Proc. Assoc. Advmt. Breed. Genet. 21: 61-64

USING PROTEIN QTL TO DISENTANGLE VARIANTS EFFECTING PROTEIN PERCENTAGE IN MILK NEAR THE CASEIN COMPLEX

K.E. Kemper[1], M.J. Carrick[2] and M.E. Goddard[1,3]

[1] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Victoria, Australia
[2] Berghan Carrick Consulting, Moonee Ponds, Victoria, Australia
[3] Dept. of Economic Dev., Jobs, Transport and Resources, AgriBio, Bundoora, Victoria, Australia

SUMMARY

Casein proteins comprise about 80% of the protein in bovine milk. The casein complex occupies a 300 kb region of the genome on BTA6. To disentangle the number of variants affecting protein percentage in bovine milk near the casein complex, we use single variant regressions with imputed full sequence variants and quantified α_{S1}-, β- and κ-casein protein levels in 444 Holstein Friesian cows. We find 2 variants, located near CSN3 (coding κ-casein) and within CSN1S1 (which codes for α_{S1}-casein), with independent effects on P% and which affect concentration of their corresponding casein gene products. Previously described protein polymorphisms in the casein genes were sometimes associated with the quantity of their respective proteins but it seems unlikely that these variants are causing variation in casein concentrations.

INTRODUCTION

Four types of casein proteins (α_{S1}-, α_{S2}-, β- and κ-casein) constitute about 80% of the protein in milk and the genes encoding the casein proteins are located in a 300kb region on *Bos taurus* autosome (BTA) 6 (Table 1). Polymorphisms in the amino acid sequence of these proteins have been known for many years and have been found to be associated with milk protein yield and concentration. However, the associations have not been consistent across studies perhaps because the mutations causing variation in amount of casein are not the same as those causing differences in amino acid sequence although they may be in linkage disequilibrium with them (reviewed by Goddard & Wiggans 1999).

Table 1. Genomic location of the casein genes on *Bos taurus* autosome (BTA) 6[*]

Gene description	Symbol	location (bp)	protein product
Bos taurus casein alpha-S1, mRNA (+)	CSN1S1	87,141,556-87,159,096	α_{S1}-casein
Bos taurus casein beta, mRNA (-)	CSN2	87,179,502-87,188,025	β-casein
Bos taurus casein alpha-S2, mRNA (+)	CSN1S2	87,262,457-87,280,936	α_{S2}-casein
Bos taurus casein kappa, mRNA (+)	CSN3	87,378,398-87,392,750	κ-casein

[*]Other genes also located in the region. Locations from UMD3.1 (www.ensembl.org/Bos_taurus/). The forward (+) or reverse (-) orientations for transcription are indicated after the gene description.

Kemper *et al.* (2015) identified a sequence variant (Chr6:87296809) as affecting protein content (P%) from a multi-trait meta analysis of Holstein and Jersey cattle. This variant was located within an intergenic region, closest to the *Bos taurus* casein alpha-S2 coding region (CSNS2). However, this analysis did not exclude the possibility that this variant was associated with cumulative effects of several different underlying P% causal variants in the region. The simplest hypothesis is that mutations in the regulatory region of each casein gene cause variation in the amount of that casein produced and therefore in the amount of total protein. The aim of this study is to disentangle the P% QTL observed in Holstein cattle near the casein complex by using phenotypes consisting of P% and quantification of three casein proteins (α_{S1}-, β- and κ-casein).

MATERIALS AND METHODS

Overview. The paper aims to identify variants (either causal variants or variants in strong LD with the causal variant) underlying α_{S1}-, β- and κ-casein concentrations in a small dataset and then to test whether or not these variants can explain the variation in P% due to the QTL near the casein complex. The methods are detailed below and consist of single variant regression for α_{S1}-, β- and κ-casein concentration, followed by conditional single regression analysis for P% using imputed whole genome sequence data.

