Genetic analysis of protein composition of bovine milk

Thesis Committee

Thesis supervisor
Prof. Dr. Ir. Johan A.M. van Arendonk
Professor of Animal Breeding and Genetics
Wageningen University

Thesis co-supervisors
Dr. Ir. Marleen H.P.W. Visker
Research Associate, Animal Breeding and Genetics Group
Wageningen University

Dr. Ir. Henk Bovenhuis
Assistant Professor, Animal Breeding and Genetics Group
Wageningen University

Other members
Dr. Paolo Carnier (University of Padova, Italy)
Prof. Dr. Fred A. van Eeuwijk (Wageningen University, the Netherlands)
Prof. Dr. Ir. Gert Jan Hiddink (Wageningen University, the Netherlands)
Dr. Dirk-Jan de Koning (The Roslin Institute, United Kingdom)

This research was conducted under the auspices of the Graduate school of Wageningen Institute of Animal Sciences (WIAS).

Genetic analysis of protein composition of bovine milk

Ghyslaine C. B. Schopen

Thesis

Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 2 June 2010
at 4 p.m. in the Aula.

Ghyslaine C. B. Schopen (2010)
Genetic analysis of protein composition of bovine milk,
189 pages

Thesis Wageningen University, Wageningen, the Netherlands

With references, with summary in English and Dutch

ISBN: 978-90-8585-645-0

Abstract

This thesis is part of the Dutch Milk Genomics Initiative, and the general aim was to obtain more insight into the genetic background of bovine milk protein composition. Morning milk samples from roughly 2000 cows were analyzed for the six major milk proteins (α_{S1}-casein, α_{S2}-casein, β-casein, κ-casein, α-lactalbumin and β-lactoglobulin) using capillary zone electrophoresis.

The estimated genetic parameters for milk protein composition showed that there was considerable genetic variation for milk protein composition and that the genetic correlations among the six major milk proteins were low. There was a strong negative genetic correlation between β-lactoglobulin and total casein in milk. The presence of genetic variation justified the performance of in-depth genetic analyses such as linkage and association mapping. A linkage study was performed to screen the whole bovine genome to identify chromosomal regions affecting milk protein composition. This study resulted in ten chromosomal regions, of which regions on BTA6, 11 and 14 showed the largest effect on milk protein composition. The confidence intervals of these regions were large, in general. Therefore, an association study was performed to narrow down these chromosomal regions and to detect new chromosomal regions affecting milk protein composition. The association study resulted in four main regions on BTA5, 6, 11 and 14, and also new regions were detected. These new regions may, in addition to the four main regions, play a role in the genetic regulation of milk protein synthesis.

The milk protein composition is important for technological properties of milk. An increase in casein index is preferable for the cheese production. Therefore, four scenario's, to increase casein index in milk, were discussed. The first scenario has been termed genetic differentiation, the second scenario was genetic selection based on estimated breeding values, the third scenario was genetic selection based on genotypes, and the last scenario was genomic selection. These four scenarios illustrated that there are opportunities to utilize genetic variation in milk protein composition.

Contents

1

General introduction

Milk

Milk, especially cow's milk is consumed as a food product in many cultures and it is a natural source of a whole range of nutrients essential for growth, development and maintenance of the human body. Milk provides protein, fat, carbohydrates, vitamins and minerals. For many years, cow's milk has been processed into dairy products such as butter, yoghurt and cheese. The suitability of milk for the production of different dairy products depends upon the composition of milk. For example, for cheese production it is important that milk protein contains a high proportion of casein (Wedholm *et al.*, 2006). The last decades, cows have been selected mainly for high milk, fat and protein production. It is not known, however, what the consequences of this selection are on e.g. the composition of the milk fat and the milk protein. In this thesis, the focus is on milk protein composition.

Milk protein composition

Research on milk proteins started around 1814, when the first paper was published by J.J. Berzelius (Fox, 2003). In 1838, J.G. Mulder described a method for the preparation of protein from milk by acid precipitation (Fox, 2003). This acid precipitated protein is referred to as casein. About fifty years later, whey proteins were separated in soluble and insoluble fractions by Seblein (1885; Fox, 2003). At that time two kinds of milk proteins were distinguished: caseins and whey proteins. The caseins are insoluble and precipitate at pH 4.6 whereas the whey proteins remain soluble at this pH (Fox, 2003). The distinction between caseins and whey proteins is still in use, however, since then the subdivision of these two main categories has been further refined. The caseins can be divided in α_{S1}-casein, α_{S2}-casein, β-casein and κ-casein, and the whey proteins in α-lactalbumin and β-lactoglobulin. These are the six major milk proteins in bovine milk and represent ±90% of the total milk protein content. The remaining 10% consist of minor proteins, like bovine serum albumin, γ-casein, immunoglobulins, lactoferrin and many proteins that appear in low concentrations (Farrell *et al.*, 2004).

Detailed milk protein composition can be determined using different methods, e.g. high-performance liquid chromatography (HPLC), polyacrylamide gel electrophoresis (PAGE) and capillary zone

electrophoresis (CZE). The quantification of proteins in different milk samples gives a detailed view of the variation in milk protein composition between individual cows. However, the quantification of milk proteins in individual milk samples is rarely done, because it is laborious and costly. For the research described in this thesis, milk protein composition of a resource population of nearly 2000 cows was determined using CZE, as described by Heck *et al.* (2008). More detailed information about CZE is given in Text box 1.

Variants of milk proteins

The six major milk proteins originate from their corresponding milk protein genes. Genes consist of DNA sequences which are transcribed into messenger-RNA. Messenger-RNA (and DNA) consists of four different bases and three consecutive bases form a codon. Each codon of the messenger-RNA is translated into an amino acid. There are 20 different amino acids, and the sequence of amino acids determines the properties of the protein. One difference in the amino acid sequence, due to mutations, can give the protein different properties. An example of different variants of milk proteins was first described for the whey protein β-lactoglobulin. Aschaffenburg and Drewry (1955) discovered that β-lactoglobulin protein exists in two variants, A and B, which differ from each other by two amino acids changes. The variant occurring in the milk of an animal is genetically controlled and may be AA, AB or BB depending on the DNA sequence of the animal. In subsequent years, variants were detected for most of the milk proteins. Only a few studies have examined the effects of genetic variants of milk proteins on milk protein composition (e.g., Ng-Kwai-Hang *et al.* 1987; Bobe *et al.* 1999 and Heck *et al.* 2009). These authors showed that variants in β-CN, κ-CN and β-LG are associated with milk protein composition and with total casein in milk.

Genetic variation, heritability and genetic correlation

Variation has been found for many traits investigated in livestock species. Part of this variation is due to genetic factors (heritability). Under the infinitesimal model it is assumed that genetic differences are caused by many genes, each with a small effect. The infinitesimal model forms the

Text box 1 Capillary Zone electrophoresis

Capillary zone electrophoresis (CZE) is a technique by which proteins can be separated based on their size and mainly on their charge. A protein sample is injected into a capillary that is filled with a liquid, and proteins are separated by applying an electric field.

After injection at the anode, the proteins with the highest positive charges are moving with the highest speed through the capillary to the cathode. Large molecules will move with a lower speed through the capillary than smaller molecules, but the charge of the molecule is the most important factor determining the speed through the capillary. At the end of the capillary individual proteins are detected by UV absorption. Figure 1 shows a schematic representation of a CZE.

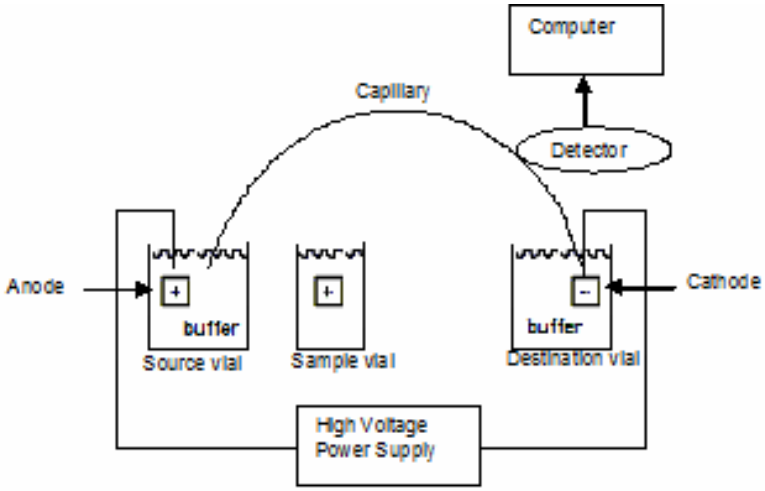

Figure 1 Capillary zone electrophoresis diagram

basis of quantitative genetics which is at the heart of present day selective breeding.

The opportunities to change a trait through breeding depend upon the amount of genetic variation. The heritability (h^2) of a trait expresses the genetic variation as a proportion of the phenotypic variation. A low heritability means that only a small fraction of the differences observed between animals is due to genetics.

To determine whether it is possible to alter the composition of milk protein through selective breeding, it is important to know its heritability. The heritability for total milk protein production (kg per cow per lactation) is about 0.26 (e.g. Chauhan and Hayes, 1991; Lund *et al.*, 1999). The heritability for milk protein content is about 0.50 (e.g. Hayes *et al.*, 1984; Ikonen *et al.*, 1999), while for casein content it is about 0.30 (Hayes *et al.*, 1984; Ikonen *et al.*, 2004). Only a few studies have estimated heritabilities for the major milk

proteins (e.g. Ikonen *et al.*, 1997; Bobe *et al.*, 1999; Graml and Pirchner, 2003), but no studies have reported genetic correlations among the major milk proteins. The limited number of studies is a reflection of the technological difficulties of quantifying the major milk proteins simultaneously on a large number of cows and daughters of bulls, which is a prerequisite for estimating their genetic parameters.

Genetic variation is needed to gain genetic improvement. However, the genetic correlation between traits is also important. Different traits, e.g. protein production and casein content, might be partially influenced by the same genes. This would mean that selection on protein production would lead to both a genetic change in protein production (direct effect) and a genetic change in casein content (correlated effect). The size of the correlated response depends on the so-called genetic correlation between traits (r_g). A positive genetic correlation between two traits means that both traits are positively affected and will speed up the genetic improvement. However, it is also possible that one trait is positively affected and the other trait negatively affected (negative genetic correlation). In this case the genetic correlation will slow down the genetic improvement.

QTL mapping in dairy cattle

When it has been established that a trait is influenced by genetic factors, it is of interest to identify polymorphisms in genes that contribute to the genetic variation. For the milk proteins, some variants (e.g. β-lactoglobulin variants A and B) are already known. Bobe *et al.* (1999) showed that the β-lactoglobulin variants A and B explain a major part of the genetic variation of β-lactoglobulin fraction. Bobe *et al.* (1999) and Heck *et al.* (2009) showed that variants of β-casein, κ-casein and β-lactoglobulin are associated with the genetic variation of milk protein composition. However, there might be more genes contributing to the genetic variation in milk protein composition. One method to identify chromosomal regions affecting the trait of interest (quantitative trait loci (QTL)) is linkage mapping. Bovenhuis and Schrooten (2002) and Khatkar *et al.* (2004) have given an overview of QTL for milk production traits in dairy cattle (milk, fat and protein yield, and fat and protein percentage). There are no publications reporting QTL for milk protein composition.

Molecular markers

Until recently, microsatellites were the primary type of markers used for QTL mapping in livestock. They are abundant, multi-allelic and occur randomly throughout the genome, and, because they are highly polymorphic, they are highly informative. Because of these marker characteristics, their use over the years has been extremely valuable. The increasing availability of Single Nucleotide Polymorphisms (SNPs) provides an alternative (Schaid *et al.*, 2004). SNPs are more abundant than microsatellites and also occur randomly throughout the genome. A disadvantage of SNPs relative to microsatellites is that they have only two alleles. As a result SNPs are less informative and, therefore, more SNPs are required to achieve the same level of information as compared to using microsatellites (Schaid *et al.*, 2004; Kruglyak, 1997). However, the major advantage of SNPs is their suitability for high-throughput genotyping. Therefore, at present, SNPs are the most abundantly used markers in genetic studies of livestock species.

Fine mapping

QTL regions obtained by linkage mapping are generally large and, thus, will contain hundreds of genes. It is impractical to consider hundreds of genes as potential candidates for the QTL effect. Therefore, the chromosomal region associated with the trait should be narrowed, i.e., the region should be fine mapped. To fine map a chromosomal region, more genetic markers are needed. The development of new techniques that enable genotyping of thousands of SNPs per individual has greatly facilitated fine mapping.

In April 2009 the bovine genome has been sequenced (The bovine genome sequencing and analysis consortium *et al.*, 2009), which is of interest for genetics. The sequence of the bovine genome results in the availability of the location of many genes and genetic markers. Therefore, the sequence of bovine genome makes it easier to point out candidate genes in chromosomal regions detected in the linkage study or association study. If no sequenced genome would be available, results from the linkage study and association study can be used for marker assisted selection. However, pointing out candidate genes will than still be a great challenge.

Text box 2 Resource Population

The Dutch Milk Genomics Initiative resource population consisted of 2000 first lactation cows from 400 herds distributed throughout the Netherlands. These cows descended from 5 proven and 50 test bulls, which resulted in 5 large paternal half-sib families of about 200 cows each and 50 small paternal half-sib families of about 20 cows each. This setup was chosen, in order to use the 5 large families for QTL analyses and the 50 small families for the estimation of the genetic parameters. In total, three morning milk samples were taken from each cow. The first sample, taken during the winter of 2005, was used to determine detailed milk protein composition. Blood samples of the cows and semen samples of the bulls were used to extract DNA.

Aim and outline of the thesis

The research described in this thesis is part of the Dutch Milk Genomics Initiative. The aim of the Dutch Milk Genomics Initiative is to identify opportunities to use natural genetic variation to improve milk quality e.g. milk fat composition and milk protein composition. More details on the resource population used in the Dutch Milk Genomics Initiative are given in Text box 2. The general aim of this thesis was to obtain more insight into the genetic background of bovine milk protein composition.

Chapter 2 describes the estimates of heritabilities and genetic correlations of bovine milk protein composition. The presence of genetic variation was a prerequisite for performing in-depth genetic analyses such as QTL and association mapping. Chapter 3 describes the comparison of microsatellite and SNP markers for their use in genetic studies. SNPs turned out to be the markers of choice for the subsequent linkage and association studies that were performed to identify chromosomal regions affecting milk protein composition, as described in chapters 4 and 5. Chapter 6 describes the comparison between a single SNP and a multiple SNP association analysis. In the general discussion (chapter 7), the results of the research described in this thesis are put into a broader perspective and options for practical application are presented.

References

Aschaffenburg R., and J. Drewry. 1955. Occurrence of different beta-lactoglobulins in cow's milk. Nature 30: 218–219.

Bobe, G., D.C. Beitz, A.E. Freeman, and G. L. Lindberg. 1999. Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. J. Dairy Sci. 82: 2797-2804.

Bovenhuis, H. and C. Schrooten. 2002. Quantitative trait loci for milk production traits in dairy cattle. Proceedings 7th World Congress on Genetics Applied to Livestock Production, Montpellier. 19-23 August. Vol. 31: 27-34.

Chauhan, V.P.S., and J.F. Hayes. 1991. Genetic parameters for first lactation milk production and composition traits for Holsteins using multivariate restricted maximum likelihood. J. Dairy Sci. 74: 603–610.

Farrell, H.M., R. Jimenez-Flores, G.T. Bleck, E.M. Brown, J.E. Butler, L.K. Creamer, C.L. Hicks, C.M. Hollar, K.F. Ng Kwai Hang and H.E. Swaisgood. 2004. Nomenclature of the proteins of cows' milk - Sixth revision. J. Dairy Sci. 87: 1641-1674.

Fox, P.F. 2003. Milk proteins: general and historical aspects. Pages 1-48 in Advanced dairy chemistry: proteins. Vol. 1 P.F. Fox and P.L.H. McSweeney, ed. Kluwer Academic/Plenum Publishers, New York.

Graml, R., and F. Pirchner. 2003. Effects of milk protein loci on content of their proteins. Arch. Tierz. Dummerstorf 46: 331–340.

Hayes, J.F., K.F. Ng-Kwai-Hang, and J.E. Moxley. 1984. Heritability of milk casein and genetic and phenotypic correlations with production traits. J. Dairy Sci. 67: 841-846.

Heck, J.M.L., C. Olieman, A. Schennink, H.J.F. van Valenberg, M.H.P.W. Visker, R.C.R. Meuldijk, and A.C.M. van Hooijdonk. 2008. Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis. Int. Dairy J. 18: 548-555.

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

Ikonen, T., M. Ojala, and E.-L. Syväoja. 1997. Effects of composite casein and β-lactoglobulin genotypes on renneting properties and composition of bovine milk by assuming an animal model. Agric. Food Sci. Finl. 6: 283–294.

Ikonen, T., K. Ahlfors, R. Kempe, M. Ojala, and O. Ruottinen. 1999. Genetic parameters for the milk coagulation properties and prevalence of noncoagulating milk in Finnish dairy cows. J. Dairy Sci. 82: 205–214.

Ikonen, T., S. Morri, A.-M. Tyrlsevä, O. Ruottinen, and M. Ojala. 2004. Genetic and phenotypic correlations between milk coagulation properties, milk production traits, somatic cell count, casein content, and pH of milk. J. Dairy Sci. 87: 458-467.

Khatkar M.S., P.C. Thomson, I. Tammen, and H.W. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet. Sel. & Evol. 36: 163–190.

Lund, M.S., J.Jensen, and P.H.Petersen. 1999. Estimation of genetic and phenotypic parameters for clinical mastitis, somatic cell production deviance, and protein yield in dairy cattle using gibbs sampling. J. Dairy Sci. 82: 1045-1051.

Kruglyak L. 1997. The use of a genetic map of bi-allelic markers in linkage studies. Nat. Genet. 17: 21-24.

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

Schaid D.J., J.C. Guenther, G.B. Christensen, S. Hebbring, C. Rosenow, C.A. Hilker, S.K. McDonnell, J.M. Cunningham, S.L. Slager, M.L Blute & S.N. Thibodeau. 2004. Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. Am. J. Hum. Genet. 75: 948-965.

The Bovine Genome Sequencing and Analysis Consortium, C.G. Elsik, R.L. Tellam, K.C. Worley**.** 2009. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324: 522-528.

Wedholm, A., L.B. Larsen, H. Lindmark-Månsson, A.H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making

properties of milk from individual dairy cows. J. Dairy Sci. 89: 2396-3305.

2

Genetic parameters for major milk proteins in Dutch Holstein-Friesians

G. C. B. Schopen, J. M. L. Heck, H. Bovenhuis, M. H. P. W. Visker,
H. J. F. van Valenberg, and J. A. M. van Arendonk

Published in Journal of Dairy Science (2009) 92: 1182 – 1191

Abstract

The objective of this study was to estimate genetic parameters for major milk proteins. One morning milk sample was collected from 1,940 first-parity Holstein-Friesian cows in February or March 2005. Each sample was analyzed with capillary zone electrophoresis to determine the relative concentrations of the six major milk proteins. The results show that there is considerable genetic variation in milk protein composition. The intraherd heritability for the relative protein concentrations was high and ranged from 0.25 for β-casein to 0.80 for β-lactoglobulin. The intraherd heritability for the summed whey fractions (0.71) was higher than that for the summed casein fractions (0.41). Further, there was relatively more variation in the summed whey fraction (CV was 11% and SD was 1.23) as compared to the summed casein fractions (CV was 2% and SD was 1.72). For the caseins and α-lactalbumin, the proportion of phenotypic variation explained by herd was approximately 14%. For β-lactoglobulin, the proportion of phenotypic variation explained by herd was considerably lower (5%). Eighty percent of the genetic correlations among the relative protein concentrations were between -0.38 and +0.45. The genetic correlations suggest that it is possible to change the relative proportion of caseins in milk. Strong negative genetic correlations were found for β-lactoglobulin with the summed casein fractions (-0.76), and for β-lactoglobulin with casein index (-0.98). This study suggests that there are opportunities to change the milk protein composition in the cow's milk using selective breeding.

Introduction

Bovine milk represents a unique source of bioactive components and nutrients, which include proteins. The major milk proteins are α_{S1}-casein (α_{S1}-CN), α_{S2}-casein (α_{S2}-CN), β-casein (β-CN), κ-casein (κ-CN), α-lactalbumin (α-LA), and β-lactoglobulin (β-LG). The protein composition of milk plays an important role in the profitability of the dairy industry. Specific proteins contribute to the production of specific milk products. Caseins, for example, are important for cheese yield, milk coagulation time, and curd firmness (Wedholm *et al.*, 2006), whereas β-LG is important for the heat stability of milk (Feagan, 1979). To explore the possibilities of altering milk protein composition by selective breeding, genetic parameters, such as heritability and genetic covariance, are needed. Although many studies have reported the genetic variation for protein percentages and protein yields (Hayes *et al.*, 1984; Bobe *et al.*, 1999; Ikonen *et al.*, 2004), only a few studies have estimated the magnitude of the genetic variation of milk proteins (Renner and Kosmack, 1975; Kroeker *et al.*, 1985; Ikonen *et al.*, 1997; Bobe *et al.*, 1999; Graml and Pirchner, 2003). Furthermore, these studies estimated the heritability of the major milk proteins, but no studies have reported genetic correlations among the major milk proteins. The limited number of studies is a reflection of the technological difficulties of quantifying the six major bovine milk proteins simultaneously on a large number of cows and daughters of bulls, which is a pre-requisite for estimating their genetic parameters.

In the present study, capillary zone electrophoresis (CZE) was used to separate the major milk proteins. This technique provides rapid separation of the proteins, high resolution, and is reproducible (Heck *et al.*, 2008a). Heck *et al.* (2008a) showed that the protein composition of milk varies substantially among cows at the phenotypic level. However, it is not known to what extent this variation arises from genetic factors.

The objective of this study was to estimate the heritability of milk protein composition, and to estimate the genetic and phenotypic correlations among the major milk proteins and of milk protein composition with milk production traits in a population of 1,940 Dutch Holstein-Friesian cows.

Materials and methods

Animals

As part of the Dutch Milk Genomics Initiative, information was collected on 1,940 first-parity cows, distributed over 398 commercial herds throughout the Netherlands. At least three cows were selected per herd, and each cow was at least 87.5 percent Holstein-Friesian. The cows descended from one of five proven bulls (899 cows), from one of 50 test bulls (849 cows), or from one of 15 other proven bulls (192 cows). The last group of cows ensured sampling of at least three cows per herd. The pedigree of the cows was supplied by the NRS (Arnhem, the Netherlands). The cows were milked twice daily; and each cow was between day 63 and day 282 of lactation at the time of sampling. Almost all animals have also been used in previous studies for the genetic analysis of urea (Stoop *et al.*, 2007) and milk fatty acid composition (Schennink *et al.*, 2007; Stoop *et al.*, 2008). A morning milk sample was collected from each cow during February and March 2005, which is the winter period, to be used in the analysis of the major milk proteins.

Phenotypes

Observations of the test-day morning milk yield were obtained from the NRS. True protein, fat, and lactose percentages were determined by infrared spectroscopy using a Fourier-transformed interferogram (MilkoScan FT 6000, Foss Electric, Denmark) at the milk control station laboratory (Zutphen, the Netherlands). Protein, fat, and lactose yield were calculated by multiplying the respective percentages by the observed milk yield. Morning milk yields were missing for 147 cows; therefore, only 1,793 records were analyzed for protein, fat and lactose yield.

The relative concentrations of the six major milk proteins were determined by CZE, which is a technique used to separate proteins based on differences in size and charge. Using this method, we quantified α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG. They were expressed as a percentage of the total protein fraction. Heck *et al.* (2008a) provides a detailed description of the CZE technique used in this study.

The milk protein κ-CN, as determined in our study only consisted of κ-CN-1P (non-glycosylated, mono-phosphorylated state) (Heck *et al.*, 2008a). Sum casein (Σcasein) was defined as the sum of the percentages of α_{S1}-CN, α_{S2}-CN, β-CN, and κ-CN. Sum whey (Σwhey) was calculated by adding the percentages of β-LG and α-LA. Furthermore, casein yield was calculated by multiplying Σcasein by total protein yield. The casein index was calculated as:

$$\text{casein index} = \frac{\Sigma\text{casein}}{\Sigma\text{casein} + \Sigma\text{whey}} \text{ x } 100$$

Genotypes

Blood samples of cows for DNA isolation were collected. Genotypes for the κ-CN C5309T, κ-CN A5345C and κ-CN A5365G (the latter 3 to enable genotyping of κ-CN variants A, B and E) polymorphisms had been genotyped using a SNaPshot assay (Applied Biosystems, Foster City, CA) (Schennink *et al.*, 2008; Heck *et al.*, 2008b). Genotypes for κ-CN were missing for 208 cows, because no DNA sample was available or the DNA sample could not be genotyped unambiguously. The β-CN and β-LG genotypes were determined by CZE and confirmed by genotyping two β-CN polymorphisms and one β-LG polymorphism for 849 genotyped cows by the Illumina Golden Gate assay (Illumina, San Diego, CA) (Heck *et al.*, 2008b).

Statistical analysis

To estimate the genetic parameters and variance components, ASReml was used (Gilmour *et al.*, 2002). The following animal model was used in the analyses:

$$y_{ijklmn} = \mu + b_1*\text{lactst}_i + b_2*e^{-0.05*\text{lactst}_i} + b_3*ca_j + b_4*ca_j^2 + \text{season}_k + \text{scode}_l + \text{animal}_m + \text{herd}_n + e_{ijklmn}, \quad [1]$$

where y_{ijklmn} was the observation for animal m in herd n with sire-code l, season k, calving age j, and lactation day i for the trait of interest. The overall mean of the trait was μ, lactst_i was a covariate describing the effect

of day i of lactation, ca_j was a covariate describing the effect of age at first calving in j days, $season_k$ was the fixed effect of the k^{th} class of calving season (three classes: summer [June-August 2004], autumn [September-November 2004], and winter [December 2004-February 2005]), $scode_l$ was the fixed effect of the l^{th} class of the three different sire groups, $animal_m$ was the random additive genetic effect of animal m, $herd_n$ was a random herd effect of the n^{th} herd, and e_{ijklmn} was the random residual effect. Effects of the β-CN, κ-CN and β-LG polymorphisms were estimated using the same animal model as described above and including a milk protein genotype as a fixed effect in the animal model. Ungenotyped animals were included as a separate class.

The variance-covariance structure of the additive genetic effects was Var(animal) = $A\sigma_a^2$, where A was a matrix of additive genetic relationships among individuals and σ^2_a was the additive genetic variation. The variance-covariance structure of the herd effects was Var(herd) = $I\sigma^2_{herd}$, where I was the identity matrix and σ^2_{herd} was the herd variation. Univariate analyses were used to estimate the intraherd heritability, which was defined as:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \qquad [2]$$

where σ^2_a was the additive genetic variation and σ^2_e was the residual variation.

The proportion of the total phenotypic variation due to differences among herds was defined as:

$$h_{herd} = \frac{\sigma_{herd}^2}{\sigma_{herd}^2 + \sigma_a^2 + \sigma_e^2} \qquad [3]$$

where σ^2_{herd} was the herd variation, σ^2_a was the additive genetic variation, and σ^2_e was the residual variation.

For estimating genetic and phenotypic correlations among the different milk proteins and of milk proteins with milk production traits, bivariate analyses were performed using model [1].

Results

Mean, standard deviation, and coefficient of variation

The means, SD and CV for the protein composition of milk and traits of milk production are in Table 1. The percentage of protein in the 1,940 morning milk samples averaged 3.5%. The six major milk proteins evaluated in this study made up about 86% of the total protein fraction (Table 1). The remaining 14% consisted of glycosylated and multi-phosphorylated κ-CN, bovine serum albumin (BSA), γ-caseins, proteose peptones, immunoglobulins, lactoferrin, and numerous other proteins that occur in very low concentrations. Although BSA can be well separated using CZE, it is difficult to quantify with CZE due to sticking to the capillary. The other proteins are very heterogeneous which could not be quantified with an acceptable reproducibility. The glycosylated and multi-phosphorylated κ-CN form partly co-migrates with β-CN, which leads to a less accurate estimation of the total amount of β-CN (Heck *et al.*, 2008).

From total protein, 75% was made up of the caseins (Σcasein). The main caseins were α_{S1}-CN and β-CN, which made up 34% and 27% of the total protein, respectively. Four percent of total protein fraction was comprised of κ-CN, which consisted of only κ-CN in the mono-phosphorylated form. The CV for α_{S1}-CN was 5% and for β-CN was 6%. There was little variation in Σcasein, the SD was 1.72 and the CV was 2%. The SD for κ-CN was about one-third that of the other three caseins. The major whey protein was β-LG, which made up 8% of the total protein fraction, and Σwhey was 11% of the total protein. The CV for Σwhey was 11%, nearly five times higher than that of Σcasein (2%). The SD for α-LA was about one-fourth that of β-LG. A low CV was found for the casein index (2%). We found that 90% of the cows had a casein index between 85 and 90. Milk yield averaged 13.5 kg based on a test-day morning milk sample (Table 1).

Intraherd heritability

The intraherd heritability is in Table 2. For the relative contribution of the proteins to the total milk protein, the intraherd heritability was moderate to high and ranged from 0.25 for β-CN to 0.80 for β-LG. Notably, the intraherd heritability for α_{S1}-CN (0.47) was almost twice that for β-CN; but the

intraherd heritability for α_{S2}-CN (0.73) was similar to the intraherd heritability for κ-CN (0.64). The intraherd heritability for β-LG was higher than that for α-LA (0.55).

The extent to which single milk protein polymorphisms (β-CN, κ-CN or β-LG) could explain the additive genetic variation in milk protein fractions (Table 2) was explored. Accounting for β-CN genotypes reduced the polygenic, additive genetic variance for β-CN concentration from 0.54 to 0.47. Accounting for κ-CN genotypes reduced the polygenic additive genetic variance for κ-CN concentration from 0.19 to 0.12, and accounting for β-LG genotypes reduced the polygenic additive genetic variance for β-LG from 1.14 to 0.11. Further, milk protein genotypes had a substantial effect on the estimated polygenic genetic variance for Σcasein and casein index.

For the traits of milk production, the intraherd heritability was 0.66 for protein percentage and 0.24 for protein yield. The intraherd heritability for lactose yield was similar to the intraherd heritability for milk yield.

Proportion of phenotypic variation explained by herd

The proportion of phenotypic variation explained by herd is also given in Table 2. For the caseins, the proportion of phenotypic variation was approximately 14%. For β-LG, the proportion of phenotypic variation explained by herd was 5%; but the variation of α-LA (16%) was similar to that of the caseins.

For the milk production traits, the proportion of phenotypic variation explained by herd ranged from 6% for lactose percentage to 36% for protein yield.

To compare the proportions of variation due to genetics and due to herd, the ratio of additive genetic variation and herd variation was calculated (Table 2). For protein yield and casein yield, herd variation was larger than additive genetic variation; but for the other milk proteins and milk production traits, additive genetic variation was similar or larger than herd variation.

Genetic correlations among the milk proteins

Phenotypic correlations were similar to the genetic correlations (Table 3), indicating that environmental correlations are similar to genetic correlations.

We will focus on the genetic correlations. Among the relative contributions of the major milk proteins to total milk protein, 80% of the genetic correlations ranged from -0.38 to +0.45. The genetic correlations among the four caseins were low to moderate. The strongest genetic correlations among the caseins were between α_{S1}-CN and α_{S2}-CN (-0.49), and between α_{S1}-CN and κ-CN (-0.56). The strongest genetic correlations among all milk proteins were found among Σcasein, Σwhey, and the casein index. A strong negative correlation was found for Σcasein with β-LG (-0.76) or Σwhey (-0.70), but Σcasein was strongly positively correlated with the casein index (0.77). A strong positive correlation was observed between Σwhey and β-LG (0.98), but Σwhey was strongly negatively correlated with the casein index (-1.00). The casein index was strongly negatively correlated with β-LG (-0.98).

Adjusting the data for β-LG genotypes gave similar correlations to those reported in Table 3, in most cases. The most important changes were the genetic correlation between Σcasein and Σwhey which increased from -0.70 to -0.28, the genetic correlation between Σwhey and α-LA which changed from -0.14 to 0.20, the genetic correlation between Σwhey and α_{S1}-CN which changed from -0.07 to 0.35, and the genetic correlation between casein index and α_{S1}-CN which changed from 0.10 to -0.25.

Genetic correlations among individual milk proteins and milk production traits

Table 4 has the genetic correlations between the different milk proteins and milk production traits. The protein percentage was negatively correlated with α_{S1}-CN (-0.61) and α-LA (-0.55), but positively correlated with κ-CN (0.55). For fat percentage, the genetic correlations with the major milk proteins were similar to protein percentage. Protein yield was positively correlated with α_{S1}-CN (0.29) and negatively correlated with κ-CN (-0.31). Milk yield was positively correlated with α_{S1}-CN (0.52) and negatively correlated with κ-CN (-0.52). Casein yield was positively correlated with α_{S1}-CN (0.32) and Σcasein (0.35), but negatively correlated with κ-CN (-0.29). The genetic correlations for protein percentage or fat percentage with the major milk proteins were different from the genetic correlations of lactose

Table 1 Means, SD, CV, and 5% and 95% quantiles for milk protein composition and milk production traits, measured on test-day morning milk samples from 1,940 first-lactation cows.

Trait	Mean	SD	CV(%)	5% quantile	95% quantile
Milk protein composition[1]					
α_{S1}-Casein	33.62	1.70	5	30.90	36.13
α_{S2}-Casein	10.38	1.41	14	8.03	12.59
β-Casein	27.17	1.60	6	24.51	29.70
κ-Casein[2]	4.03	0.58	14	3.10	4.98
α-Lactalbumin	2.44	0.32	13	1.94	2.95
β-Lactoglobulin	8.35	1.20	14	6.29	10.29
Σcasein[3]	75.20	1.72	2	72.46	77.76
Σwhey[4]	10.79	1.23	11	8.73	12.78
Casein index[5]	87.45	1.40	2	85.19	89.79
Casein yield[6] (kg)	0.35	0.07	20	0.24	0.47
Milk production traits					
Milk yield[7] (kg)	13.46	2.73	20	9.00	18.10
Protein (%)	3.51	0.30	9	3.04	4.01
Fat (%)	4.36	0.71	16	3.33	5.48
Lactose (%)	4.64	0.14	3	4.41	4.85
Protein yield[7] (kg)	0.47	0.09	19	0.32	0.61
Fat yield[7] (kg)	0.58	0.11	19	0.40	0.76
Lactose yield[7] (kg)	0.62	0.13	21	0.42	0.84

[1] Expressed as percentage of the total protein fraction (w/w), except casein yield

[2] Only κ-casein in the mono-phosphorylated form

[3] Σcasein = α_{S1}-casein + α_{S2}-casein + β-casein + κ-casein

[4] Σwhey = α-lactalbumin + β-lactoglobulin

[5] Casein index = Σcasein / (Σcasein + Σwhey) * 100

[6] Casein yield = Σcasein * protein yield

[7] Based on 1793 morning milk samples

percentage with the major milk proteins. The genetic correlations for lactose yield or milk yield with the major milk proteins were similar. Except for a few correlations, adjusting the data for β-LG genotypes gave similar correlations to those reported in Table 4. The most important changes in genetic correlations were observed between protein percentage and β-LG which increased from 0.07 to 0.27. Further, the genetic correlation between protein yield and β-LG decreased from -0.04 to -0.31, and between protein yield and casein index increased from 0.09 to 0.49.

Discussion

This study reports the heritability and the genetic and phenotypic correlations for the protein composition of milk. Until now, limited information on these parameters was available in the literature. In this study, we determined milk protein composition for a large number of cows using CZE.

Milk samples

In this study, only the morning milk sample for cows were analyzed to decrease the transport time from the farm to the laboratory. However, milk production data are usually analyzed by mixing the morning and evening milk sample. Using only the morning sample could have affected our results. McLaren *et al.* (1998) showed that the β-CN, α-LA and β-LG concentration of cows kept on unrestricted pasture did not significantly differ between morning and evening samples. Although, McLaren *et al.* (1998) found a significant difference in β-CN and β-LG concentration between morning and evening samples when the cows had restricted pasture intake.

Capillary zone electrophoresis

CZE has the capacity to simultaneously quantify the caseins and whey proteins. The reproducibility for CZE was reported by Heck *et al.* (2008a) and varied between 1.5% for α_{S1}-CN and 5.7% for α_{S2}-CN. These reproducibility values for the relative protein fractions were better than the repeatability values obtained in previous studies (Bobe *et al.*, 1998; Ortega

Table 2 Phenotypic variance ($_p^2$), intraherd heritability[1] (h^2), proportion of variance explained by herd[2] (h_{herd}^2), the ratio of additive genetic variation to herd variation ($_{a\ herd}^2$), and additive genetic variance without accounting for milk protein genotypes ($_a^2$), with accounting for single β-CN ($_{aBCN}^2$), -CN ($_{aKCN}^2$), or β-LG ($_{aBLG}^2$) genotypes for milk protein composition and milk production traits, measured on test-day morning milk samples from 1,940 first-lactation cows.

Trait	σ^2_p	h^2	h_{herd}	$\sigma^2_a / \sigma^2_{herd}$	σ^2_a	σ^2_{aBCN}	σ^2_{aKCN}	σ^2_{aBLG}
Milk protein composition[3]								
α_{S1}-CN	2.58	0.47	0.12	3.5	1.20	1.20	1.06	1.17
α_{S2}-CN	1.81	0.73	0.13	4.7	1.32	1.19	1.33	1.23
β-CN	2.14	0.25	0.16	1.4	0.54	0.47	0.49	0.54
κ-CN[4]	0.30	0.64	0.12	4.9	0.19	0.20	0.12	0.19
α-LA	0.09	0.55	0.16	2.8	4.80E-02	4.84E-02	4.07E-02	4.65E-02
β-LG	1.42	0.80	0.05	13.9	1.14	1.15	1.21	0.11
∑casein[5]	2.68	0.41	0.11	3.4	1.10	1.07	1.11	0.62
∑whey[6]	1.45	0.71	0.07	9.0	1.03	1.06	1.10	0.10
Casein index[7]	1.88	0.70	0.07	9.0	1.31	1.36	1.40	0.14
Casein yield[8] (kg)	3.01E-03	0.26	0.35	0.5	7.71E-04	7.05E-04	7.75E-04	7.65E-04
Milk production traits								
Milk yield (kg)	5.01	0.41	0.28	1.1	2.05	1.94	2.00	2.06
Protein (%)	7.17E-02	0.66	0.19	2.8	4.72E-02	4.71E-02	3.97E-02	4.71E-02
Fat (%)	0.47	0.50	0.08	5.8	0.24	0.23	0.24	0.24
Lactose (%)	1.95E-02	0.62	0.06	9.6	1.21E-02	1.20E-02	1.19E-02	1.21E-02
Protein yield (kg)	5.07E-03	0.24	0.36	0.4	1.19E-03	1.08E-03	1.19E-03	1.21E-03
Fat yield (kg)	9.15E-03	0.39	0.24	1.2	3.60E-03	3.61E-03	3.53E-03	3.61E-03
Lactose yield (kg)	1.11E-02	0.43	0.28	1.1	4.74E-03	4.55E-03	4.58E-03	4.76E-03

[1]SE between 0.08 and 0.12.
[2]SE between 0.02 and 0.03.
[3]Expressed as percentage of the total protein fraction (wt/wt), except for casein yield.
[4]Only κ-CN in the nonglycosylated mono-phosphorylated form.
[5]∑casein = α_{S1}-CN + α_{S2}-CN + β-CN + κ-CN.
[6]∑whey = α-LA + β-LG.
[7]Casein index = ∑casein/(∑casein + ∑whey) x 100.
[8]Casein yield = ∑casein x protein yield.

et al., 2003). Moreover, Bobe *et al.* (1998) could not separate α-LA and BSA.

Major milk proteins

In our study, α_{S1}-CN and β-CN were the major caseins, and α_{S2}-CN and κ-CN were less abundant, which is the pattern seen in most ruminant species (Bevilacqua *et al.*, 2006). The average relative protein concentration of the major milk proteins was in the range of those previously reported for cattle (Walstra and Jenness, 1984; Bobe *et al.*, 1998), with the exception of κ-CN. The mean for κ-CN (4.03) was lower than previously reported (10.7 and 16.9) (Walstra and Jenness, 1984; Bobe *et al.*, 1998). Only about 50% of the κ-CN was measured in this study; we measured κ-CN in the mono-phosphorylated form, which constitutes a major fraction of κ-CN, without the minor κ-CN fractions that occur because of different glycosylation or phosphorylation (Heck *et al.*, 2008a). We assumed that the relative concentration of κ-CN in the mono-phosphorylated form was a good indicator of the relative concentration of κ-CN as a whole. We ignored the effect of variation in κ-CN phosphorylation and glycosylation between cows when estimating the intraherd heritability and the genetic and phenotypic correlations.

Intraherd heritability

In the present study, we modeled herd as a random effect. Including herd as a fixed effect into the model did not influenced the heritability estimates for the milk protein composition. The intraherd heritability for the protein composition of milk ranged from 0.25 to 0.80 in this study and indicated that it is feasible to alter the milk protein composition using selective breeding. The heritability estimates in this study were similar or higher than those previously reported for the protein composition of milk from dairy cattle. In particular, 0.25 for β-CN in our study compared to 0.03 (Kroeker *et al.*, 1985) or 0.33-0.40 (Ikonen *et al.*, 1997), 0.55 for α-LA in our study compared to 0.27 (Renner and Kosmack, 1975) or 0.00-0.27 (Ikonen *et al.*, 1997) or 0.00 (Bobe *et al.*, 1999), and 0.73 for α_{S2}-CN in our study compared to 0.00–0.31 (Ikonen *et al.*, 1997) or 0.17 (Graml and Pirchner, 2003).

