>0.53</td>
<td>4.54</td>
<td>rs38260219 7</td>
<td>intergenic</td>
<td>2nd SNP downstream insertion (rs439352216).</td>
</tr>
<tr>
<td>12:4175651</td>
<td>26</td>
<td>0.53</td>
<td>4.93</td>
<td>rs21042560 3</td>
<td>intergenic</td>
<td>4th SNP downstream insertion (rs730672163), and 7th SNP upstream insertion (rs475572632).</td>
</tr>
<tr>
<td>12:27333601</td>
<td>27</td>
<td>0.58</td>
<td>5.65</td>
<td>rs110868614</td>
<td>intergenic</td>
<td>3rd SNP downstream insertion (rs516566149), and 1st SNP upstream insertion EPO_Low_coverage,p-value=7.27e-38).</td>
</tr>
<tr>
<td>12:37703706</td>
<td>25</td>
<td>0.59</td>
<td>5.01</td>
<td>rs51749220 6</td>
<td>intergenic</td>
<td>6th SNP downstream insertion (rs466179177).</td>
</tr>
<tr>
<td>12:37703710</td>
<td>22</td>
<td>0.56</td>
<td>4.54</td>
<td>rs79707440 5</td>
<td>intergenic</td>
<td>7th SNP downstream insertion (rs466179177).</td>
</tr>
<tr>
<td>12:37703717</td>
<td>21</td>
<td>0.62</td>
<td>4.95</td>
<td>rs79990455 4</td>
<td>intergenic</td>
<td>8th SNP downstream insertion (rs466179177).</td>
</tr>
<tr>
<td>12:56760225</td>
<td>23</td>
<td>0.63</td>
<td>5.10</td>
<td>rs51585447 3</td>
<td>intergenic</td>
<td>Deletion (not SNP!). 3rd SNP upstream a deletion, and 3rd SNP downstream an insertion/deletion.</td>
</tr>
<tr>
<td>12:70406611</td>
<td>26</td>
<td>0.58</td>
<td>5.52</td>
<td>rs38317462 3</td>
<td>intergenic</td>
<td>Located in a CNV with 9 features.</td>
</tr>
<tr>
<td>12:72035259</td>
<td>13</td>
<td>0.78</td>
<td>4.53</td>
<td>rs47353351 5</td>
<td>intergenic</td>
<td>Located in a CNV with 3 features.</td>
</tr>
<tr>
<td>12:73690071</td>
<td>26</td>
<td>0.60</td>
<td>5.63</td>
<td>rs52663731 2</td>
<td>intron</td>
<td>ENSBTAG00000046041 gene located in a CNV with 35 features.</td>
</tr>
<tr>
<td>12:84632742</td>
<td>25</td>
<td>0.61</td>
<td>5.63</td>
<td>rs516248211</td>
<td>intergenic</td>
<td>Downstream of insertion (rs135337902).</td>
</tr>
</tbody>
</table>

1Chr:BP means the chromosome is followed by the position in base-pairs based on the UMD3.1 bovine assembly.
Figure 1. Quantile-Quantile plots showing the inflated P-values in gray, and the corrected P-values in blue.
Figure 2. Integrative Genome Viewer plots showing results for BTA12. Specifically, the pairwise allele-frequency differences measured by FST ranging from 0 to 1; and the significance of the allele frequency differences ranging from $0 \leq -\log_{10}(P\text{-value}) \leq 8.92$. 
Figure 3. Integrative Genome Viewer plots showing the pairwise comparisons between the pooled DNA sequences between Swedish Reds cows with non-coagulating (NC) and well-coagulating (WC) milk for variant rs515854473. NC pool showing 20 cows with G-alleles in orange followed by 14 insertions (in purple), and WC showing 24 cows with C-alleles only.
Figure 4. Integrative Genome Viewer plots showing the pairwise comparisons between the pooled DNA sequences between Swedish Reds cows with non-coagulating (NC) and well-coagulating (WC) milk for variant rs516248211. NC pool showing only one insertion, and WC showing 21 insertions (in purple).