Phenotypes and genotypes. There are two datasets. The first dataset consists of genotypes and phenotypes for 444 cows measured for α_{S1}-, β- and κ-casein concentration (mg/g) using capillary zone electrophoresis (Kanning, Casella & Oliman 1993) on combined morning and afternoon milking at two sampling days, approximately 6 weeks apart. A model with fixed (mean concentration, breed, 4^{th} order polynomials for age & days-in-milk) and random (herd, permanent environment (PE), animal) effects was fitted to the data and trait-deviations for animals constructed as the average of PE, animal and residual effects for animals with two measurements. Genotypes were available for the 50K bovine single nucleotide polymorphism (SNP) chip and these genotypes were imputed to the high-density array (632,002 SNP) following Erbe *et al.* (2012). Protein types for α_{S1}-, β- and κ-casein were determined using gel electrophoresis following Ng-Kwai *et al.* (1984).

The second dataset, described by Kemper *et al.* (2015), consists of P% phenotypes and genotypes of 632,002 (real and imputed) high-density SNP for 8478 Holstein cows. Unlike Kemper *et al.* (2015), this analysis uses only Holstein animals.

Sequence variants. Sequence variants consisted of SNP and small INDEL from a 5 Mb region centred on the casein complex (BTA6: 84.5-89.5 Mb). Data were obtained from run 4 of the 1000 bull genomes project (Daetwyler *et al.* 2014).

Imputation and the association study. Sequence variants were imputed into the 2 datasets for the target region on BTA6 using Minimac (Fruchberger, Abecasis and Hinds 2015) and 260 sequenced Holstein animals as the reference population. The association study for each phenotype used EMMAX (Kang et al. 2010) following Kemper *et al.* (2015). Multi-allelic protein polymorphisms were treated as a series of contrasts (i.e. A1 & B types vs. A2 for β-casein). The conditional analysis for P% was also conducted using EMMAX.

RESULTS AND DISCUSSION

Variants associated with individual casein concentrations. The most significant results were obtained for κ-casein, followed by β- and α_{S1}-casein concentrations (Table 2). The variant most highly associated with κ-casein concentration was Chr6:87405588, located about 13 kb downstream of CSN3 ($P = 7.7x10^{-12}$). Similarly for β- and α_{S1}-casein, the most significant variants were outside the coding regions for the genes, where Chr6:87098077 is 43 kb upstream of CSN1S1 and Chr6:87206907 is 19 kb upstream of CSN2 (N.B. that CSN2 is transcribed on the reverse strand). However, there are a number of other sequence variants which are also highly associated with the casein concentrations and any one of these could be the causal mutation (Figure 1).

Variants associated with protein variants. The frequency of the protein polymorphisms varied widely between the casein genes, with β-casein A1 and A2 variants being of intermediate frequency (0.44 & 0.51 respectively), the κ-casein B variant having a relatively high frequency compared to the C variant (0.78 vs. 0.22) and the β-casein B having a low frequency (0.04). There were few observations of the C allele for α_{S1}-casein and the A3 allele at β-casein, effectively rendering the α_{S1}-casein protein type monomorphic.

The missense mutations causing the known protein polymorphisms were associated with the protein variants as expected. In each case, there were a number of other sequence variants that

also associated with the protein type due to the high degree of LD in the region and the small sample size (Table 2). These sequence variants were 20-30 kb away from the variants identified as influencing the quantity of these proteins. Both the κ-casein B/C protein polymorphism (Chr6:87390576; P = $4.3x10^{-11}$) and the β-casein B protein polymorphism (Chr6:87181453; P = $1.3x10^{-4}$) were strongly associated with the concentrations of their respective proteins. However, these variants were more than 1 $\log_{10}$ unit from the most significant variant and it seems likely that they are in LD with variants affecting the protein concentrations.