The discrepancy between our results and those reported by Kroeker *et al.* (1985) is especially remarkable. Kroeker *et al.* (1985) concluded, based on their estimates, that alteration of the detailed composition of the casein fraction would not be feasible using conventional selection methods. Their study included a data set of over 11,000 test-day records, which suggests that their heritability estimates are accurate. Heritability estimates might differ between studies for several reasons, one of them being the analytical methods used to quantify milk protein composition. We used CZE and Kroeker *et al.* (1985) used polyacrylamide gel electrophoresis combined with densitometry. Our CZE method had a superior reproducibility which will decrease the random error variance, and subsequently increase the heritability estimates in our study. Heritability estimates reported by Renner and Kosmack (1975), Ikonen *et al.* (1997), Bobe *et al.* (1999) and Graml and Pirchner (2003) were also based on analytical methods different from the ones used in the present study. The difference in heritability estimates among these studies could be from differences in breeds, in population or in allele frequencies. In addition, Ikonen *et al.* (1997) estimated heritabilities for only 174 samples from 59 Finnish Ayrshire and 155 samples from 55 Finnish Friesian. Bobe *et al.* (1999) reported heritability estimates based on 592 milk samples from 233 cows on a single farm, and therefore, the standard errors of the estimates were relatively large. Graml and Pirchner (2003) reported heritability estimates which are closer to our heritability estimates, though estimated in Fleckvieh and Braunvieh cattle for roughly 2000 cows per breed. Graml and Pirchner (2003) combined heritability estimates derived from a sire model and a daughter to dam regression for both breeds, whereas our heritabilities were derived from an animal model in which we accounted for all family relationships among animals. The polygenic additive genetic variance of the milk protein fractions decreased after adjusting for differences in known β-CN, κ-CN or β-LG polymorphisms. For α_{S1}-CN, α_{S2}-CN and β-CN, standard errors of estimates were high (0.29, 0.26 and 0.17 respectively). The κ-CN and β-LG genotypes had no effect on the polygenic additive genetic variance for α_{S1}-CN (Table 2), whereas Bobe *et al.* (1999) found that κ-CN and β-LG genotypes explained a significant part of the genetic control of α_{S1}-CN. The decrease in polygenic additive genetic variance for β-LG fraction from 1.14 to 0.11 is in

Table 3 Genetic (below diagonal) and phenotypic (above diagonal) correlations[1] among the milk proteins,[2] measured on test-day morning milk samples from 1,940 first-lactation cows.

	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN	α-LA	β-LG	∑casein	∑whey	Casein index
α_{S1}-CN		–0.50 (0.03)	–0.06 (0.03)	–0.39 (0.03)	0.20 (0.03)	–0.13 (0.04)	0.39 (0.03)	–0.08 (0.04)	0.14 (0.03)
α_{S2}-CN	–0.49 (0.12)		–0.32 (0.03)	0.15 (0.04)	–0.02 (0.04)	–0.27 (0.04)	0.10 (0.03)	–0.28 (0.03)	0.26 (0.03)
β-CN	0.01 (0.20)	–0.30 (0.16)		0.01 (0.03)	0.03 (0.03)	–0.20 (0.03)	0.57 (0.02)	–0.19 (0.03)	0.27 (0.03)
κ-CN[3]	–0.56 (0.12)	0.11 (0.14)	–0.04 (0.19)		–0.07 (0.04)	–0.15 (0.04)	0.08 (0.03)	–0.17 (0.04)	0.16 (0.04)
α-LA	0.35 (0.15)	0.12 (0.15)	0.06 (0.19)	–0.34 (0.14)		–0.08 (0.04)	0.19 (0.03)	0.17 (0.04)	–0.12 (0.04)
β-LG	–0.13 (0.15)	–0.38 (0.12)	–0.19 (0.17)	–0.10 (0.14)	–0.34 (0.14)		–0.58 (0.02)	0.97 (0.00)	–0.97 (0.00)
∑casein[4]	0.29 (0.16)	0.43 (0.14)	0.35 (0.17)	–0.08 (0.16)	0.45 (0.15)	–0.76 (0.08)		–0.53 (0.02)	0.65 (0.02)
∑whey[5]	–0.07 (0.15)	–0.38 (0.12)	–0.18 (0.18)	–0.19 (0.14)	–0.14 (0.15)	0.98 (0.01)	–0.70 (0.09)		–0.99 (0.00)
Casein index[6]	0.10 (0.15)	0.40 (0.12)	0.21 (0.17)	0.16 (0.14)	0.18 (0.15)	–0.98 (0.01)	0.77 (0.08)	–1.00 (0.00)	

[1]SE in parentheses.
[2]Expressed as percentage of the total protein fraction (wt/wt).
[3]Only κ-CN in the nonglycosylated mono-phosphorylated form.
[4]Σcasein = α_{S1}-CN + α_{S2}-CN + β-CN + κ-CN.
[5] Σwhey = α-LA + β-LG.
[6]Casein index = ∑casein/(∑casein + ∑whey) x100.

agreement with Bobe *et al.* (1999) who concluded that the genetic control of β-LG fraction is nearly complete by β-LG genotypes. Especially for some milk protein fractions (κ-CN, β-LG, Σcasein and casein index), the milk protein polymorphisms explained a considerable part of the genetic variance. However, there is still genetic variation in the rest genome to change the relative proportions of milk proteins by selective breeding. Bobe *et al.* (1999) indicated that there is no genetic variation in the rest genome to change the relative proportions of milk proteins. Moreover, Bobe *et al.* (1999) used 592 milk samples from only 233 cows on a single farm.

For both protein percentage (0.66) and protein yield (0.24), the intraherd heritability was in the range previously reported for Holstein cattle: 0.53

(Hayes *et al.*, 1984), 0.61 (Chauhan and Hayes, 1991) and 0.48 (Ikonen *et al.*, 1999) for protein percentage and 0.12 (Hayes *et al.*, 1984) and 0.25 (Chauhan and Hayes, 1991) for protein yield. The intraherd heritability for lactose yield was similar to the intraherd heritability for milk yield, which is in agreement with a previous study by Miglior *et al.* (2007).

Table 4 Genetic correlations[1] of milk protein composition traits[2] with milk production traits, measured on test-day morning milk samples from 1,940 first-lactation cows

	Percentage (%)			**Yield[3] (kg)**				
Trait	**Protein**	**Fat**	**Lactose**	**Protein**	**Fat**	**Lactose**	**Milk**	**Casein[4]**
α_{S1}-CN	–0.61	–0.60	0.22	0.29	–0.08	0.54	0.52	0.32
	(0.12)	(0.13)	(0.15)	(0.20)	(0.19)	(0.15)	(0.15)	(0.19)
α_{S2}-CN	0.20	0.17	–0.05	0.15	0.21	0.00	–0.00	0.21
	(0.14)	(0.15)	(0.14)	(0.19)	(0.16)	(0.16)	(0.16)	(0.18)
β-CN	–0.03	0.11	0.10	–0.18	–0.02	–0.12	–0.13	–0.11
	(0.19)	(0.20)	(0.18)	(0.24	(0.21)	(0.21)	(0.21)	(0.24)
κ-CN[5]	0.55	0.45	–0.21	–0.31	–0.09	–0.57	–0.52	–0.29
	(0.11)	(0.14)	(0.14)	(0.20)	(0.17)	(0.14)	(0.15)	(0.19)
α-LA	–0.55	–0.36	0.40	–0.07	–0.08	0.29	0.22	0.02
	(0.12)	(0.14)	(0.13)	(0.21)	(0.18)	(0.16)	(0.17)	(0.20)
β-LG	0.07	0.25	–0.04	–0.04	0.13	–0.07	–0.07	–0.18
	(0.14)	(0.14)	(0.14)	(0.19)	(0.16)	(0.16)	(0.16)	(0.18)
∑casein[6]	–0.07	–0.15	0.17	0.19	0.09	0.21	0.19	0.35
	(0.17)	(0.17)	(0.16)	(0.22)	(0.19)	(0.18)	(0.18)	(0.20)
∑whey[7]	–0.05	0.19	0.05	–0.07	0.11	–0.02	–0.03	–0.20
	(0.15)	(0.15)	(0.14)	(0.20)	(0.17)	(0.17)	(0.17)	(0.19)
Casein index[8]	0.04	–0.19	–0.02	0.09	–0.08	0.04	0.06	0.23
	(0.15)	(0.15)	(0.14)	(0.20)	(0.17)	(0.17)	(0.17)	(0.19)

[1]SE given in parentheses.
[2]Expressed as percentage of the total protein fraction (ww%).
[3]Based on 1,793 morning milk samples.
[4]Casein yield = casein x protein yield.
[5]Only κ-CN in the nonglycosylated mono-phosphorylated form.
[6]∑casein = α_{S1}-CN + α_{S2}-CN + β-CN + κ-CN.
[7]∑whey = α-LA + β-LG.
[8]Casein index = ∑casein/(∑casein + ∑whey) x 100.

Proportion of phenotypic variation explained by herd

The proportion of phenotypic variation of the major milk proteins explained by herd was relatively small and much lower than that of the individual milk fatty acids, which was estimated for the same population of cows and ranged between 0.16 and 0.64 (Stoop *et al.*, 2008). This suggests that herd has a smaller influence on the milk protein composition than it has on milk fat composition. A herd effect may arise from differences in housing, management, and feeding between herds, though we expect that a herd effect mainly reflects differences in feeding. Similarly, Sutton (1989) reported that the scope of changing the milk protein concentration by dietary effects is far smaller than changing the milk fat concentration. Our results support the conclusion that feeding will not have an important effect on the protein composition of milk and confirms results from Coulon *et al.* (1998), who concluded that the proportion of caseins in cow's milk depends mostly on genetic factors. In addition, Walker *et al.* (2004) reported that nutrition appears to have little effect on the major milk proteins.

Genetic correlations among the major milk proteins

The four casein genes are clustered within a 250 kb region of chromosome 6 in the following order: α_{S1}-CN, β-CN, α_{S2}-CN, and κ-CN (Threadgill and Womack, 1990; Bevilacqua *et al.*, 2006). There is homology between the promoter region of all the Ca^{2+} sensitive casein (α_{S1}-CN, α_{S2}-CN and β-CN) genes (Groenen *et al.*, 1993). Based on these findings, one might expect strong genetic correlations among caseins. Surprisingly, we found the correlations among the caseins to be relatively low, except between α_{S1}-CN and α_{S2}-CN (-0.49), and between α_{S1}-CN and κ-CN (-0.56). Bevilacqua *et al.* (2006) showed that the transcription of the casein genes occurs at the same level, but the translation efficiency of the casein messengers is different for the each of the four genes. This suggests that there is a general regulation of casein gene expression; but there is a differential post-transcriptional regulation, which might lead to low genetic correlations.

Genetic correlations of casein proteins with whey proteins were relatively low. The strongest genetic correlations were found between α_{S2}-CN and β-LG (-0.38), and between α_{S1}-CN and α-LA (0.35). These two correlations support the suggestion that the regulation of casein and whey genes will, to

some extent, involve the same co-factors, hormones, and transcription factors that are involved in the synthesis of milk proteins (Groenen and van der Poel, 1994). The genetic correlations between Σcasein and individual whey proteins were stronger than the genetic correlations between Σcasein and individual caseins. This confirms results obtained in previous studies which reported a negative relationship between β-LG and casein concentration (van den Berg *et al.*, 1992; Wedholm *et al.*, 2006).
Large amounts of casein increase cheese yield and are, therefore, profitable for the dairy industry. The strong negative genetic correlation between the relative β-LG concentration and the relative proportion of casein in milk is, therefore, of importance for the cheese production. Ng-Kwai-Hang *et al.* (1987) and Bobe *et al.* (1999) showed that genetic variants of β-LG and κ-CN affect the protein composition of milk, which may explain part of the genetic relation that is found. The B-variant of β-LG is associated with a lower β-LG concentration (Ng-Kwai-Hang *et al.*, 1987; Bobe *et al.*, 1999), with a higher casein content (van den Berg *et al.*, 1992), and with a somewhat longer renneting time and less heat stability (van den Berg *et al.*, 1992). Boland and Hill (2001) showed in a feasibility study that the selection for the B-variant of β-LG increased the milk casein and cheese yield per kilogram of milk protein. Thus, selection for the B-variant of β-LG will result in more casein in milk, which leads to more cheese production, without large influences on cheese properties.

Genetic correlations of major milk proteins with milk production traits
For the last few decades, breeding and payment schemes for the dairy industry have been focused on increasing protein yield (Boland *et al.*, 2001). The average milk protein yield in the Netherlands has more than doubled from 148 kg in 1960 to 320 kg in 2006 per lactation per cow (NRS, 2007). Selection for protein yield will have a negligible effect on the relative protein concentration of the major milk proteins because the genetic correlations are low to very low (Table 4). This result confirms results reported by Bobe *et al.* (2007), who concluded that selection for milk yield has little effect on the milk protein composition. Selection for protein percentage, however, can be expected to have a small effect on the milk protein composition by increasing the relative protein concentration of κ-CN

and decreasing the relative protein concentrations of α_{S1}-CN and α-LA. Selection for milk yield is expected to have a small effect on the relative protein concentration of the major milk proteins, which is the opposite of protein percentage, by decreasing the relative protein concentration of κ-CN and increasing the relative protein concentrations of α_{S1}-CN and α-LA. The protein α-LA is also positively correlated with lactose yield and lactose percentage. This might be a consequence of the fact that the amount of lactose in milk is influenced by the capacity of α-LA to maximize its synthesis (Walstra and Jenness, 1984).

Conclusions

The heritability for protein composition was moderate to high. Most of the genetic correlations among the major milk proteins were low. The relative β-LG concentration was strongly negatively correlated with the relative proportion of casein in milk, which is of importance for the cheese production. Our results suggest interesting possibilities to change the cow's milk protein composition using selective breeding.

Acknowledgements

This study is part of the Milk Genomics Initiative, funded by Wageningen University, NZO (Dutch Dairy Organization), CRV (cooperative cattle improvement organization), and technology foundation STW. The authors would like to thank the owners of the herds for their help in collecting the data, the Milk Control Station (Zutphen, the Netherlands) for analyzing the milk samples, and CRV (Arnhem, the Netherlands) for supplying pedigrees and milk production data.

References

Bevilacqua, C., J.C. Helbling, G. Miranda, and P. Martin. 2006. Translational efficiency of casein transcripts in the mammary tissue of lactating ruminants. Reprod. Nutr. Dev. 42: 567–578.

Bobe, G., D.C. Beitz, A.E. Freeman, and G.L. Lindberg. 1998. Separation and quantification of bovine milk proteins by reversed-phase high-performance liquid chromatography. J. Agric. Food Chem. 46: 458-463.

Bobe, G., D.C. Beitz, A.E. Freeman, and G.L. Lindberg. 1999. Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. J. Dairy Sci. 82: 2797-2804.

Bobe, G., G.L. Lindberg, A.E. Freeman, and D.C. Beitz. 2007. Short Communication: Composition of milk protein and milk fatty acids is stable for cows differing in genetic merit for milk production. J. Dairy Sci. 90: 3955-3960.

Boland, M., A. MacGibbon, and J. Hill. 2001. Designer milks for the new millennium. Livest. Prod. Sci. 72: 99-109.

Boland, M., J. Hill. 2001. Genetic selection to increase cheese yield- the Kaikoura experience. Austr. J. Dairy Tech. 56: 171-176.

Chauhan, V.P.S., and J.F. Hayes. 1991. Genetic parameters for first lactation milk production and composition traits for Holsteins using multivariate restricted maximum likelihood. J. Dairy Sci. 74: 603-610.

Coulon, J-B., C. Hurtaud, B. Remond, and R. Verite. 1998. Factors contributing to variation in the proportion of casein in cow's milk true protein: a review of recent INRA experiments. J. Dairy Res. 65: 375-387.

Feagan, J.T. 1979. Factors affecting protein composition of milk and their significance to dairy processing. Aust. J. Dairy Technol. 34: 77-81.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, S.J. Welham, and R. Thompson. 2002. Asreml user guide. Release 1.0. VSN International Ltd., Hemel Hempstead, UK.

Graml, R., and F. Pirchner. 2003. Effects of milk protein loci on content of their proteins. Arch. Tierz. Dummerstorf 46: 331-340.

Groenen, M.A.M., R.J.M. Dijkhof, A.J.M. Verstege, and J.J. van der Poel. 1993. The complete sequence of the gene encoding bovine αs2-casein. Gene 123: 187-193.

Groenen, M.A.M., and J.J. van der Poel. 1994. Regulation of expression of milk protein genes: a review. Livest. Prod. Sci. 38: 61-78.

Hayes, J.F., K.F. Ng-Kwai-Hang, and J.E. Moxley. 1984. Heritability of milk casein and genetic and phenotypic correlations with production traits. J. Dairy Sci. 67: 841-846.

Heck, J.M.L., C. Olieman, A. Schennink, H.J.F. van Valenberg, M.H.P.W. Visker, R.C.R. Meuldijk, and A.C.M. van Hooijdonk. 2008a. Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis. Int. Dairy J. 18: 548-555.

Heck, J.M.L., A. Schennink, H.J.F. van Valenberg, H. Bovenhuis, M.H.P.W. Visker, J.A.M. van Arendonk, and A.C.M. van Hooijdonk. 2008b. Effects of milk protein variants on the protein composition of bovine milk. J. Dairy Sci. 92: 1192-1202.

Ikonen, T., M. Ojala, and E-L. Syväoja. 1997. Effects of composite casein and β-lactoglobulin genotypes on renneting properties and composition of bovine milk by assuming an animal model. Agric. Food Sci. Finnl. 6: 283-294.

Ikonen, T., K. Ahlfors, R. Kempe, M. Ojala, and O. Ruottinen. 1999. Genetic parameters for the milk coagulation properties and prevalence of noncoagulating milk in Finnish dairy cows. J. Dairy Sci. 82: 205-214.

Ikonen, T., S. Morri, A.-M. Tyrisevä, O. Ruottinen, and M. Ojala. 2004. Genetic and phenotypic correlations between milk coagulation properties, milk production traits, somatic cell count, casein content, and pH of milk. J. Dairy Sci. 87: 458-467.

Kroeker, E.M., K.F. Ng-Kwai-Hang, J.F. Hayes, and J.E. Moxley. 1985. Heritabilities of relative percentages of major bovine casein and serum proteins in test-day milk samples. J. Dairy Sci. 68: 1346-1348.

McLaren, R.D., M.J. Auldist, and C.G. Prosser. 1998. Diurnal variation in the protein composition of bovine milk. Proc. N.Z. Soc. Anim. Prod. 58: 49-51.

Miglior, F., A. Sewalem, J. Jamrozik, J. Bohmanova, D.M. Lefebvre, and R.K. Moore. 2007. Genetic analyses of milk urea nitrogen and lactose and their relationships with other production traits in Canadian Holstein cattle. J. Dairy Sci. 90: 2468-2479.

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

NRS. 2007. Year Statistics 2006. NRS, Arnhem, the Netherlands.

Ortega, N., S.M. Albillos, and M.D. Busto. 2003. Application of factorial design and response surface methodology to the analysis of bovine caseins by capillary zone electrophoresis. Food Control 14: 307-315.

Renner, E., and U. Kosmack. 1975. Genetische aspekte zum eiweißgehalt und zu den eiweißfraktionen in der Milch. II. Eiweißfraktionen. Züchtungskunde 47: 441-457.

Schennink, A., W.M. Stoop, M.H.P.W. Visker, J.M.L. Heck, H. Bovenhuis, J.J. van der Poel, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2007. DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. Anim. Genet. 38: 467-473.

Schennink, A., J.M.L. Heck, H. Bovenhuis, M.H.P.W. Visker, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2008. Milk fatty acid unsaturation: genetic parameters and effects of stearoyl-CoA desaturase (SCD1) and acyl CoA:diacylglycerol acyltransferase 1 (DGAT1). J. Dairy Sci. 91: 2135-2143.

Stoop, W.M., H. Bovenhuis, and J.A.M. van Arendonk. 2007. Genetic parameters for milk urea nitrogen in relation to milk production traits. J. Dairy Sci. 90: 1981-1986.

Stoop, W.M., J.A.M. van Arendonk, J.M.L. Heck, H.J.F. van Valenberg, and H. Bovenhuis. 2008. Genetic parameters for milk fatty acids and milk production traits of Dutch Holstein-Friesians. J. Dairy Sci. 91: 385-394.

Sutton, J.D. 1989. Altering milk composition by feeding. J. Dairy Sci. 72: 2801-2814.

Threadgill, D.W., and J.E. Womack. 1990. Genomic analysis of the major milk protein genes. Nucl. Acids Res. 18: 6935-6942.

Van den Berg, G., J.T.M. Escher, P.J. de Koning, and H. Bovenhuis. 1992. Genetic polymorphism of κ-casein and β-lactoglobulin in relation to milk composition and processing properties. Neth. Milk Dairy J. 46: 145-168.

Walker, G.P., F.R. Dunshea, and P.T. Doyle. 2004. Effects of nutrition and management on the production and composition of milk fat and protein: a review. Aust. J. Agric. Res. 55: 1009-1028.

Walstra, P., and R. Jenness, ed. 1984. Protein composition of milk. Dairy Chemistry and Physics. Wiley, New York

Wedholm, A., L.B. Larsen, H. Lindmark-Månsson, A.H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. J. Dairy Sci. 89: 2396-3305.

3

Comparison of information content for microsatellites and SNPs in poultry and cattle

G. C. B. Schopen, H. Bovenhuis, M. H. P. W. Visker and
J. A. M. van Arendonk

Published as short communication in Animal Genetics (2008) 39: 451 – 453

Abstract

The objective of this study was to compare the information content of microsatellites and single nucleotide polymorphisms (SNPs) in commercial poultry populations and in cattle populations. Data was available for 12 microsatellites and 29 SNPs for one poultry chromosome, and for 34 microsatellites and 36 SNPs for one cattle chromosome. The microsatellites and SNPs were compared for their information content. Stochastic permutation was used to determine the number of SNPs needed to obtain the same average information content as a given number of microsatellites for different marker densities. By using all available microsatellites and SNPs, the 12 poultry microsatellites provided an average information content of 0.71 compared with 0.72 of the 29 poultry SNPs. The 34 cattle microsatellites provided an average information content of 0.92 compared with 0.79 of the 36 cattle SNPs. For poultry, stochastic permutation showed that the number of SNPs needed per microsatellite to obtain the same average information content increased with increasing average information content required. The number of SNPs needed per microsatellite varied between 1 and 2.3 SNPs per microsatellite. For cattle, stochastic permutation showed that the number of SNPs needed per microsatellite to obtain the same average information content fluctuated around 3. This study, therefore, indicates that 3 SNPs per microsatellite are needed to obtain the same average information content.

Introduction

Several types of molecular markers are available for researchers interested in mapping and utilisation of quantitative trait loci (QTL). Microsatellite (MS) markers are extremely valuable for linkage analysis because they are highly polymorphic, and appear frequently throughout the genome, and because techniques are available for large-scale genotyping (Kruglyak 1997). More recently single nucleotide polymorphism (SNP) markers have become available as a result of large genome sequencing projects in a number of species, e.g. in poultry (Wong *et al.* 2004). SNPs are bi-allelic, but they appear more frequently throughout the genome than MS (Vignal *et al.* 2002; Schaid *et al.* 2004), and they can be genotyped with high-throughput methods.

In humans, several studies have compared the value of MS and SNPs for genome scans to detect QTL. Kruglyak (1997) showed that 2 to 3 SNPs are needed per MS to obtain the same information content. Entropy, a measure of information content, for 10.423 SNPs was 0.75; for 3.300 SNPs was 0.65; and for 360 M was 0.57 (John *et al.* 2004). In simulated nuclear human families (two parents with n children), 1 MS with 9 equally frequent alleles had the same information rate as 4 to 5 SNPs (Lindholm *et al.* 2004).

Studies comparing the use of MS and SNPs for genome scans are not available for livestock, and it might not be possible to translate results directly from human to livestock populations due to differences in population history and family structure. Large numbers of SNPs are available or will soon become available for most livestock species, and SNPs are obvious candidates to replace MS in QTL mapping. The objective of this study, therefore, was to compare the information content of MS and SNP markers in poultry and in cattle. For this purpose, we used a two-generation design, i.e. a design where parents and their offspring are genotyped while phenotypes are collected on the offspring generation only (e.g. Van der Beek *et al.* 1995).

Materials and Methods

Information content

The value of MS and SNPs was evaluated based on the information content (IC) per centimorgan (cM) for the chromosomal region under study. The IC is defined as the variance of the probability that an offspring at a specific position inherited a given allele from its parent (Spelman *et al.* 1996). The IC quantifies how accurately the transmission of alleles from parent to offspring can be reconstructed. The IC at position k of a chromosome (IC_k) was computed as follows:

$$IC_k = 4 \cdot var(prob_k) = 4 \cdot \left[\frac{\sum_{i=1}^{N} prob_{i,k}^2 - \frac{\left[\sum_{i=1}^{N} prob_{i,k}\right]^2}{N}}{N-1} \right]$$

where $prob_{i,k}$ is the probability that at position k, offspring i has inherited allele A (alleles of the parent are arbitrarily named A and B) from the parent, var($prob_k$) is the variance of the probabilities for N offspring at position k.

The probability that the offspring inherits a given allele from its parent was calculated based on the genotypic information of marker genotypes of the parent and the offspring, flanking position k, and the recombination fraction between the flanking informative markers (Knott *et al.* 1994). Probabilities were calculated assuming the linkage map, i.e. the order of markers and the recombination fraction between markers, is known. The linkage phase in the parents was inferred based on the data, and the most likely linkage phase was assumed to be the true linkage phase (Knott *et al.* 1994).

The IC was computed at positions that were 1 cM apart for the chromosomal region under study. The average IC was computed for the chromosomal region under study by summing the IC over all cM and dividing by the length of the region (in cM).

Polymorphism information content

The polymorphic information content (PIC) was used to compute the degree of polymorphism for each MS and SNP polymorphism, and was computed as (Botstein *et al.* 1980):

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2 p_i^2 p_j^2$$

where p_i is the frequency of allele i, p_j is the frequency of allele j, and n is the number of alleles.

Data

Both MS and SNP data were available for one chromosomal region in poultry and for one chromosomal region in cattle. For poultry, data consisted of two full-sib (FS) families with a total of 96 offspring: 42 in one family, and 54 in the other family. Genotypes were available for 12 MS and 29 SNPs, which were distributed over a region from 10-102 cM of chromosome 10, based on the chicken consensus genetic linkage map (Groenen *et al.* 2000). The identification and design of the poultry MS have been described in detail by Crooijmans *et al.* (1993) and Cheng & Crittenden (1994). The poultry SNP assays were developed with the support of the USDA Agricultural Research Service (USDA, ARS) and the USDA-CSREES National Research Initiative Competitive Grants Program, and through the efforts of Hans Cheng, William Muir, Gane Wong, Martien Groenen and Huanmin Zhang due to their work on the USDA-CSREES-NRICGP proposal no. 2004-05434. Besides the 29 SNPs, information from 10 additional SNPs was available, but these SNPs were not segregating in our two poultry families, therefore, these SNPs were not included in the calculations. Genotypes were available on sires, dams, and the offspring.

For cattle, data consisted of 29 Dutch Holstein Friesian half-sib (HS) families with a total of 1599 offspring. Average number of offspring per sire was 55 and ranged from 21 to 118. Genotypes were available for 34 MS and 36 SNPs, which were distributed over a region from 68-106 cM of chromosome 18, based on the international society for animal genetics (ISAG) cattle map (Ihara *et al.* 2004). The cattle MS were chosen from the ISAG genetic map. The cattle SNPs were traced in the same region as the

microsatellites, however, no genetic of physical map of the SNPs was available. Genotypes were available on sires, and the offspring.

Permutations

To determine the number of SNPs needed to obtain the same average IC as a given number of MS for different markers densities, we used stochastic permutation. In each permutation, a predefined number of markers was randomly selected, without replacement, out of the available markers. For increasing number of MS and SNPs, from 1 to all available MS and SNPs, permutations of alternatives were performed. For each alternative we performed 1000 permutations, and for each permutation we calculated the average IC. For a given number of MS, the average IC was compared with the average IC just below and just above the average IC for the given number of SNPs. The corresponding number of SNPs required to obtain the same average IC for the given number of MS was computed by interpolation.

The marker density in poultry was lower than the marker density in cattle. To compare the IC at the same marker density, the average number of MS and SNPs per cM was computed for both species. For poultry, there were 12 MS, 29 SNPs, and a chromosomal region of 92 cM. The minimum marker density, therefore, was 1/92 = 0.01 marker per cM and the maximum marker density was 29/92 = 0.32 markers per cM.

For cattle, there were 34 MS, 36 SNPs, and a chromosomal region of 38 cM. The minimum marker density, therefore, was 1/38 = 0.03 markers per cM and the maximum marker density was 36/38 = 0.95 markers per cM.

Results

Polymorphism information content

For poultry (Figure 1A), the PIC for the 12 MS averaged 0.45, and ranged between 0.19 and 0.66. The PIC for the 29 SNPs averaged 0.30, and ranged between 0.19 and 0.38. There were no poultry markers with PIC values between 0 and 0.1, or between 0.2 and 0.3. The average PIC of the SNPs was two-third that of the MS.

For cattle (Figure 1B), the PIC for the 34 MS averaged 0.52, and ranged between 0.18 and 0.83. The PIC for the 36 SNPs averaged 0.28, and ranged between 0.06 and 0.52. The average PIC of the SNPs was about half that of the MS.
The PIC for MS and for SNPs showed more variation for cattle than for poultry.

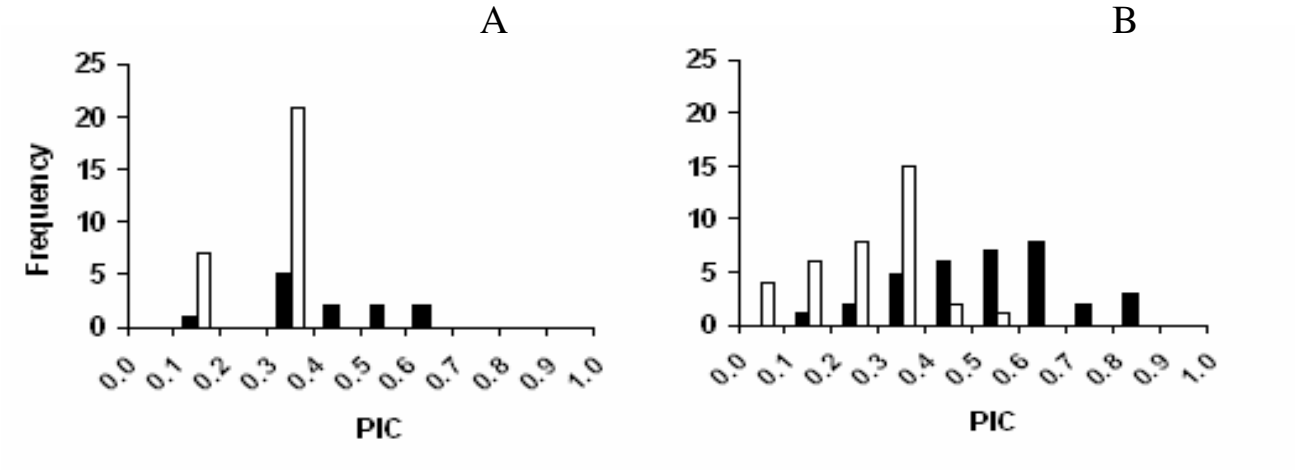

Figure 1 Histogram of the PIC for MS (■) and SNPs (□), for poultry (A) and forcattle (B).

Information content

An important factor determining the IC is the heterozygosity of the markers in the parents. For poultry, a parent was heterozygous for 7 of the 12 MS and for 13 of the 29 SNPs, on average. Genotyping was successful for all markers in all poultry parents. Genotyping in poultry was not successful for 16 of the 96 offspring per MS, and for 17 of the 96 offspring per SNP, on average, which were randomly distributed over the markers, and over the animals. Of the 29 selected SNPs, 12 were segregating in only one of the two families.
For cattle, the 29 sires were heterozygous for 19 of the 34 MS and for 11 of the 36 SNPs, on average. For 6 of the 29 sires, genotyping was not successful for 2 MS, on average, and for 5 of the 29 sires, genotyping was not successful for 3 SNPs, on average. Genotyping was also not successful for 293 of the 1599 offspring per MS and 248 of the 1599 offspring per SNP, on average. All unsuccessful genotyping were randomly distributed

over the markers, and over the animals. In contrast with poultry, there were no markers that were homozygous in all of the 29 cattle sires.
The IC for the chromosomal regions under study for poultry and cattle based on all available MS and SNPs are in Figure 2. For poultry (Figure 2A), the IC averaged 0.71 for 12 MS and 0.72 for 29 SNPs. For cattle

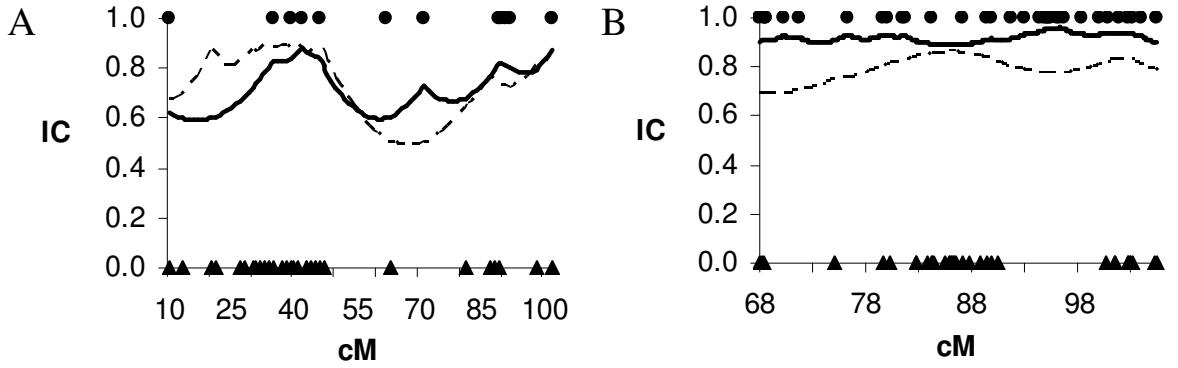

Figure 2 The IC of the MS (—) and SNPs (- - -) for poultry (A) and for cattle (B). The dots (•) at the top indicate the positions of the MS and the triangles (▲) at the bottom indicate the positions of the SNPs.

(Figure 2B), the IC averaged 0.92 for 34 MS and 0.79 for 36 SNPs. The IC showed more variation across the region under study in poultry than in cattle. In cattle, the IC for both marker types was fairly constant and high for the whole region.

Permutations

Permutations were used to determine the number of SNPs needed to obtain the same average IC as a given number of MS (Figure 3). For poultry, 1 SNP per MS was needed to obtain the same average IC of 0.21. To obtain the same average IC of 0.71, however, 2.3 SNPs per MS were needed. The number of SNPs needed per MS to obtain the same average IC was close to 1 when up to an average IC of 0.41. The number of SNPs needed per MS increased from about 1 when the average IC increased more than 0.41 to 2.3 when 12 MS were used.

For cattle, the number of SNPs needed per MS to obtain the same average IC in the studied chromosomal region fluctuated around 3. Although the number of SNPs needed per MS decreased from 3.3 to 2.7

when increasing the number of MS from 2 to 3, the number of SNPs needed per MS increased from 2.7 to 3.0 as the number of MS increased to 12, which corresponds to an average IC of 0.80.

Figure 4 shows that the average IC increased with increasing marker density. The average IC increased most at low marker densities. For poultry, for example, the average IC increased from 0.21 to 0.31 as the average number of MS per cM increased from 0.01 to 0.02, whereas, the average IC increased from 0.47 to 0.51 as the average number of MS per cM increased from 0.05 to 0.07. The average IC of the MS was equal to the average IC of the SNPs until about 0.04 markers per cM. From about 0.04 markers per cM onwards, the average IC increased more for MS than for SNPs (Figure 4A). For cattle (Figure 4B), the average IC of the MS was always greater than the average IC of the SNPs. From about 0.3 markers per cM onwards, the difference in average IC of the MS and the SNPs becomes smaller. For the same marker density, the poultry markers (Figure 4A) had greater average IC than the cattle markers (Figure 4B). For 0.1 marker per cM, for example, the poultry MS provided an average IC of 0.63, whereas the cattle MS provided an average IC of 0.58.

Discussion

In this study, we compared the IC of MS and SNP markers for using linkage in a two-generation design in poultry and in cattle.

PIC

For poultry, the PIC for the 12 MS averaged 0.45, which corresponds closely to results of Zhu *et al.* (2001), who found an average PIC of 0.46 in a commercial broiler sire line and 0.44 in a commercial broiler dam line. The PIC for the 29 poultry SNPs averaged 0.30; to our knowledge, no previous PIC values for SNPs in poultry have been reported.

For cattle, the PIC for the 34 MS averaged 0.52 and ranged between 0.18 and 0.83. This range of PIC is larger than those reported previously in cattle: 0.35 – 0.86 in Belgian Holstein Friesian (HF) (Peelman *et al.* 1998), 0.54 – 0.85 in Polish HF (Radko *et al.* 2005), and 0.37 – 0.82 in Galloway cattle (Herráez *et al.* 2005). These studies, however, used MS that were selected from a kit recommended by the international society for animal

genetics (ISAG). These selected MS were highly polymorphic and, therefore, had greater PIC values.

For cattle SNPs, the PIC for the 36 SNPs averaged 0.28. This PIC value is lower than those reported previously in cattle: 0.31 in Australian Holstein

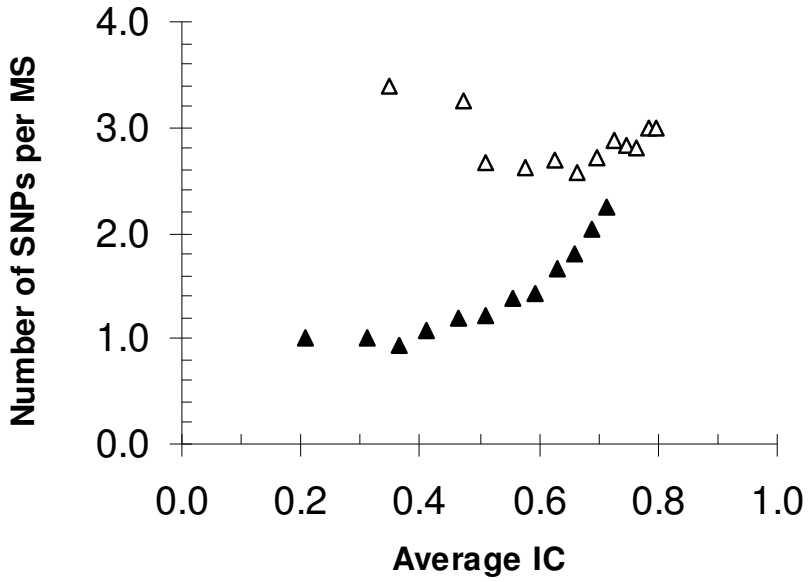

Figure 3 The number of SNPs per MS needed to obtain the same average IC for poultry (▲) and for cattle (△) with increasing number of MS from 1 (left) to 12 (right).

Friesian (Zenger *et al.* 2006), and 0.35 in Angus and Hereford cattle (Van Eenennaam *et al.* 2007). Van Eenennaam *et al.* 2007, however, used beef cattle.

In our study, there were only 2 FS poultry families compared with 29 HS cattle families. By chance, therefore, the range of PIC for the poultry markers might be lower than that for the cattle markers (Figure 1).

Differences in marker selection and populations studied contribute to differences in average PIC of markers reported in different studies. It can be concluded that the PIC for MS and for SNPs in our study are in line with reported values in the literature.

Difference between poultry and cattle

From our results (Figure 3), the number of SNPs needed per MS to obtain the same average IC was greater for cattle than for poultry. For poultry, the

number of SNPs needed per MS increased with increasing average IC required, whereas, for cattle, the number of SNPs needed per MS fluctuated around 3 independent of the average IC required. The average IC was higher in poultry than in cattle with the same marker density (Figure 4). A number of factors might have contributed to these differences: the two datasets differed in family structure (FS poultry families versus HS cattle families), the way in which the markers were selected, the distribution of the markers across the chromosomal region under study, and the PIC for the markers.

We theoretically derived the consequences of having a FS or HS family structure for the IC. For simplicity, we assumed a population in Hardy Weinberg-equilibrium and equi-frequent marker alleles. The theoretical calculations showed that the IC in a FS family was greater than in a HS family. For SNPs, the IC was 1.5 times greater in a FS family than in a HS family. For a MS with 5 equi-frequently alleles the IC was 1.2 times greater in a FS family than in a HS family. The results of the theoretical calculations showed that the family structure had an effect on the IC, and that the number of alleles has a different effect on HS and FS families.

A genetic map of MS was available for each species. The poultry MS were identified and designed as described by Crooijmans *et al.* (1993) and Cheng *et al.* (1994), whereas the cattle MS were chosen from the ISAG genetic map (Ihara *et al.* 2004). The cattle MS were chosen, based on their position and on the minor allele frequency (>0.05) as reported in the USDA MARC cattle reference families (Bishop *et al.* 1994), which is expected to result in a higher IC of the cattle MS than the poultry MS. For SNPs, a physical map was available for poultry (Wong *et al.* 2004), whereas such a map was not available for cattle. The poultry SNPs were selected at equal physical distance, although, the recombinant rate between adjacent poultry SNPs might still differ. The used selection procedure for the SNPs might have an effect on the outcome of this study. Ten additional SNPs that were monomorphic in all poultry parents, were excluded from the analysis. Had they been included, the average heterozygosity of poultry SNPs would be reduced from 45% to 33%, which is similar to the average heterozygosity of 31% of cattle SNPs, the average IC of 0.72 would have stayed equal, only for 39 SNPs in stead of 29 SNPs, and the number of SNPs needed per MS

to obtain the same average IC would have increased by a factor of 1.34. Opportunities to exclude monomorphic SNPs, however, are not possible when standard arrays are being applied for genotyping SNPs. The SNPs, however, used for standard arrays will be common SNPs which are expected to be informative in many populations. The genotyping of the selected MS and SNPs in each species was not always successful, however, the number of unsuccessful genotypes for poultry and for cattle was quite similar.

The distribution of the MS and of the SNPs over the studied chromosomal region was different for poultry, whereas the distribution of the MS and of the SNPs over the studied chromosomal region was quite similar for cattle. The poultry MS were further apart and less equally distributed than the poultry SNPs. The difference in distribution of the markers might have an effect on the increasing number of SNPs needed per MS to obtain the same average IC with increasing average IC for poultry.

The PIC for the markers is influenced by the effective population size. The effective population size will have an effect on the number of alleles present in the population. In our study the effective population size, and the number of alleles were smaller for poultry than for cattle. This might explain the smaller range of PIC for poultry than for cattle.

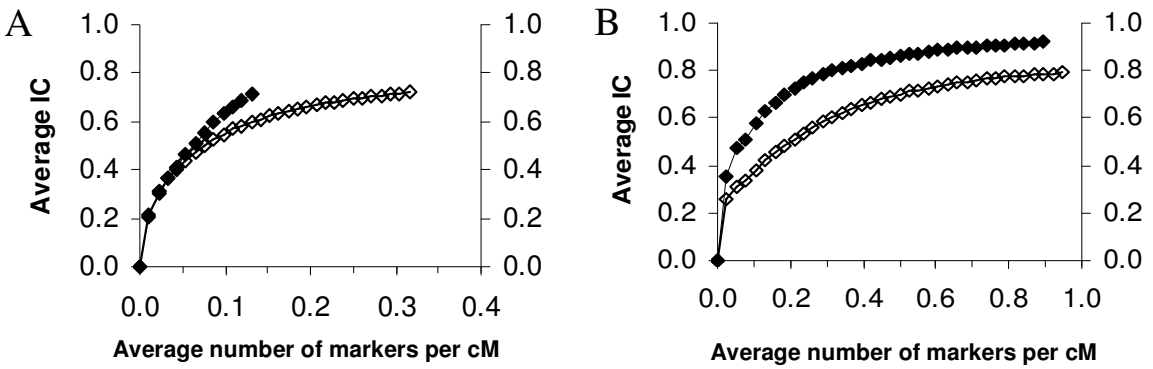

Figure 4 The average IC for increasing marker density (average number of markers per cM) for MS (♦) and SNPs (◊), for poultry (A) and for cattle(B).

IC and permutations

For poultry, the IC of the 12 MS averaged 0.71, which was similar to the IC of the 29 SNPs (0.72). For cattle, the IC of the 34 MS averaged 0.92, which

was 0.13 greater than that of the 36 SNPs (0.79). There was more variation in IC across the chromosomal region for poultry than for cattle (Figure 2), due to the distribution of the markers and the marker density. The poultry markers covered a larger chromosomal region than the cattle markers and consequently the average distances between markers were larger for poultry than for cattle. A large distance between markers decreases the IC between those markers, and, therefore, decreases the average IC of the chromosomal region.

To determine the number of SNPs needed for a given number of MS to obtain the same average IC, we used permutations based on the two data sets. The permutations of each species showed that, with equal marker density, the MS had a greater average IC than the SNPs (Figure 4) due to higher polymorphisms in MS than in SNPs. For QTL mapping, therefore, it would be better to use MS, however, there are limited MS available for performing a whole genome scan, especially for fine mapping. There is a decreasing benefit in average IC when adding more markers, and if even enough markers are available, it is difficult to get an average IC of 1 (Figure 4).

For poultry, the number of SNPs needed to obtain the same average IC as a given number of MS increased from 1 to 2.3 (Figure 3). With increasing marker density, the average IC increased more for the MS than for the SNPs (Figure 4A), which resulted in higher number of SNPs per MS needed to obtain the same average IC. We expect that the number of SNPs needed per MS to obtain the same average IC will increase further when the average required IC increase further.

For cattle, the number of SNPs needed to obtain the same average IC as a given number of MS fluctuated around 3 (Figure 3). The average IC increased more for the MS than for the SNPs until 0.3 markers per cM (Figure 4B). When using more than 0.3 markers per cM, the average IC increased equally for the MS and for the SNPs. Based on paternity exclusion, Herrázes *et al.* (2005) found that 2.65 SNPs were needed per MS for Galloway cattle. This is comparable to our result. In human studies the number of SNPs needed for 1 MS is 2.25-2.5 (Kruglyak 1997), 1.7 (Goddard & Wijsman 2002), 2-3 (John *et al.* 2004), 2.76 (Matise *et al.* 2003), and 4-5 (Lindholm *et al.* 2004). Lindholm *et al.* (2004), however,

used MS with 9 equally frequent alleles. Our poultry results show very clearly that the number of SNPs needed per MS to obtain the same average IC depends on the average IC required, which makes it difficult to compare directly between studies of different species. The general trends reported in our study are expected, however, to apply to other species.
Based on our study, we indicate that 3 SNPs are needed per MS to obtain the same amount of information about the inheritance of chromosomal segments from parents to offspring to perform a whole genome scan.

Acknowledgements

This study is part of the Milk Genomics Initiative, funded by the Dutch Dairy association (NZO), CRV (cooperative cattle improvement organization), Wageningen University, and The Technology Foundation STW. The authors thank Erik Mullaart for providing the MS/SNP data for cattle, and John Bastiaansen for his suggestions and discussion. The poultry SNP assays were developed with the support of the USDA Agricultural Research Service (USDA, ARS) and the USDA-CSREES National Research Initiative Competitive Grants Program, and through the efforts of Hans Cheng, William Muir, Gane Wong, Martien Groenen and Huanmin Zhang due to their work on the USDA-CSREES-NRICGP proposal no. 2004-05434.

References

Bishop M.D., Kappes S.M., Keele J.W., Stone R.T., Sunden S.L.F., Hawkins G.A., Toldo S.S., Fries R., Grosz M.D., Yoo J. & Beattie C.W. (1994) A genetic linkage map for cattle. Genet. 136: 619-639.

Botstein D., R.L. White, M. Skolnick, and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32: 314-331.

Cheng H.H., and L.B. Crittenden. 1994. Microsatellite markers for genetic mapping in the chicken. Poult. Sci. 73: 539-546.

Crooijmans R.P.M.A., A.J.A. van Kampen, J.J. van der Poel, and M.A.M Groenen. 1993. Highly polymorphic microsatellite markers in poultry. Anim. Genet. 24: 441-443

Goddard K.A.B. , and E.M. Wijsman. 2002. Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. Genet. Epidemiol. 22: 205-220.

Groenen M.A.M., H.H. Cheng, N. Bumstead, B.F. Benkel, W.E. Briles, T. Burke, D.W. Burt, L.B. Crittenden, J. Dodgson, J. Hillel, S. Lamont, A.P de Leon, M. Soller, H. Takahashi, and A. Vignal. 2000. A consensus linkage map of the chicken genome. Genome Res. 10: 137-147.

Herráez D.L., H. Schäfer, J. Mosner, H.-R. Fries, and M. Wink. 2005. Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population. Zeitschrift für Naturforschung 60c: 637-643.

Ihara N., A. Takasuga, K. Mizoshita, H. Takeda, M. Sugimoto, Y. Mizoguchi, T. Hirano, T. Itoh, T. Watanabe, K.M. Reed, W.M. Snelling, S.M. Kappes, C.W. Beattie G.L. Bennett, and Y. Sugimoto. 2004. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. Genome Res. 14: 1987-1998.

John S., N. Shephard, G. Liu, E. Zeggini, M. Cao, W. Chen, N. Vasavda, T. Mills, A. Barton, A. Hinks, S. Eyre, K.W. Jones, W. Ollier, A. Silman, N. Gibson, J. Worthington, and G.C. Kennedy. 2004. Whole-genome scan, in a complex disease, using 11.245 single-nucleotide polymorphisms: comparison with microsatellites. Am. J. Hum. Genet. 75: 54-65.

Knott S.A., J.-M. Elsen, and C.S. Haley. 1994. Multiple marker mapping of quantitative trait loci in half-sib populations. Proceedings of the 5th World Congress on Genetics Applied to livestock Production, Guelph, ON, CANADA XXI: 33.

Kruglyak L. 1997. The use of a genetic map of bi-allelic markers in linkage studies. Nat. Genet. 17: 21-24.

Lindholm E., S.E. Hodge, and D.A. Greenberg. 2004. Comparative informativeness for linkage of multiple SNPs and single microsatellites. Hum. Hered. 58: 164-170.

Matise T.C., R. Sachidanandam, A.G. Clark, L. Kruglyak, E. Wijsman, J. Kakol, S. Buyske, B. Chui, P. Cohen, C. de Toma, M. Ehm, S. Glanowski, C. He, J. Heil, K. Markianos, I. McMullen, M.A.

Pericak-Vance, A. Silbergleit, L. Stein M. Wagner, A.F. Wilson, J.D. Winick, E.S. Winn-Deen, C.T. Yamashiro H.M. Cann, E. Lai, and A.L. Holden. 2003. A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am. J. Hum. Genet. 73: 271-284.

Peelman L.J., F. Mortiaux, A. Van Zeveren, A. Dansercoer, G. Mommens, F. Coopman, Y. Bouquet, A. Burny, R. Renaville,and D. Portetelle. 1998. Evaluation of the genetic variability of 23 bovine microsatellite markers in four Belgian cattle breeds. Anim. Genet. 29: 161-167.

Radko A., A. Żyga, T. Ząbek, and E. Słota. 2005. Genetic variability among Polish Red, Hereford and Holstein-Friesian cattle raised in Poland based on analysis of microsatellite DNA sequences. J. Appl. Genet. 46: 89-91.

Schaid D.J., J.C. Guenther, G.B. Christensen, S. Hebbring, C. Rosenow, C.A. Hilker, S.K. McDonnell, J.M. Cunningham, S.L. Slager, M.L. Blute, and S.N. Thibodeau. 2004. Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. Am. J. Hum. Genet. 75: 948-965.

Spelman R.J., W. Coppieters, L. Karim, J.A.M. van Arendonk, and H. Bovenhuis. 1996. Quantitative trait loci analysis for five milk production traits on chromosome s*ix* in the Dutch Holstein-Friesian population. Genetics 144: 1799-1808.

Van der Beek S., J.A.M. van Arendonk, and A.F. Groen. 1995. Power of two and three-generation QTL mapping experiments in an outbred population containing full-sib or half-sib families. Theor. Appl. Genet. 91: 1115-1124.

Van Eenennaam A.L., R.L. Weaberg, D.J. Drake, M.C.T. Penedo, R.L. Quass, D.J. Garrick, and E.J. Pollak. 2007. DNA-based paternity analysis and genetic evaluation in a large commercial cattle ranch setting. J. Anim. Sci., online published jas.2007-0284v1.

Vignal A., D. Milan, M. SanCristobal, and A. Eggen. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 34: 275-305.

Wong G.K., B. Liu, J. Wang, Y. Zhang, X. Yang, Z. Zhang, Q. Meng, J. Zhou, D. Li, J. Zhang, P. Ni, S. Li, L. Ran, H. Li, J. Zhang, R. Li, S. Li, H. Zheng, W. Lin, G. Li, X. Wang, W. Zhao, J. Li, C. Ye, M. Dai, J. Ruane, Y. Zhou, Y. Li, X. He, Y. Zhang, J. Wang, X. Huang, W. Tong, J. Chen, J. Ye, C. Chen, N. Wei, G. Li, L. Dong, F. Lan, Y. Sun, Z. Zhang, Z. Yang, Y. Yu, Y. Huang, D. He, Y. Xi, D. Wei, Q. Qi, W. Li, J. Shi, M. Wang, F. Xie, J. Wang, X. Zhang, P. Wang, Y. Zhao, N. Li, N. Yang, W. Dong, S. Hu, C. Zeng, W. Zheng, B. Hao, L.W. Hillier, S.P. Yang, W.C. Warren, R.K. Wilson, M. Brandström, H. Ellegren, R.P. Crooijmans, J.J van der Poel, H. Bovenhuis, M.A.M. Groenen, I. Ovcharenko, L. Gordon, L. Stubbs, S. Lucas, T. Glavina, A. Aerts, P. Kaiser, L. Rothwell, J.R. Young, S. Rogers, B.A. Walker, A. van Hateren, J. Kaufman, N. Bumstead, S.J. Lamont, H. Zhou, P.M. Hocking, D. Morrice, D.-J. de Koning, A. Law, N. Bartley, D.W. Burt, H. Hunt, H.H. Cheng, U. Gunnarsson, P. Wahlberg, L. Andersson, E. Kindlund, M.T. Tammi, B. Andersson, C. Webber, C.P. Ponting, I.M. Overton, P.E. Boardman, H. Tang, S.J. Hubbard, S.A. Wilson, J. Yu, J. Wang, and H. Yang. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432: 717–722.

Zenger K.R., M.S. Khatkar, J.A.L Cavanagh, R.J. Hawken, and H.W. Raadsma. 2006. Genome-wide genetic diversity of Holstein Friesian cattle reveals new insight into Australian and global population variability, including impact of selection. Anim. Genet. 38: 7-14.

Zhu J.J., H.S. Lillehoj, H.H. Cheng, D. Pollock, M. Sadjadi, and M.G. Emara. 2001. Screening for highly heterozygous chickens in outbred commercial broiler lines to increase detection power for mapping quantitative trait loci. Poult. Sci 80: 6-12.

4

Whole genome scan to detect quantitative trait loci for bovine milk protein composition

G. C. B. Schopen, P. D. Koks, J. A. M. van Arendonk, H. Bovenhuis and M. H. P. W. Visker

Published in Animal Genetics (2008) 40: 524 – 537

Abstract

The objective of this study was to perform a whole genome scan to detect quantitative trait loci (QTLs) for milk protein composition in 849 Holstein-Friesian cows originating from seven sires. One morning milk sample was analyzed for the major milk proteins using capillary zone electrophoresis. A genetic map was constructed with 1,341 single nucleotide polymorphisms, covering 2,829 centimorgans (cM) and 95% of the cattle genome. The chromosomal regions most significantly related to milk protein composition (P_{genome}<0.05) were found on *Bos taurus* autosome (BTA) 6, 11 and 14. The QTL on BTA6 was found at about 80 cM, and affected α_{S1}-casein, α_{S2}-casein, β-casein and κ-casein. The QTL on BTA11 was found at 124 cM, and affected β-lactoglobulin, and the QTL on BTA14 was found at 0 cM, and affected protein percentage. The proportion of phenotypic variance explained by the QTL was 3.6% for β-casein and 7.9% for κ-casein on BTA6, 28.3% for β-lactoglobulin on BTA11, and 8.6% for protein percentage on BTA14. The QTLs affecting α_{S2}-casein on BTA6 and 17 showed a significant interaction. We investigated the extent to which the detected QTLs affecting milk protein composition could be explained by known polymorphisms in β-casein, κ-casein, β-lactoglobulin, and DGAT1 genes. Correction for these polymorphisms decreased the proportion of phenotypic variance explained by the QTLs previously found on BTA6, 11, and 14. Thus, several significant QTLs affecting milk protein composition were found, of which some QTLs could partially be explained by polymorphisms in milk protein genes.

Introduction

Bovine milk protein is important because of its nutritional value for humans. Bovine milk protein is composed primarily (±90%, w/w) of six major proteins; α_{S1}-casein (α_{S1}-CN), α_{S2}-casein (α_{S2}-CN), β-casein (β-CN), κ-casein (κ-CN), α-lactalbumin (α-LA) and β-lactoglobulin (β-LG).

Aschaffenburg & Drewry (1955) were the first to report that the whey protein β-LG is polymorphic. Later, polymorphisms for β-CN (Aschaffenburg 1961), α_S-CN (Thompson *et al.* 1962) and κ-CN (Neelin 1964) were identified. Since then, several genetic variants have been described; eight for α_{S1}-CN, four for α_{S2}-CN, 12 for β-CN, 11 for κ-CN, three for α-LA and 11 for β-LG (Farrell *et al.* 2004). The four casein genes have been located next to each other at *Bos taurus* autosome (BTA) 6 (Threadgill & Womack 1990), α-LA has been located at BTA5 (Hayes *et al.* 1993) and β-LG has been located at BTA11 (Hayes & Petit 1993).

Several studies examining the effects of milk protein polymorphisms found that polymorphisms in β-CN are associated with concentration of β-CN, polymorphisms in κ-CN are associated with concentration of κ-CN, and polymorphisms in β-LG are associated with concentration of β-LG. Furthermore, polymorphisms in β-CN, κ-CN and β-LG are associated with other milk proteins, and with total casein in milk (e.g., Ng-Kwai-Hang *et al.* 1987; Bobe *et al.* 1999 and Heck *et al.* 2009). It has also been shown that polymorphisms in milk proteins are related to manufacturing properties of milk. Cheese-making properties, e.g., milk coagulation time and curd firmness, are associated with polymorphisms in milk proteins (e.g., Marziali & Ng-Kwai-Hang 1986; Mayer *et al.* 1997; Wedholm *et al.* 2006). Moreover, the B variant of β-LG has been associated with an increase in cheese yield (Schaar *et al.* 1985; van den Berg *et al.* 1992; Wedholm *et al.* 2006).

Schopen *et al.* (2009) found that there was substantial genetic variation among cows with regard to milk protein composition. However, the genes contributing to this variation are largely unknown. Many quantitative trait loci (QTL) studies have been conducted for protein percentage and protein yield; QTLs for protein percentage on BTA6 (e.g., Spelman *et al.* 1996; Zhang *et al.* 1998; Kühn *et al.* 1999) and BTA14 (e.g., Coppieters *et al.* 1998) have been reported (for a review, see Khatkar *et al.* 2004). However, no whole genome scan for milk protein composition has been conducted thus far.

Therefore, the objective of our study was to screen the whole bovine genome to detect QTLs for milk protein composition (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA and β-LG) in Dutch Holstein-Friesians.

Materials and methods

Animals

As part of the Dutch Milk Genomics Initiative, phenotypic data on 1,912 first parity cows distributed among 398 commercial herds throughout the Netherlands were collected. The cows were descended from one of five proven sires representing five large half-sib families (873 cows), from one of 50 test sires representing 50 small half-sib families (848 cows), or from 15 other proven sires (191 cows). The last group of 191 cows ensured sampling of at least three cows per herd. The pedigrees of the cows were supplied by the CRV (Arnhem, The Netherlands). Each cow was at least 87.5% Holstein-Friesian. Further details of the animals used in this study are provided by Stoop *et al.* (2007).

Phenotypes

Cows were milked twice a day and from each cow, a morning milk sample was collected between February and March 2005, which is the winter period when cows are kept indoors. Each cow was between day 63 and 282 of their first lactation at the time of sampling.

Protein percentage was determined by infrared spectroscopy using a Fourier-transformed interferogram (MilkoScan FT 6000, Foss Electric, Denmark) at the milk control station laboratory (Zutphen, The Netherlands). Protein yield was calculated by multiplying protein percentage with the test-day morning milk yield (Table 1), which was obtained from the official milk recording organization CRV (Arnhem, The Netherlands). Morning milk yields were missing for 141 cows; therefore, only 1,771 records were analyzed for protein yield. Milk protein composition was determined by capillary zone electrophoresis (CZE), as described by Heck *et al.* (2008). Using CZE, we quantified α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA and β-LG. All six major milk protein fractions were expressed as the weight-proportion of

total protein weight (100%). The protein fraction α_{S1}-CN comprised α_{S1}-CN with eight phosphate groups (α_{S1}-CN-8P) and nine phosphate groups (α_{S1}-CN-9P). The protein fraction α_{S2}-CN comprised α_{S2}-CN with ten phosphate groups (α_{S2}-CN-10P), 11 phosphate groups (α_{S2}-CN-11P) and 12 phosphate groups (α_{S2}-CN-12P). In our study, we measured κ-CN in the mono-phosphorylated non-glycosylated form only; this constitutes the largest single fraction of κ-CN but does not include the minor fractions that

Table 1 Means and standard deviations (SD) for milk protein composition and milk production traits, measured on test-day morning milk samples from 1,912 first-lactation Holstein-Friesian cows.

Trait	Mean	SD
Milk protein composition[1]		
α_{S1}-casein	33.62	1.70
α_{S2}-casein	10.38	1.41
β-casein	27.17	1.59
κ-casein[2]	4.03	0.58
α-lactalbumin	2.44	0.32
β-lactoglobulin	8.34	1.21
Σcasein[3]	75.20	1.72
Σwhey[4]	10.79	1.24
Casein index[5]	87.46	1.40
Casein yield[6] (kg)	0.35	0.07
Milk production traits		
Protein (%)	3.51	0.30
Protein yield[7] (kg)	0.47	0.09

[1]Expressed as percentage of the total protein fraction (ww%), except for casein yield.

[2]κ-casein in the mono-phosphorylated form only.

[3]Σcasein = α_{S1}-casein + α_{S2}-casein + β-casein + κ-casein.

[4]Σwhey = α-lactalbumin + β-lactoglobulin.

[5]Casein index = Σcasein / (Σcasein + Σwhey) x 100.

[6]Casein yield = Σcasein x protein yield.

[7]Based on 1,771 morning milk samples.

occur due to differences in glycosylation or phosphorylation. We measured about 50% of κ-CN (Heck *et al.* 2008). Sum of casein (Σcasein) was defined as the sum of the percentages of α_{S1}-CN, α_{S2}-CN, β-CN, and κ-CN. Sum of whey (Σwhey) was calculated by adding the percentages of α-LA and β-LG. Casein yield was computed by multiplying Σcasein with protein yield. Furthermore, the casein index was calculated as:

$$\text{casein index} = \frac{\Sigma\text{casein}}{\Sigma\text{casein} + \Sigma\text{whey}} \text{ x } 100$$

The means and standard deviations (SD) for milk protein composition, protein percentage and protein yield are given in Table 1, and are described in more detail by Schopen *et al.* (2009).

Genotypes

DNA was isolated from blood samples of cows and semen samples of sires. For this study DNA was available for the five large paternal half-sib families (199, 188, 180, 179 and 100 cows) and for two smaller paternal half-sib families (29 and 24 cows).

Genotypes for a set of 3,072 single nucleotide polymorphisms (SNPs) for four of the five proven sires were obtained from CRV and were used as a starting point for the selection of the SNPs included in the present study. We aimed for an even distribution of the SNPs, which were evaluated based on their position on the Bosmap composite map (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20060815–freeze/) and on the bovine genome assembly (build 3.1; ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20060815–freeze/ReadMeBovine.3.1.txt). Further, the information content of the SNPs was evaluated based on heterozygosity in these four sires. Gaps between adjacent markers larger than about 10 cM were filled with markers from the dbSNP database (*N*=276; http://www.ncbi.nlm.nih.gov/projects/SNP). SNPs from the dbSNP database were filtered for availability of frequency data in Holsteins contained in the database. The resulting set of 1,536 SNPs was genotyped with the Golden Gate assay (Illumina, San Diego, CA, USA).

In addition, all 1912 animals were genotyped for a number of polymorphisms in candidate genes. The DGAT1 K232A polymorphism was determined using an allelic discrimination assay (Schennink *et al.* 2007).

The SCD1 A293V, κ-CN C5309T, κ-CN A5345C and κ-CN A5365G polymorphisms were genotyped using a SNaPshot assay (Schennink *et al.* 2008; Heck *et al.* 2009). The latter three polymorphisms enabled genotyping of the κ-CN variants A, B and E. The β-CN genotypes (A^1, A^2, A^3 and B) and β-LG genotypes (A and B) were determined by CZE and confirmed by genotyping two β-CN polymorphisms and one β-LG polymorphism for the 849 successfully genotyped cows using the Golden Gate assay (Heck *et al.* 2009). All of the genotypes for all of the SNPs were independently scored by two researchers and discrepancies were resolved.

Genetic maps

Genotypes were used to construct linkage maps for all 29 autosomes with CriMap (version 2.4; Green *et al.* 1990). The paternal half-sib design of our study did not allow us to construct linkage maps for the sex chromosomes. Markers were assigned to linkage groups based on pair-wise LOD > 3.0 scores obtained from the TWO–POINT option. Assigned markers were ordered within each linkage group using multipoint linkage analysis. Marker order was established with the BUILD option, in which markers were added in order of decreasing numbers of informative meioses. Alternative marker orders were evaluated with the FLIPS option, analyzing up to nine markers at a time. Genetic distances were estimated in the BUILD option, using the Kosambi mapping function. When CriMap positioned two or more markers at the same position, markers were placed 0.1 cM apart to ensure that all markers were included in the subsequent QTL analysis.

The BLAT sequence alignment tool (Kent 2002) was used to trace back SNP DNA sequences to reference SNP accession numbers (rs#) in the dbSNP database, and was used to determine the position of the SNPs on the genome (physical map). This was done for the Baylor College of Medicine BTAU4 version (fftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20070913–freeze/README.Btau20070913.txt).

QTL analysis

QTL analysis was performed based on seven half-sib families consisting of 849 daughters. To detect QTLs, we used a multimarker regression approach for half-sib families, as described by Knott *et al.* (1996). The

regression analysis was performed for each trait on each chromosome using the following model:

$$Y_{ij} = \mu + s_i + b_{ik}X_{ijk} + e_{ijk}$$

where Y_{ij} is the phenotype adjusted for systematic environmental effects: day of lactation, age at first calving, season of calving, and herd (Schopen *et al.* 2009) for daughter *j* nested within sire *i*, μ is the overall mean, s_i is the fixed effect of sire *i*, b_{ik} is the allele substitution effect of sire *i* at position *k*, X_{ijk} is the probability of daughter *j* inheriting gamete 1 from sire *i* at position *k* and e_{ijk} is the random residual effect for daughter *j*.

The systematic environmental effects were estimated using an animal model in ASReml (Gilmour *et al.* 2002) on all 1,912 cows (1,771 for protein yield) with phenotypes as described by Schopen *et al.* (2009). The adjusted phenotypes of the 849 successfully genotyped cows were subsequently used for the QTL analysis. The multimarker regression approach contained several steps. In brief, the paternal alleles of every offspring were identified for all informative (i.e., heterozygous) sire SNPs and the most likely linkage phase in the parent was inferred based on the frequency of recombination events in the offspring. The probability that the offspring inherited a given allele from its parent was calculated conditional on the informative marker genotypes of the parent and the offspring flanking position *k*, and the recombination fraction between the flanking informative markers based on our genetic map. A test statistic across families was calculated every 0.1 cM of the chromosome to test for the presence of a QTL. The most likely position for a QTL was the position with the highest test statistic.

Based on the probabilities that a daughter inherited gamete 1 from its sire, the information content (IC) of the SNPs, defined as the variance of the probability that an offspring at a specific position inherited a given allele from its parent, was calculated (Spelman *et al.* 1996).

In the present study, a sire was considered to be heterozygous for the QTL if the estimated allele substitution effect at the most likely position divided by its standard error had a value higher than 1.96 or lower than −1.96. These thresholds correspond to a *P* value < 0.05 when assuming a *t*-distribution. The proportion of phenotypic variance explained by a significant QTL was calculated as the difference in R^2 (proportion of variance explained by the model) between the model with and without a

QTL ($b_{ik}X_{ijk}$) at the best position using PROC GLM procedure in SAS 9.1 (2002). In the case of multiple significant QTLs for one trait, we calculated the proportion of phenotypic variance explained by multiple QTLs using a model without QTLs and a model with multiple QTLs (($b_{ik1}X_{ijk1}$) + ($b_{ik2}X_{ijk2}$) + ..). In addition, we used the model with multiple QTLs, in which we also modeled interaction effects for all pairwise QTL combinations at their most likely positions.

It has been shown that polymorphisms in β-CN, κ-CN, and β-LG are associated with milk protein composition, and that a polymorphism in DGAT1 is associated with milk protein percentage and milk protein yield. Therefore, we performed additional analyses in which we examined whether these polymorphisms could (partially) explain variation assigned to chromosomal regions identified in the genome scan. In this analysis, the phenotypes were corrected for systematic environmental effects and additionally corrected for known β-CN genotypes (A^1A^1, A^1A^2, A^2A^2, A^2A^3, A^1B and A^2B), κ-CN genotypes (AA, AB, BB, EE, AE and BE), β-LG genotypes (AA, AB and BB), and DGAT1 genotypes (AA, AK and KK) by including these genotypes as a fixed effect in the animal model. The κ-CN genotypes were missing for 180 out of the 1912 cows (Heck *et al.* 2008) and DGAT1 genotypes were missing for 153 out of the 1912 cows (Schennink *et al.* 2008). We conducted analyses in which we corrected for single genotypes and one analysis in which we corrected for all these genotypes simultaneously. Subsequently, the QTL regression was repeated for all traits and all 29 autosomes.

Significance thresholds and confidence intervals

Significance thresholds were obtained empirically by permutations as described by Churchill & Doerge (1994). A total of 10,000 permutations of the phenotypic data within sire families were performed to estimate chromosomewise *P* values. The genomewise *P* values were calculated by applying the Bonferroni correction using the formula (de Koning *et al.* 1999):

$$P_{\text{genomewise}} = 1 - (1 - P_{\text{chromosomewise}})^{1/r}$$

where *r* is the contribution of a chromosome to the total length of the genome. This was obtained as the length of a specific chromosome divided

by the length of the whole genome. In this study, a 5% genomewise significance threshold was used to indicate significant QTLs, and a 5% chromosomewise significance threshold was used to indicate suggestive QTLs. To estimate the 95% confidence intervals (CI) of the location of the QTL, 1,000 bootstraps were performed.

Results

Genotypes

Of the 1,536 SNPs genotyped with the Golden Gate assay, 121 could not be called in any of the samples or were not polymorphic in our population. The remaining 1,415 SNPs were called in over 95% of all samples, 1,390 SNPs were called in over 99% of all samples, and 1,301 SNPs were called in all samples.

In total, 946 samples were genotyped with the Golden Gate assay, of which 40 (the same four on each 96-well plate) were included to check technical repeatability. Based on the results of these 40 samples, the error rate was 0.03%. Of the remaining 906 samples, 50 were excluded: 26 because they could not be called for more than 1% of the 1,415 SNPs and 24 because they showed pedigree errors for more than 1% of the 1,415 SNPs. For the final 856 samples genotypes were called for over 99% of all 1,415 SNPs, of which 500 samples were called for all 1415 SNPs. Out of the final 856 samples, the number of missing genotypes was 5 for DGAT1 K232A, 22 for SCD1 A293V, 2 for κ-CN C5309T, 3 for κ-CN A5345C and 4 for κ-CN A5365G. The latter three polymorphisms enabled genotyping of the κ-CN variants A, B and E.

The final dataset contained genotypes for 1,420 SNPs from 856 animals (cows and sires). Samples represented 193, 179, 170, 166, and 91 cows from the five large half-sib families, 29 and 21 cows from the two small half-sib families, and the seven sires. Average heterozygosity of the seven sires for the 1,420 SNPs was 50%. Only 1.5% of the SNPs showed a minor allele frequency of less than 0.1%, which shows that the SNP selection process worked well. The average minor allele frequency was 0.34.

Genetic maps

Pair-wise analysis to assign markers to chromosomes resulted in the exclusion of 41 markers that did not meet the LOD > 3.0 score. Another 38 markers were removed because these were typed in duplo and yielded identical genotypes. The remaining 1,341 markers were ordered within the 29 autosomes and covered 2,829 cM (supporting information available online at Animal Genetics 40: 524-537). The length of the linkage groups ranged from 44 cM (BTA27) to 145 cM (BTA1), and the average length of marker intervals varied between 2.6 cM (BTA27) and 5.8 cM (BTA19). On average, each linkage group contained one marker interval larger than 10 cM. This demonstrates that our attempts to fill these gaps with markers from the dbSNP database failed in a number of cases. Many of these markers ended up at positions surrounding the gaps rather than within the gaps.

The correlation between marker order in our genetic map and the bovine physical map (BTAU4) ranged from 0.99 to 1.00 for all chromosomes. The genetic map covered 95% of the expected length of the bovine genome (2410/2545 Mbp assigned to linear scaffolds).

The IC across all 29 autosomes averaged 0.83 per cM across half-sib families with a minimum value of 0.54 and a maximum value of 0.97 (supporting information available online at Animal Genetics 40: 524-537). In small chromosomal regions on chromosomes 6, 9, 14, and 19, the IC was lower than 0.60.

Detected QTLs

In total, ten distinct significant chromosomal regions on BTA1, 3, 5, 6, 9, 10, 11, 14, 15 and 17 were found (Table 2). Some of these chromosomal regions affected multiple traits. Ten traits out of the 12 traits analyzed showed significant evidence for the presence of one or more QTLs affecting milk protein composition ($P_{genome} < 0.05$). No significant QTLs affecting protein yield and casein yield were found. Most of the significant chromosomal regions affecting multiple traits were found on BTA6, 11 and 14. Figure 1A–1C shows the graphs of the test statistics for the significant traits on BTA6, 11, and 14.

On BTA6 at a position of about 80 cM, we found significant evidence for a QTL affecting α_{S1}-CN, α_{S2}-CN, β-CN, and κ-CN fractions and protein percentage (Table 2). The number of sires which were identified as being heterozygous for the QTL on BTA6 affecting α_{S1}-CN was four, affecting α_{S2}-CN was three, affecting β-CN was two, and affecting κ-CN was four (Table 3). For four of the five large half-sib families, the QTL affecting α_{S1}-CN on BTA6 had a positive allele substitution effect, whereas the same QTL affecting α_{S2}-CN, β-CN and κ-CN had a negative allele substitution effect. For half-sib family 1, the allele substitution effects of the QTL affecting α_{S1}-CN and β-CN were the opposite compared to the other four large half-sib families (Table 3). The proportion of phenotypic variance explained by the QTL on BTA6 was 6.8% for α_{S1}-CN, 6.7% for α_{S2}-CN, 3.6% for β-CN, 7.9% for κ-CN, and 3.6% for protein percentage (Table 4).

On BTA11, at position 124 cM, we found significant evidence for a QTL affecting β-LG fraction, Σcasein, Σwhey, and casein index (Table 2). The QTL affecting β-LG fraction, Σcasein, Σwhey, and casein index on BTA11 significantly segregated within large half-sib family 1, 3, and 5 (Table 3). The QTL affecting β-LG and Σwhey on BTA11 had a negative allele substitution effect for half-sib family 1 and 5, and a positive allele substitution effect for half-sib family 3 and 6, which is the opposite of the allele substitution effects for casein index (Table 3). The proportions of phenotypic variance explained by the QTL on BTA11 were very high: 28.3% for β-LG, 10.4% for Σcasein, 25.0% for Σwhey, and 25.3% for casein index (Table 4).

On BTA14 at position 0 cM, we found significant evidence for a QTL affecting protein percentage (Table 2). The QTL affecting protein percentage on BTA14 significantly segregated within large half-sib family 1, 2, 3, and 4 (Table 3). The QTL affecting protein percentage on BTA14 had a negative allele substitution effect for half-sib family 1, 2, 3 and 4, and a positive allele substitution effect for half-sib family 6 (Table 3). The proportion of phenotypic variance explained by the QTL affecting protein percentage on BTA14 was 8.6% (Table 4).

In addition to these three chromosomal regions affecting milk protein composition, we also found significant evidence for QTLs affecting α_{S1}-CN on BTA9, α_{S2}-CN on BTA1, 10, and 17, β-CN on BTA3, α-LA on BTA5, and

protein percentage on BTA15 (Table 2). On BTA5, the best position of the QTL affecting α-LA fraction was 41 cM, which is close to the position of the α-LA gene on BTA5.

In addition to the ten significant chromosomal regions, we found 11 chromosomal regions with suggestive evidence for QTLs affecting milk protein composition ($P_{chromosomewise}$<0.05) (Table 2). We detected suggestive or significant QTLs affecting milk protein composition on most of the chromosomes, except for BTA2, 4, 8, 18, 20, 21, 22 and 29. For each of the six major milk proteins, we found more than one chromosomal region with significant or suggestive evidence for a QTL; there were five regions affecting the α_{S1}-CN fraction, seven regions affecting α_{S2}-CN fraction, two regions affecting the β-CN fraction, three regions affecting the κ-CN fraction, five regions affecting the α-LA fraction, and three regions affecting the β-LG fraction (Table 2).

Total variance explained by detected QTL

The proportion of phenotypic variance explained in a multiple QTL analysis for α_{S1}-CN fraction on BTA6 and 9 (9.1%) was very similar to the sum of the single QTL analyses (9.6%) (Table 5). Also for β-CN fraction on BTA3 and 6, the proportion of phenotypic variance explained in a multiple QTL analysis (6.4%) was similar to the sum of the single QTL analyses (7.2%) (Table 5). The proportion of phenotypic variance explained in a multiple QTL analysis for α_{S2}-CN fraction on BTA1, 6, 10, and 17 was 14.4%, which is 2.0% lower than the sum of the single QTL analyses (16.4%) (Table 5). Analyses revealed that this difference of 2.0% can be explained by the significant interaction (P = 0.04) between the QTLs affecting α_{S2}-CN on BTA6 and 17. This suggests a negative covariance between these two QTLs.

Accounting for known polymorphisms in β-CN, κ-CN, β-LG and DGAT1 genes

We investigated whether polymorphisms in the β-CN, κ-CN, β-LG, and DGAT1 genes could (partially) explain the detected QTLs affecting milk protein composition. Figure 2A–2C shows the profiles of the test statistics for the β-CN and κ-CN fractions on BTA6, and for the β-LG fraction on

BTA11 without and with corrections for known β-CN, κ-CN, or β-LG genotypes.

Both β-CN and κ-CN genes are located on BTA6, the β-LG gene is located on BTA11, and the DGAT1 gene is located on BTA14. Therefore, we focused on the consequences of the corrections for the known β-CN, κ-CN, β-LG, or DGAT1 genotypes on BTA6, 11, and 14.

On BTA6, the correction for known β-CN genotypes resulted in elimination of the previously significant QTL affecting β-CN fraction (Figure 2A). Correction for known β-CN genotypes decreased the proportion of phenotypic variance explained by the QTL affecting the β-CN fraction on BTA6 from 3.6% to 1.8% (Table 4). A previously suggestive QTL affecting Σwhey on BTA6 became significant after correcting for β-CN genotypes. In addition, correction for known β-CN genotypes resulted in detection of one new significant QTL affecting β-CN on BTA7, a previously suggestive QTL became significant (Σcasein on BTA1), and five previously significant QTLs became suggestive (α_{S1}-CN on BTA9, α_{S2}-CN on BTA10 and 17, β-CN on BTA3 and α-LA on BTA5). Correction for known κ-CN genotypes on BTA6 resulted in the remaining of the previously significant QTL affecting κ-CN on BTA6 (Figure 2B). Correction for known κ-CN genotypes decreased the proportion of phenotypic variation explained by the QTL affecting the κ-CN fraction on BTA6 from 7.9% to 3.2% and also decreased the proportion of phenotypic variance explained by the QTL affecting the α_{S1}-CN and α_{S2}-CN fraction (Table 4). With correction for known κ-CN genotypes, one previously suggestive QTL became significant (Σcasein on BTA1), and three previously significant QTLs became suggestive (α_{S1}-CN on BTA9, α_{S2}-CN on BTA10, and α-LA on BTA5).

On BTA11, correction for known β-LG genotypes resulted in remaining of the significant QTL affecting the β-LG fraction on BTA11 (Figure 2C). Correction for known β-LG genotypes enormously decreased the proportion of phenotypic variance explained by the QTL affecting the β-LG fraction from 28.3% to 4.5%, and consequently also enormously decreased the proportion of phenotypic variance explained by the QTL affecting Σwhey and casein index (Table 4). In addition, correction for known β-LG genotypes resulted in the elimination of one previously significant QTL (α-LA on BTA5, four previously suggestive QTLs became significant (β-LG,

Σwhey and casein index on BTA6, and α_{S2}-CN on BTA14), and one previously significant QTL became suggestive (β-CN on BTA3).
On BTA14, correction for known DGAT1 genotypes resulted in elimination of all of the significant and suggestive QTLs previously found on that chromosome. The correction for known DGAT1 genotypes decreased the proportion of phenotypic variance explained by the QTL affecting protein percentage from 8.6% to 1.6% (Table 4). In addition, correction for known DGAT1 genotypes resulted in detection of three new suggestive QTLs (α_{S2}-CN on BTA9, α_{S2}-CN on BTA27, and α-LA on BTA17), and one previously significant became suggestive (protein percentage on BTA6).

Discussion

This study reports on QTLs affecting the milk protein fractions (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA and β-LG) of dairy cattle. This is, to our knowledge, the first time that the results of a genome wide scan to detect QTLs for casein and whey composition have been reported.

Genetic maps

The genetic map that we constructed to enable the genome scan described in this paper comprised 1,341 markers and covered 2,829 cM. This length corresponds with previously published bovine genetic maps based on comparable numbers of markers (Kappes *et al.* 1997; Ihara *et al.* 2004; Snelling *et al.* 2005). However, comparisons with other genetic maps revealed that the length of the genetic maps of individual chromosomes was relatively short for BTA12 and 27. This is supported by the relatively low percentage of coverage of the bovine physical map for BTA12 (87%) and 27 (74%). Thus, the resolution of our genome scan may have been somewhat low at both ends of BTA12 and 27. Further comparisons with the bovine physical map also showed relatively low percentages of coverage of the bovine physical map for our genetic maps for BTA15 (87%), 19 (88%), and 25 (88%). The marker order in our genetic map is in good accordance with the marker order of the bovine physical map. The few differences may be due either to undetected poor marker quality in our study, or to imperfections in the physical map. The high average heterozygosity (50%) and consequently high average information content (0.83) in this study

Table 2 Significant ($P_{genomewise}$ < 0.05) and suggestive ($P_{chromosomewise}$ < 0.05) QTL per chromosome (BTA: Bos taurus autosome) for the six milk proteins, Σcasein[1], Σwhey[2], casein index[3], casein yield[4], protein yield and protein percentage with their F-value, best QTL location (cM), 95% confidence interval (CI) for location QTL and genome-wise P-value before (P_{genome}) and after correction for β-casein, κ-casein, β-lactoglobulin and DGAT1 genotypes simultaneously (P_{all}). A genome-wise P-value indicates that the QTL is at least chromosome-wise significant.