Table 2. Most significant sequence variants for casein concentrations and protein types, where the variants within 1 $\log_{10}$ unit (number; location range, bp) assesses level of confidence in the top variant

phenotype	top variant (P value)	additional variants within 1 $\log_{10}$ unit
κ-casein conc.	Chr6:87405588 ($7.7x10^{-12}$)	132 (87,333,107 – 87,407,175)
α_{S1}-casein conc.	Chr6:87098077 ($9.2x10^{-6}$)	18 (87,085,525 – 87,154,594)
β-casein conc.	Chr6:87206907 ($8.3x10^{-6}$)	4 (87,090,414 – 87,115,771)
κ-casein B vs. C protein type[1]	Chr6:87393434 ($4.7x10^{-14}$)	122 (87,363,855 – 87,405,868)
β-casein A1/B vs. A2 protein[2]	Chr6:87184548 ($2.2x10^{-167}$)	8 (87,169,673 – 87,184,548)
β-casein A1/A2 vs. B protein[3]	Chr6:87186827 ($1.8x10^{-319}$)	2 (87,185,552 – 87,189,903)

[1]Chr6:87390576 is the mutation causing Ile>Thr (ref>alt) substitution in κ-casein and was ranked 29th in the analysis; [2]Chr6:87181619 is the mutation causing the His>Pro substitution in β-casein and was ranked 5th; [3]Chr6:87181453 is the mutation causing the Ser>Arg substitution in β-casein and was ranked 3rd.

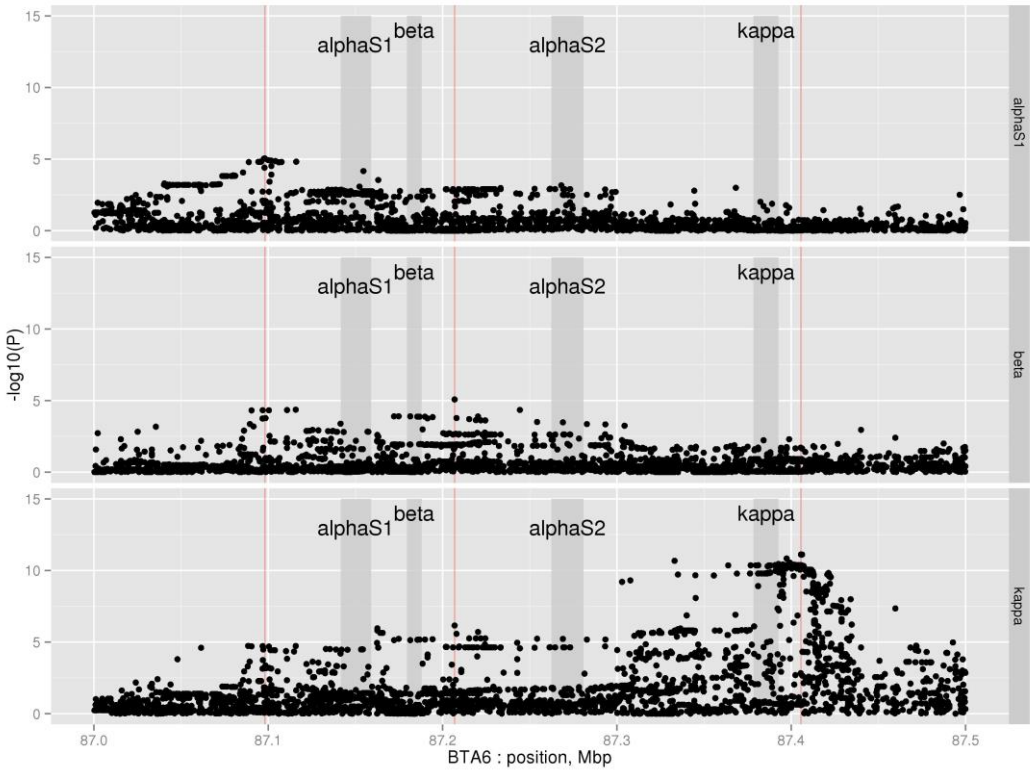

Figure 1. Association study in the casein region for quantity of α_{S1}-, β-, and κ-casein. The most significant variant for each trait (Chr6:87098077, Chr6:87206907 & Chr6:87405588) is highlighted with vertical lines.