BTA	Trait	F-value	Location QTL (cM)	CI for location QTL	P_{genome}	P_{all}
1	α_{S2}-casein	4.63	129.1	58.8–141.1	0.012	0.012
1	Σcasein	4.42	57.2	21.6–133.6	0.070	0.432
1	κ-casein	2.11	75.2	0.0–144.7	–	0.523
3	β-casein	4.49	48.2	13.5–125.9	0.038	0.220
3	β-lactoglobulin	2.70	11.0	0.0–125.9	–	0.446
3	Casein index	2.52	11.0	0.0–125.9	–	0.584
5	α-lactalbumin	4.46	40.7	20.2–123.4	0.043	0.034
5	Protein yield	3.56	73.6	16.1–114.1	0.455	0.278
5	Casein yield	3.59	87.0	14.2–112.8	0.476	0.408
6	α_{S1}-casein	10.37	81.8	76.0–99.7	0.000	0.041
6	α_{S2}-casein	9.27	81.5	74.9–97.5	0.000	0.047
6	β-casein	4.66	98.5	50.6–122.3	0.023	–
6	κ-casein	12.89	80.3	77.1–89.5	0.000	0.321
6	α-lactalbumin	3.93	71.0	29.9–80.2	0.122	–
6	β-lactoglobulin	3.24	80.5	9.9–111.1	0.584	–
6	Σwhey	4.13	80.5	10.0–91.6	0.116	–
6	Casein index	3.65	80.5	9.9–96.9	0.282	–
6	Protein (%)	4.20	79.5	39.9–83.6	0.041	–
7	β-casein	2.98	0.9	0.0–114.3	–	0.038
7	Casein index	3.18	28.2	0.0–125.4	0.642	–
9	α_{S1}-casein	4.40	71.8	21.8–103.5	0.027	0.168
9	α_{S2}-casein	2.99	0.0	0.0–90.7	–	0.626
10	α_{S1}-casein	3.72	70.1	3.4–113.3	0.153	0.168
10	α_{S2}-casein	4.41	60.3	31.6–113.3	0.027	0.051
11	α_{S2}-casein	3.01	123.6	20.4–123.6	0.590	–
11	β-lactoglobulin	57.61	123.6	123.6–123.6	0.000	0.005
11	Σcasein	15.18	123.6	122.6–123.6	0.000	0.000
11	Σwhey	48.4	123.6	123.6–123.6	0.000	0.007
11	Casein index	48.8	123.6	123.6–123.6	0.000	0.011
12	α_{S2}-casein	2.81	44.9	0.0–78.0	0.843	–
13	α-lactalbumin	2.88	7.8	0.1–89.0	0.734	–
13	Σcasein	3.25	77.5	11.9–107.8	0.674	–
13	Protein (%)	3.65	62.2	11.6–78.3	0.168	0.617
14	α_{S1}-casein	3.57	14.5	0.0–75.9	0.216	–
14	α_{S2}-casein	4.16	0.0	0.0–64.9	0.051	–
14	Protein (%)	11.90	0.0	0.0–0.0	0.000	–

Table 2 Continued

BTA	Trait	F-value	Location QTL (cM)	CI for location QTL	P_{genome}	P_{all}
15	Protein (%)	4.66	44.4	36.0–60.7	0.018	0.000
16	α-lactalbumin	3.03	105.4	1.2–105.4	0.624	–
16	Protein (%)	3.46	29.6	1.2–104.0	0.296	0.541
17	α_{S2}-casein	4.45	37.8	0.0–97.9	0.017	0.157
17	κ-casein	2.45	83.4	0.0–97.9	–	0.447
17	Protein (%)	3.02	98.0	0.0–98.0	0.538	0.704
19	Casein yield	3.50	103.9	0.0–103.9	0.491	0.262
19	Protein yield	3.63	103.9	0.0–103.9	0.325	0.132
23	β-casein	2.62	14.0	0.0–75.9	–	0.575
23	κ-casein	3.00	40.3	1.7–56.4	0.738	–
23	α-lactalbumin	3.8	35.7	18.1–54.2	0.158	0.307
23	Σwhey	2.97	15.3	0.0–72.9	0.796	–
23	Protein (%)	2.95	43.0	9.5–72.9	0.750	–
24	κ-casein	2.82	9.8	0.2–68.3	0.861	0.794
24	β-lactoglobulin	3.35	0.0	0.0–59.4	0.521	0.465
24	Σcasein	3.26	20.2	0.0–43.4	0.678	–
25	Protein yield	3.08	23.6	0.0–64.2	0.838	0.850
26	α-lactalbumin	2.38	19.9	2.5–67.6	–	0.536
26	Protein (%)	3.42	28.7	5.8–64.3	0.282	0.807
27	α_{S1}-casein	2.70	44.3	0.0–44.3	0.890	0.236
27	α_{S2}-casein	2.52	31.6	2.2–44.3	–	0.226
27	β-casein	2.38	0.0	0.0–44.3	–	0.948
27	κ-casein	2.00	44.3	0.0–44.3	–	0.741
28	Casein yield	3.69	20.3	11.1–65.7	0.402	0.440
28	Protein yield	3.69	18.9	10.6–65.9	0.395	0.449
29	β-lactoglobulin	2.49	8.4	0.0–65.7	–	0.719
29	Σwhey	2.23	8.4	0.0–65.7	–	0.831
29	Casein index	2.42	8.4	0.0–65.7	–	0.849

– = $P_{chromosomewise} > 0.05$.

Genome-wise significant QTL are in bold.

[1]Σcasein = α_{S1}-casein + α_{S2}-casein + β-casein + κ-casein.

[2]Σwhey = α-lactalbumin + β-lactoglobulin.

[3]Casein index = Σcasein/(Σcasein + Σwhey) × 100.

[4]Casein yield = Σcasein × protein yield.

were most likely due to our knowledge of the heterozygosity of four of the five proven sires that we used in the selection of our SNPs.

Power of design used in present study

For a heritability of 25%, QTLs explaining 5% – 7.5% of the phenotypic variance can be detected with a power higher than 0.80 for a daughter design of 5 large half-sib families consisting of 200 daughters each. The two small families in our study will not contribute much to this power. Schopen *et al.* (2009) showed that the heritabilities for the six major milk proteins ranged from 25% for β-CN to 80% for β-LG. The proportion of phenotypic variance explained by our significant QTLs ranged from 2.7% to 28.3% (Table 4). The power calculations suggested that we could not detect QTLs explaining a small fraction (< 3%) of the total phenotypic variance. However, the detected QTL for α_{S2}-CN on BTA10 explained only 2.7% of the phenotypic variation, and indicates that our design to detect QTLs affecting milk protein composition was adequate.

To examine whether the β-CN, κ-CN, β-LG and DGAT1 genotypes partly explain the variance of the significant QTLs, we included the four genotypes simultaneously as a fixed effect in the animal model. This reduces the residual variation in our dataset. Decreasing residual variance is expected to result in a higher power of detection of other (smaller) QTLs (de Koning *et al.* 2001). In our study, correction for β-CN known genotypes resulted in one new significant QTL affecting β-CN on BTA7, which previously was not even suggestive. Correction for known DGAT1 genotypes resulted in three new suggestive QTLs affecting α_{S2}-CN on BTA9 and 27, and α-LA on BTA17. However, for some proteins, the test statistic decreased.

Protein percentage

Significant evidence for the presence of a QTL affecting protein percentage was detected on BTA6, 14, and 15. On BTA6 and 14, the significant QTL affecting protein percentage is in agreement with previously reported QTLs affecting protein percentage on BTA6 (e.g., Spelman *et al.* 1996; Zhang *et al.* 1998; Kühn *et al.* 1999; Khatkar *et al.* 2004) and on BTA14 (Coppieters *et al.* 1998). The chromosomal region affecting protein percentage on

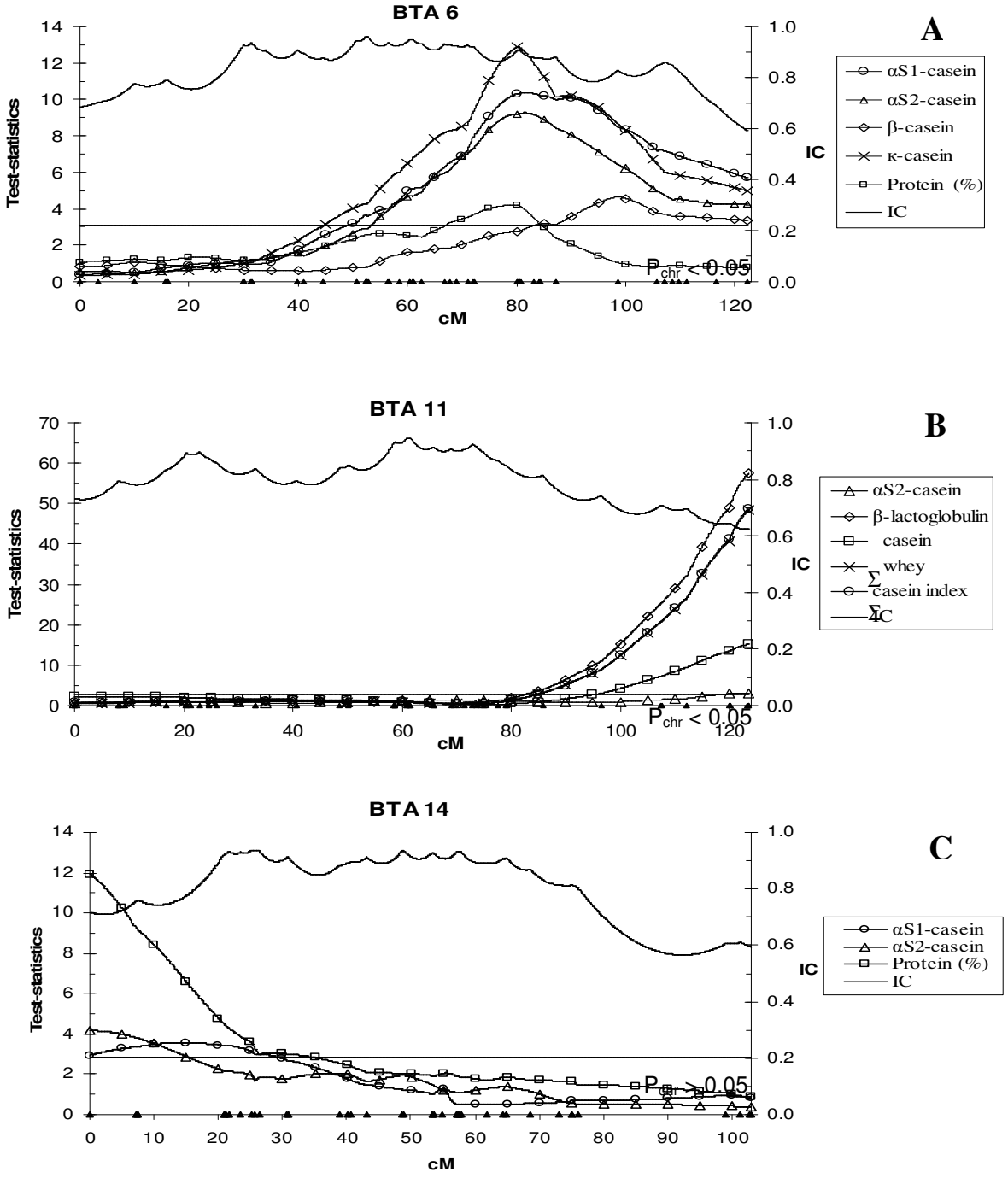

Figure 1 Test statistics and information content (IC) for each cM of the genetic map for *Bos taurus* autosome (BTA) 6 (A), 11 (B), and 14 (C) from the across-family analysis for the traits with $P_{chromosomewise} < 0.05$. The triangles on the x-axis indicate the position of the SNPs. ΣCasein = α_{S1}-casein + α_{S2}-casein + β-casein + κ-casein, Σwhey = α-lactalbumin + β-lactoglobulin, and casein index = Σcasein / (Σcasein + Σwhey) x 100.

BTA15 has not been reported before, while Khatkar *et al.* (2004) reported strong evidence of a QTL affecting protein percentage on BTA3, and 20. Suggestive evidence for the presence of a QTL affecting protein percentage was detected on BTA13, 16, 17, 23, and 26. The chromosomal region on BTA23 was detected previously by Olsen *et al.* (2002), whereas the other four chromosomal regions that were found to affect protein percentage were not detected in previously performed genome scans for protein percentage (e.g. Ashwell *et al.* 2001; Mosig *et al.* 2001; Olsen *et al.* 2002; Schrooten *et al.* 2004).

Six major milk proteins

There was one chromosomal region on BTA6 that was significantly related to the four casein fractions and one chromosomal region on BTA11 that was significantly related to the β-LG fraction, Σcasein, Σwhey, and casein index.

On BTA6, the position of the significant QTLs affecting the α_{S1}-CN, α_{S2}-CN, and κ-CN fractions was 80cM; and for the β-CN fraction, the position was 99 cM (Table 2). The CIs suggest that the same locus may underlie the QTLs for all four casein fractions. The CI of the QTL includes the location of the casein locus on BTA6. Correction of the data for known β-CN genotypes decreased the proportion of phenotypic variance explained by the detected QTL on BTA6 affecting the β-CN fraction by 50% (Table 4). This suggests that the genotypes for β-CN are associated with the significant QTL affecting the β-CN fraction on BTA6. Correction of the data for known κ-CN genotypes decreased the proportion of phenotypic variance explained by the detected QTL on BTA6 affecting the κ-CN fraction by 59% (Table 4). This suggests that the genotypes for κ-CN are associated with the significant QTL affecting the κ-CN fractions on BTA6. However, after correction for known β-CN or κ-CN genotypes, the QTLs affecting α_{S1}-CN and α_{S2}-CN remained significant and there was still a significant QTL detected for κ-CN after correction for known κ-CN genotypes. Therefore, in this study, we found evidence for the presence of other polymorphisms, besides the β-CN A^1, A^2, $A^3$and B, and the κ-CN A, B and E polymorphisms, in the chromosomal region on BTA6 that have an effect on

Table 3 Residual phenotypic standard deviation (σ_p), and significant allele substitution effects with standard errors within each sire for the significant QTL ($P_{genomewise} < 0.05$), expressed in phenotypic standard deviation, after adjusting for systematic environmental effects (BTA: Bos taurus autosome).

BTA	Location (cM)	Trait	σ_p	Sire1 (n = 193)	Sire2 (n = 179)	Sire3 (n = 170)	Sire4 (n = 166)	Sire5 (n = 91)	Sire6 (n = 29)	Sire7 (n = 21)
1	129.1	α_{S2}-casein	1.15	$-0.32_{0.14}$	–	$0.48_{0.15}$	$-0.29_{0.16}$	$0.48_{0.21}$	–	$-1.22_{0.47}$
3	48.2	β-casein	1.31	$-0.41_{0.15}$	–	$0.41_{0.15}$	–	$0.58_{0.21}$	$0.86_{0.37}$	–
5	40.7	α-lactalbumin	0.25	$0.51_{0.16}$	–	–	$0.55_{0.16}$	–	–	–
6	81.8	α_{S1}-casein	1.42	$-0.43_{0.13}$	$0.49_{0.13}$	$0.68_{0.14}$	$0.69_{0.14}$	–	–	–
6	81.5	α_{S2}-casein	1.15	–	$-0.70_{0.14}$	–	$-0.73_{0.15}$	$-0.49_{0.22}$	–	–
6	98.5	β-casein	1.31	$0.66_{0.15}$	–	$-0.48_{0.17}$	–	–	–	–
6	80.3	κ-casein	0.51	$-0.86_{0.14}$	–	$-0.45_{0.14}$	$-0.72_{0.14}$	–	$0.80_{0.39}$	–
6	79.5	Protein (%)	0.23	$-0.69_{0.13}$	–	–	–	–	–	–
9	71.8	α_{S1}-casein	1.42	–	$0.28_{0.14}$	–	$-0.37_{0.14}$	$-0.66_{0.19}$	–	–
10	60.3	α_{S2}-casein	1.15	–	$-0.40_{0.15}$	–	$0.55_{0.15}$	$0.49_{0.21}$	–	–
11	123.6	β-lactoglobulin	1.11	$-1.71_{0.13}$	–	$1.38_{0.13}$	–	$-1.50_{0.18}$	$1.74_{0.32}$	–
11	123.6	Σcasein[1]	1.44	$1.13_{0.15}$	–	$-0.74_{0.16}$	–	$0.95_{0.21}$	–	–
11	123.6	Σwhey[2]	1.13	$-1.66_{0.12}$	–	$1.27_{0.13}$	–	$-1.36_{0.19}$	$1.54_{0.34}$	–
11	123.6	Casein index[3]	1.29	$1.69_{0.12}$	–	$-1.27_{0.14}$	–	$1.38_{0.18}$	–	$-1.50_{0.33}$
14	0.0	Protein (%)	0.23	$-0.52_{0.17}$	$-0.60_{0.17}$	$-0.82_{0.17}$	$-0.86_{0.17}$	–	$1.21_{0.56}$	–
15	44.4	Protein (%)	0.23	$0.47_{0.13}$	–	–	$-0.47_{0.17}$	$0.60_{0.22}$	–	–
17	37.8	α_{S2}-casein	1.15	–	$-0.62_{0.15}$	–	$0.48_{0.17}$	$0.49_{0.23}$	–	–

–No significant evidence that the sire is heterozygous for the detected QTL.

[1]Σcasein = α_{S1}-casein + α_{S2}-casein + β-casein + κ-casein.

[2]Σwhey = α-lactalbumin + β-lactoglobulin.

[3]Casein index = Σcasein/(Σcasein + Σwhey) × 100.

α_{S1}-CN, α_{S2}-CN, and κ-CN. This chromosomal region on BTA6 seems to be involved in the regulation of all four casein fractions. Besides the casein genes, a possible BTA6 candidate gene is osteopontin (OPN), which is associated with milk protein percentage (Schnabel *et al.* 2005; Leonard *et al.* 2005, Olsen *et al.* 2007) and is located in the CI of the detected QTLs for the four casein fractions. Nemir *et al.* (2000) showed that mice, which suppress OPN production in the mammary epithelia, significantly reduce

Table 4 Percentage of phenotypic variance explained by the QTL without correction (V_{QTL}) and with correction for β-casein genotypes ($V_{QTL-\beta}$), κ-casein genotypes ($V_{QTL-\kappa}$), β-lactoglobulin genotypes ($V_{QTL-\beta lg}$), and DGAT1 genotypes (V_{QTL-D}) individually and simultaneously ($V_{QTL-All}$) for all significant QTL ($P_{genomewise} < 0.05$) across half-sib families (BTA: Bos taurus autosome).

BTA	Trait	V_{QTL} (%)	$V_{QTL-\beta}$ (%)	$V_{QTL-\kappa}$ (%)	$V_{QTL-\beta lg}$ (%)	V_{QTL-D} (%)	$V_{QTL-All}$ (%)
1	α_{S2}-casein	3.3	3.4	3.5	3.6	3.2	3.5
3	β-casein	3.6	3.0	3.2	3.1	3.4	2.9
5	α-lactalbumin	3.4	3.4	3.3	3.7	3.5	3.5
6	α_{S1}-casein	6.8	6.1	4.3	6.9	6.9	2.8
6	α_{S2}-casein	6.7	5.8	4.5	6.3	7.4	3.1
6	β-casein	3.6	1.8	3.7	3.4	3.5	1.8
6	κ-casein	7.9	6.2	3.2	8.5	8.2	2.4
6	Protein (%)	3.6	3.5	1.5	3.2	2.8	1.1
9	α_{S1}-casein	2.8	2.6	2.5	2.8	3.2	2.4
10	α_{S2}-casein	2.7	3.1	2.8	3.3	3.2	3.1
11	β-lactoglobulin	28.3	28.1	27.8	4.5	28.2	4.3
11	Σcasein[1]	10.4	10.6	10.3	1.9	10.5	10.8
11	Σwhey[2]	25.0	24.9	24.5	3.6	25.0	3.6
11	Casein index[3]	25.3	25.2	24.8	3.7	25.2	3.7
14	Protein (%)	8.6	8.6	8.7	8.6	1.6	1.7
15	Protein (%)	3.6	3.6	3.6	3.5	3.6	3.8
17	α_{S2}-casein	3.7	2.8	3.3	3.1	3.2	2.8

[1]Σcasein = α_{S1}-casein + α_{S2}-casein + β-casein + κ-casein.
[2]Σwhey = α-lactalbumin + β-lactoglobulin.
[3]Casein index = Σcasein/(Σcasein + Σwhey) × 100.

the synthesis of β-CN. This indicates that OPN could be a possible candidate gene underlying the QTL for the four caseins on BTA6.

On BTA11, the best position of the QTL affecting the β-LG fraction, Σcasein, Σwhey, and casein index was 124 cM, which is very close to the position of the β-LG gene at 123 cM on BTA11. Correction for β-LG genotypes decreased the proportion of phenotypic variance explained by the detected QTL on BTA11 affecting β–LG fraction with 84% (Table 4). This result is in agreement with Heck *et al.* (2009), who reported that the β-LG genotype explained 90% of the total genetic variation in the β-LG

Table 5 The percentage of phenotypic variance explained by multiple significant QTL ($P_{genomewise} < 0.05$) for one trait, with single QTL analysis and multiple QTL analysis across half-sib families (BTA: Bos taurus autosome).

		Detected QTL					
		Single QTL analysis[1]					
Trait	**BTA**	**1**	**2**	**3**	**4**	**Sum[2]**	**Multiple QTL analysis[3]**
α_{S1}-casein	6, 9	6.8	2.8	–	–	9.6	9.1
α_{S2}-casein	1, 6, 10, 17	3.3	6.7	2.7	3.7	16.4	14.4
β-casein	3, 6	3.6	3.6	–	–	7.2	6.4
Protein (%)	6, 14, 15	3.6	8.6	3.6	–	15.8	14.5

[1]The QTL numbers correspond to the BTA numbers in the second column.
[2]Sum is the phenotypic variance explained by single QTL 1 + single QTL 2 + single QTL 3 + single QTL 4.
[3]For multiple QTL analysis, the proportion of phenotypic variance explained by the multiple QTL was calculated by using the model with multiple QTL $[(b_{ik1}X_{ijk1}) + (b_{ik2}X_{ijk2}) + ...]$ for that trait.

fraction. The effect of correction for the β-LG genotype on Σcasein, Σwhey, and casein index (Table 4) is in agreement with the strong genetic correlation between β-LG and Σcasein (–0.76), between β-LG and Σwhey (0.98), and between β-LG and casein index (–0.98), as reported by Schopen *et al.* (2009). These results suggest that the detected QTL on BTA11 affecting β-LG can, to a large extent, be attributed to the A and B variant of β-LG. However, the QTL affecting β-LG remained significant after correction for the known β-LG genotypes, and therefore, we found evidence that this chromosomal region contains additional polymorphisms, besides the β-LG A and B polymorphisms, with an effect on the β-LG fraction. This

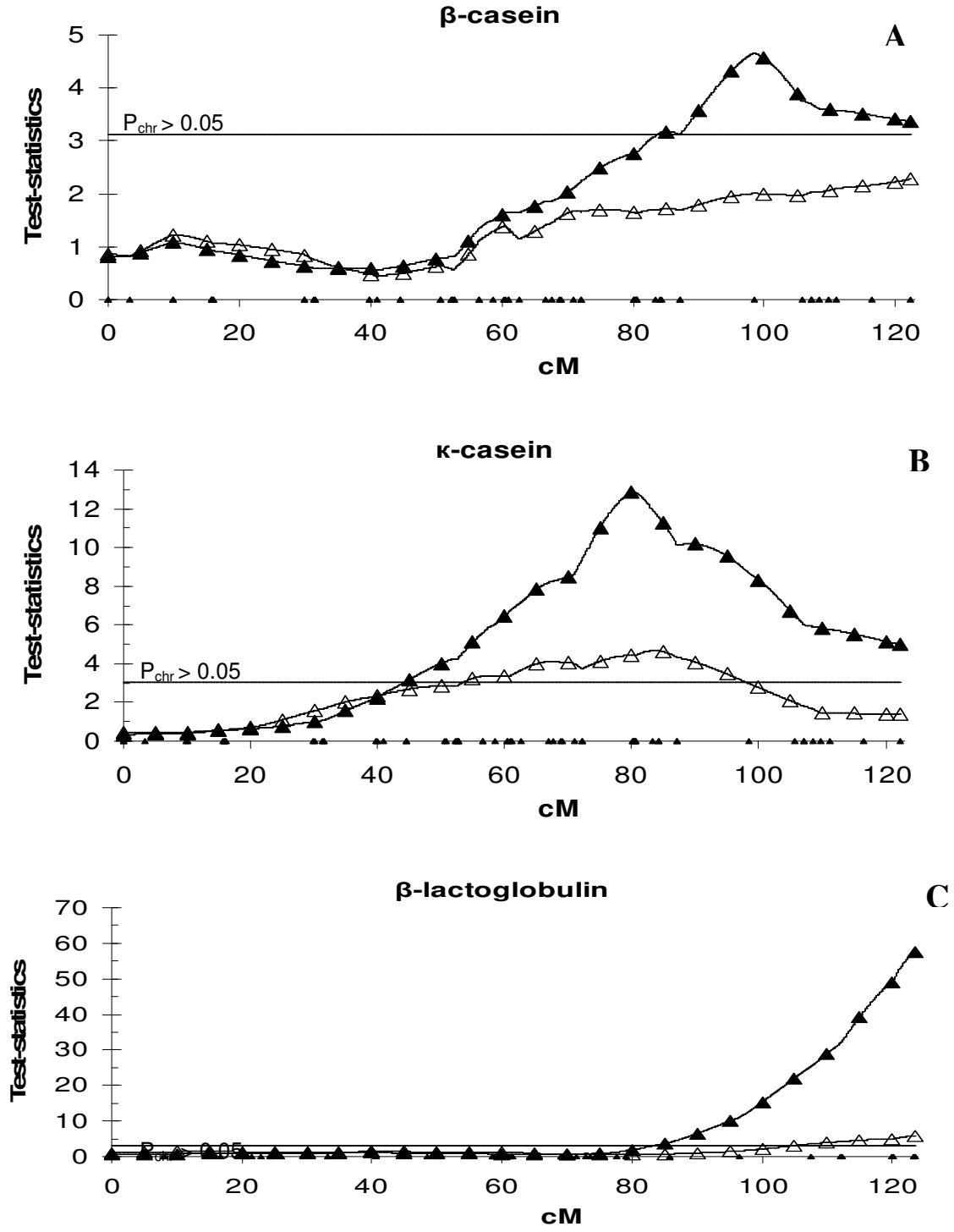

Figure 2 Test statistics for each cM of the genetic map for *Bos taurus* autosome (BTA) 6 for β-casein (A) and κ-casein (B), and for BTA11 for β-lactoglobulin (C) before (▲) and after (Δ) correction for β-casein or κ-casein or β-lactoglobulin genotypes, respectively, from the across-family analysis. The triangles on the x-axis indicate the positions of the SNPs.

confirms recent findings by Ganai *et al.* (2009), who reported 33 new polymorphisms in the coding and promoter regions of the β-LG gene. We did not detect any significant chromosomal regions affecting the β-LG fraction other than the chromosomal region on BTA11. Thus, the chromosomal region on BTA11 appears to be the most important region for controlling the β-LG fraction content of cow's milk.

Significant QTLs affecting all four casein fractions were detected on BTA6, and one QTL affecting the β-LG fraction was detected on BTA11. The other significant QTLs affecting the six major milk protein were mainly located on different chromosomes: BTA1, 3, 5, 9, 10, and 17. These results suggest that other than the "master control" chromosomal region on BTA6, there is remarkably little overlap in the regions controlling the expression of the different casein fractions. For example, for κ-CN, we detected a significant region on BTA6 only, whereas for α_{S2}-CN, we detected significant chromosomal regions on BTA1, 6, 10, and 17. These results suggest that, for the most part, different genes are involved in the regulation of the six major milk proteins, which is in agreement with the low genetic correlations among the six major milk proteins, as reported by Schopen *et al.* (2009), except for α_{S1}-CN and α_{S2}-CN. Two chromosomal regions (BTA10 and 14) affected both the α_{S1}-CN and the α_{S2}-CN fraction. The allele substitution effects of the QTL affecting α_{S1}-CN and α_{S2}-CN on both BTA10 and 14 were, in general, opposite from one another, which is in agreement with the genetic correlation of –0.49 between the α_{S1}-CN and α_{S2}-CN fractions (Schopen *et al.* 2009).

Conclusions

We detected ten significant chromosomal regions affecting milk protein composition. The QTLs detected on BTA6 that affected the α_{S1}-CN, α_{S2}-CN, β-CN, and κ-CN fractions were partially explained by the polymorphisms in the casein genes. The QTL detected on BTA11 that affected the β-LG fraction and Σcasein, Σwhey, and casein index were partially explained by the polymorphism in the β-LG gene. The QTLs detected on BTA14 that affected the α_{S1}-CN and α_{S2}-CN fractions, and protein percentage were mostly explained by the polymorphism in the DGAT1 gene. The regions on BTA6, 11, and 14 as well as regions on other chromosomes where we

detected QTLs affecting major milk proteins, will be a fruitful source of data in the future. Fine mapping will be required to further narrow the chromosomal regions in order to find the underlying genes causing the QTLs.

Acknowledgements

This study is part of the Milk Genomics Initiative, funded by Wageningen University, NZO (Dutch Dairy Organization), CRV (cooperative cattle improvement organization), and the Dutch technology foundation STW. The authors thank the owners of the herds for their help in collecting the data, Ruben van 't Slot and Bert Dibbits for their assistance in generating the genotypes, and Richard Spelman for his work on writing the QTL program in Fortran.

References

Aschaffenburg R., and J. Drewry. 1955. Occurrence of different beta-lactoglobulins in cow's milk. Nature 176: 218–219.

Aschaffenburg R. 1961. Inherited casein variants in cow's milk. Nature 192: 431–432.

Ashwell M.S., C. P. Van Tassell, and T.S. Sonstegard. 2001. A genome scan to identify quantitative trait loci affecting economically important traits in a US Holstein population. J. Dairy Sci. 84: 2535–2542.

van den Berg G., J.T.M. Escher, P.J. de Koning, and H. Bovenhuis. 1992. Genetic polymorphism of κ-casein and β-lactoglobulin in relation to milk composition and processing properties. Neth. Milk Dairy J. 46: 145–168.

Bobe G., D.C. Beitz, A.E. Freeman, and G.L. Lindberg. 1999. Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. J. Dairy Sci. 82: 2797–2804.

Churchill G.A., and R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971.

Coppieters W., J. Riquet, J.J.. Arranz, P. Berzi, N. Cambisano, B. Grisart, L. Karim, F. Marcq, L. Moreau, C. Nezer, P. Simon, P. Vanmanshoven, D. Wagenaar, and M. Georges. 1998. A QTL with

major effect on milk yield and composition maps to bovine chromosome 14. Mamm. Genome 9: 540–544.

Farrell H.M., R. Jimenez-Flores, G.T. Bleck, E.M. Brown, J.E. Butler, L.K. Creamer, C.L. Hicks, C.M. Hollar, K.F. Ng-Kwai-Hang, and H.E. Swaisgood. 2004. Nomenclature of the proteins of cows's milk – sixth revision. J. Dairy Sci. 87: 1641-1674.

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

Gilmour A. R., B.J. Gogel, B.R. Cullis, S.J. Welham, and R. Thompson. 2002. Asreml user guide. Release 1.0. VSN International Ltd., Hemel Hempstead, UK.

Green P., K. Falls, and S. Crooks. 1990. Cri–map documentation, version 2.4. Washington University School of Medicine, St. Louis, MO, USA.

Hayes H.C., and E.J. Petit. 1993. Mapping of the β-lactoglobulin gene and of an immunoglobulin M heavy chain-like sequence to homoeologous cattle, sheep, and goat chromosomes. Mamm. Genome 4: 207-210.

Hayes H.C., P. Popescu, and B. Dutrillaux. 1993. Comparative gene mapping of lactoperoxidase, retinoblastoma, and α-lactalbumin genes in cattle, sheep, and goats. Mamm. Genome 4: 593-597.

Heck J.M.L., C. Olieman, A. Schennink, H.J.F. van Valenberg, M.H.P.W. Visker. R.C.R. Meuldijk, and A.C.M. van Hooijdonk. 2008. Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis. Int. Dairy J. 18: 548–555.

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

Ihara N., A. Takasuga, K. Mizoshita, H. Takeda, M. Sugimoto, Y. Mizoguchi, T. Hirano, T. Itoh, T. Watanabe, K.M. Reed, W.M. Snelling. S.M. Kappes, C.W. Beattie, G.L. Bennett, and Y. Sugimoto. 2004. A comprehensive genetic map of the cattle

genome based on 3802 microsatellites. Genome Res. 14: 1987–1998.

Kappes S.M., J.W. Keele, R.T. Stone, R.A. McGraw, T.S. Sonstegard, T.P.L. Smith, N.L. Lopez–Corrales, and C.W. Beattie. 1997. A second-generation linkage map of the bovine genome. Genome Res. 7: 235–249.

Kent W.J. (2002) BLAT – the BLAST-like alignment tool. Genome Res. 12: 656–664.

Khatkar M.S., P.C. Thomson, I. Tammen, and H.W. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet. Sel. Evol. 36: 163–190.

Knott S.A., J.M. Elsen, and C.S. Haley. 1996. Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. Theor. Appl. Genet. 93: 71–80.

De Koning D.J., L.L.G. Janss, A.P. Rattink, P.A.M. van Oers, B.J. de Vries, M.A.M. Groenen, J.J. van der Poel, P.N. de Groot, E.W. Brascamp, and J.A.M. van Arendonk. 1999. Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (*Sus scrofa*). Genetics 152: 1679–1690.

De Koning D.J., N.F. Schulmant, K. Elo, S. Moisio, R. Kinos, J. Vilkki, and A. Maki-Tanila. 2001. Mapping of multiple quantitative trait loci by simple regression in half-sib designs. J. Anim. Sci. 79: 616–622.

Kühn Ch., G. Freyer, R. Weikard, T. Goldammer, and M. Schwerin. 1999. Detection of QTL for milk production traits in cattle by application of a specifically developed marker map of BTA6. Anim. Genet. 30: 333–340.

Leonard S., H. Khatib, V. Schutzkus, Y.M. Chang, and C. Maltecca. 2005. Effects of the osteopontin gene variants on milk production traits in dairy cattle. J. Dairy Sci. 88: 4083–4086.

Marziali A.S., a nd K.F. Ng-Kwai-Hang. 1986. Effects of milk composition and genetic polymorphism on coagulation properties of milk. J. Dairy Sci. 69: 1793–1798.

Mayer H.K., M. Ortner, E. Tschager, and W. Ginzinger. 1997. Composite milk protein phenotypes in relation to composition and cheesemaking properties of milk. Int. Dairy J. 7: 305–310.

Mosig M.O., E. Lipkin, G. Khutoreskaya, E. Tchourzyna, M. Soller, and A. Friedmann. 2001. A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics 157: 1683–1698.

Neelin J.M. 1964. Variants of κ-casein revealed by improved starch gel electrophoresis. J. Dairy Sci. 47: 506–509.

Nemir M., D. Bhattacharyya, X. Li, K. Singh, A.B. Mukherjee, and B.B. Mukherjee. 2000. Targeted inhibition of osteopontin expression in the mammary gland causes abnormal morphogenesis and lactation deficiency. J Biol. Chem. 275: 969–976.

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

Olsen H.G., L. Gomez–Raya, D.I. Våge, I. Olsaker, H. Klungland, M. Svendsen, T. Ådnøy, A. Sabry, G. Klemetsdal, N. Schulman, W. Krämer, G. Thaller, K. Rønningen, and S. Lien. 2002. A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle. J. Dairy Sci. 85: 3124–3130.

Olsen H.G., H. Nilsen, B. Hayes, P.R. Berg, M. Svendsen, S. Lien, and T. Meuwissen. 2007. Genetic support for a quantitave trait nucleotide in the ABCG2 gene affecting milk composition of dairy dattle. BMC Genet. 8: 32.

Schaar J., B. Hansson, and H.E. Pettersson. 1985. Effects of genetic variants of κ-casein and β-lactoglobulin on cheesemaking. J. Dairy Res. 52: 429–437.

Schennink A., W.M. Stoop, M.H.P.W. Visker, J.M.L. Heck, H. Bovenhuis, J.J. van der Poel, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2007. DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. Anim. Genet. 38: 467–473.

Schennink A., J.M.L Heck, H. Bovenhuis, M.H.P.W. Visker, H.J.F. van Valenberg , and J.A.M. van Arendonk. 2008. Milk fatty acid unsaturation: genetic parameters and effects of stearoyl-CoA desaturase (SCD1) and acyl CoA: diacylglycerol acyltransferase 1 (DGAT1). J. Dairy Sci. 91: 2135–2143.

Schnabel R.D., J.-J. Kim, M.S. Ashwell, T.S. Sonstegard, C.P. van Tassell, E.E. Connor, and J. F. Taylor. 2005. Fine-mapping milk production quantitative trait loci on BTA6: Analysis of the bovine osteopontin gene. Proceedings of the National Academy of Sciences 102: 6896–6901.

Schopen G.C.B., J.M.L Heck, H. Bovenhuis, M.H.P.W. Visker, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. J. Dairy Sci. 92:1182-1191

Schrooten C., M.C.A.M. Bink, and H. Bovenhuis. 2004. Whole genome scan to detect chromosomal regions affecting multiple traits in dairy cattle. J. Dairy Scie. 87, 3550–3560.

Snelling W.M., E, Casas, R.T. Stone, J.W. Keele, G.P. Harhay, G.L. Bennett, and T.P.L. Smith. 2005. Linkage mapping bovine EST-based SNP. BMC Genomics 6: 74.

Spelman R.J., W. Coppieters, L. Karim, J.A.M. van Arendonk, and H. Bovenhuis. 1996. Quantitative trait loci analysis for five milk production traits on chromosome s*ix* in the Dutch Holstein-Friesian population. Genetics 144: 1799–1808.

Stoop W.M., H. Bovenhuis, and J.A.M. van Arendonk. 2007. Genetic parameters for milk urea nitrogen in relation to milk production traits. J. Dairy Sci. 90: 1981–1986.

Thompson M.P., C.A. Kiddy, L. Pepper, and C.A. Zittle. 1962. Variations in the α_s-casein fraction of individual cow's milk. Nature 195: 1001–1002.

Threadgill D. W., and J.E. Womack. 1990. Genomic analysis of the major milk protein genes. Nucleic Acids Res. 18: 6935-6942.

Wedholm A., L.B. Larsen, H. Lindmark–Månsson, A.H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. J. Dairy Sci. 89: 3296–3305.

Zhang Q., D. Boichard, I. Hoeschele, C. Ernst, A. Eggen, B. Murkve, M. Pfister–Genskow, L.A. Witte, F.E. Grignola, P. Uimari, G. Thaller, and M.D. Bishop. 1998. Mapping quantitative trait loci for milk

production and health of dairy cattle in a large outbred pedigree. Genetics 149: 1959–1973.

5

Whole genome association study for milk protein composition in dairy cattle

G. C. B. Schopen, M. H. P. W. Visker, P.D. Koks, E. Mullaart, J.A.M. van Arendonk and H. Bovenhuis

Submitted

Abstract

Our objective was to perform a genome-wide association study for milk protein composition, casein index, protein percentage, and protein yield using a 50K single nucleotide polymorphism (SNP) chip in 1,713 Dutch Holstein-Friesian cows. DNA was isolated from the blood samples of cows, and a total of 49,643 SNPs were used in a two-step procedure. The first step involved a general linear model while a mixed model was used in the second step. Chromosomal regions with SNPs significantly associated with milk protein composition were distributed over 15 bovine autosomes. The main regions with SNPs significantly associated with milk protein composition were found on *Bos taurus* autosomes (BTAs) 5, 6, 11, and 14. The number of chromosomal regions with SNPs significantly associated with a trait ranged from two for β-casein (CN) to nine for α_{S2}-CN. Two regions (on BTA6 and 11) were significantly associated with the caseins and at least with one of the whey proteins. However, some regions were significantly associated with SNPs that were unique for α_{S1}-CN (region 13_1), α_{S2}-CN (regions 1_2, 9_2, 10_2, and 17_2), κ-CN (regions 13_2 and 21), α-lactalbumin (regions 1_1, 5_1, and 26), and β-lactoglobulin (LG) (region 14). The proportion of genetic variance explained by the SNP most significantly associated with a trait ranged from 1.1% for α_{S2}-CN on BTA29 to 65.8% for β-LG on BTA11. The proportion of genetic variance explained by known genotypes ranged from 1.7% for *κ-CN* genotypes for β-CN and β-LG fraction on BTA6 to 64.9% for *β-LG* genotypes for the β-LG fraction on BTA11. The results indicate that in addition to the four main regions on BTA 5, 6, 11 and 14, also other regions play a role in the genetic regulation of milk protein synthesis.

Introduction

Because of their nutritional value, dairy products (e.g., milk and cheese) make a significant contribution to human diets. In the dairy industry, the protein composition of milk is important; for example, high casein content in milk results in a higher cheese yield per kilogram of milk protein. This implicates that it is possible to produce more cheese from the same amount of milk protein with high casein content. In the Netherlands, the protein content of milk is of relevance to dairy farmers, who are paid based on the amount of protein in milk.

Several studies have examined the effects of milk protein variants on milk protein composition (e.g., Ng-Kwai-Hang *et al.*, 1987; Bobe *et al.*, 1999; Heck *et al.*, 2009). Heck *et al.* (2009) showed that variants of the *β-CN* and *κ-CN* genes, which are both located on *Bos taurus* autosome (BTA) 6, were associated with milk protein composition (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG). Variants of the *β-LG* gene, which is located on BTA11, also affect milk protein composition. A clear example is the positive effect of the *β-LG* B variant on cheese yield (Heck *et al.*, 2009).

In addition to the known genetic variants of milk proteins that are associated with milk protein composition, the results of a genome-wide linkage study that we previously reported revealed significant QTLs on BTA3, 5, 9, 10, 15, and 17 that affect milk protein composition (Schopen *et al.*, 2009b). However, which genes are responsible for these detected QTLs remains unknown. The confidence intervals of the QTLs reported by Schopen *et al.* (2009b) were in general large, but they can be reduced by using high-density SNP genotyping per individual, which has recently become available. Moreover, other chromosomal regions with small effects on milk protein composition can be detected due to the higher power offered by high-density SNP genotyping. Thus, these high-density SNP arrays can facilitate identification of new candidate genes that affect milk protein composition. The objective of this study, therefore, was to perform a genome-wide association analysis for milk protein composition (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG), casein index, protein percentage, and protein yield using a 50K SNP chip in Dutch Holstein-Friesian cows.

Materials and methods

Phenotypes

As part of the Dutch Milk Genomics Initiative, the phenotypic data of 1,912 first lactation Holstein-Friesian cows from 398 commercial herds throughout the Netherlands were collected. Details about the animals used in this study are available in Schopen *et al.* (2009a).

Protein percentage was determined by infrared spectroscopy. To calculate protein yield, we multiplied protein percentage by the morning milk yield. Milk yields were missing for 141 cows, leaving only 1,771 cows available for protein yield determination. Milk protein composition was evaluated using capillary zone electrophoresis (CZE), as described by Heck *et al.* (2008). Using CZE, we quantified α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG. All six major milk proteins were expressed as the weight-proportion of total protein (ww%). Furthermore, the casein index was calculated as follows:

$$\text{casein index} = \frac{\Sigma\text{casein}}{\Sigma\text{casein} + \Sigma\text{whey}} \times 100 \qquad [1]$$

where Σcasein was defined as the sum of the percentages of α_{S1}-CN, α_{S2}-CN, β-CN, and κ-CN, and Σwhey was calculated by adding the percentages of α-LA and β-LG. Table 1 gives the mean, phenotypic standard deviation, and intraherd heritability for milk protein composition, casein index, protein percentage, and protein yield.

Genotypes

DNA was isolated from the blood samples of cows. For this study, DNA was available from cows in five large paternal half-sib families (214, 187, 175, 174, and 97 cows) and from cows in 53 small paternal half-sib families (with 9–29 cows). A 50K SNP chip was designed by CRV (cooperative cattle improvement organization, Arnhem, the Netherlands) and obtained from Illumina and used to genotype all animals with the Infinium assay (Illumina, San Diego, CA, USA). Charlier *et al.* (2008) provide more information about the 50K SNP chip. This approach resulted in 50,856 technically successful SNPs, which were mapped using the bovine genome assembly (BTAU4.0, Liu *et al.*, 2009). Of the 50,856 SNPs, a total of 778

were not mapped to any of the 29 bovine autosomes, and 589 SNPs were mapped to chromosome X. The SNPs on chromosome X were not included

Table 1 Mean, phenotypic standard deviation[1] (σ_p), and intraherd heritability (h^2) for the six major milk proteins[2], casein index, protein percentage, and protein yield for 1,912 Dutch Holstein-Friesian cows in their first lactation.

Trait	Mean	σ_p	h^2
α_{S1}-CN	33.62	1.59	0.47
α_{S2}-CN	10.38	1.34	0.73
β-CN	27.17	1.46	0.26
κ-CN[3]	4.03	0.55	0.63
α-LA	2.44	0.29	0.57
β-LG	8.34	1.19	0.80
Casein index[4]	87.46	1.37	0.69
Protein (%)	3.51	0.27	0.66
Protein (kg)	0.47	0.07	0.25

[1]Phenotypic standard deviation after adjusting for systematic environmental effects: day of lactation, age at first calving, season of calving, and herd.
[2]Expressed as a percentage of the total protein fraction (ww%).
[3]κ-CN in the monophosphorylated form only.
[4]Casein index = Σcasein / (Σcasein + Σwhey) * 100.

in the association study, and chromosome null is defined as the chromosome that contains the unmapped SNPs. Of the remaining 50,267 SNPs, 231 SNPs were monomorphic and 393 SNPs had a genotyping rate <80%. The final dataset consisted of 49,643 SNPs for use in the association study.

A total of 1,868 animals were genotyped and formed a large part of a subset of the 1,912 animals with phenotypes. Thus, not all phenotyped animals had genotypes, and not all genotyped animals had phenotypes. Of the 1,868 genotyped animals, 155 animals did not have phenotypes for milk protein composition, casein index, protein percentage, and protein yield. The dataset that was used in the whole association study consisted of 1,713 animals, and 130 of these had no record for protein yield.

Genotypes for polymorphisms in the *β-CN*, *κ-CN*, *β-LG*, and *DGAT1* genes were known. The polymorphisms in *κ-CN*, *β-LG*, and *DGAT1* genes were included on the 50K SNP chip; however, for the *β-CN* gene, not all polymorphisms were on the chip. The genotypes for polymorphisms in the *β-CN* gene were determined using a SNaPshot assay (Visker *et al.*, 2010).

Whole genome association study

The whole genome association study was performed using a two-step procedure. In the first step, a single SNP analysis was performed using the SNPassoc package (González *et al.*, 2007) in R using the following general linear model:

$$Y_{ij} = Sire_i + SNP_j + e_{ij}, \qquad [2]$$

where Y_{ij} was the phenotype adjusted for systematic environmental effects: day of lactation, age at first calving, season of calving, and herd; $Sire_i$ was the fixed effect of sire i; SNP_j was the fixed effect of the jth class of the SNP; and e_{ij} was the random residual effect ($e_{ij} \sim N(0,\sigma^2_e)$).

The systematic environmental effects, which were used to adjust the phenotypes, were estimated by using an animal model in ASReml (Gilmour *et al.*, 2002) for all 1,912 cows with phenotypes, as described by Schopen *et al.* (2009a). Furthermore, the sire effect was included in the general linear model to account for a family effect.

Additional genetic relationships among individuals might exist that were not accounted for in the general linear model, possibly leading to false-positive associations. To reduce the number of false-positive associations (Kennedy *et al.*, 1992), we accounted for all genetic relationships among individuals in the second step of the whole genome association study. For this step, we analyzed regions containing SNPs that were significantly (false discovery rate, FDR < 0.01) associated with one of the traits using the linear model. In this analysis, a single SNP was simultaneously adjusted for systematic environmental effects and for all genetic relationships among individuals in ASReml (Gilmour *et al.*, 2002) by using the following animal model:

$$y_{ijklmno} = \mu + b_1*lactst_i + b_2*e^{-0.05*lactst}{}_i + b_3*ca_j + b_4*ca_j^2 + season_k + scode_l + SNP_o + animal_m + herd_n + e_{ijklmn}, \quad [3]$$

where $y_{ijklmno}$ was the dependent variable; μ was the overall mean; $lactst_i$ was a covariate describing the effect days in lactation; ca_j was a covariate describing the effect of age at first calving; $season_k$ was the fixed effect with three classes for calving season (June–August 2004, September–November 2004, and December 2004–February 2005); $scode_l$ was the fixed effect accounting for possible differences in genetic level between proven bull daughters and young bull daughters; SNP_o was the fixed effect of the SNP; $animal_m$ was the random additive genetic effect of animal m; $herd_n$ was a random herd effect; and e_{ijklmn} was the random residual effect. The variance-covariance structure of the additive genetic effects was $Var(animal) = A\sigma^2_{a,}$ where A was a matrix of additive genetic relationships among individuals and σ^2_a was the additive genetic variation.

Significance thresholds and SNP variance

Significance thresholds were obtained by calculating the FDR based on the qvalue package (Storey and Tibshirani, 2003) in R. The FDR was calculated based on the *P* values obtained from the general linear model for all 49,643 SNPs for each trait separately. SNPs with an FDR < 0.01 were considered to be significantly associated with the traits.

The proportion of genetic variance explained by an SNP was calculated from the estimated genotype effects and the observed genotype frequencies. The result was expressed as a percentage of the phenotypic variance.

Results

Genotypes

In total, 50,267 SNPs for 1,713 animals were distributed over 29 bovine autosomes; of these, 231 SNPs were monomorphic and 393 SNPs had a genotyping rate <80%. The number of monomorphic SNPs per chromosome ranged from 1 on BTA27 to 20 on BTA2, and the number of

SNPs with a genotyping rate <80% varied from 4 on BTA29 to 25 on BTA1 and 6 (Table 2).

Detected associations using the general linear model

Using a general linear model in the association study resulted in significant associations of SNPs (FDR < 0.01) with at least one of the traits on all 29 bovine autosomes (Figure 1). The main regions of SNPs significantly associated with the six major milk proteins were found on BTA5, 6, 11, and 14. On BTA6, significant associations of SNPs were found with all traits, except for protein yield. The regions of SNPs significantly associated with casein index and β-LG were similar.

On the null chromosome (unmapped SNPs), SNPs were significantly associated with all traits, except with β-CN and protein yield. Only one SNP was significantly associated with α_{S1}-CN, κ-CN, and α-LA; three were associated with β-LG, casein index, and protein percentage; and four were associated with α_{S2}-CN. There was overlap in the SNPs significantly associated with the traits and, therefore, nine different SNPs were significantly associated with the milk protein composition on chromosome null.

The regions of SNPs with an FDR < 0.01 were selected for further analyses using a mixed model in addition to the general linear model. All SNPs (also SNPs with FDR > 0.01) located within the selected regions were analyzed. For each chromosome, the same region was used for all traits. For the null chromosome, the nine SNPs were analyzed for all traits using the mixed model. Ultimately, a total of 3,655 SNPs were distributed over 32 regions on 22 bovine autosomes. Table 3 gives the start and end positions and the number of SNPs for each region.

Detected associations using the mixed model

Using a mixed model in the association study resulted in SNPs that were significantly associated (FDR < 0.01) with at least one trait for 20 out of 32 regions distributed over 15 chromosomes (Table 4). For 15 regions (regions 1_1, 1_2, 5_1, 5_2, 5_3, 9_2, 10_2, 13_1, 13_2, 15, 17_2, 20, 21,

Table 2 Map length[1], number of markers, number of monomorphic markers, and number of markers with a genotyping rate <80% (#GenotRate) for all 29 *Bos taurus* autosomes (BTAs).

BTA	Length (Mbp)	# Markers	# Monomorphic	#GenotRate
NULL	-	778	1	8
1	160.91	3,012	4	25
2	140.64	2,451	20	19
3	127.13	2,342	12	11
4	124.09	2,300	11	20
5	125.78	2,215	5	18
6	122.51	2,844	12	25
7	111.67	2,017	8	14
8	116.93	2,131	14	16
9	108.05	1,860	14	20
10	106.10	1,911	14	11
11	110.01	2,193	18	12
12	85.22	1,512	6	14
13	84.00	1,689	11	12
14	81.29	2,122	9	17
15	84.23	1,446	5	7
16	77.83	1,455	6	14
17	76.40	1,561	5	14
18	66.04	1,282	3	8
19	65.13	1,452	5	7
20	75.41	1,479	2	9
21	69.08	1,246	5	15
22	61.75	1,256	4	10
23	53.27	1,169	5	12
24	64.93	1,296	9	14
25	43.44	1,256	6	9
26	51.00	1,131	5	7
27	48.73	933	1	15
28	46.01	899	5	6
29	51.78	1,029	6	4

[1]Based on bovine physical map BTAU4.0.

24, and 26), only one trait showed a significant association with SNPs, whereas for the other five regions (regions 6, 10_1, 11, 14, and 29), multiple traits showed a significant association with SNPs. The region on BTA6 was significantly associated with all traits, except for protein yield (Table 4).

For all investigated traits in this study, except for protein yield, we found more than one chromosomal region with significantly associated SNPs. The number of chromosomal regions with SNPs significantly associated with a trait ranged from two for β-CN to nine for α_{S2}-casein (Table 4). There were similarities between the chromosomal regions with SNPs significantly associated with the six major milk proteins. In addition, two regions (regions 6 and 11) were significantly associated with the caseins and at least one of the whey proteins. However, some regions with significantly associated SNPs were unique for α_{S1}-CN (region 13_1), α_{S2}-CN (regions 1_2, 9_2, 10_2, and 17_2), κ-CN (regions 13_2 and 21), α-LA (regions 1_1, 5_1 and 26), and β-LG (region 14) (Table 4).

For protein percentage, four regions (regions 6, 10_1, 14, and 29) were significantly associated with at least one of the six major milk proteins. However, some regions had significant associations with SNPs that were unique for protein percentage (regions 5_2, 5_3, 15, and 20) (Table 4).

On the null chromosome (unmapped SNPs), SNPs were significantly associated with α_{S1}-CN, α_{S2}-CN, κ-CN, α-LA, β-LG, casein index, and protein percentage (Table 5).

Six major milk proteins

α_{S1}-CN. The number of chromosomes with SNPs significantly associated with a trait was different for each of the six major milk proteins. SNPs on BTA6, 7, 10, 11, 12, 13, 14, 16, 22, and 29 (Fig. 1) and one unmapped SNP were significantly associated with α_{S1}-CN (FDR < 0.01). On BTA12 and 22, the SNP significantly associated with α_{S1}-CN involved only one animal in one of the genotype classes. The phenotype of the animal on BTA12 was also an outlier for α_{S1}-CN. Removing the genotype of this single animal and rerunning the mixed model (a kind of sensitivity test) for the SNPs on BTA1

Table 3 The regions on *Bos taurus* autosomes (BTAs) analyzed using the mixed model in the association study with the number of regions, the name of each region, the length of each region, and the total number of SNPs located in each region[1]

BTA	# Regions	Name_Region	Length (Mbp)	# SNPs
Null	-	Null		9
1	2	1_1	109.7 – 110.2	10
		1_2	146.6 – 150.3	90
2	1	2	67.4 – 67.4	5
5	3	5_1	13.5 – 42.9	438
		5_2	75.3 – 75.3	1
		5_3	97.5 – 98.5	16
6	1	6	61.1 – 97.7	672
9	2	9_1	17.4 – 17.4	1
		9_2	80.6 – 80.6	1
10	2	10_1	51.4 – 52.4	17
		10_2	91.8 – 91.8	1
11	1	11	84.3 – 110.2	472
13		13_1	38.2 – 38.2	1
		13_2	60.5 – 60.5	1
14	1	14	0.0 – 13.6	842
15	1	15	41.4 – 61.6	330
16	1	16	53.6 – 53.6	5
17	2	17_1	19.4 – 19.4	1
		17_2	29.6 – 36.0	116
19	1	19	32.9 – 39.3	110
20	1	20	27.3 – 39.1	200
21	1	21	47.1 – 47.1	1
22	2	22_1	40.0 - 40.1	6
		22_2	52.3 – 52.3	1
23	1	23	25.2 – 25.2	1
24	1	24	35.7 – 35.7	1
25	2	25_1	9.4 – 9.4	1
		25_2	35.0 – 39.6	90
26	1	26	33.1 – 33.1	1
28	2	28_1	18.6 – 26.1	120
		28_2	43.2 – 43.2	1
29	1	29	39.4 – 45.5	93
				3,655

[1]On some chromosomes, three regions were defined: _1 is the first region, _2 is the second region, and _3 is the third region on the chromosome.

and 22 resulted in elimination of significance for the SNPs in this BTA12 region (*P* value from 1.8E-06 to 0.132) and on BTA22 (*P* value from 4.1E-05 to 4.0E-04). On BTA7, 10, 16, and 29, no SNPs were significantly associated with α_{S1}-CN after the running of the mixed model. The remaining SNPs significantly associated with α_{S1}-CN were located on BTA6, 11, 13, and 14 (Table 4) and the null chromosome (Table 5). The proportion of genetic variance explained by the SNP most significantly associated with α_{S1}-CN on these five chromosomes varied from 1.2% on BTA13 to 7.7% on BTA6 (Table 6). Including these five SNPs simultaneously in the mixed model resulted in elimination of significance of the unmapped SNP. The sum of the proportion of the genetic variance explained by the remaining four SNPs was 13.9%, which was similar to the proportion of the genetic variance (14.4%) explained by these four SNPs when they were simultaneously included (multiple SNP analysis) in the mixed model (Table 7).

In addition, we investigated whether polymorphisms in the *β-CN* and *κ-CN* genes on BTA6, the *β-LG* gene on BTA11, and the *DGAT1* gene on BTA14 could explain the detected association for milk protein composition on BTA6, 11, and 14, respectively. Therefore, we performed additional single SNP analyses including the known genotypes as an extra fixed effect in the mixed model. Accounting for known *DGAT1* genotypes on BTA14 resulted in elimination of the significance for SNPs previously associated with α_{S1}-CN on BTA14. The proportion of genetic variance of α_{S1}-CN explained by known genotypes varied from 2.8% for *DGAT1* genotypes to 14.7% for *β-CN* genotypes (Table 8).

α_{S2}-CN. For α_{S2}-CN, SNPs on BTA1, 6, 9, 10, 11, 14, 17, 19, 28, and 29 (Fig. 1) and four unmapped SNPs were significantly associated (FDR < 0.01). On BTA19 and 28, no SNPs were significantly associated with α_{S2}-CN after the running of the mixed model. The remaining SNPs were located on BTA1, 6, 9, 10, 11, 14, and 17 (Table 4) and the null chromosome (Table 5). The proportion of genetic variance explained by the SNP most significantly associated with α_{S2}-CN on these eight chromosomes ranged from 1.1% on BTA29 to 11.1% on BTA6 (Table 6). Including these 11 SNPs (four SNPs on the null chromosome) simultaneously in the mixed model resulted in elimination of significance of the four unmapped SNPs. The sum of the

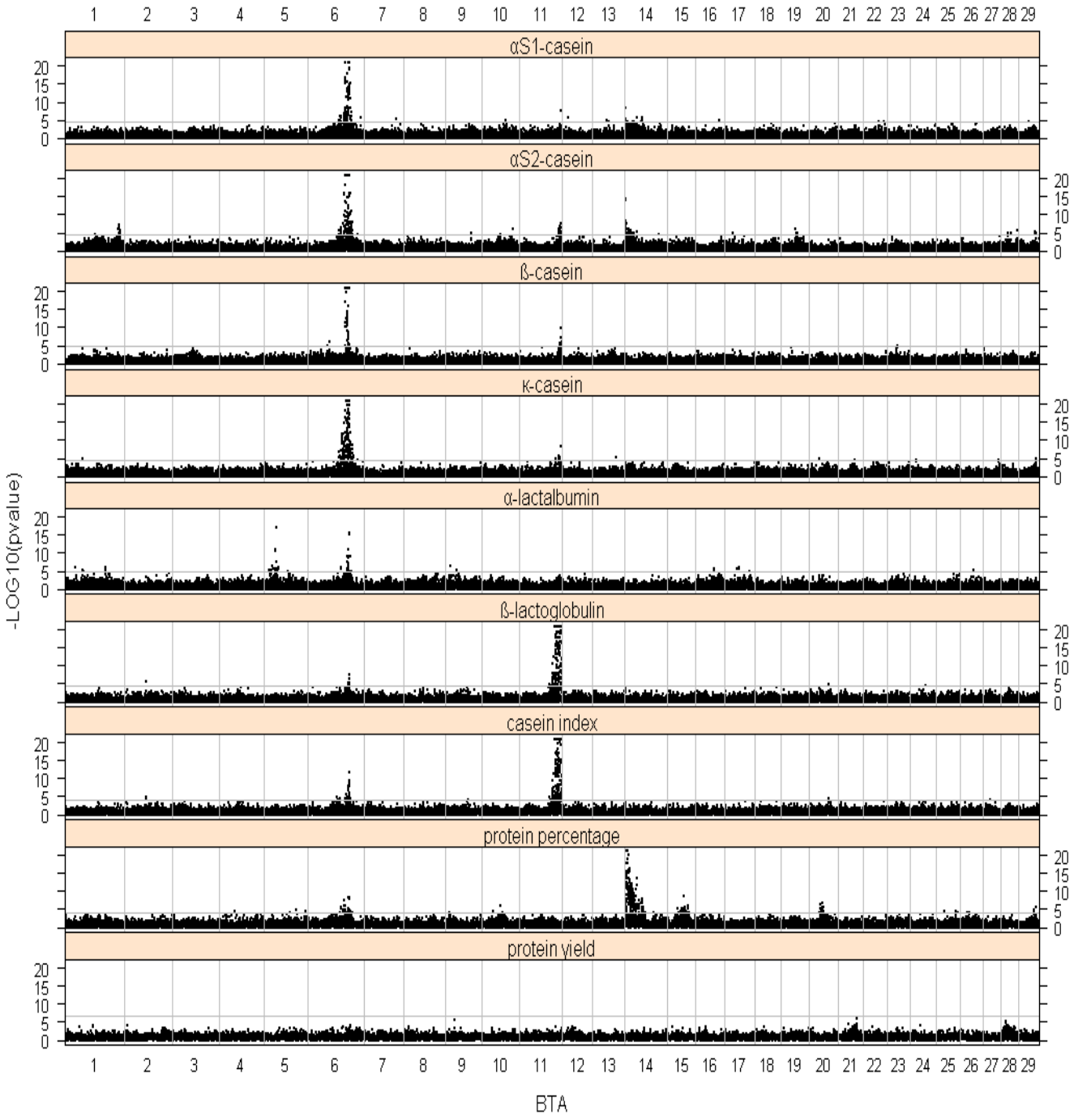

Figure 1 The $-\log_{10}$(*P*-values) for all 48.874 SNPs, for all 29 *Bos taurus* autosomes (BTA) for the six major milk proteins, casein index, protein percentage, and protein yield. The horizontal grey line is the threshold level where FDR = 0.01. A $-\log_{10}$(*P* value) > 21 was set to 21 to retain an overview of the SNPs that exceeded the threshold level. The $-\log_{10}$(*P* value) was based on the general linear model in the association study using R.

proportion of the genetic variance explained by the remaining seven SNPs was 25.7%, which was higher than the proportion of the genetic variance (21.8%) explained by the multiple SNP analysis (Table 7). Accounting for known *DGAT1* genotypes on BTA14 resulted in elimination of significance of the SNPs previously associated with α_{S2}-CN on BTA14. The proportion of genetic variance of α_{S2}-CN explained by known genotypes varied from 2.0% for *κ-CN* genotypes to 12.6% for *β-CN* genotypes (Table 8).

β-CN. For β-CN, SNPs on BTA6, 11, and 23 (Fig. 1) were significantly associated (FDR < 0.01); however, on BTA23, no SNPs were significantly associated with β-CN after the running of the mixed model. The remaining SNPs were located on BTA6 and 11 (Table 4). The proportion of genetic variance explained by the SNP most significantly associated with α_{S2}-CN on BTA6 was 24.7%, and on BTA11, it was 3.0% (Table 6). The sum of the proportion of the genetic variance explained by these two SNPs was 27.7%, which was similar to the proportion of the genetic variance (26.8%) explained by the multiple SNP analysis (Table 7). Accounting for known *β-CN* genotypes on BTA6 resulted in a decrease in significance of SNPs previously associated with β-CN on BTA6. The proportion of genetic variance for β-CN explained by known genotypes varied from 1.7% for *κ-CN* genotypes to 26.6% for *β-CN* genotypes (Table 8).

κ-CN. For κ-CN, SNPs on BTA1, 6, 11, 13, 20, 21, 24, and 29 (Fig. 1) and one unmapped SNP were significantly associated (FDR < 0.01). On BTA1 and 24, the SNP significantly associated with κ-CN contained fewer than 10 animals in one of the genotype classes. Setting the genotype of these animals to missing and rerunning the mixed model resulted in elimination of the significance of the SNP association on BTA1 (*P* value from 2.8E-05 to 1.1E-02) and on BTA24 (*P* value from 1.39E-05 to 1.61E-04). On BTA20, no SNPs were significantly associated with κ-CN after the running of the mixed model. The remaining SNPs significantly associated with κ-CN were located on BTA6, 11, 13, 21, and 29 (Table 4) and the null chromosome (Table 5). The proportion of genetic variance explained by the SNP most significantly associated with κ-CN on these six chromosomes varied from 1.2% on BTA21 to 16.4% on BTA6 (Table 6). The sum of the proportion of the genetic variance explained by these six SNPs was 27.0%, which was similar to the

Table 4 SNPs with the most significant association with the six major milk proteins, casein index, protein percentage, and protein yield with the position, the name, and the $-\log_{10}(P$ value) for each trait in each region using the mixed model in the association study, and the false discovery rate (FDR)[1,2]

Region	Trait	Position (Mbp)	SNP_Name	-log10 (*P* value)	FDR
1_1	α-LA	109.7	ULGR_BTA-120188	5.96	1.62E-03
1_2	α_{S2}-CN	149.2	ARS-BFGL-NGS-8140	6.30	2.85E-05
5_1	α-LA	34.4	ULGR_SNP_U63109_1966[3]	21.45	4.39E-13
5_2	Protein (%)	78.5	ULGR_rs29012209	4.45	1.25E-02
5_3	Protein (%)	98.5	ULGR_BTA.15561	4.72	1.98E-03
6	α_{S2}-CN	83.6	ULGR_BTC-053514	38.09	5.34E-36
6	α_{S1}-CN	88.1	ULGR_BTC-043582	26.87	5.43E-20
6	β-CN	88.3	ULGR_BTC-060550[4]	100.69	1.66E-105
6	Casein index	88.5	ULGR_SNP_X14908_5345[5]	10.19	5.79E-10
6	Protein (%)	88.5	ULGR_SNP_X14908_5345	11.28	3.08E-06
6	β-LG	88.5	ULGR_SNP_X14908_5345	6.31	6.56E-06
6	κ-CN	88.5	ULGR_SNP_X14908_5345	63.88	4.12E-57
6	α-LA	88.5	ULGR_rs29024684	16.36	6.92E-12
9_2	α_{S2}-CN	80.6	ULGR_AAFC03001453_132183	4.40	2.79E-03
10_1	Protein (%)	51.6	ULGR_AAFC03042309_74455	4.60	3.55E-04
10_1	α_{S2}-CN	52.4	ARS-BFGL-NGS-91094	4.23	2.75E-02
10_2	α_{S2}-CN	91.8	ULGR_BTA-109153	4.42	3.85E-04
11	Casein index	107.2	ULGR_SNP_X14710_1740[6]	134.37	7.61E-304
11	α_{S1}-CN	107.2	ULGR_SNP_X14710_1740	8.04	1.88E-05
11	α_{S2}-CN	107.2	ULGR_SNP_X14710_1740	9.00	1.41E-05
11	β-CN	107.2	ULGR_SNP_X14710_1740	9.10	2.40E-07

Table 4 Continued

Region	Trait	Position (Mbp)	SNP_Name	-log10 (*P* value)	FDR
11	β-LG	107.2	ULGR_SNP_X14710_1740	153.23	7.63E-304
11	κ-CN	107.2	ULGR_SNP_X14710_1740	7.91	2.92E-06
13_1	α_{S1}-CN	38.2	BTA-32346-no-rs	4.65	3.60E-03
13_2	κ-CN	60.5	ULGR_BTA-33109	5.47	1.47E-03
14	α_{S1}-CN	0.4	ULGR_SNP_AJ318490_1c[7]	10.19	2.35E-06
14	α_{S2}-CN	0.4	ULGR_SNP_AJ318490_1c	15.23	5.08E-12
14	Protein (%)	0.4	ULGR_SNP_AJ318490_1c	45.23	5.41E-42
15	Protein (%)	51.9	ARS-BFGL-NGS-107234	8.21	1.08E-06
17_2	α_{S2}-CN	19.4	ULGR_AAFC03022572_96396	4.26	3.51E-03
20	Protein (%)	38.3	ULGR_BTA-50418	6.88	3.83E-04
21	κ-CN	47.1	BTB-00821654	4.44	6.88E-03
24	β-LG	35.7	ULGR_rs29016076	5.01	9.73E-03
26	α-LA	33.1	ULGR_BTA-61176	5.08	6.17E-03
29	α_{S2}-CN	45.2	ULGR_SNP_CAPN1-AF248054-7813	4.51	3.71E-03
29	Protein (%)	45.5	ULGR_rs29026584	5.08	5.59E-04
29	κ-CN	45.9	ULGR_BTA-65731	5.87	3.18E-03

[1]FDR was calculated based on the general linear model in the association study.
[2]On some chromosomes, three regions were defined: _1 is the first region, _2 is the second region, and _3 is the third region on the chromosome.
[3]SNP is located within the promoter region of the *α-LA* gene.
[4]SNP is located in intron 7 of the *β-CN* gene and is in full linkage disequilibrium with an SNP in exon 7, which causes the protein variants A_1 and A_2 for β-CN.
[5]SNP causes the protein variants A and B for κ-CN.
[6]SNP is located in the promoter of the *β-LG* gene and is in full linkage disequilibrium with another SNP. Both SNPs gave similar results and yield the protein variants A and B for β-LG.
[7]SNP is in full linkage disequilibrium with another SNP. Both SNPs gave similar results and are located in the *DGAT1* gene.

proportion of the genetic variance (24.5%) explained by the multiple SNP analysis (Table 7). Accounting for known *κ-CN* genotypes on BTA6 resulted in decreased significance for SNPs previously associated with κ-CN on BTA6. The proportion of genetic variance of κ-CN explained by known genotypes varied from 2.0% for *β-LG* genotypes to 16.3% for *κ-CN* genotypes (Table 8).

Table 5 Unmapped SNPs with significant association with all traits, except with β-CN and protein percentage with the name, the $-\log_{10}(P$ value), the false discovery rate (FDR),[1] and the proportion of the genetic variance[2] (Var_{SNP}) for each trait.

Trait	SNP_Name	-log10 (*P* value)	FDR	Var_{SNP}
α_{S1}-CN	ULGN_SNP_AJ318490_2	9.70	5.89E-06	2.6
α_{S2}-CN	ULGN_SNP_AJ318490_2	14.90	1.13E-11	4.1
α_{S2}-CN	ULGR_MARC_31463_612	9.91	1.72E-07	2.9
α_{S2}-CN	ULGR_rs29024688	5.59	3.07E-03	1.5
α_{S2}-CN	ULGR_rs29017638	4.67	6.33E-03	1.3
κ-CN	ULGR_MARC_31463_612	5.93	2.48E-04	1.8
α-LA	BTA.117471.no.rs	7.34	1.65E-03	2.1
β-LG	ULGR_MARC_12075_173	27.33	5.52E-22	7.2
β-LG	ULGR_rs29018273	19.01	3.84E-13	5.3
β-LG	ULGR_BTA.15485	13.26	3.10E-10	3.7
Casein index	ULGR_MARC_12075_173	23.10	1.73E-18	6.4
Casein index	ULGR_rs29018273	15.00	2.52E-10	4.4
Casein index	ULGR_BTA.15485	13.17	4.87E-10	3.9
Protein (%)	ULGN_SNP_AJ318490_2	45.30	6.38E-42	12.5
Protein (%)	ULGR_rs29024688	19.38	6.62E-17	5.4
Protein (%)	ULGR_BTC.049604	7.84	2.99E-06	2.3

[1]FDR was calculated based on the general linear model in the association study.
[2]This value is expressed as the phenotypic variance after adjusting for the systematic environmental effects: day of lactation, age at first calving, season of calving, and herd.

1.3% on BTA26 to 6.2% on BTA5 (Table 6). The sum of the proportion of the genetic variance explained by these five SNPs was 13.6%, which was similar

α-LA. For α-LA, SNPs on BTA1, 5, 6, 9, 16, 17, and 26 (Fig. 1) and one unmapped SNP were significantly associated (FDR < 0.01). On BTA9, 16, and 17, no SNPs were significantly associated with α-LA after the running of the mixed model. The remaining SNPs significantly associated with α-LA were located on BTA1, 5, 6, and 26 (Table 4) and the null chromosome (Table 5). The proportion of genetic variance explained by the SNP most significantly associated with α-LA on these five chromosomes varied from to the proportion of the genetic variance (12.0%) explained by the multiple SNP analysis (Table 7). The proportion of genetic variance of α-LA explained by known genotypes varied from 2.9% for *β-CN* genotypes to 4.4% for *κ-CN* genotypes (Table 8).

β-LG. For β-LG, SNPs on BTA2, 6, 11, 20, and 24 (Fig. 1) and three unmapped SNPs were significantly associated (FDR < 0.01). On BTA2, the SNP significantly associated with β-LG had a minor allele frequency <2%, and on BTA20, no SNPs were significantly associated with β-LG after the running of the mixed model. The remaining SNPs significantly associated with β-LG were located on BTA6, 11, and 24 (Table 4) and the null chromosome (Table 5). The proportion of genetic variance explained by the SNP most significantly associated with β-LG on these four chromosomes varied from 1.3% on BTA24 to 65.8% on BTA11 (Table 6). Including these six SNPs (three SNPs on the null chromosome) simultaneously in the mixed model resulted in elimination of the association significance of the three unmapped SNPs. The sum of the proportion of the genetic variance explained by the remaining three SNPs was 68.8%, which was similar to the proportion of the genetic variance (65.5%) explained by the multiple SNP analysis (Table 7). Accounting for known *β-LG* genotypes on BTA11 resulted in a decrease in significance of SNPs associated with β-LG on BTA11. The proportion of genetic variance of β-LG explained by known genotypes varied from 1.7% for *κ-CN* genotypes to 64.9% for *β-LG* genotypes (Table 8).

Discussion

This study reports on associations between SNPs and the major milk proteins (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG), casein index, protein percentage, and protein yield of dairy cattle. This work is, to our

Table 6 The proportion of genetic variance[1] explained by the SNP most significantly associated with the six major milk proteins, casein index, protein percentage, and protein yield for each region[2]

Region	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN	α-LA	β-LG	Casein index	Protein (%)
1_1					1.6			
1_2		1.7						
5_1					6.2			
5_2								1.0
5_3								1.4
6	7.7	11.1	24.7	16.4	4.5	1.6	2.8	3.2
9_2		1.2						
10_1		1.2						2.2
10_2		1.3						
11	2.3	2.6	3.0	2.4		65.8	60.6	
13_1	1.2							
13_2				1.6				
14	2.8	4.2						12.5
15								2.7
17_2		1.3						
20								2.1
21				1.2				
24						1.3		
26					1.3			
29		1.1		3.6				1.5

[1]This value is expressed as the phenotypic variance after adjusting for the systematic environmental effects: day of lactation, age at first calving, season of calving, and herd.

[2]On some chromosomes, three regions were defined: _1 is the first region, _2 is the second region, and _3 is the third region on the chromosome.

knowledge, the first to describe the results of a genome-wide association study for milk protein composition.

The whole genome association was performed using a two-step procedure. In the first step, a general linear model was used, and in the second step, a mixed model was used. The *P* values of the SNPs for each trait using the general linear model were on average lower than those of the SNPs for each trait using the mixed model, except for β-CN and κ-CN. The average difference in *P* values between the general linear model and the mixed model ranged from -2.73E-02 for β-LG to 4.66E-03 for κ-CN, and the correlation ranged from 0.85 for α-LA to 0.92 for κ-CN. This result suggests that the use of a mixed model in the association study decreased the number of SNPs significantly associated with milk protein composition. This finding is in agreement with Kenney *et al.* (1992) and Yu *et al.* (2005), who both showed that the use of a mixed model will decrease the number of false positives.

SNPs significantly associated (FDR < 0.01) with milk protein composition were found on 15 bovine autosomes (BTA1, 5, 6, 9, 10, 11, 13, 14, 15, 17, 20, 21, 24, 26, and 29), which is in agreement with our previously reported linkage study performed on a subset of the same population (Schopen *et al.*, 2009b), except for BTA20 and 21. The identification of these two new regions on BTA20 and 21 illustrates the increased power that we achieved in this association study (because of the inclusion of around 1000 additional animals from roughly 50 families) compared to the linkage study. On BTA20, the SNPs identified as being significantly associated with protein percentage are in agreement with previously reported QTLs affecting protein percentage on BTA20 (e.g., Georges *et al.*, 1995; Blott *et al.*, 2003; Boichard *et al.*, 2003).

The inclusion in the mixed model of the unmapped SNPs for a trait together with the SNP most significantly associated with the same trait in each region resulted in the elimination of significance of the unmapped SNPs for all traits, except for κ-CN and α-LA. This outcome suggests that the unmapped SNPs are located in regions that were defined in this study and that the unmapped SNPs are in linkage disequilibrium with at least one SNP located in the defined regions. Therefore, we performed a BLAST

Table 7 The proportion of genetic variance[1] explained by multiple significantly associated SNPs for one trait in different regions, with the sum of the single SNP analysis and the multiple SNP analysis[2].

Trait	Region	Sum[3]	Multiple SNP analysis[4]
α_{S1}-CN	6, 11, 13_1, 14	13.9	14.4
α_{S2}-CN	1_2, 6, 9_2, 10_1, 10_2, 11, 14, 17_2, 29	25.7	21.8
β-CN	6, 11	27.7	26.8
κ-CN	Null, 6, 11, 13_2, 21, 29	27.0	24.5
α-LA	Null, 1_1, 5_1, 6, 26	15.7	11.7
β-LG	6, 11, 24	68.8	65.5
Casein index	6, 11	63.4	61.1
Protein (%)	5_2, 5_3, 6, 10_1, 14, 15, 20, 29	26.5	25.9

[1]This value is expressed as the phenotypic variance after adjusting for the systematic environmental effects: day of lactation, age at first calving, season of calving, and herd.

[2]On some chromosomes, three regions were defined: _1 is the first region, _2 is the second region, and _3 is the third region on the chromosome.

[3]Sum is sum of the genetic variance explained by the SNP most significantly associated with each trait for each region.

[4]For multiple SNP analysis, the proportion of genetic variance explained by multiple SNPs was calculated by including the most significantly associated SNP for each trait for each region simultaneously as a fixed effect in the model for that trait.

analysis, comparing the 50K SNP chip to the new physical bovine map (UMD3.0). As a result, the nine unmapped SNPs could be mapped to positions on BTA5, 6, 11, and 14, which were located in the regions defined in this study on BTA5 (5_1), 6, 11, and 14.

BTA5

Region 5_1 on BTA5 starts at 13.5 Mbp and ends at 42.9 Mbp, and the SNP most significantly associated with α-LA was located at 34.4 Mbp in region 5_1, at the same position as the *α-LA* gene on BTA5. The SNP most significantly associated with α-LA is located within the promoter region of the *α-LA* gene; therefore, the *α-LA* gene is a good candidate gene for the

Table 8 The proportion of genetic variance[1] explained by known polymorphisms in the genes for β-CN and κ-CN on *Bos taurus* autosome (BTA)6 (BTA6_BCN and BTA6_KCN), β-LG on BTA11 (BTA11_BLG), and DGAT1 on BTA14 (BTA14_DGAT1) for the traits that were significantly associated with SNPs on BTA6, 11, and 14

Trait	BTA6_BCN	BTA6_KCN	BTA11_BLG	BTA14_DGAT
α_{S1}-CN	14.7	4.8	2.6	2.8
α_{S2}-CN	12.6	2.0	2.1	4.1
β-CN	26.6	1.7	3.0	-
κ-CN	9.7	16.3	2.0	-
α-LA	2.9	4.4	-	-
β-LG	3.0	1.7	64.9	-
Casein index	4.8	2.7	60.2	-
Protein (%)	2.6	2.8	-	12.2

[1]This value is expressed as the phenotypic variance after adjusting for the systematic environmental effects: day of lactation, age at first calving, season of calving, and herd.

- = no SNP was significantly associated with this trait.

detected association of α-LA on BTA5. The protein α-LA is one of the two proteins involved in lactose synthesis, and a high concentration of α-LA is necessary to ensure maximum synthesis of lactose (Fitzgerald *et al.*, 1970). Menzies *et al.* (2009a) reported that lactose, cultured in mammary explants, increased with increased *α-LA* gene expression. An additional analysis, therefore, was performed for lactose percentage on BTA5. Results showed that 18 SNPs were associated ($P < 0.01$) with lactose percentage. The SNP most associated (P = 1.74E-04) with lactose percentage was located at 33.0 Mbp and had a significant effect on lactose percentage (0.0352% less lactose for AA animals compared to GG animals). This SNP was located close to the *α-LA* gene and to the SNP most significantly associated with the α-LA fraction. Therefore, these results confirm the results from Menzies *et al.* (2009a), who reported that lactose is related to *α-LA* gene expression. In addition to the *α-LA* gene, two other possible candidate genes, the *lysosyme* gene and *Socs2* gene, are located close to the SNP most significantly associated with α-LA at 34.4 Mbp on BTA5. The *lysosyme*

gene is located at 48 Mbp on BTA5 and has strong similarities with the sequence of the *α-LA* gene (Kumagai *et al.*, 1992; Qasba and Kumar, 1997). *Socs2* is a gene belonging to the group of prolactin-regulated genes and is located at 26 Mbp on BTA5. A *Socs2* deficiency in mice results in recovery of α-CN and β-CN production (Harris *et al.*, 2009), which suggests that *Socs2* has an effect on milk protein composition and might be a candidate gene for the SNPs significantly associated with α-LA on BTA5.

BTA6

Region 6 on BTA6 starts at 61.1 Mbp and ends at 97.7 Mbp, and SNPs located within this region were significantly associated with all traits except for protein yield. For β-CN, Fig. 1 shows two peaks; one around 88 Mbp (casein locus) and one around 42 Mbp. The *osteopontin* (*OPN*) gene and the *ABCG2* gene are located around 38 Mbp on BTA6, possibly suggesting that *OPN* or *ABCG2* might be a candidate gene for the first peak for β-CN and the casein locus for the second peak for β-CN on BTA6. An additional analysis, therefore, was performed in which the SNP most significantly associated with β-CN in the first and in the second peaks was simultaneously included in the mixed model. This approach eliminated the significance of the SNP that was previously most significantly associated with β-CN in the first peak (*P* value from 6.57E-22 to 0.07). The result suggests that the SNP most significantly associated with β-CN in the first peak and in the second peak might be in linkage disequilibrium and that there is only one QTL for β-CN, around 88 Mbp on BTA6.