Dissecting the P% QTL observed in the casein region. The variant identified by Kemper *et al.* (2015), Chr6:87296809, was highly significant for P% ($6.7x10^{-14}$) but was not within a $\log_{10}$ unit of the top variant for any of the three individual casein concentrations. Thus the analyses aimed to discover independent variants for P%, based on the most significant κ-casein variant, followed by any variants remaining significant ($P < 1x10^{-6}$) after a conditional analysis on the κ-casein variant. It was found that fitting two variants (Chr6:87405588 and Chr6:87154594) reduced all other variants to $P > 1x10^{-6}$ (Figure 2). The Chr6:87154594 variant was located within an intron of CSN1S1 and was within 1 $\log_{10}$ unit of the top variant for α_{S1}-casein concentration ($P = 6.8x10^{-5}$). The most significant variant for β-casein concentration (Chr6:87206907) was not significantly associated with P% (P = 0.82) after adjusting for Chr6:87405588. This suggests there is not an independent effect of β-casein concentration on P%. It is possible that Chr6:87206907 is capturing the effect of a haplotype which affects both κ- and β-casein concentrations.

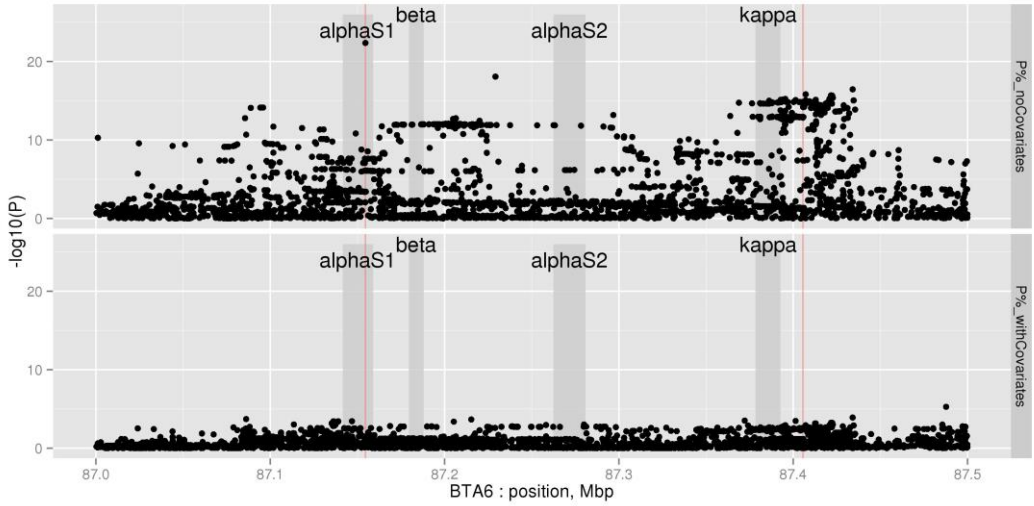

Figure 2. Association study in the casein region for P% without fitting covariates (top), and after fitting covariates of Chr6:87405588 and Chr6:87154594 (bottom, variants indicated by vertical lines).

We conclude that the significant result for Chr6:87296809 (Kemper *et al.* 2015) was likely due to the cumulative effects of at least two variants affecting P% in the casein complex. Our results suggest 2 independent variants that influence κ- and α_{S1}-casein concentrations and therefore cause variation in P%. It seems that these variants are distinct (but sometimes associated with) the known mutations causing the protein polymorphisms for α_{S1}-, κ- and β-casein.

REFERENCES

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., *et al.* (2014) *Nat Genet* **46:**858.
Erbe, M., Hayes, B.J., Matukumalli, L.K., Gosawami, S., *et al.* (2012) *J Dairy Sci* **95:**4114.
Fruchberger, C., Abecasis, G.R., Hinds, D.A. (2015) *Bioinformatics* **31:**782.
Goddard, M.E. and Wiggans, G.R. (1999) In 'The genetics of cattle', pp. 522-525, editor R. Fries & A. Ruvinsky, CABI Publishing.
Kang, H.M., Sul J.H., Service S.K., Zaitlen N.A, *et al.* (2010) *Nat Genet* **42:**348.
Kanning, M., Casella, M. and Olieman, C. (1993) *LC-GC Int* **6:**701.
Kemper K.E., Hayes B.J., Daetwyler, H.D. & Goddard M.E. (2015) *J Anim Breed Genet.* **132:**121.
Ng-Kwai-Hang, K.F., Hayes, J.F., Moxley, J.E. and Monardes, H.G. (1984) *J Dairy Sci* **67:**835.