BTA14

Region 14 on BTA14 starts at 0.0 Mbp and ends at 13.9 Mbp, and the SNP most significantly associated with α_{S1}-CN, α_{S2}-CN, and protein percentage was located at 0.4 Mbp, at the same position as the *DGAT1* gene on BTA14. Accounting for known *DGAT1* genotypes eliminated the significance for SNPs previously associated with α_{S1}-CN, α_{S2}-CN, and protein percentage. This result suggests that *DGAT1* affects protein percentage and, surprisingly, also milk protein composition, influencing α_{S1}-CN and α_{S2}-CN. The effect of *DGAT1* on milk protein composition showed that the A allele of *DGAT1* was associated with higher α_{S1}-CN (0.7464

ww%) and lower α_{S2}-CN (-0.7431 ww%) and protein percentage (-0.2561%).

BTA15

Region 15 on BTA15 starts at 41.4 Mbp and ends at 61.6 Mbp, and the SNP most significantly associated with protein percentage was located at 51.9 Mbp on BTA15, close to the *Elf5* gene and *FOLR1* gene. The *Elf5* gene is a single transcription factor located at 65 Mbp on BTA15. Harris *et al.* (2009) showed that an *Elf5* deficiency in mice (partly) rescued lactation failure induced by prolactin. Moreover, Sheehy (2008) showed that an inhibition of the *Elf5* gene decreased levels of Elf5 mRNA, consequently increasing expression of β-CN and κ-CN by twofold. Furthermore, *Elf5* is a mechanism by which insulin regulates milk protein synthesis (Menzies *et al.*, 2009a). Next to *Elf5*, insulin also facilitates the *FOLR1* gene, which is located at 51 Mbp on BTA15; expression of the *FOLR1* gene increases rapidly during lactation and regulates milk protein synthesis in mammary gland (Menzies *et al.*, 2009b). Although SNPs were not significantly associated with the six major milk proteins on BTA15, some SNPs were associated ($P < 0.01$) with the six major milk proteins around 54 Mbp on BTA15 (data not shown). *Elf5* and *FOLR1*, therefore, might be possible candidate genes for the detected SNPs significantly associated with protein percentage on BTA15.

Regulation of the six major milk proteins

The results of the whole genome association study showed that the six major milk proteins are associated with SNPs in more than one chromosomal region. For the casein fractions, the number of chromosomal regions with SNPs exhibiting a significant association ranged from two for β-CN to nine for α_{S2}-casein. Although the casein genes are sequentially arranged on BTA6, the difference in regions with SNPs significantly associated with the four casein fractions suggest that regulation of the expression of the casein genes is complex. In addition, some regions had significantly associated SNPs that were unique for each of the four casein fractions. For the whey proteins, the number of chromosomal regions with SNPs significantly associated with α-LA was four; for β-LG, it was three. In

addition, some regions had significantly associated SNPs that were unique for α-LA (regions 1_1, 5_1, and 26) and β-LG (region 24).

The different number of chromosomal regions and the unique regions associated with a specific milk protein suggest that some regions of the bovine genome contain genes that are more involved in casein composition and some regions contain genes that are more involved in whey composition. These results showed that the regulation of milk protein synthesis is complex and that the number of genes to be considered as candidates involved in this regulation is therefore greater. A better understanding of the regulation of milk protein synthesis, even by further clarification of the detected regions to identify the actual mutation, might be achieved by studying haplotypes.

Conclusions

In total, 15 bovine autosomes contained SNPs that were significantly associated with milk protein composition. One chromosomal region on BTA6 contained SNPs that were significantly associated with all six major milk proteins. For the other chromosomal regions, the number of major milk proteins significantly associated with SNPs differed. This variation in the number of chromosomal regions with SNPs significantly associated with one of the six major milk proteins implies that next to the “master regulator” on BTA6 also other regions are important. The results suggest instead that different genes are involved in the regulation of what appears to be the complex process of milk protein synthesis.

Acknowledgements

This study is part of the Milk Genomics Initiative, funded by Wageningen University, NZO (Dutch Dairy Association), CRV (cooperative cattle improvement organization), and the Dutch technology foundation, STW. The authors thank the owners of the herds for their help in collecting the data and Sylvia Kinders for her assistance in generating the genotypes.

References

Bennewitz J., N. Reinsch, S. Paul, C. Looft, B. Kaupe, C. Weimann, G. Erhardt, G. Thaller, Ch. Kühn, M. Schwerin, H. Thomsen, F.

Reinhardt, R. Reents, and E. Kalm. 2004. The DGAT1 k232a mutation is not solely responsible for the milk production quantitative trait locus on the bovine chromosome 14. J. Dairy Sci. 87: 431–442.

Blott S., J.-J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, D. Johnson, L. Karim, P. Simon, R. Snell, R. Spelman, J. Wong, J. Vilkki, M. Georges, F. Farnir, and W. Coppieters. 2003. Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163: 253–266.

Bobe G., D.C. Beitz, A.E. Freeman, and G.L. Lindberg. 1999. Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. J. Dairy Sci. 82: 2797–2804.

Boichard D., C. Grohs, F. Bourgeois, F. Cerqueiria, R. Faugeras, A. Neau, R. Rupp, Y. Amigues, M. Y. Boscher, and H. Levéziel. 2007. Detection of genes influencing economic traits in three French dairy cattle breeds. Genet. Sel. Evol. 35: 77–101.

Charlier C., W. Coppieters, F. Rollin, D. Desmecht, J.S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, J.-C. Frennet, R. Hanset, X. Hubin, C. Jorgensen, L. Karim, M. Kent, K. Harvey, B.R. Pearce, P. Simon, N. Tama, H. Nie, S. Vandeputte, S. Lien, M. Longeri, M. Fredholm, R.J. Harvey, and M. Georges. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat. Genet. 40: 449–454.

Fitzgerald D.K., U. Brodbeck, I. Kiyosawa, R. Mawal, B. Colvin, and K.E. Ebner. 1970. α-Lactalbumin and the lactose synthetase reaction. J. Biol. Chem. 245: 2103–2108.

Georges M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto, A.T. Pasquino, L.S. Sargeant, A. Sorensen, M.R. Steele, X. Zhao, J.E. Womack, and I. Hoeschele. 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics 139: 907–920.

Gilmour A.R., B.J. Gogel, B.R. Cullis, S.J. Welham, and R. Thompson. 2002. ASReml user guide. Release 1.0. VSN International Ltd., Hemel Hempstead, UK.

González J.R., L. Armengol, X. Solé, E. Guinó, J.M. Mercader, X. Estivill, and V. Moreno. 2007. SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23: 644-650.

Harris J., P.M. Stanford, K. Sutherland, S.R. Oakes, M.J. Naylor, F.G. Robertson, K.D. Blazek, M. Kazlauskas, H.N. Hilton, S. Wittlin, W.A. Alexander, G.J. Lindeman, J.E. Visvader, and C.J. Ormandy. 2006. Socs2 and Elf5 Mediate Prolactin-Induced Mammary Gland Development Mol. Endocrinol. 20: 1177–1187.

Heck J.M.L., C. Olieman, A. Schennink, H.J.F. van Valenberg, M.H.P.W. Visker, R.C.R. Meuldijk, and A.C.M. van Hooijdonk. 2008. Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis. Int. Dairy J. 18: 548–555.

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

Kennedy B.W., Q. Quinton, and J.A.M. van Arendonk. 1992. Estimation of effects of single genes on quantitative traits. J. Dairy Sci. 70: 2000–2012.

Kumagai I., S. Takeda, and K. Miura. 1992. Functional conversion of the homologous proteins α-lactalbumin and lysozyme by exon change. Proc. Natl. Acad. Sci. USA 89: 5887–5891.

Liu Y., X. Qin, X.-Z. H. Song, H. Jiang, Y. Shen, K. J. Durbin, S. Lien, M. P. Kent, M. Sodeland, Y. Ren, L. Zhang, E. Sodergren, P. Havlak, K.C. Worley, G.M. Weinstock, and R.A. Gibbs. 2009. *Bos taurus* genome assembly. BMC Genomics 10: 180–191.

Menzies K.K., C. Lefèvre, J.A. Sharp, K.L. Macmillan, P.A. Sheehy, and K.R. Nicholas. 2009b. A novel approach identified the FOLR1 gene, a putative regulator of milk protein synthesis. Mamm. Genome 20: 498–503.

Menzies K.K., C. Lefèvre, K.L. Macmillan, and K.R. Nicholas. 2009a. Insulin regulates milk protein synthesis at multiple levels in the bovine mammary gland. Funct. Integr. Genom. 9: 197–217.

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

Qasba P., and S. Kumar. 1997. Molecular divergence of lysozymes and α-lactalbumin. Clin. Rev. Biochem. Mol. Biol. 32: 255–306.

Schopen G.C.B., J.M.L. Heck, H. Bovenhuis, M.H.P.W. Visker, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. J. Dairy Sci. 92: 1182–1191.

Schopen G.C.B., P.D. Koks, J.A.M. van Arendonk, H. Bovenhuis, and M.H.P.W. Visker. 2009. Whole genome scan to detect quantitative trait loci for bovine milk protein composition. Anim. Genet. 40: 524–537.

Storey J.D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA 100: 9440–9445.

Sheehy P. 2008. Elf5 expression in the bovine mammary gland during the lactation cycle and its role in milk protein gene expression in vitro. 5th Int. Symposium on Milk Genomics & Human Milk in Sydney (Australia).

Visker M.H.P.W, B. Dibbits, S. Kinders, H.J.F. van Valenberg, J.A.M. van Arendonk, and H. Bovenhuis. 2010. Effects of bovine β-casein protein variant I on milk production and milk protein composition. Submitted.

Yu J., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J. . Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

6

Single and multiple SNP genome wide association analysis in dairy cattle

G. C. B. Schopen, M.P.L. Calus, M. H. P. W. Visker, J.A.M. van Arendonk and H. Bovenhuis

Concept

Abstract

The objective of this study was to compare the SNPs showing the most significant effects, the location, and the fraction of variance explained by these SNPs between single SNP analysis and multiple SNP analysis in the Dutch Holstein-Friesian population for the relative concentrations of the six major milk proteins. In total, 1713 cows with genotypes and phenotypes were available. DNA was isolated from blood samples of cows. In total, 45,999 SNPs distributed across 29 bovine autosomes were used in the single and multiple SNP analyses. The same main four chromosomal regions on BTA5, 6, 11, and 14 were detected in the single and multiple SNP analysis. The proportion of genetic variance explained by each of the SNPs in the single SNP analysis was higher compared to the SNP with the highest posterior probability in the multiple SNP analysis, except for β-CN. For β-CN, the proportion of genetic variance explained by the most significantly associated SNP in the single SNP analysis explained 42.7 %, whereas the SNP with the highest posterior probability in the multiple SNP analysis explained only 1.88%. Summing up the proportion of genetic variance explained by adjacent SNPs next to the SNP with the highest posterior probability in the multiple SNP analysis, resulted in an increase of the genetic variance explained similar to the SNP most significantly associated in the single SNP analysis, except for β-CN. There was one additional region on BTA7 detected in the multiple SNP analysis. The number of SNPs with effects is considerably lower in the multiple SNP analysis as compared to the single SNP analysis. These results indicate that multiple SNP analysis result in higher power and in higher mapping precision to detect QTL as compared to single SNP analysis.

Introduction

Single SNP genome wide association analyses have been widely used (e.g. Aulchenko *et al.*, 2007; Kolbehdari *et al.*, 2008; Daetwyler *et al.*, 2008). The main advantage of a single SNP analysis is the ease of implementation and calculation. However, there are a number of drawbacks associated with the single SNP models. The first one is that the calculation of the phenotypic variance explained by the QTL is not straightforward which is partly due to differences in estimated residual variances across models. The second one is that there is no simultaneously adjustment for all the genetic variation that is captured by all SNPs which decreases the power to detect (smaller) QTL. The third one is the issue of multiple testing, meaning that without correcting the significance threshold, the number of false positive QTLs is high.

These difficulties can at least to some extent be overcome by fitting all SNPs simultaneously into a multiple SNP analysis. With multiple SNP analysis there is simultaneously adjustment for all the genetic variation that is captured by all SNPs. This will reduce the residual genetic variance, which is expected to result in higher power to detect other SNPs associated with the trait of interest. The increase in power is similar to the principles of multiple QTL mapping (e.g. Jansen, 1993; Zeng, 1994; De Koning *et al.*, 2001). The increase in power can lead to detection of new chromosomal regions associated with the trait of interest, which will not be detected in a single SNP analysis. Furthermore, two simulation studies (Sillanpää and Arjas, 1998, and Uleberg and Meuwissen, 2007) showed that the likelihood peaks became smaller when information from all QTL positions was used in multiple QTL analysis as compared to single QTL analysis.

Schopen *et al.* (2010) performed a genome wide association study for milk protein composition using 50K SNPs in 1713 Holstein-Friesian cows. They used single SNP analysis which resulted in many regions which affected one or more of the six major milk proteins. The use of real data to compare single SNP analysis with an analysis where thousands of SNPs across the whole genome are simultaneously used to detect QTL would be a good addition to the simulation studies, and such a comparison has not yet been performed. The objective of this study, therefore, was to compare the SNPs showing the most significant effects, the location, and the fraction of

variance explained by these SNPs between single SNP analysis and multiple SNP analysis in the Dutch Holstein-Friesian population for the relative concentrations of the six major milk proteins.

Materials and methods

Animals

This study is part of the Dutch Milk Genomics Initiative. Phenotypic data of 1,912 first lactation Holstein-Friesian cows were collected. The cows were daughters of five proven sires (873 cows), 50 test sires (848 cows) or 15 other proven sires (191 cows). The full pedigree of the cows was supplied by CRV (Arnhem, the Netherlands). Further details of the animals used in this study are provided by Schopen *et al.* (2009).

Phenotypes

Milk protein composition was determined by capillary zone electrophoresis (CZE), as described by Heck *et al.* (2008). Using CZE, we quantified α_{S1}-casein (α_{S1}-CN), α_{S2}-casein (α_{S2}-CN), β-casein (β-CN), κ-casein (κ-CN), α-lactalbumin (α-LA) and β-lactoglobulin (β-LG). All six major milk protein fractions were expressed as a percentage of the total protein fraction (ww%).

The mean, phenotypic variance and intraherd heritability for the relative concentrations of the six major milk proteins are given in Table 1.

Genotypes

DNA was isolated from blood samples of cows. A 50K SNP chip was designed by CRV and obtained from Illumina, and was used to genotype the animals with the Infinium assay (Illumina, San Diego, CA, USA). This assay resulted in 50,856 technically successful SNPs which were mapped using the bovine genome assembly (BTAU4.0, Liu *et al.*, 2009). In total, 4,857 SNPs (6% of all available SNPs) were excluded because one of the following criteria: percentage of missing genotypes across animals > 5%, minor allele frequency < 2%, monomorphic, gene calling score (provided by Beadstudio software, Illumina) < 0.20, gene train score (provided by

Beadstudio software, Illumina) < 0.55, deviation from Hardy-Weinberg equilibrium (Hardy-Weinberg χ2 value > 600, Hayes *et al.*, 2009), mapped

Table 1 The mean, phenotypic variance[1] (σ^2_p), and intraherd heritability (h^2) for the six major milk proteins[2] for 1912 first-lactation Dutch Holstein-Friesian cows.

Trait	Mean	σ^2_p	h^2
α_{S1}-casein	33.62	2.53	0.47
α_{S2}-casein	10.38	1.80	0.73
β-casein	27.17	2.13	0.26
κ-casein	4.03	0.30	0.63
α-lactalbumin	2.44	0.08	0.57
β-lactoglobulin	8.34	1.42	0.80

[1]Phenotypic variance after adjusting for systematic environmental effects: day of lactation, age at first calving, season of calving, and herd.
[2]Expressed as percentage of the total protein fraction (ww%).

on the X-chromosome or not mapped on any of the 29 autosomes (Table 2). Most SNPs were excluded because of a minor allele frequency lower than 2%. The final dataset for the SNP analysis resulted in 45,999 SNPs distributed across 29 bovine autosomes.
The dataset, which was used in the single and multiple SNP analysis consisted of 1,713 animals with both phenotypic and genotypic information.

Systematic environmental effects

To account for systematic environmental effects, the phenotypes of 1,912 cows were adjusted for day of lactation, age at first calving, season of calving, and herd. These systematic environmental effects were estimated using an animal model in ASReml (Gilmour *et al.* 2002) for all 1,912 cows with phenotypes, as described by Schopen *et al.* (2009). The adjusted phenotypes of cows were subsequently used for both the single and multiple SNP analyses.

Single SNP analysis

The single SNP analysis was performed using the SNPassoc package (González *et al.*, 2007) in R using the following general linear model:

$$Y_{ij} = Sire_i + SNP_j + e_{ij}, \quad [1]$$

where Y_{ij} was the phenotype adjusted for systematic environmental effects, $Sire_i$ was the fixed effect of sire i, SNP_j was the fixed effect of the jth class of the SNP, and e_{ij} was the random residual effect ($e_{ij} \sim N(0,\sigma^2_e)$).

Sire was included in the general linear model to account for family effects.

The SNPs most significantly associated with one of the six major milk proteins using the linear model were consecutively analyzed in ASReml, to account for all genetic relationships among animals, using the following animal model:

$$Y_{ij} = \mu + SNP_j + animal_i + e_{ij}, \quad [2]$$

where Y_{ij} was the phenotype adjusted for systematic environmental effects, μ was the overall mean, SNP_j was the fixed effect of the SNP, $animal_i$ was the random additive genetic effect of animal *i*, and e_{ij} was the random residual effect. The variance-covariance structure of the additive genetic effects was $Var(animal) = A\sigma^2_{a,}$ where A was a matrix of additive genetic relationships among individuals and σ^2_a was the additive genetic variance.

Table 2 Criteria for the SNP quality with the threshold and the number of SNPs excluded for each criterium.

Criterium	Threshold	Number of excluded SNPs
% of missing genotypes	>5%	541
Minor allele frequency	<2%	2724
Monomorphic		193
Gen Calling-score[1]	<0.20	10
Gen Train-score[1]	<0.55	75
Hardy-Weinberg χ2 values	>600	15
Mapped on X-chromosome		563
Not mapped on BTAU4.0		736
Total removed SNPs		4857

[1] provided by Beadstudio software (Illumina)

Multiple SNP analysis

The multiple SNP analysis was performed using the following model (Meuwissen and Goddard, 2004):

$$Y_i = \mu + \sum_{j=1}^{45999} (q_{ij1} + q_{ij2}) v_j + animal_i + e_i \qquad [3]$$

where Y_i was the phenotype adjusted for systematic environmental effects, μ was the overall mean, v_j was the scale parameter of the QTL effect of the SNP at putative QTL position j, q_{ij1} (q_{ij2}) was the size of the QTL effect for the paternal (maternal) allele of animal i at SNP j drawn from a standard normal distribution $N(0,1)$, animal was the random polygenic effect of animal i (Var(animal) = $A\sigma^2_a$, where A was a matrix of additive genetic relationships among individuals and σ^2_a was the residual polygenic variance), and e_i was the random residual effect ($e_i \sim N(0,\sigma^2_e)$). Per SNP, q_{ij1} and q_{ij2} had three categories: one category for each of the two segregating alleles, and one additional category in which all missing genotypes were combined.

The multiple SNP analysis was performed using a Markov chain Monte Carlo method using Gibbs sampling to obtain posterior estimates for all the effects in the model. The scale parameter of a putative QTL at SNP j, v_j, was sampled from a normal distribution $N(0,\sigma^2_v)$, if a QTL was present at SNP j, whereas v_j was sampled from $N(0,\sigma^2_v/100)$ if no QTL was present at SNP j. The variance of v_j, σ^2_v, was sampled from a scaled inverse chi-square distribution with a prior variance. This prior variance was calculated as the additive genetic variance, divided by 58, i.e. assuming 58 additive and independent QTL affecting the trait, across the 29 chromosomes. The number 58 reflects a prior assuming two QTL on each chromosome.

The presence of a QTL at SNP j was sampled from a Bernoulli distribution with probability equal to $\frac{P(v_j \mid \sigma^2_v) \times Pr_j}{P(v_j \mid \sigma^2_v) \times Pr_j + P(v_j \mid \sigma^2_v/100) \times (1 - Pr_j)}$, where $P(v_j \mid \sigma^2_v)$ is the probability of v_j from $N(0,\sigma^2_v)$, i.e. $\frac{1}{\sqrt{2\pi\sigma^2_v}} e^{-\frac{v_j^2}{2\sigma^2_v}}$, and Pr_j is the prior probability of the presence of a QTL at SNP j. More details on

the prior distributions and the full conditional distributions can be found in Meuwissen and Goddard (2004) and Calus *et al.* (2008). Visual inspection showed that there was mixing of the posterior probabilities. The Gibbs sampler, therefore, was run for all models for 30,000 iterations and 2,000 iterations were removed as burn-in. The posterior probability and estimates for allelic effects for each SNP used in this study was the average over the 28,000 post burn-in iterations.

Significance threshold

Significance threshold for the single SNP analysis was obtained by calculating the false discovery rate (FDR) based on the qvalue package (Storey & Tibshirani, 2003) in R. The FDR was calculated based on the p-values obtained from the single SNP analysis for all 45,999 SNPs for each trait separately. SNPs with a FDR < 0.05 were considered to be significantly associated with the trait.

For the multiple SNP analysis, a posterior probability level of > 0.05 was used as an arbitrary threshold to determine whether a SNP was associated with a trait.

Variance explained by SNP and linkage disequilibrium

The proportion of genetic variance explained by a SNP in the single SNP analysis was calculated from the estimated genotype effects obtained from ASReml and the observed genotype frequencies. In this way, the additive genetic variance and the dominance variance are taken into account.

For the multiple SNP analysis, the proportion of additive genetic variance explained by a SNP was calculated using the following formula (Falconer and MacKay, 1996):

$$Var_{SNP} = 2 * p * q * a^2 \qquad [4]$$

where *p* was the allele frequency of one allele of the SNP, *q* was the allele frequency of the other allele of the SNP and *a* was the allele substitution effect. The allele substitution effect was calculated by the difference in estimated effects of both alleles (v_j*q_{ij1} - v_j*q_{ij2}).

Using formula 3 it is assumed that there is no dominance at the SNP under consideration.

To determine the level of linkage disequilibrium (LD), the pairwise r^2 was calculated and plotted using Haploview version 4.1 (Barret *et al.*, 2005) for some regions.

Results

Identified SNPs in single and multiple SNP analysis

Single SNP analysis resulted in four main regions (FDR < 1.00E-05) with SNPs significantly associated with one or more of the six major milk proteins. These main regions were located on *Bos taurus* autosomes (BTA) 5, 6, 11 and 14. The SNPs most significantly (FDR < 0.05 in the single SNP analysis) associated within a region with each of the six major milk proteins are given in Table 3.

Multiple SNP analysis resulted in nearly the same four main regions containing SNPs with high posterior probabilities for the six major milk proteins (Table 3). On BTA6, SNPs had a posterior probability > 0.05 for all milk proteins, except for β-LG (Table 3). The posterior probability for β-LG was 0.042 (Table3). Besides the four main regions, multiple SNP analysis resulted in an additional chromosomal region on BTA7. Further, on BTA27, an association with α-LA was detected in the multiple SNP analysis which was not significant in the single SNP analysis.

Figure 1 shows the graphs of SNPs associated with each of the six major milk proteins on BTA6 for the single and multiple SNP analysis. For some chromosomal regions (e.g. on BTA6), the most significant SNP in the single SNP analysis was not always the same as the SNP with the highest posterior probability in the multiple SNP analysis. For example on BTA6, the SNP (ULGR_BTC-053514) most significantly associated with α_{S2}-CN in the single SNP analysis was located at 83.6 Mbp whereas the SNP (ULGR_BTC-060527) with the highest posterior probability in the multiple SNP analysis was located at 88.3 Mbp (Table 3). For α_{S2}-CN, the posterior probability of SNP ULGR_BTC-05351 was 2.10E-04 and the FDR of SNP ULGR_BTC-060527 was 0.43. The difference in position of the SNPs between single and multiple SNP analysis was highest for κ-CN and β-LG on BTA20 (Table 3). In the single SNP analysis the SNP (ULGR_BTA-50132) most significantly associated with κ-CN on BTA20 was located at

28.6 Mbp whereas in the multiple SNP analysis the SNP (ULGR_ BTA-27776) with highest posterior probability was located at 6.0 Mbp (Table 3). For κ-CN on BTA20, the posterior probability of SNP ULGR_BTA-50132 was 2.07E-03 and the FDR of SNP ULGR_BTA-27776 was 0.06.

Variance explained by SNP

The proportion of genetic variance explained by each of the SNPs with the strongest association with one of the six major milk proteins in the four main regions was higher in the single SNP analysis than in the multiple SNP analysis (Table 4). There is an enormous difference in genetic variance explained by the SNP with the strongest association with β-CN on BTA6 in the multiple SNP analysis (1.88%) compared to the single SNP analysis (42.70%). On BTA11, a similar result for β-CN was obtained; 5.11% in the multiple SNP analysis as compared to 0.42% in the single SNP analysis. In addition, the additive genetic variance explained by the SNP with the highest posterior probability for κ-CN on BTA7 (2.17E-02), BTA13 (1.31E-01), BTA20 (7.73E-03) and BTA27 (3.01E-03), for α-LA on BTA27 (1.15E-03), and for β-LG on BTA20 (5.11E-02) was very low (Table 4).

In the single SNP analysis each SNP gets the possibility to explain all of the genetic variance, whereas in the multiple SNP analysis, the genetic variance is divided across several SNPs due to LD of several SNPs with the QTL. Therefore, the proportions of additive genetic variance explained by each of the adjacent SNPs on the right and left side of the SNP with the highest posterior probability were added to the proportion of additive genetic variance explained by the SNP with the highest posterior probability. When the proportion of genetic variance of more adjacent SNPs on the right and left side were added up to the proportion of genetic variance explained by the SNP with the highest posterior probability, the proportion of additive genetic variance increased or stayed equal (Table 4). For example for α_{S2}-CN on BTA6, the proportion of additive genetic variance explained by the SNP with highest posterior probability was 3.40% and increased to 11.64% when the proportions of additive genetic variance of 50 adjacent SNPs (25 SNPs left and 25 SNP right from the SNP with the highest posterior probability) were added. However, for β-CN on BTA6 and

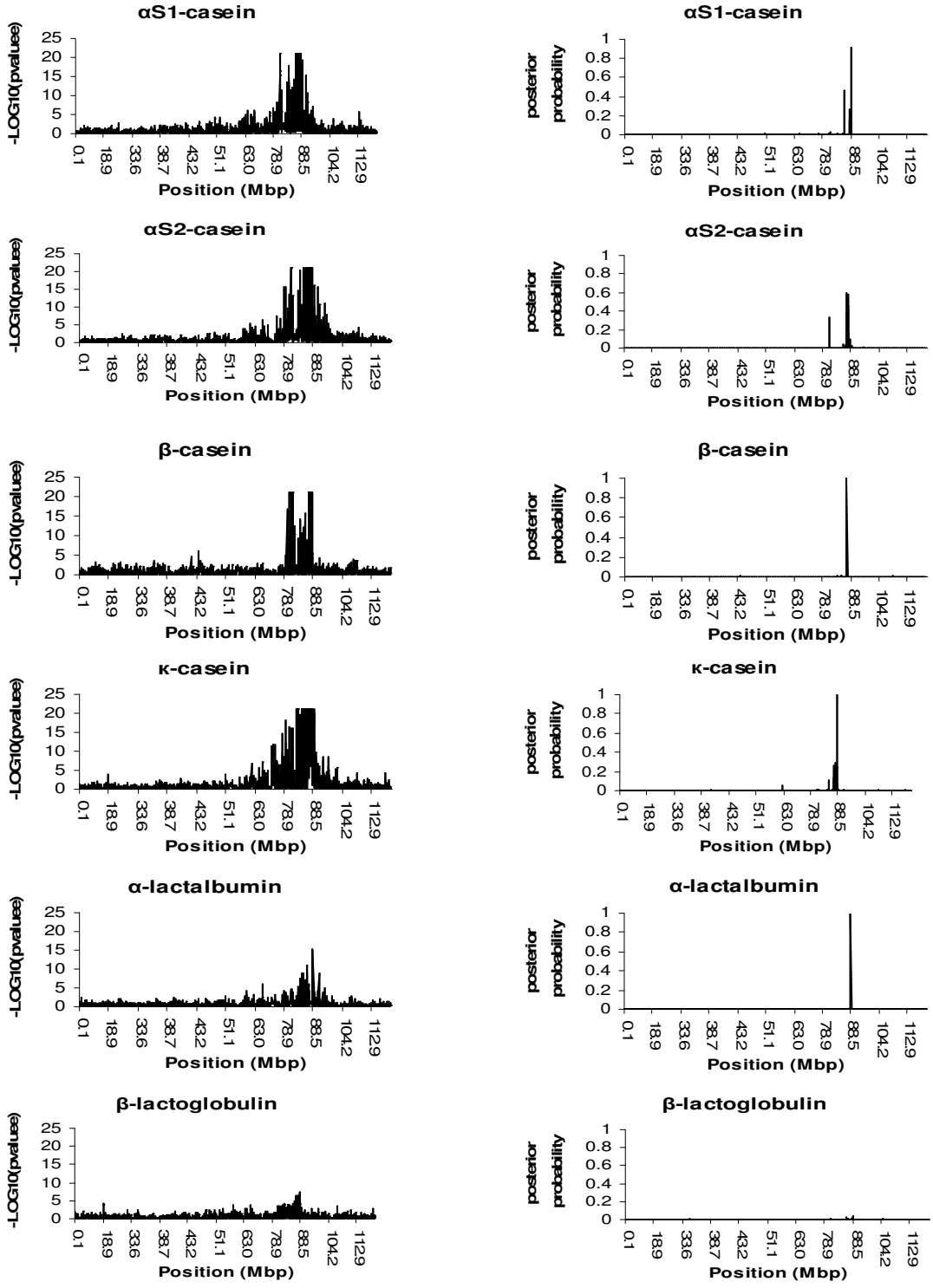

Figure 1 Graphical overview of the single SNP analysis[1] (left column) and the multiple SNP analysis (right column) for *Bos taurus* autosome 6 for each of the six major milk proteins. A $-\log_{10}(\text{pvalue}) > 21$ was set to 21 to retain an overview of the SNPs that exceeded the threshold level.

11, the enormous difference in proportion of additive genetic variance between single SNP analysis and the sum of proportion of adjacent SNPs in the multiple SNP analysis remained, even after adding up the variances of all SNPs for β-CN on BTA6 (1.89%) and BTA11 (0.67%).

Linkage disequilibrium

The multiple SNP analysis (Figure 1) indicates that there might be two QTLs for α_{S2}-CN; one at 83.6 Mbp and one at 88.3 Mbp. Therefore, pairwise LD between adjacent SNPs in the chromosomal region on BTA6 was calculated and given in Figure 2. In this chromosomal region on BTA6 there are blocks of SNPs which show higher LD (black spots in Figure 2A) than others. In Figure 2B, one of these blocks is enlarged. This enlarged block starts at 88.35 Mbp (SNP ULGR_BTC-060507) and ends at 88.43 Mbp (SNP ULGR_BTC-072885), and contained 21 SNPs. There is high pairwise LD between these 21 SNPs (Figure 2B), and within this region several SNPs were significantly associated with the six major milk proteins. The r^2 between the two SNPs located at 83.6 Mbp and 88.3 Mbp was 0.24.

Discussion

This study compares results from a single and multiple SNP genome wide association study to detect QTL for milk protein composition in Holstein Friesian cows. Until now, limited information on such a comparison using real data was available in the literature.

Identified SNPs

The same four main regions on BTA5, 6, 11 and 14 showing associations with the six major milk proteins were detected in the single and multiple SNP analyses. Additional associations on BTA7 and 27, however, were detected in the multiple SNP analysis compared to the single SNP analysis. The SNP with the highest posterior probability for κ-CN on BTA7 had a FDR of 0.44, and for β-LG on BTA27 had a FDR of 0.55 in the single SNP analysis (Table 4). This suggests that multiple SNP analysis has a greater power to detect QTL than single SNP analysis.

On BTA6, 11, 14 and 20, the SNP most significantly associated with a trait

Table 3 The SNP most significantly[1] associated with each of the six major milk proteins in the single SNP analysis and the SNP with the highest posterior probability[2] for each of the six major milk proteins in the multiple SNP analysis, with the position (Mbp), the name[3] (SNP_Name), the false discovery rate (FDR) for the single SNP analysis and the posterior probability for the multiple SNP analysis. *Bos taurus* autosomes (BTA).

		Detected SNPs					
		Single SNP analysis			**Multiple SNP analysis**		
BTA	**Trait**	**Position (Mbp)**	**SNP_Name**	**FDR**	**Position (Mbp)**	**SNP_Name**	**Posterior probability**
5	α-lactalbumin	34.4	SNP_U63109_1966	4.39E-13	34.4	SNP_U36109_1966	1.00
6	α_{S1}-casein	88.1	BTC-043582	5.43E-20	88.5	AAFC03044644_5880	0.92
6	α_{S2}-casein	83.6	BTC-053514	5.34E-36	88.3	BTC-060527	0.59
6	β-casein	88.3	BTC-060550	1.66E-105	88.4	BTC-060513	1.00
6	κ-casein	88.5	SNP_X14908_5345	4.12E-57	88.1	SNP_X14908_5345	1.00
6	α-lactalbumin	88.5	rs29024684	6.92E-12	88.5	rs29024684	1.00
6	β-lactoglobulin	88.5	SNP_X14908_5345	6.56E-06	-	SNP_X14908_5345	0.04
7	κ-casein	-	BTA-79042	4.42E-01	11.6	BTA-79042	0.31
11	α_{S1}-casein	107.2	SNP_X14710_1740	1.88E-05	107.2	SNP_X14710_3984	1.00
11	α_{S2}-casein	107.2	SNP_X14710_1740	1.41E-05	107.2	SNP_X14710_1740	1.00
11	β-casein	107.2	SNP_X14710_1740	2.40E-07	107.2	BTA-116267	0.44
11	κ-casein	107.2	SNP_X14710_1740	2.40E-07	107.2	SNP_X14710_1740	0.84
11	β-lactoglobulin	107.2	SNP_X14710_1740	7.63E-304	107.2	SNP_X14710_1740	1.00
13	κ-casein	60.5	BTA-33109	1.47E-03	60.5	BTA-33109	0.69
14	α_{S1}-casein	0.4	SNP_AJ318490_1c	2.35E-06	0.4	SNP_AJ318490_1c	1.00
14	α_{S2}-casein	0.4	SNP_AJ318490_1c	5.08E-12	0.4	SNP_AJ318490_1c	1.00
14	κ-casein	11.1	BTC-059786	4.67E-02	0.4	SNP_AJ318490_1c	0.29

Table 3 Continued

		Detected SNPs					
		Single SNP analysis		Multiple SNP analysis			
BTA	**Trait**	**Position (Mbp)**	**SNP_Name**	**FDR**	**Position (Mbp)**	**SNP_Name**	**Posterior probability**
20	κ-casein	28.6	BTA-50132	2.83E-03	6.0	BTA-27776	0.17
20	β-lactoglobulin	53.3	BTA-111508	4.15E-03	36.4	BTA-50240	0.31
27	κ-casein	42.0	BTA-98604	1.05E-02	42.0	BTA-98604	0.13
27	α-lactalbumin	-	AAFC03016002_47078	5.52E-01	12.6	AAFC03016002_47078	0.14

[1]The SNP was considered significant if FDR < 0.05 in the single SNP analysis.

[2] The posterior probability had to be > 0.05 in the multiple SNP analysis.

[3] In front of all SNP names 'ULGR_' should be placed, except for the SNP names starting with rs (e.g. rs29024684)

- means that no SNP was significantly associated

Table 4 The proportion of genetic variance (Var_{SNP}) explained by the SNPs (1) most significantly[1] associated with each of the six major milk proteins for the single SNP analysis, and the proportion of genetic variance explained by the SNPs (1) with the highest posterior probability[2] and 10, 20 and 50 adjacent[3] SNPs together with the length of the chromosome covered (Mbp) for each of the SNP with the highest posterior probability[3] for the six major milk proteins in the multiple SNP analysis. *Bos taurus* autosomes (BTA).

		Single	**Multiple**						
		Var_{SNP}	**Var_{SNP}**						
BTA	**Trait**	**1**	**1**	**10**	**(Mbp)**	**20**	**(Mbp)**	**50**	**(Mbp)**
5	α-lactalbumin	0.44	0.34	0.34	1.48	0.34	1.85	0.34	2.84
6	α_{S1}-casein	17.02	5.72	5.72	0.88	5.72	1.59	6.34	2.32
6	α_{S2}-casein	16.37	3.40	3.40	0.04	8.90	0.07	11.64	0.53
6	β-casein	42.70	1.88	1.88	0.04	1.88	0.07	1.88	0.33
6	κ-casein	4.11	4.70	4.72	0.26	4.73	0.64	4.86	1.95

Table 4 Continued

		Single	**Multiple**						
		Var_{SNP}	**Var_{SNP}**						
BTA	**Trait**	**1**	**1**	**10**	**(Mbp)**	**20**	**(Mbp)**	**50**	**(Mbp)**
6	α-lactalbumin	0.33	0.26	0.26	0.55	0.26	1.14	0.26	2.13
6	β-lactoglobulin	2.15	-	-	-	-	-	-	-
7	κ-casein	**-**	2.17E-02	0.03	1.08	0.03	1.76	0.03	3.17
11	α_{S1}-casein	4.80	4.14	4.14	0.58	4.14	1.09	4.14	2.46
11	α_{S2}-casein	3.87	2.96	2.96	0.56	2.96	1.04	2.96	2.58
11	β-casein	5.11	0.42	0.42	0.49	0.65	1.21	0.66	2.72
11	κ-casein	0.58	0.26	0.26	0.56	0.26	1.04	0.26	2.58
11	β-lactoglobulin	83.36	79.6	79.56	0.56	80.25	1.04	80.25	2.58
13	κ-casein	0.39	1.31E-01	0.13	0.66	0.13	1.17	0.13	3.11
14	α_{S1}-casein	6.02	3.77	3.77	0.64	3.77	0.73	3.77	1.18
14	α_{S2}-casein	6.14	4.16	4.16	0.64	4.16	0.73	4.16	1.18
14	κ-casein	-	0.02	0.02	0.64	0.02	0.73	0.02	1.18
20	κ-casein	0.27	7.72E-03	0.01	0.26	0.01	0.70	0.01	2.36
20	β-lactoglobulin	1.39	5.11E-02	0.05	0.62	0.05	1.31	0.05	2.98
27	κ-casein	-	3.01E-03	4.33E-03	0.60	4.35E-03	1.05	4.42E-03	2.61
27	α-lactalbumin	-	1.15E-03	1.18E-03	0.51	1.19E-03	1.27	1.22E-03	3.47

[1]The SNP was considered significant if FDR < 0.05 in the single SNP analysis.

[2] The posterior probability had to be > 0.05 in the multiple SNP analysis.

[3] The SNP with the highest posterior probability was located in the middle

- means that no SNP significantly associated.

in the single SNP analysis was not always the same SNP with the highest posterior probability in the multiple SNP analysis. The SNPs with the highest posterior probabilities for α_{S1}-CN and β-CN on BTA6, for β-CN on BTA11, and for κ-CN on BTA14, however, were also significant in the single SNP analysis (FDR was 1.58E-16 for α_{S1}-CN and 4.58E-40 β-CN on BTA6, 1.05E-4 for β-CN on BTA11, and 3.16E-02 for κ-CN on BTA14). The SNPs with the highest posterior probabilities for α_{S2}-CN on BTA6, and for κ-CN and β-LG on BTA20 were not significant in the single SNP analysis (FDR was 0.43 for α_{S2}-CN on BTA6, and 0.06 for κ-CN and 0.44 for κ-CN on BTA20). However, for κ-CN on BTA20, the SNP most significantly associated was at the border of significance (as determined in this study). For α_{S2}-CN on BTA6, the SNP most significantly associated in the single SNP analysis was located at 83.6 Mbp, whereas the SNP with the highest posterior probability in the multiple SNP analysis was located at 88.3 Mbp. The position of 88.3 Mbp is close to the casein cluster, especially to the *α_{S2}-CN* gene which is located at 88.4 Mbp on BTA6 (NCBI, Map viewer). There could, however, also be two QTL for α_{S2}-CN on BTA6. This is supported by the multiple SNP analysis (Figure 1) which indicates that there might be one QTL at 83.6 Mbp (ULGR_BTC-053514) and one QTL at 88.3 Mbp (ULGR_BTC-060527). The r^2 between the SNP at 83.6 Mbp and the SNP at 88.3 Mbp was 0.24. However, it is unlikely that the SNP at 83.6 Mbp is due to LD with the SNP at 88.3 Mbp, because in this region (83.6 – 88.3 Mbp) other SNPs have a higher LD value with the SNP at 88.3 Mbp (e.g. ULGR_BTC-043266 with r^2=0.82 and ULGR_BTC-043259 with r^2=0.82) than the SNP at 83.6 Mbp (Figure 3). In addition, pre-adjusting α_{S2}-CN for the SNP at 88.3 Mbp and rerun the multiple SNP model resulted in the remaining of the significance of the SNP (posterior probability = 0.29436) at 83.6 Mbp. However, when pre-adjusting α_{S2}-CN for the SNP at 83.6 Mbp and rerun the multiple SNP model resulted in the elimination of the significance of the SNP (posterior probability = 0.00121) at 88.3 Mbp This suggests that there is only 1 QTL for α_{S2}-CN on BTA6.

The region on BTA6 in which the SNPs, showing association with the six major milk proteins in the multiple SNP analysis, were located ranged from 88.1 Mbp to 88.5 Mbp and in this region there is high LD among the SNPs (Figure 2). Due to the high LD among SNPs in the region on BTA6, one

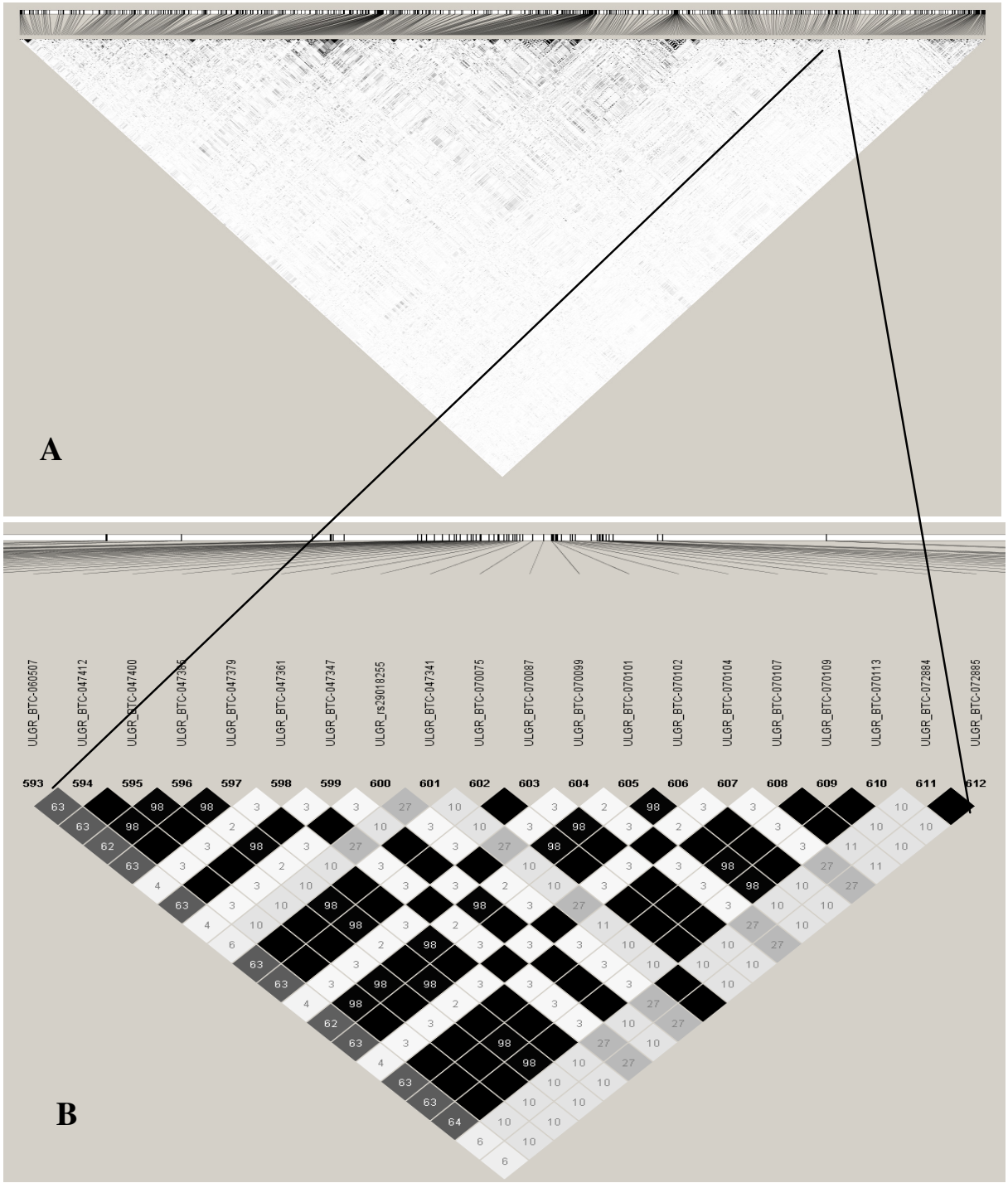

Figure 2 Graphical overview of pairwise LD for a region (54.5 Mbp – 104.3 Mbp) on *Bos taurus* autosome 6 (A) and for a region (88.3 – 88.4 Mbp) with the SNPs most significantly associated with the milk proteins on *Bos taurus* autosome 6 (B). Deeper black colours mean more LD between the SNPs; dark black means that two SNPs are in complete LD (r^2=1.0). The values in the graph are the r^2 values times 100.

SNP may explain variation which can not be explained by another linked SNPs. Once one SNP has a large effect in a region, the posterior probability of this SNP may then not change anymore in the following iterations. These results suggest that the LD in a chromosomal region might explain differences in position between the SNP most significantly associated in the single SNP analysis and the SNP with the highest posterior probability in the multiple SNP analysis. Note that for the identification of candidate genes, the location of the SNP most significantly associated or with the highest posterior probability is very important. A difference in the position may lead to wrong identification of candidate genes. Therefore, pointing out candidate genes will than still be a great challenge.

In the multiple SNP analysis, the genetic variance explained by one SNP can not be explained anymore by another SNP. This is not the case in the single SNP analysis. Therefore, the number of SNPs with effects is considerably lower in the multiple SNP analysis than the single SNP analysis. This suggests that the number of genes located within the chromosomal region of interest is lower by using multiple SNP analysis in stead of using single SNP analysis. The lower number of genes in a region is preferable because this might more easily lead to identification of possible candidate genes underlying the QTL.

Comparison between FDR and posterior probability

Differences in position of the SNP most significantly associated in the single SNP analysis and the SNP with the highest posterior probability in the multiple SNP analysis could be due to differences in threshold levels. In the single SNP analysis the FDR was used, whereas in the multiple SNP analysis the posterior probability was used. The FDR controls the expected proportion of false positives among all significant hypotheses of having a QTL (type I errors), whereas the posterior probability of having a QTL is based on the combination of the likelihood of the QTL and the prior probability (Meuwissen and Goddard, 2004). The results of both analyses using FDR (single) and posterior probability (multiple) are hard to compare. In this study, however, the FDR was calculated using the q-value package in R in which the q-value was defined as the minimum positive FDR at

which a hypothesis can be called significant (Storey & Tibshirani, 2003). Previous studies showed that the positive FDR (conditioned on at least one positive finding; Storey, 2003) and FDR (Efron *et al.*, 2001) can be written as a posterior probability.

When plotting the FDR of the SNPs against the posterior probabilities (Figure 3), SNPs with the lowest FDR in the single SNP analysis had the highest posterior probability in the multiple SNP analysis, in general. For α_{S2}-CN, κ-CN, α-LA and β-LG, there were some SNPs with a FDR of about 0.50 in the single SNP analysis whereas their posterior probability in the multiple SNP analysis was about 0.40. However, in general, an increase in the FDR of the SNPs resulted in an exponential decrease in the posterior probability of the SNPs. This suggests that the multiple SNP analysis does partly solved the problem of multiple testing.

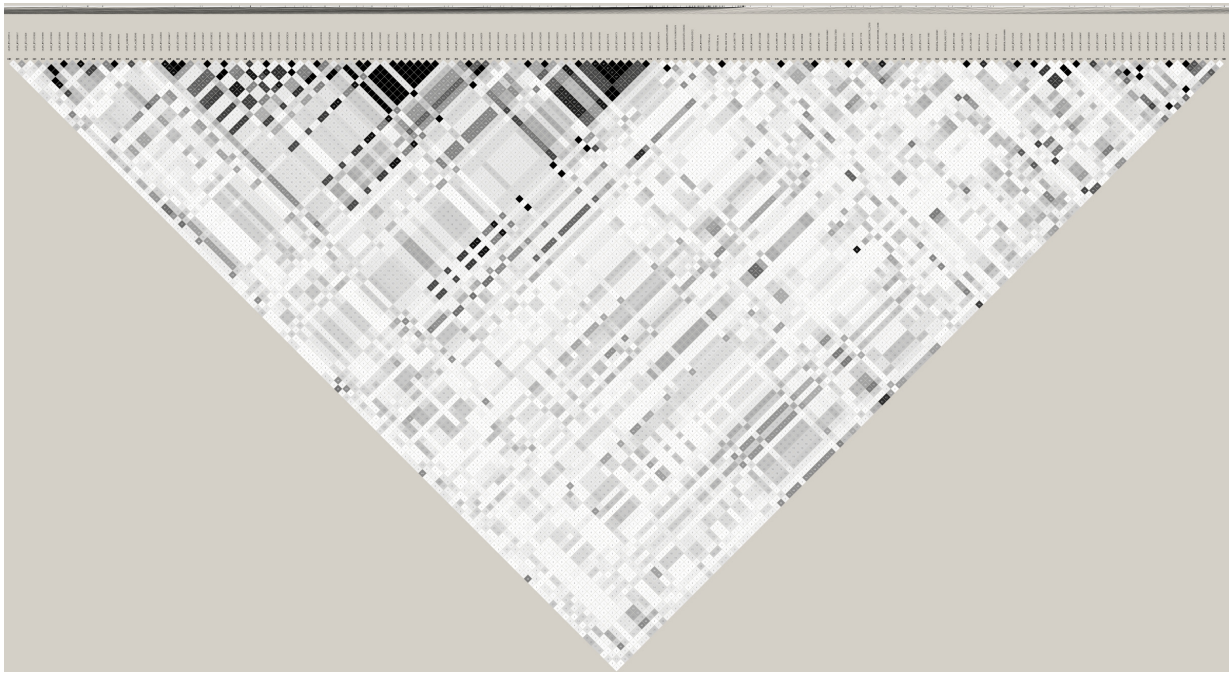

Figure 3 Graphical overview of pairwise LD for the region from 83.6 Mbp (ULGR_BTC-053514) to 88.3 Mbp (ULGR_BTC-060527) on *Bos taurus* autosome 6. Deeper black colours mean more LD between the SNPs; dark black means that two SNPs are in complete LD (r^2=1.0). The values in the graph are the r^2 values times 100.

Prior distribution

A prior QTL variance was specified for each trait. The prior distribution will not completely match the true QTL distribution. However, Verbyla *et al.* (2009) showed that this will not affect results from models assuming unequal variances across SNPs. In this study, the multiple SNP analysis was based on a Bayesian approach which also assumed unequal variances across SNPs. For β-LG, using different priors for the QTL distribution resulted each time in the same SNP (ULGR_ SNP_X14710_1740) on BTA11 with the highest posterior probability of 1. For the SNP ULGR_ SNP_X14710_1740 for β-LG on BTA11, the evidence in the data was already so large that the prior did not have any influence anymore. However, using different priors for the QTL distribution for β-LG on BTA20 resulted not always in the same SNP having the highest posterior probability. Using a QTL variance of 0.05 resulted in another SNP (ARS-BFGL-NGS-98321) with the highest probability for β-LG than the SNP (ULGR_BTA-50240) with the highest probability using a QTL variance of 1.32E-03 or 2.132E-02. However, SNP ARS-BFGL-NGS-98321 was only 2 Mbp away from SNP ULGR_BTA-50240. In all three analyses, SNP ULGR_BTA-50240 still had a posterior probability > 0.05. These results suggest that the choice of the prior distribution is more important when the data itself is less informative.

In addition, Meuwissen *et al.* (2001) reported that an uncorrected prior hardly affects the accuracy of estimated breeding values. This illustrates that when the interest is not on individual SNP effects, but on the sum of SNP effects, the choice of prior QTL distribution is also less important.

Variance explained by SNP

The proportion of genetic variance explained by a SNP most significantly associated with a trait in the single SNP analysis was higher than the proportion of additive genetic variance explained by the SNP with the highest posterior probability in the multiple SNP analysis. In particular for β-CN on BTA6: 42.70% in the single SNP analysis compared to 1.88% in the multiple SNP analysis. The difference in genetic variance explained by the SNP between single and multiple SNP analysis could be explained by important differences between these two analyses. One difference is that in

the single SNP analysis the additive genetic variance and the dominance variance were taken into account, whereas in the multiple SNP analysis only the additive genetic variance was taken into account. Heck *et al.* (2009) and Visker *et al.* (2010) showed that dominance is important in associations with *β-CN* variants. This might explain part of the difference in proportion of genetic variance between single and multiple SNP analysis. A second difference is that in the single SNP analysis each SNP gets the possibility to explain all of the genetic variance, whereas in the multiple

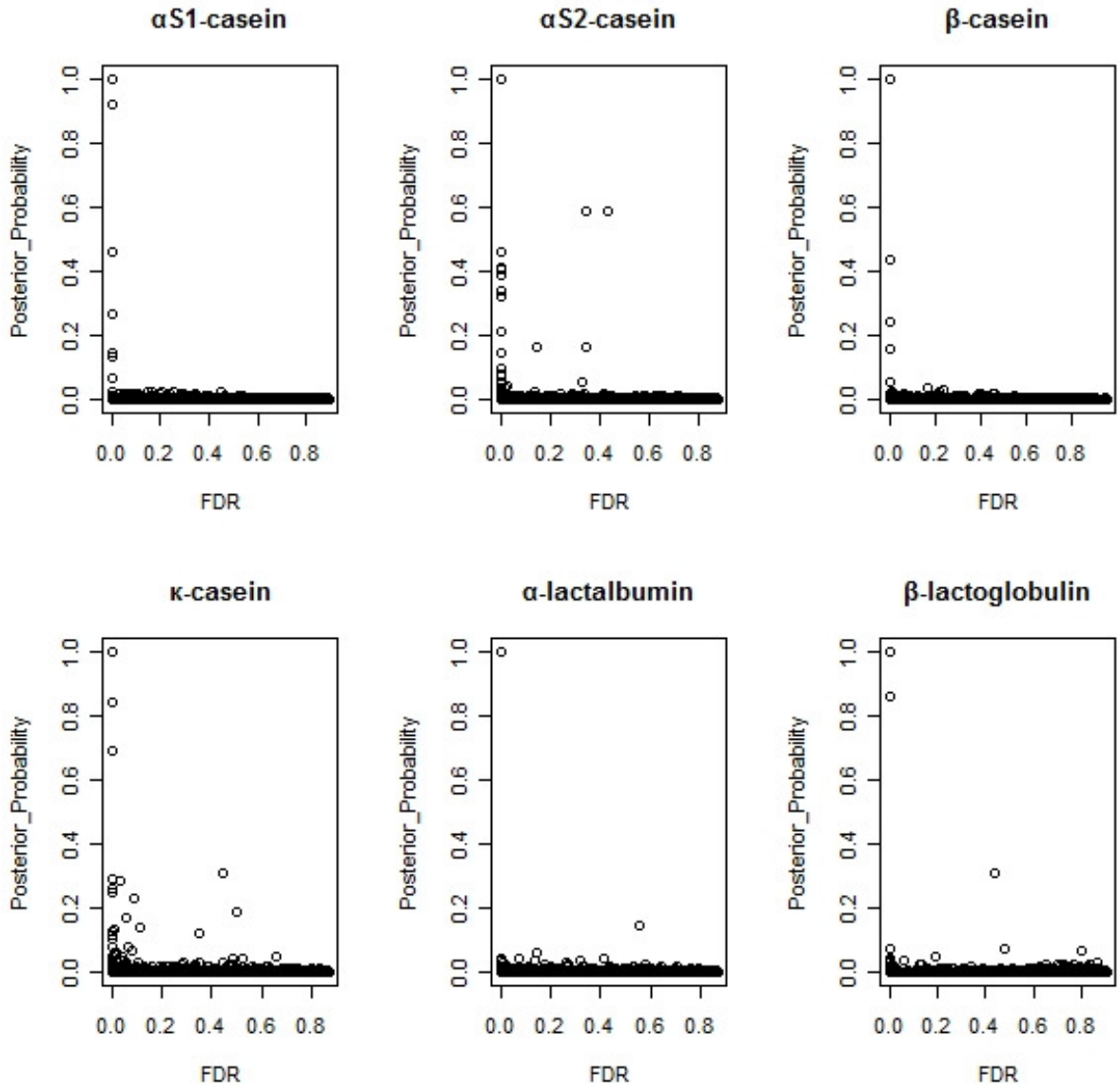

Figure 4 Graphical overview of the relation between the false discovery rate (FDR) and the poster probability for each of the six major milk proteins.

SNP analysis, the genetic variance can only be explained once. The genetic variance explained by one SNP can not be explained anymore by another SNP in the multiple SNP analysis. In the multiple SNP analysis, some SNPs (e.g. the SNP ULGR_ SNP_X14710_1740 for β-LG on BTA11) have an enormous effect on a trait. Adjacent SNPs on the right and left side of the SNP with the highest posterior probability, however, will due to LD still explain part of the additive genetic variance, as was illustrated in Table 4. When more adjacent SNPs on the right and left side of the SNP with the highest posterior probability were added up to the proportion of genetic variance explained by the SNP with the highest posterior probability, the proportion of additive genetic variance increased to the level of genetic variance explained by a single SNP in the single SNP analysis, except for β-CN on BTA6 and 11.

A third difference between the single and multiple SNP analysis was that in the single analysis the SNP was fitted as a fixed effect, whereas in the multiple analysis the SNP was fitted as a random effect. To test whether this affected the results, additional analyses were performed in which some of the SNPs most significantly associated in the single SNP analysis on BTA5, 6, 11 and 14 were separately fitted in the MCMC analysis. The proportion of genetic variance explained by each of these SNPs was calculated and results (data not shown) were similar to those from the single SNP analysis, except for β-CN on BTA6. The SNP most significantly associated with β-CN on BTA6 explained 42.70% of the genetic variance in the single SNP analysis. Running this SNP separately in the MCMC analysis, the proportion of the genetic variance explained by this SNP was 1.89%, which is similar to the 1.88% from the multiple SNP analysis. Fitting the SNP most significantly associated with β-CN as a random effect in the ASReml analysis resulted in similar proportion of genetic variance explained by this SNP as compared to fitting the SNP as a fixed effect. Thus making the results more comparable actually resulted in most cases in the same proportion of genetic variance explained by the SNP.

Conclusions

The same four main regions on BTA5, 6, 11 and 14 showing association with the six major milk proteins were detected in the single SNP analysis

and in the multiple SNP analysis. The multiple SNP analysis, however, identified a limited number of SNPs showing an effect as compared to the single SNP analysis. Furthermore, additional associations on BTA7 and 27 were detected with multiple SNP analysis compared to single SNP analysis. Thus, multiple SNP analysis results in higher power and in higher mapping precision to detect QTL as compared to the single SNP analysis.

Acknowledgements

This study is part of the Milk Genomics Initiative, funded by Wageningen University, NZO (Dutch Dairy Organization), CRV (cooperative cattle improvement organization), and the Dutch technology foundation STW. The authors thank the owners of the herds for their help in collecting the data.

References

Aulchenko, Y.S., D.-J. de Koning, and C. Haley. 2007. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait SNP association analysis. Genetics 177: 577-585.

Barret, J.C. B. Fry, J. Maller and M.J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-265.

Calus, M.P.L. T.H.E. Meuwissen, A.P.W. de Roos and R.F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics: 178: 553-561.

Daetwyler, H.D., F.S. Schenkel, M. Sargolzaei, and J.A. Robinson. 2008. A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. J. Dairy Sci. 91: 3225-3236.

De Koning, D.J., N.F. Schulman, K. Elo, S. Moisio, R. Kinos, J. Vilkki, and A. Mäki-Tanila. 2001. Mapping of multiple quantitative trait SNP by simple regression in half-sib designs. J. Anim. Sci. 79: 616-622.

Efron, B., R. Tibshirani, J.D. Storey and V. Tusher. 2001. Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association 96: 1151-1160.

Falconer, D.S and T.F.C. Mackay. 1996. Introduction to quantitative genetics. Fourth Edition: 122-144.

Gilmour A.R., B.J. Gogel, B.R. Cullis, S.J. Welham and R. Thompson. 2002. Asreml user guide. Release 1.0. VSN International Ltd., Hemel Hempstead, UK.

González J.R., L. Armengol, X. Solé, E. Guinó, J.M. Mercader, X. Estivill, and V. Moreno. 2007. SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23: 644-650.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433-443.

Heck J.M.L., C. Olieman, A. Schennink, H.J.F. van Valenberg, M.H.P.W. Visker, R.C.R. Meuldijk and A.C.M. van Hooijdonk. 2008. Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis. Int. Dairy J. 18: 548–555.

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

Jansen, R.C. 1993. Interval mapping of multiple quantitative trait SNP. Genetics 135: 205–211.

Kolbehdari, D., Z. Wang, J.R. Grant, B. Murdoch, A. Prasad, Z. Xiu, E. Marques, P. Stothard, and S.S. Moore. 2008. A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein bulls. J. Dairy Sci. 91: 2844-2856.

Liu Y., X. Qin, X.-Z. H. Song, H. Jiang, Y. Shen, K.J. Durbin, S. Lien, M.P. Kent, M. Sodeland, Y. Ren, L. Zhang, E. Sodergren, P. Havlak, K.C. Worley, G.M. Weinstock and R.A. Gibbs. 2009 Bos taurus genome assembly. BMC Genomics. 10: 180-191.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic gain value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Meuwissen, T.H.E. and M.E. Goddard. 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. 2004. Genet. Sel. Evol. 36: 261-279.

Sillanpää, M.J. and E. Arjas. 1998. Bayesian Mapping of Multiple Quantitative Trait SNP From Incomplete Inbred Line Cross Data. Genetics 148: 1373-1388.

Schopen G.C.B., J.M.L. Heck, H. Bovenhuis, M.H.P.W. Visker, H.J.F. van Valenberg and J.A.M. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. J Dairy Sci. 92: 1182-1191.

Schopen G.C.B., M.H.P.W. Visker, P.D. Koks, E. Mullaart, J.A.M. van Arendonk and H. Bovenhuis. 2010. Whole genome association study for milk protein composition in dairy cattle. Submitted.

Storey J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA 100: 9440–9445.

Storey, J.D. 2003. The positive false discovery rate: a bayesian interpretation and the q-value. Ann stat. 31: 2013-2035.

Uleberg, E. and T.H.E. Meuwissen. 2007. Fine mapping of multiple QTL using combined linkage and linkage disequilibrium – A comparison of single QTL and multi QTL methods. Genet. Sel. Evol. 39: 285-299.

Verbyla, K.L., P.Bowman, B.J. Hayes and M.E. Goddard. 2009. Sensitivity of genomic selection to using different prior distributions. Proc. of QTL MAS workshop in Wageningen.

Visker M.H.P.W, B. Dibbits, S. Kinders, H.J.F. van Valenberg, J.A.M. van Arendonk, and H. Bovenhuis. 2010. Effects of bovine β-casein protein variant I on milk production and milk protein composition. Submitted.

Zeng, Z.B. 1994. Precision mapping of quantitative trait SNP. Genetics 136: 1457–1468.

7

General discussion

Introduction

The general aim of this thesis was to study the extend in which bovine milk protein composition is determined by genetic factors, and to look for opportunities to utilize this genetic variation to improve milk protein composition. For this thesis, milk protein composition was determined in morning milk samples from nearly 2000 cows. In addition, blood samples of cows and semen samples of bulls were taken for DNA analysis.

Heritabilities and genetic correlations were estimated for the relative concentration of the six major milk proteins (α_{S1}-casein (α_{S1}-CN), α_{S2}-casein (α_{S2}-CN), β-casein (β-CN), κ-casein (κ-CN), α-lactalbumin (α-LA) and β-lactoglobulin (β-LG)), and the milk production traits (Chapter 3). The results show that there is considerable genetic variation in relative concentration of the six major milk proteins. Genetic correlations among the six major milk proteins were low, in general. The latter indicates that genes affecting two different milk protein fractions only partly overlap. A genome-wide screen was used to identify chromosomal regions associated with milk protein composition (Chapters 4 and 5). Important chromosomal regions have been detected on chromosomes 5, 6, 11 and 14. Some of these regions harbor genes that have been reported in literature to influence milk protein composition. We, for example, confirmed the important role of the casein genes on BTA6 and the β-LG gene on BTA11.

In this final chapter, the general discussion, the first section deals with the composition of milk and dairy products, the second section deals with the impact of using relative or absolute concentrations for milk protein composition, and the third section concentrates on casein index. In the last section, the opportunities to use the genetic variation of the milk protein composition are explored.

Milk and dairy products

Bovine milk is the basis for a large variety of consumer products, like liquid milk, fermented milk (e.g. yoghurt), cheese (many varieties), butter, creams and condensed milk (coffee milk) as well as dairy ingredients. The consumption of dairy products is an important contribution to the nutrient supply, especially for calcium, vitamin B2 and B12, protein, zinc and to less extend for magnesium, phosphor, vitamin A, B1, B6, B11 and D, and

selenium. Dairy products provide about 15 % of the energy intake. This high nutrient density and relatively low energy supply are preferable for the prevention of obesity (Kok, 2009). The Dutch centre for food products, therefore, recommends consuming 450 – 650 ml milk and 20 to 30 gram cheese each day for people who are at least 19 years old (Kok, 2009).
Although some people are oversensitive (cow's milk allergy) to milk protein, milk has a high nutritional value because the milk proteins contain many essential amino acids. The most important essential amino acids are cysteine, methionine, and tryptophan (de Wit, 1998). People need amino acids for a number of essential biological functions, i.e. to grow, to maintain their body, and to build and repair muscles. Next to the proteins, milk is also a very important and unique source of vitamins and minerals. An important mineral is calcium which is needed for development of bones and teeth. Calcium, protein, vitamin B2 and vitamin B12 contribute significantly to the high nutrient richness score of milk, cheese and yoghurt (Steijns *et al.*, 2008). Thus bovine milk contains a surplus of major nutritional components for the consumer.
Calcium and milk protein composition also have an effect on the technological properties of milk. Calcium is important for the coagulation of milk during cheese manufacturing. Furthermore, several studies (e.g. Schaar *et al.*, 1985; van den Berg *et al.*, 1992; St-Gelais and Haché, 2005; Wedholm *et al.*, 2006) have shown that casein composition has an effect on the renneting process of milk and on the technological properties of cheese. For example, milk with a high content of β-CN will result in poorer coagulation and, as moisture content decreases, harder cheeses (St-Gelais and Haché, 2005) Another example is milk with a high content of κ-CN which has a better coagulation and a firmer curd which increases cheese yield (Wedholm *et al.*, 2006) The protein κ-casein stabilizes the surface of the casein micelle while the other three caseins form the core. The amount of κ-CN determines the size of the casein micelle and a higher amount of κ-CN results in smaller casein micelles (Dalgleish *et al.*, 1989) Smaller casein micelles may increase the ability to entrap milk constituents (Niki *et al.*, 1994; Walsh *et al.*, 1998), aggregate more rapidly and have a higher rennet gelation rate than large casein micelles (Park *et al.*, 1999)

Whey protein is a high quality protein in cow's milk. Whey protein is the richest source of essential amino acids like methionine and cysteine and, therefore, has a high nutritional value (de Wit, 1998) Whey protein products are used to replace egg proteins in confectionery and bakery products, and as milk replacers in products, such as ice cream (de Wit, 1998). Whey consists mainly of α-LA and β-LG. The biological function of α-LA is to support the synthesis of lactose (Walstra and Jennes, 1984), whereas the biological function of β-LG is a transporter of retinol (provitamin A) from the cow to the calf (de Wit, 1998). This function of β-LG might be less important for human babies (de Wit, 1998), which might be an explanation for the fact that human milk does not contain β-LG. The protein β-LG is a rich source of cysteine, an essential amino acid that appears to stimulate glutathione synthesis. Although the protein composition is different between bovine and human milk, bovine milk is the most used nutrient source in infant milk. Bovine α-LA is a rich source of cysteine and tryptophan, and therefore used to increase the amount of cysteine and tryptophan in infant milk.
For most dairy products heat treatment is part of the production process. However, heat treatments of proteins lead to fouling of the heating equipment. Elofsson *et al.* (1996) showed that a lower β-LG concentration in bovine milk is associated with a lower fouling rate of the heating equipment. This will reduce the costs to clean the heating equipment, which is interesting from an economical point of view.
In summary, different applications of milk require different milk protein compositions. From a nutritional (infant milk) and economical (reduce cleaning cost of heating equipment) point of view, the optimal milk protein composition is less casein and a whey protein composition with more α-LA and less β-LG. For some technological properties, however, the optimal milk protein composition is not that straightforward because the different caseins have different effects on the technological properties. For more cheese production, however, the optimal milk composition is more casein and less whey protein. This suggests that there is no single milk composition that suits each product direction. Therefore, an optimum needs to be found between nutritional value, technological value and economical value as a driver to change the milk protein composition.

Relative versus absolute concentrations of major milk proteins

We used capillary zone electrophoresis (CZE) to obtain relative concentrations of the different proteins. CZE was used because this method can simultaneously separate the casein and whey proteins, including some protein variants (Heck *et al.*, 2008). Furthermore, protein percentage is routinely recorded. This information can be used to calculate the relative as well as the absolute concentration of the different milk proteins. For the dairy industry and selective breeding, protein percentage and the relative protein composition are complementary information sources. We have chosen to use the relative protein composition in our analysis, i.e. the amount of a certain protein expressed as a fraction of the total amount of milk protein. A disadvantage of using relative concentrations is the fact that correlations are introduced by calculating relative concentration. When the total protein fraction consists of two independent proteins A and B, the relative concentrations of A and B are fully dependent. The relative concentration of B is equal to 1 minus the relative concentration of A and consequently, the correlation between these two relative concentrations is -1. In this thesis, however, six instead of two protein fractions were used. This reduces the auto-correlations among the relative concentrations. When the concentration of one of the six proteins decreases, this does not mean that the concentration of only one other protein increases with the same magnitude, but it can be divided over some or all of the other five major milk proteins. The autocorrelations complicates the interpretations of the relative concentrations, a problem that does not occur when using absolute concentrations.

From a biological point of view, absolute as well as relative concentrations of individual proteins are of interest. A disadvantage of using absolute concentrations is that relationships between different protein fractions are a combined effect of the total protein production and the relative composition. Using absolute concentrations or relative concentrations for the six major milk proteins have both advantages and disadvantages. However, the question rises, what is the difference between using relative or absolute protein concentrations.

The absolute concentrations were calculated by multiplying the relative concentrations of the six major milk proteins with protein percentage. The

phenotypic correlation between relative and absolute concentrations for the six major milk proteins ranged from 0.29 for α_{S1}-CN to 0.91 for κ-CN (Table 1). Especially, the phenotypic correlations for α_{S1}-CN and for β-CN were low. The genotypic correlation between relative and absolute concentrations ranged from 0.41 for β-CN to 0.93 for κ-CN (Table 1). This demonstrates that ranking of cows based on absolute and relative concentrations are different to very different. This holds for both ranking on phenotypes as well as ranking on genotypes. The difference is caused by the variation between cows in protein percentage.

The heritability for absolute β-CN concentration was twice as high as the heritability for the relative β-CN concentration, whereas the heritabilities for the other five milk proteins were similar for absolute and relative concentrations (Table 2). This indicates that genetic factors are more important for the absolute β-CN concentration than the relative β-CN concentration. Therefore, different genes might affect the absolute β-CN concentration compared to the relative β-CN concentration, which is also supported by the low genetic correlation between the absolute and relative β-CN concentration. The differences in genetic parameters (Table 1 and 2) between relative and absolute concentrations of the six major milk proteins illustrate that the two concentrations are not the same trait.

Table 1 Phenotypic (r_P) and genetic (r_G) correlations between relative concentration (ww%) and absolute concentration (g/L) for the six major milk proteins.

Trait	$\mathbf{r_P}$	$\mathbf{r_G}$
α_{S1}-casein	0.29 (0.03)	0.64 (0.14)
α_{S2}-casein	0.90 (0.01)	0.89 (0.03)
β-casein	0.49 (0.03)	0.41 (0.16)
κ-casein	0.92 (0.01)	0.93 (0.08)
α-lactalbumin	0.81 (0.01)	0.80 (0.00)
β-lactoglobulin	0.88 (0.01)	0.92 (0.02)

To illustrate that there is a difference in using relative or absolute concentrations, I repeated the association study for only BTA6 using absolute concentrations for the six major milk proteins. For α_{S2}-CN, β-CN,

κ-CN, the most significantly associated SNP was identical to the most significantly associated SNP using the relative concentrations on BTA6. However, for α_{S1}-CN, α-LA and β-LG, additional associations were found. The most significantly associated SNP using absolute concentrations was located at 85.8 Mbp for α_{S1}-CN and at 92.0 Mbp for α-LA, whereas the most significantly associated SNP using relative concentrations was located at 88.1 Mbp for α_{S1}-CN and at 88.5 Mbp for α-LA. For β-LG, the SNP most significantly associated using absolute concentrations was only 0.1 Mbp away from the SNP most significantly associated using relative concentrations.

Table 2 Heritabilities for the relative concentration and absolute concentration of the six major milk proteins.

Trait	relative	absolute
α_{S1}-casein	0.47	0.42
α_{S2}-casein	0.73	0.66
β-casein	0.25	0.59
κ-casein	0.64	0.66
α-lactalbumin	0.55	0.38
β-lactoglobulin	0.80	0.81

The above described results of the relative and absolute concentrations of the six major milk proteins suggests that performing calculations using relative protein concentrations is not the same as performing calculations using absolute protein concentrations. Analysing both the relative as well as the absolute concentration will help to increase our understanding of the genetic background of differences in protein composition.

Casein index

In this thesis, casein index is used as a parameter for the amount of cheese that can be produced out of milk. In literature, the casein number is often used. Casein index as defined in this thesis is the total caseins in milk divided by the sum of total caseins and total whey proteins in milk, whereas casein number is defined as the difference between protein content in milk and protein content in whey. There are two main differences between

casein index and casein number. The first one is the difference in determination method. Casein index was calculated based on the six major milk proteins which were determined using CZE. Using CZE, the amount of total nitrogen is not measured, whereas determination of casein number is based on the method of Rowland (1938), which involves precipitation of casein at pH=4.6 and measuring the Kjeldahl nitrogen in milk and in whey. Using the Kjeldahl method, the amount of total nitrogen, including non protein nitrogen (NPN), is determined. The amount of NPN varies between 0.1% and 0.3% (Heinrichs *et al.*, 1997) and including NPN will increase the casein number. However, casein number is neither a perfect indicator for cheese production efficiency because the κ-CN tails are included in the casein number, whereas during cheese production, the κ-CN tails end up in the whey.

The second difference is that the six major milk proteins as evaluated in this research made up about 86% of the total protein fraction. The remaining 14% consists of proteins which for a large part belong to the whey proteins (e.g. BSA, lactoferrin, immunoglobulins), which were now not taken into account with the calculation of the casein index. The casein index (87.45, chapter 3) as defined in this research, therefore, is higher than the casein number as reported in literature (e.g. 81.5, Coulon *et al.*, 1998; 78.2, Lindmark-Månsson *et al.*, 2003).

Opportunities to increase casein index using genetic selection

In this thesis (chapter 3) we showed that there is substantial genetic variation in milk protein composition. Furthermore, we have shown that known protein variants are associated with quantitative variation in milk protein composition. For example, the B allele of β-LG is associated with a lower relative concentration of β-LG protein. Heck *et al.* (2009) found that the B allele of β-LG is also associated with an increased casein index. In chapter 4, we reported a whole genome scan for milk protein composition in which we found a number of genomic regions that contribute to genetic variation in milk protein composition. These findings open opportunities for selection aimed at increasing the casein index in milk.

For the Dutch dairy industry, cheese production is very important. The casein content of milk has a direct effect on cheese properties; cheese

yield, milk coagulation time, and curd firmness (Wedholm *et al.*, 2006). The milk payment scheme for the farmers in the Netherlands is based on protein (and fat) yield. Protein yield is positively correlated with casein index (chapter 3) and, therefore, selection on protein yield is in an indirect way to improve the amount of casein in their milk. In this section, I will explore a number of scenarios to improve the casein index in milk.

There are several opportunities to use genetic variation to increase the casein index of milk. The first option that I will explore is differentiation (scenario 1, (§7.5.1). In this scenario, farmers use the 10% bulls with the highest estimated breeding value (EBV) for casein index to breed a subset of cows to produce the next generation. In this scenario, the breeding program is not changed but the variation between bulls in casein index is exploited.

In the second scenario the breeding program is focused on improving casein index, i.e. casein index is the only trait in the breeding goal. In this scenario, the bulls with the highest merit for casein index are used to produce the next generation of bulls which will lead to genetic improvement of the population over time. Genetic selection of bulls can be based on their EBVs calculated from performances on female relatives (scenario 2, §7.5.2), or based on their genotypes for protein genes (scenario 3, §7.5.3). Finally, in scenario 4 (§7.5.4), I have explored the consequences of using genomic selection in which bulls are selected based on their genomic breeding values. In all four scenarios, the genetic parameters as calculated in chapter 3 were used.

Scenario 1: Differentiation of bulls based on EBV

In the first scenario, 10% bulls with the best EBV for casein index are used to breed a subset of cows to produce the next generation. It is assumed that a limited number of farms will only use these 10% best bulls for casein index. In this way, these farms will improve their herd average for casein index and produce milk more suitable for cheese production. The genetic herd levels of the participating farms in this scenario will change over time. However, there will be no change in genetic level for casein index on the non-participating farms because it is assumed that the breeding program for the population is not changed. This scenario capitalizes on the variation

between bulls and uses the bulls with high EBV for casein index on the specialized group of farms.

The 10% best bulls can be selected within the group of progeny tested bulls, based on their estimated breeding values (EBV) for casein index. To quantify the consequences of this scenario, I used the dataset of the Dutch Milk Genomics Initiative. In this dataset, the EBV for 50 young test bulls was estimated using an animal model. The 10% best bulls for casein index have an average EBV of 1.84 (Table 3). This means that their offspring on average will have ½ * 1.84 is 0.92 higher casein index than the herd average (a casein index of 88.4 instead of 87.5). Due to genetic correlations between traits, these 10% best bulls also had positive EBVs for α_{S2}-CN, β-CN, α-LA and milk yield, and negative EBVs for α_{S1}-CN, κ-CN, β-LG and slightly lower EBV for protein percentage (Table 3). The EBV for the four casein fractions summed up to an increase of 1.095 ww% casein in milk and the two whey proteins summed up to an decrease of -1.668 ww% whey in milk.

Table 3 Estimated breeding values (EBVs) of 10% young bulls in the Dutch Milk Genomics dataset with the best EBV for casein index and its correlated responses.

Trait	**EBV**
Casein index	+1.84
Correlated responses	
α_{S1}-casein (ww%)	-0.31
α_{S2}-casein (ww%)	+0.64
β-casein (ww%)	+0.80
κ-casein (ww%)	-0.04
α-lactalbumin (ww%)	+0.04
β-lactoglobulin (ww%)	-1.71
Protein (%)	-0.05
Milk[1] (kg)	+0.92

[1] Test-day morning milk yield

The example illustrated with Table 3 is based on a relatively small group of bulls (10% of 50 bulls) with a relative small number of offspring each.

Furthermore, the genetic level will increase over time when the farmer keeps on using bulls with superior EBV for casein index. However, the genetic improvement will level off and the asymptotic improvement is equal to the selection differential of the 10% best bulls. Figure 1 shows the increase in herd level for casein index when each year one third of 100 cows are replaced by the offspring of 10% best bulls with the highest EBV for casein index.
As showed in Figure 1, it will take about 3 years before the first daughters will start producing milk with a higher casein index.

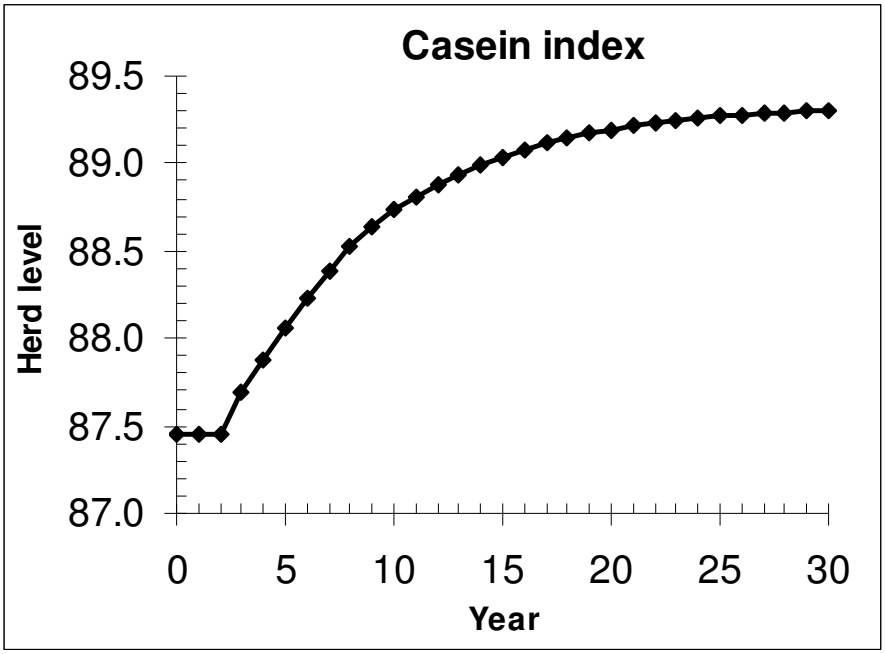

Figure 1. Herd level for casein index when each year one third of the cows are replaced by offspring of bulls with 10% highest EBV for casein index produced from cows in the herd.

The asymptote of Figure 1 is equal to the maximum level of casein index which can be reached by selecting 10% best bulls on a herd each year. The accuracy (r_{IH}) of the EBV for casein index was calculated using 100 progeny with information on casein index and a heritability of 0.70 for casein index (chapter 3). The expected breeding values of the 10% best bulls for casein index (selection differential) is equal to $i * r_{IH} * \sigma_a$, where i is the selection intensity corresponding to selected fraction of 10%, r_{IH} is the

accuracy and σ_a is the additive genetic standard deviation. For casein index this selection differential becomes 1.755 * 0.98 * 1.14 = + 1.96%. In 10 years of time, casein index will increase with 1.29 % (from 88.74 to 87.45, Figure 1) which means that 73% of the maximum level has been reached.
Assuming a farm with 100 cows producing 900.000 kg milk each year and using the average protein percentage of 3.51% as found in the Dutch Milk Genomics Initiative, this farm will produce 31.500 kg protein each year. With current population level for casein index of 87.45%, this farm produces 27.547 kg casein and 3.953 kg whey. Under scenario 1, the maximum improvement in casein index is 1.96% which corresponds to an improvement of 31.500 * 0.0196 = 617 kg of casein and a decrease of 617 kg whey. In this calculation, correlated response in other traits, as observed in Table 2 were ignored. At this moment, 1 kg of casein has a value of €6.00 and 1 kg of whey has a value of €2.50 (Heck, personal communication). Using these prices, selecting the 10% bulls with highest EBV for casein index will result in an increase in profit of 2160 euro each year. When 50% of the farmers will select the 10% bulls with the highest EBV for casein index, this will lead to an increase in profit of 2.7 million each year.
This scenario can be relatively easily be implemented by a group of interested farmers, which want to produce milk with higher casein content. This interested group of farmers could decide to specifically start producing milk more suited for cheese production. This differentiation could be implemented by farmers that produce cheese themselves or by a cheese producing company who stimulates farms to pay more attention to protein composition, for example by changing the milk payment scheme. Note that this scenario assumes that EBVs of bulls are available. To estimate these breeding values, phenotypes on casein index need to be collected on offspring of bulls.

Scenario 2: New breeding programme for casein index

In the second scenario, a new breeding program is developed which has as its only objective to increase the casein index of milk. In this breeding program, the cows and bulls with highest EBV for casein index are selected to produce the next generation. This breeding program will lead to

continuous genetic improvement over generations. The computer programme SelAction (Rutten *et al.*, 2002) was used to calculate the selection response resulting from this breeding programme. The heritabilities, phenotypic variance, and phenotypic and genetic correlations as estimated in chapter 3 were used as input parameters for SelAction. More information about the parameters used in SelAction is given in Text box 1. Simulation showed that the breeding program with selection for an increased casein index will result in a genetic improvement of casein index of 0.22% per year (Table 4). Single trait selection on casein index might lead to correlated response in other traits due to genetic correlations with casein index. Correlated response shows that selection for casein index will result in an increase of five of the major milk protein fractions and a decrease in β-LG content (Table 4). Selection for a higher casein index hardly affects protein percentage and milk yield.

Table 4 Expected genetic change of the population with selection for an increased casein index and its correlated response.

Trait	**Response per year**
Casein index	+0.22
Correlated response	
α_{S1}-casein (ww%)	+0.02
α_{S2}-casein (ww%)	+0.09
β-casein (ww%)	+0.03
κ-casein (ww%)	+0.01
α-lactalbumin (ww%)	+0.01
β-lactoglobulin (ww%)	-0.20
Protein (%)	+0.00
Milk (kg)	+0.02

[1] Test-day morning milk yield

In the Netherlands, we produce 12 billion kg milk each year, of which 50% is used for cheese making. The average protein percentage in the Dutch Milk Genomics population was 3.51% (chapter 3). Thus each year, there is $12 \cdot 10^9$ * ½ * 0.0351 = $210{,}6 \cdot 10^6$ kg protein. The annual genetic improvement for casein index was +0.22, which corresponds to an annual increase of 0.0022 * $210.6 \cdot 10^6$ = $4.63 \cdot 10^5$ kg casein and an annual

decrease of $4.63 \cdot 10^5$ kg whey. Using the same economic values for casein and whey as described in scenario 1, the genetic improvement for casein index will result in an increase in annual profit of $1.6 \cdot 10^6$ euro each year. Genetic progress in this scenario is cumulative.

Text box 1 Input parameters SelAction
Simulated population: 25 bulls x 200 dams

Information sources Index

age class	male candidates	male information source	female candidates	female information
1	0	BLUP	0	BLUP
2	0	BLUP	0	BLUP
3	0	BLUP, HS4	100	BLUP,HS3,OP
4	0	BLUP, HS4	95	BLUP,HS3,OP
5	100	BLUP, HS4, Progeny	90	BLUP,HS3,OP
6	99	BLUP, HS4, Progeny	85	BLUP,HS3,OP
7	98	BLUP, HS4, Progeny	80	BLUP,HS3,OP
8	97	BLUP, HS4, Progeny	75	BLUP,HS3,OP

[1]HS4 = Half sib group with 4 female half sibs; HS3 = Half sib group with 3 female half sibs; OP = Own performance; Progeny = Progeny group with 100 female progeny; BLUP = Best Linear Unbiased Prediction.

To implement this scenario, breeding values of cows and bulls have to be estimated which requires the collection of information on casein index. In this scenario, the whole breeding scheme is changed and consequently all participating farms (also the farms which produce liquid milk products) will produce milk with a modified milk protein composition.
In this scenario, I used casein index as the only trait in the breeding goal. This scenario resulted in an annual genetic improvement of 0.22% increase for casein index. This response should be regarded as upper bound of what can be achieved by selection based on phenotypes collected for casein index. In real life, however, selection of animals will never be done on only one trait but animals will be selected for several traits. Therefore, more traits will be included in the breeding goal and the genetic progress for casein index will become smaller.

Scenario 3: Selection based on known genotypes.

Information at DNA level can be collected. In the third scenario, selection is based on DNA information on milk protein genes. I assumed that the protein variants are the casual mutations in the protein genes and used the effects of the protein variants estimated by Heck *et al.* (2009). Animals were selected based on their known genotypes for the milk protein genes which will lead to an increase of the frequency of the alleles with favorable effects on milk protein composition.

In this scenario, I concentrated on genotypes for κ-CN and β-LG proteins. Genotypes for κ-CN and β-LG were chosen because variants for these proteins are associated with total casein in milk (Ng-Kwai-Hang *et al.*, 1987; Bobe *et al.*, 1999; Heck *et al.*, 2009), with casein yield milk (Schaar *et al.*, 1985; Van Den Berg *et al.*, 1992; Wedholm *et al.*, 2006), and with cheese-making properties (Marziali and Ng-Kwai-Hang 1986; Mayer *et al.*, 1997; Wedholm *et al.*, 2006).

For κ-CN, the B allele has favorable effect on casein index and I determined the potential improvement in casein index when the whole population will be homozygous for the B-allele of κ-CN. A population with only κ-CN BB animals is expected to have a 0.145 ww% higher casein index in milk compared to the current population (Table 5).

Table 5 Expected genetic change in milk protein composition of population homozygous for the B- allele of κ-CN (B-allele κ-CN) compared to the current population.

Trait	Current population	Fixation at κ-CN BB
Casein index	87.45	+0.145
α_{S1}-casein (ww%)	33.62	-0.917
α_{S2}-casein (ww%)	10.38	+0.244
β-casein (ww%)	27.17	+0.059
κ-casein (ww%)	4.03	+0.536
α-lactalbumin (ww%)	2.44	-0.125
β-lactoglobulin (ww%)	8.35	-0.029
Protein (%)	3.51	+0.060
Milk[1] (kg)	13.46	-0.134

[1] Test-day morning milk yield

For β-LG, the favorable allele for casein index is the B allele. Selection for only β-LG BB animals is expected to increase casein index with 1.927% (Table 6).
Ganai *et al.* (2009) identified two additional polymorphisms which segregate within the B-allele for β-LG: g.-462G>A and g.3748G>A. The g.-462G>A polymorphism showed the biggest effect (0.44, Ganai *et al.*, 2009) on the β-LG concentration. For animals that are already homozygous for the B-allele of B-LG, additional selection for animals that are homozygous for the A allele of the g.-462G>A polymorphism is expected to increase casein index with an additional 0.099% (Table 6).

Table 6 Potential for genetic change using animals which are homozygous for the B allele of β-LG (B-allele β-LG) and animals which are homozygous for the A allele of the g.-462G>A polymorphism in β-LG (g.-462G>A β-LG).

Trait	**Current population**	**Fixation at β-LG BB**	**Fixation at g.-462G>A**
Casein index	87.44	+1.927	+0.099
α_{S1}-casein (ww%)	33.62	+0.453	-0.046
α_{S2}-casein (ww%)	10.38	+0.386	+0.085
β-casein (ww%)	27.17	+0.386	-0.063
κ-casein (ww%)	4.03	+0.133	+0.020
α-lactalbumin (ww%)	2.44	+0.066	-0.025
β-lactoglobulin (ww%)	8.35	-1.762	-0.071
Protein (%)	3.51	-0.019	+0.006
Milk[1] (kg)	13.46	-0.077	-0.009

[1] Test-day morning milk yield

I have looked at effects of selection based on single genotypes but the effects of the three genotypes can be combined. This means that the estimated potentials can be summed. For casein index, this will results in an increase of 0.145 + 1.927 + 0.099 = 2.171 %.
The genetic improvement for casein index of 2.171% corresponds to an increase of 0.02171 * $210{,}6 \cdot 10^6$ = $4.6 \cdot 10^6$ kg casein. Using the same economic values for casein and whey as described in scenario 1, the genetic improvement for casein index will result in an increase in profit of

$1.6 \cdot 10^6$ euro. However, once all animals in the population have the favorable genotype, the genetic improvement will stop.
To implement this scenario, there is no need to collect phenotypes and selection can be based on genotypes. It is relatively easy to collect genotypes for sires and cows.

Scenario 4: Genomic selection

In dairy cattle, breeding is largely based on phenotypic records of the individual itself and its relatives. Breeding values are estimated based on the phenotypic records and pedigree relationships using Best Linear Unbiased Prediction (BLUP, Henderson 1985), like in scenario 2. Meuwissen *et al.* (2001) have shown that genetic progress can be improved by using genomic selection, a breeding scheme in which information on a large number of markers is used. Since a few years, individual cows and/or bulls can be genotyped for thousands of single nucleotide polymorphisms (SNPs), like for example the 50K bovine SNP chip (chapter 5), at relatively low costs. This has enabled the introduction of genomic selection in dairy cattle (Goddard and Hayes, 2009).
The linkage study (chapter 4) showed several significant and suggestive QTLs for milk protein composition. Some of these QTLs were confirmed in the whole genome association study using 50K SNPs (chapter5). These QTLs, therefore, add to the potential for genetic change for milk protein composition. The benefit of using QTL information depends on the amount of genetic variance explained by the QTL. Spelman *et al.* (1999) reported an increase in genetic gain ranging from 10 to 55% in different breeding structures in dairy cattle when 50% of the genetic variances can be explained.
To exploit the genetic variance explained by all QTL, genomic selection (Meuwissen *et al.*, 2001) can be used. Implementation of genomic selection proceeds in two steps. The first step is the estimation of marker effects in a reference population, which has been phenotyped and genotyped. The second step is the prediction of genomic estimated breeding values (gEBV) for bulls in subsequent generations that usually only have genotypic and no phenotypic information. The gEBVs of a group of bulls are predicted using the estimated marker effects obtained in the reference population. Using

the 2000 cows of the Dutch Milk Genomics Initiative, which have been genotyped and phenotyped, as reference population, the accuracy of gEBV can be estimated as described by Daetwyler *et al.* (2008) using the following prediction equation:

$$r_{IH} = \sqrt{Nh^2/(Nh^2+q)},$$

where *N* is the number of individuals genotyped and phenotyped (*N*=2000), h^2 is the heritability of the trait and *q* is the number of independent chromosome segments in the population. I used this prediction equation to determine the accuracy of gEBV for milk protein composition The value of q is equal to $2N_eL$, where L is the length of the genome (30 Morgan) and N_e is the effective population size (N_e = 64, de Roos *et al.*, 2008).

Table 7 The expected accuracy (r_{IH}) of genomic selection using the 2000 cows of the Dutch Milk Genomics Initiative as reference population and the maximum increase in herd level when selecting 10% bulls with the best gEBV for each of the different traits.

Trait	**r_{IH}**	**Maximum level**
α_{S1}-casein (ww%)	0.44	0.846
α_{S2}-casein (ww%)	0.52	1.048
β-casein (ww%)	0.34	0.438
κ-casein (ww%)	0.50	0.382
α-lactalbumin (ww%)	0.47	0.181
β-lactoglobulin (ww%)	0.54	1.012
Casein index	0.52	1.045
Protein (%)	0.51	0.194
Milk[1] (kg)	0.42	1.055

[1] Test-day morning milk yield

The expected accuracies of genomic selection using our data as reference population ranged from 0.34 for β-CN to 0.54 for β-LG (Table 7). The value of q was equal to the number of independent chromosome segments, whereas q can also be the number of independent segments affecting the trait. This number actually could be rather small for some of the traits, especially for β-LG where there is one gene with a major effect. Using the number of independent segments affecting the trait for the value of q will

increase the accuracy of the genomic breeding values for the six major milk proteins. VanRaden *et al.* (2009) reported reliabilities of genomic predictions for milk production traits ranging from 0.50 for fat yield to 0.72 for fat percentage. These reliabilities, however, were based on roughly 3600 bulls in a training dataset which was used to compute predictions, which were tested on roughly 1800 bulls in a test dataset. In addition, de Roos *et al.* (2009) reported reliabilities of genomic predictions for production traits ranging from 0.61 to 0.78 (fat percentage). VanRaden *et al.* (2009) and de Roos *et al.* (2009) showed that an increase in reference population also increased the reliability of the genomic predictions. To obtain an accuracy of about 0.75 for β-LG, 6000 animals with genotyped and phenotypes are needed as a reference population. This would mean that also the other two milk samples that have been collected as part of the Dutch Milk Genomics Initiative should be phenotyped.

The expected accuracies of genomic selection were calculated for the six major milk proteins, casein index, protein percentage and milk yield (Table 7). Based on these genomic accuracies, the maximum potential ($i * r_{IH} * \sigma_a$, as described in more detail in scenario 1) for each of the traits by selecting the 10% best bulls with gEBV were calculated (Table 6). For casein index this will mean a selection differential of 1.755 * 0.52 * 1.14 = + 1.05%.

To implement genomic selection, many SNP genotypes across the genome are needed. It is nowadays no problem to obtain a large number of SNP genotypes and the genotyping costs per SNP marker reduced considerably. However, for genomic selection also phenotypes are required to estimate effects of chromosome segments in the reference population.

Note that the recombination between the SNP markers and the QTL, the gEBV accuracy using chromosome segment effects obtained in the reference population will reduce. This suggests that the chromosome segments have to be re-estimated after several years. Meuwissen *et al.* (2001) showed with simulations that the gEBV accuracy decreases from 0.85 to 0.76 after three generations. This might suggest that re-estimation of the chromosome segment effects in the reference population should at least take place every 3 generations. However, the phenotypes of the 2000 cows of the Dutch Milk Genomics Initiative were collected in 2005, which is five years ago and is more or less equal to one generation. This means that

the phenotypes (and genotypes) from the 2000 cows of the Dutch Milk Genomics Initiative can be used as reference population to predict accurate gEBV for milk protein composition for a group of animals that only have genotypes.

Summarizing the scenarios

In scenario 1 it is possible to obtain a genetic improvement for casein index of 1.29% within 10 years. This increase in casein index can relatively easily be adapted by a group of interested farmers, which will differentiate from other farmers (producing e.g. liquid milk products), whereas in scenario 2, the whole population is changed into one direction. However, the annual genetic improvement for casein index is 0.249%. This means that within 10 years, it is possible to increase the casein index with 2.49%, which is twice as high as compared to scenario 1. Note, that the selection in scenario 2 was based on one trait, which is not very realistic. In both scenarios 1 and 2 phenotypes are required to estimate the breeding values.

No phenotypes are required for scenario 3, which also had the highest genetic improvement for casein index of 2.2%. However, this increase is dependent on the time required to change the allele frequency in the population from current level to fixation. The fact that only genotypes are required, which can easily be collected on cows and bulls, makes scenario 3 relatively easy to implement.

Scenario 4 is also based on genotypes, however, also phenotypes are required. Every three generations (about 15 years), phenotypes have to be collected to re-estimate the chromosome segments of the reference population, whereas in scenario 1 and 2 phenotypes have to be collected annually to estimate breeding values. For this thesis, milk protein composition has been quantified using CZE (Heck *et al.*, 2008). CZE is a good and reliable technique to separate milk proteins, however, this method is costly and time consuming. This makes CZE not a good method for routinely measuring milk protein composition. To obtain phenotypes routinely, therefore, infrared spectroscopy (IR) might be useful. Ghiroldi *et al.* (2004), showed that the use of IR method is reliable and can be used for routinely record the phenotypes for casein content in milk. In addition, Rutten *et al.* (2009) showed that it is possible to predict individual fatty

acids with a desired r^2 value > 0.6 (concentration > 2.54 g/100g) or > 0.8 (concentration > 0.19 g/dL) based on infrared spectroscopy. IR to predict milk protein composition accurately, however, is more difficult than fat composition which results in lower accuracies (Rutten, personal communication). Evaluation of accuracy and options to improve the accuracy are very important for implementation of selection for improvement of milk protein composition.

For scenario 3 and 4, DNA testing of the animals is required. Based on these DNA tests, the best cows and bulls with an improved milk protein composition are selected. Especially for bulls, selection based on genotypes is very interesting as bulls have no phenotypic information on milk protein composition. For scenario 3 and 4, therefore, the bulls can be selected at a young age without having information on the performance of the daughters. This will reduce the generation interval from around 6 years to around 2 years. This will increase the genetic improvement each year. The optimal scenario to increase casein index in milk would be a combination of scenario 3 and 4, i.e. a scenario in which the effects of known genes as well as anonymous genes (genomic selection) are used.

References

Bobe, G., D.C. Beitz, A.E. Freeman, and G.L. Lindberg. 1999. Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. J. Dairy Sci. 82: 2797-2804.

Coulon, J-B., C. Hurtaud, B. Remond, and R. Verite. 1998. Factors contributing to variation in the proportion of casein in cow's milk true protein: a review of recent INRA experiments. J. Dairy Res. 65: 375-387.

Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PlosOne 3: e3395

Dalgleish, D.G., D.S.Horne and A.J.R. Law. 1989. Size related differences in bovine casein micelles. Biochim. Biophys. Acta 991: 383–387.

Elofsson, U.M., M.A. Paulsson, P. Sellers and T. Arnebrant. 1996. Adsorption during heat treatment related to the thermal

unfolding/aggregation of β-lactoglobulins A and B. J. f Colloid Interface Sci. 183: 408-415.

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

Ghiroldi S., C. Nicoletti, A. Rossoni. 2004. Genetic parameter estimation for casein in Brown Swiss. Proceedings of the Interbull meeting SOUSSE, TUNISIA, May 29-31, pages: 125-128.

Goddard, M.E., and B.J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Genet.10: 381-391.

Heck, J.M.L., C. Olieman, A. Schennink, H.J.F. van Valenberg, M.H.P.W. Visker, R.C.R. Meuldijk, and A.C.M. van Hooijdonk. 2008. Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis. Int. Dairy J. 18: 548-555.

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

Henderson, C.R. 1985. Best linear unbiased predictions of nonadditive genetic merits in noninbred populations. Journal of Animal Science 60: 111- 117.

Heinrichs, J., C. Jones, and K. Bailey. 1997. Milk components: Understanding the causes and importance of milk fat and protein variation in your dairy herd. Dairy & Animal Science Fact Sheet 05–97: 1e-8e.

Kok, F.J. 2009. Richtlijnen goede voeding en recente ontwikkelingen. NZO-NVD Symposium, 4 December, Ede, the Netherlands.

Lindmark-Månsson, H., R. Fondén, and H.E. Pettersson. 2003. Composition of Swedish dairy milk. Int. Dairy J. 13: 409-425.

Marziali, A.S., and K.F. Ng-Kwai-Hang. 1986. Effects of milk composition and genetic polymorphism on cheese composition. Journal of Dairy Science, 69: 2533-2542.

Mayer, H.K., M. Ortner, E. Tschager, and W. Ginzinger. 1997. Composite milk protein phenotypes in relation to composition and cheesemaking properties of milk. Int. Dairy J. 7: 305-310.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic gain value using genome-wide dense marker maps. Genetics 157: 1819-1829.

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

Niki, R., K., Kohyama, Y. Sano and K. Nishinari. 1994. Rheological study on the rennet-induced gelation of casein micelles with different sizes. Polymer Geld and Network 2: 105-118.

Park, S-Y., R., Niki and Y., Sano. 1999. Size effects of casein micelles on rennet gels in the presence of β-lactoglobulin. Int. Dairy J. 9: 379-380.

de Roos, A.P.W., B.J. Hayes, R.J. Spelman and M.E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179: 1503-1512.

de Roos, A.P.W., C. Schrooten, E. Mullart, S. Van Der Beek, G. De Jong and W. Voskamp. 2009. Genomic Selection at CRV. Proceedings of the Interbull international Workshop - Genomic information in genetic evaluations, Uppsala, Sweden, January 26-29: pages 47-50.

Rowland, S.J. 1938. The precipitation of the proteins in milk. I. Casein. II. Total proteins. III. Globulin. IV. Albumin and Proteose-peptone. J. Dairy Res. 9: 30:41.

Rutten, M.J.M., P. Bijma, J.A. Woolliams, and J.A.M. van Arendonk. 2002. SelAction: software to predict selection response and rate of inbreeding in livestock breeding programs. J. Hered. 93: 456-458.

Rutten, M.J.M., H.Bovenhuis, K.A. Hettinga, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J. Dairy Sci. 92: 6202-6209.

Schaar, J., B. Hansson, and H. E. Pettersson. 1985. Effects of genetic variants of kappa casein and beta lactoglobulin on cheesemaking. J. Dairy Res. 52: 429-438.

Spelman, R.J., D.J. Garrick, and J.A.M. van Arendonk. 1999. Utilisation of genetic variation by marker assisted selection in commercial dairy cattle populations. Livest. Prod. Sci. 59: 51-60.

Steijns, J.M. 2008. Dairy products and health: Focus on their constituents or on the matrix? Int. Dairy J. 18: 425-435.

St-Gelais, D., and S. Haché. 2005. Effect of *β*-casein concentration in cheese milk on rennet coagulation properties, cheese composition and cheese ripening. Food Res. Int. 38: 523-531.

Van den Berg, G., J.T.M. Escher, P.J. de Koning, and H. Bovenhuis. 1992. Genetic polymorphism of kappa-casein and beta-lactoglobulin in relation to milk composition and processing properties. Neth. Milk Dairy J. 46: 145-168.

VanRaden, P.M., C.P. van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor and F.S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16-24.

Walsh, C.D., T.P.Guinee, W.D.Reville, D.Harrington, J.J.Murphy, B.T. O'Kennedy and R.J. Fitzgerald. 1998. Influence of κ-casein genetic variant on rennet gel microstructure, cheddar cheesemaking properties and casein micelle size. Int. Dairy J. 8: 707-714.

Walstra, P., and R. Jenness, eds. 1984. Protein composition of milk. Dairy Chemistry and Physics. Wiley, New York, NY.

Wedholm, A., L.B. Larsen, H. Lindmark-Månsson, A.H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. J. Dairy Sci. 89: 2396-3305.

de Wit, J.N. 1998. Nutritional and Functional Characteristics of Whey Proteins in Food Products. J. Dairy Sci. 81: 597 – 608.

Summary

Milk, especially cow's milk, is consumed as a food product in many cultures. Besides consumption as liquid milk, milk is also consumed in the form of processed dairy products such as butter, yoghurt and cheese. For many years dairy cows have been selected for high milk, fat and protein production. It is not known, however, what the consequences of this selection policy are on e.g. the composition of the milk fat and the milk protein. This thesis is part of the Dutch Milk Genomics Initiative which investigates the possibilities to use natural genetic variation for changing milk composition. The focus of this thesis has been on milk protein composition, and the objectives were to estimate genetic parameters for milk protein composition, to detect chromosomal regions affecting milk protein composition, and to fine map these chromosomal regions. For this purpose morning milk samples of about 2000 first lactation cows were collected in the winter of 2005. From these 2000 cows, also a blood sample was taken for DNA analyses. The morning milk samples were analyzed for the six major milk proteins (α_{S1}-casein, α_{S2}-casein, β-casein, κ-casein, α-lactalbumin and β-lactoglobulin) using capillary zone electrophoresis. Blood samples were taken from all cows to extract DNA for genotyping.

In **chapter 2** genetic parameters for milk protein composition were estimated and the relationships among the six major milk proteins as well as between milk proteins and milk production traits were studied. Results showed that there was considerable genetic variation for milk protein composition with heritabilities ranging from 0.25 for β-casein to 0.80 for β-lactoglobulin. Genetic correlations among the six major milk proteins were low, in general. Protein percentage was negatively correlated with α_{S1}-casein and α–lactalbumin and positively correlated with κ-casein. There was a strong negative genetic correlation between β-lactoglobulin and total casein in milk. The presence of genetic variation justified the performance of in-depth genetic analyses such as linkage and association mapping.

In **chapter 3**, two types of molecular markers were compared for their use in genetic studies. Microsatellites are very informative due to the large number of alleles that each microsatellite can have. Single nucleotide polymorphisms (SNPs) are less informative than microsatellites because

SNPs usually have only two alleles. However, SNPs have the major advantage of being more suitable for high-throughput genotyping. The results of this study showed that three SNPs are required to achieve the same information content as one microsatellite. In chapters 4, 5 and 6, SNPs were used as genetic markers to detect chromosomal regions that affect milk protein composition.

In **chapter 4**, a linkage study was performed to screen the whole bovine genome to identify chromosomal regions affecting milk protein composition. In total, ten significant chromosomal regions were detected. The chromosomal regions most significantly related to milk protein composition ($P_{genome} < 0.05$) were found on *Bos taurus* autosomes (BTA) 6, 11 and 14. The proportion of the phenotypic variance explained by these chromosomal regions ranged from 4% for β-casein on BTA6 to 28% for β-lactoglobulin on BTA11. Effects of these chromosomal regions could be partially explained by known polymorphisms in milk protein genes. The confidence intervals for chromosomal regions as detected in this linkage study, however, were rather large. To narrow down the chromosomal regions and to detect new chromosomal regions affecting milk protein composition, a whole genome association study using 50k SNPs was performed (**chapter 5)**. Chromosomal regions with SNPs significantly associated with milk protein composition were distributed over 15 bovine autosomes. The main regions significantly associated with milk protein composition were found on BTA5, 6, 11 and 14. The proportion of genetic variance explained by the SNP most significantly associated with on of the trait on these four chromosomes ranged from 24.7% for β-CN on BTA6 to 65.8% for β-LG on BTA11. Besides the four main regions, several other chromosomal regions affecting one of six major milk proteins were detected. Although these regions only explain a small part of the genetic variance, they might in addition to the four main regions also play a role in the genetic regulation of milk protein synthesis.

In **chapter 6**, a single SNP association study as used in chapter 5 was compared to a multiple SNP association study. Results of both analyses for the six major milk proteins were compared. The number of SNPs with effects is considerably lower in the multiple SNP analysis than in the single SNP analysis. In addition, the multiple SNP analysis detected chromosomal

regions which were not detected in the single SNP analysis. These results suggest that multiple SNP analysis has a higher power to detect associations as compared to a single SNP analysis.

The **last chapter** discusses different scenarios to increase the casein index, which is preferable for the cheese production. Four different scenarios were described. The first scenario has been termed genetic differentiation. In this scenario it is assumed that a group of farmers use the top 10% of the bulls with respect to their estimated breeding value (EBV) for casein index. In the second scenario the current breeding goal was adjusted and selection is only aimed at an increased casein index. Selection can be based on EBVs, which is scenario 2, or based on genetic information of genes known to affect the casein index, which is scenario 3. In scenario 4, genomic selection was used and bulls are selected based on genomic breeding values. These four scenarios illustrated that there are opportunities to utilize genetic variation in milk protein composition. In practice also a combination of scenario 3 and 4 might be feasible. This would imply selection based on information of known genes as well as anonymous genes (genomic selection).

The main conclusions of this thesis are:

- There is considerable genetic variation for milk protein composition and the genetic correlations among the six major milk proteins are in general low.
- Genomic regions on chromosome 5, 6, 11 and 14 are significantly associated with milk protein composition. Effects detected in these regions can be partially explained by polymorphisms in known milk protein genes.
- Besides the four main regions, several other chromosomal regions have been detected that significantly affect one of the six major milk proteins.
- The use of a multiple SNP analysis as compared to single SNP analysis results in a higher power to detect association with milk protein composition.

This thesis provides new insight in the genetic regulation of for milk protein composition and it shows that there are interesting possibilities

to change the cow's milk protein composition by means of selective breeding.

Samenvatting

Melk, in het bijzonder koemelk, wordt in veel culturen geconsumeerd als voedingsproduct. Naast consumptiemelk wordt melk ook geconsumeerd in andere melkproducten zoals boter, yoghurt en kaas. Melkkoeien zijn de afgelopen jaren sterk geselecteerd op een hoge melkproductie met een hoog vet en eiwitgehalte. Maar het is niet bekend wat de gevolgen zijn van deze selectie op de melkvet en melkeiwit samenstelling. Dit proefschrift is onderdeel van het Nederlandse Milk Genomics Initiatief wat gericht is op de mogelijkheden om de natuurlijke genetische variatie van de melksamenstelling te gebruiken om de melksamenstelling te veranderen. Dit proefschrift is gericht op melkeiwitsamenstelling en had de volgende doelen: het schatten van genetische parameters voor melkeiwitsamenstelling, het identificeren van gebieden op het DNA die een effect hebben op de melkeiwitsamenstelling en het kleiner maken van deze gebieden. Om deze doelen te bereiken zijn ochtend melkmonsters van 2000 koeien in hun eerste lactatie verzameld in de winter van 2005. De ochtend melkmonsters zijn geanalyseerd voor de zes grote melkeiwitten (α_{S1}-caseïne, α_{S2}-caseïine, β-caseïne, κ-caseïne, α-lactalbumine en β-lactoglobuline) met behulp van capillaire zone electroforese. Naast de melkmonsters zijn er ook bloedmonsters genomen van alle 2000 koeien. Uit deze bloedmonsters is DNA geïsoleerd om alle koeien te kunnen genotyperen.

In **hoofdstuk 2** zijn de genetische parameters voor melkeiwitsamenstelling geschat, zijn de relaties tussen de zes verschillende melkeiwitten bestudeerd en zijn de relaties tussen de melkeiwitten en de melkproductie kenmerken bestudeerd. De resultaten lieten zien dat er een behoorlijke genetische variatie is voor de melkeiwitten, met erfelijkheidsgraden die variëren van 0.25 voor β-caseïne tot 0.80 β-lactoglobuline. De genetische correlaties tussen de zes melkeiwitten waren in het algemeen laag. Eiwit percentage was negatief gecorreleerd met α_{S1}-caseïne en α-lactalbumine, en positief gecorreleerd met κ-caseïne. Verder was er een sterke negatieve genetische correlatie tussen β-lactoglobuline en het totaal aan caseïne in de melk. De aanwezigheid van genetische variatie was nodig om

gedetailleerde genetische analyses te doen zoals een linkage studie en een associatie studie.

In **hoofdstuk 3**, twee verschillende soorten moleculaire merkers zijn vergeleken op basis van hun gebruik in genetische analyses. Microsatellieten zijn erg informatieve merkers omdat iedere microsatelliet een groot aantal allelen bevat. Single Nucleotide Polymorfismen (SNPs) zijn minder informatief dan microsatellieten omdat SNPs gebruikelijk maar twee allelen bevatten. Maar, SNPs hebben het grote voordeel dat ze geschikter zijn voor het genotyperen van duizenden merkers. De resultaten van de deze studie laten zien dat 3 SNPs nodig zijn om dezelfde informatie te krijgen als een microsatelliet. In de hoofdstukken 4, 5 en 6 zijn SNPs als genetische merkers gebruikt om gebieden op het DNA te identificeren die een effect hebben op de melkeiwitsamenstelling.

In **hoofdstuk 4** is een linkage studie gedaan waarbij het hele koeien genoom is gescreend om gebieden op het DNA te vinden die een effect hebben op de melkeiwitsamenstelling. In totaal zijn er 10 significante gebieden op het DNA gevonden. De DNA gebieden met het grootste effect op de melkeiwitsamenstelling lagen op de *Bos taurus* autosomen (BTA) 6, 11 en 14. Het percentage van de fenotypische variatie verklaard door deze DNA gebieden varieerde van 4% voor β-caseïne op BTA6 tot 28% voor β-lactoglobuline op BTA11. De effecten van de DNA gebieden op de melkeiwitsamenstelling konden gedeeltelijk worden verklaard door bekende polymorfismen in de melkeiwit genen. De betrouwbaarheidsintervallen van de DNA gebieden zoals die in de linkage studie geïdentificeerd zijn, waren vrij lang. Om deze DNA gebieden kleiner te maken en om nieuwe DNA gebieden met een effect op de melkeiwitsamenstelling te identificeren, is er een associatie studie voor het hele koeien genoom gedaan met behulp van 50.000 SNPs (**hoofdstuk 5**). De DNA gebieden die SNPs bevatten die een effect hebben op de melkeiwitsamenstelling waren verdeeld over 15 *Bos taurus* autosomen. De DNA gebieden met de grootste effecten op de melkeiwitsamenstelling waren gevonden op BTA 5, 6, 11 en 14. Het percentage genetische variatie verklaard door de SNP die het grootste effect had op een van de kenmerken op deze vier chromosomen varieerde van 24.7% voor β-caseïne op BTA6 tot 65.8% voor β-lactoglobuline op BTA11. Naast deze vier DNA gebieden met enorme effecten, waren er ook

DNA gebieden op andere chromosomen gevonden die een effect hadden op een van de zes grote melkeiwitten. Ondanks dat deze gebieden maar een klein gedeelte van de genetische variatie verklaarde, kunnen ze als toevoeging aan de vier DNA gebieden met enorme effecten, een rol spelen bij de genetische regulatie van de melkeiwit synthese.

In **hoofdstuk 6** was de single SNP associatie studie (zoals gebruikt in hoofdstuk 5) vergeleken met een multiple SNP associatie studie. De resultaten van beide analyses voor de zes grote melkeiwitten zijn vergeleken. In zowel de single SNP analyse als de multiple SNP analyse zijn de DNA gebieden met de grootste effecten gevonden op BTA5, 6 11 and 14. Het aantal SNPs met een effect op de melkeiwitsamenstelling was aanzienlijk lager in de multiple SNP analyses dan in de single SNP analyse. Bovendien zijn in de multiple SNP analyse DNA gebieden geïdentificeerd die niet waren geïdentificeerd in de single SNP analyse. Deze resultaten suggereren dat multiple SNP analyse een hogere power heeft om associaties te vinden dan single SNP analyse.

Het **laatste hoofdstuk** bediscussieert verschillende scenario's om de caseïne index in melk te verhogen. Een verhoogde caseïne index in melk is gewenst voor de kaas productie. In totaal worden er vier scenario's besproken. Het eerste scenario is genetische differentiatie. In dit scenario wordt aangenomen dat een groep boeren de 10% beste stieren gebruiken op basis van hun fokwaarde voor caseïne index. In het tweede scenario wordt het huidige fokdoel aangepast en zal selectie alleen gebaseerd zijn op een verhoging van de caseïne index. Deze selectie kan gebaseerd zijn op fokwaardes, zoals in scenario 2, of kan gebaseerd zijn op genetische informatie van genen waarvan bekend is dat ze een effect hebben op caseïne index, zoals in scenario 3. In scenario 4, worden stieren geselecteerd op basis van genomische fokwaardes voor caseïne index. Dit wordt selectie op basis van merker informatie genoemd ('genomic selection'). Alle vier de scenario's illustreren dat er mogelijkheden zijn om de genetische variatie in melkeiwitsamenstelling te gebruiken. In de praktijk is het ook goed mogelijk om een combinatie van scenario 3 en 4 uit te voeren. Dit zou betekenen dat er selectie plaats vindt op basis van informatie van bekende genen als onbekende genen (genomic selection).

De belangrijkste conclusies van dit proefschrift zijn:

- Er is een behoorlijke genetische variatie voor melkeiwitsamenstelling en de genetische correlaties tussen de zes grote melkeiwitten waren in het algemeen laag.
- DNA gebieden op chromosoom 5, 6, 11 en 14 hebben een groot effect op de melkeiwitsamenstelling. De effecten van de SNPs in deze DNA gebieden kunnen gedeeltelijk worden verklaard door polymorfismen in bekende melkeiwit genen.
- Naast de vier DNA gebieden met de grootste effecten, zijn er ook verschillende andere DNA gebieden die een effect hebben op een van de zes grote melkeiwitten.
- Het gebruik van multiple SNP analyse zal een hogere power hebben om associaties met de melkeiwitsamenstelling te identificeren dan het gebruik van single SNP analyse.

Dit proefschrift geeft nieuwe inzichten in de genetische regulatie van de melkeiwitsamenstelling en het laat zien dat er interessante mogelijkheden zijn om de melkeiwitsamenstelling van de koe te veranderen met behulp van selectieve fokkerij.

Dankwoord

December 2005 ben ik bij de leerstoelgroep fokkerij en genetica begonnen als kwantitatieve aio op het Milk Genomics project. Het leuke, uitdagende en interessante van dit project was dat er een nauwe samenwerking was tussen de leerstoelgroep fokkerij en genetica, de leerstoelgroep zuivel en de verschillende bedrijven. Hierdoor kwamen verschillende disciplines meerdere malen bij elkaar, waardoor er van een andere hoek naar de behaalde resultaten werd gekeken. Dit leverde leuke en boeiende discussies op en leidden vaak tot extra analyses om nog verder in het diepe te duiken. Het voordeel was ook dat door de samenwerking met de leerstoelgroep zuivel en de bedrijven, het uitleggen van de resultaten soms een uitdaging was om dit begrijpbaar te maken voor de niet genetica mensen. Graag wil ik iedereen van het Milk Genomics project bedanken voor de leuke, gezellige en fijne samenwerking van de afgelopen 4 jaar.

Mijn dagelijkse begeleiders wil ik nog in het bijzonder bedanken. Henk, jouw kennis, ervaring en ideeën hebben een belangrijke bijdrage geleverd aan de tot standkoming van dit proefschrift. Hartstikke bedankt! Marleen, ook jou wil graag bedanken voor jouw bijdrage aan dit proefschrift. Naast de spel en grammatica fouten die jij eruit haalde, had jij ook inhoudelijk een goede bijdrage (zeker als je bedenkt dat jij uit de plantenwereld komt). Naast mijn dagelijkse begeleiders, wil ik ook graag Johan bedanken voor zijn ideeën en suggesties, die vaak tot extra analyses leidden die weer een positieve bijdrage aan de artikelen leverde. Naast jullie inhoudelijk en begeleidende bijdrage, wil ik jullie ook bedanken voor jullie bijdrage aan mijn persoonlijke ontwikkeling. Deze ontwikkeling zal ik zeker voort zetten in mijn huidige baan bij CRV.

Tijdens mijn promotie onderzoek heb ik een aantal maanden in Lelystad gewerkt. Ik wil Roel bedanken die mij de mogelijkheid heeft gegeven om een tijdje in Lelystad te werken om zo te ervaren hoe het is om bij een andere organisatie te werken. Ook al zit er veel overlap tussen Lelystad en Wageningen, zijn er ook veel verschillen. Zeker het carpoolen naar Lelystad kost veel tijd en is best wel vermoeiend. Mario, bedankt voor jouw steun en begeleiding tijdens mijn werkzaamheden in Lelystad.

De laatste maanden van 2009 was het soms wel stressen om alles op tijd af te krijgen, zeker het feit dat ik al een nieuwe baan bij CRV had gevonden. Graag wil ik bij deze ook CRV bedanken die mij de mogelijkheid heeft gegeven om de leesversie van mijn proefschrift zo goed als af te ronden voordat ik bij CRV aan de slag zou gaan.
Ik heb de leerstoelgroep fokkerij en genetica als een plezierige, gezellige, fijne en warme leerstoelgroep ervaren. Naast de ontspannen sfeer tijdens werktijd, waren er ook vele leuke, sportieve en gezellige momenten tijdens de vele 'uitjes', bier- en spelavonden en home-made diners. Hiervoor wil ik iedereen van de leerstoelgroep fokkerij en genetica bedanken. In het bijzonder wil ik Albart, Raoul, en Aniek bedanken die de afgelopen 4 jaar voor een leuke, gezellige en rustige sfeer op onze kamer hebben gezorgd. Albart en Patrick, ik heb het al meerdere malen gezegd, maar ik vind het super leuk dat jullie vandaag naast mij op het podium willen staan.
Zoals jullie hierboven hebben kunnen lezen, hebben veel mensen bijgedragen aan de tot standkoming van dit proefschrift en aan de leuke werksfeer van de afgelopen vier jaar, maar dit alles zou nooit hebben plaats gevonden zonder mijn ouders. Pap en mam, jullie hebben mij de mogelijkheid gegeven om te gaan studeren en die mogelijkheid heb ik met beide handen aangepakt. Ik wil jullie bedanken voor deze mogelijkheid en voor jullie steun tijdens mijn studie en promotie onderzoek. Ook de gezellige, sportieve en leuke weekendjes weg met z'n allen waren altijd goede momenten om het werk even helemaal te vergeten. Yvette, jouw steun aan mij is ook altijd aanwezig evenals jouw betrokkenheid (ook al was het niet altijd begrijpbaar wat ik deed). Bedankt! Wilbert, bedankt voor jouw interesse in mijn werk. Ook al ben jij, Maud, pas anderhalf jaar op deze aardbol, jouw vrolijkheid en ontdeuglijkheid hebben bijgedragen aan vele ontspannende momenten, zeker tijdens het afronden van dit proefschrift.
Last, but not least wil ik jou, Joshua, bedanken. Als ik weer eens op congres was, moest jij alleen voor Chika zorgen en daar heb ik jou nooit over horen klagen. Jouw steun die je mij geboden hebt om in de avonduren en in het weekend (voornamelijk in de afgelopen maanden) aan mijn proefschrift te werken, hebben een positieve bijdrage geleverd aan de tot standkoming van dit proefschrift. Net zo goed als jouw betrokkenheid en

interesse in mijn promotie onderzoek. Ik ben ook super blij met jouw creatieve bijdrage aan het ontwerp van de cover. Jij bent mijn steun en toeverlaat, bedankt voor alles!!

Ghyslaine

About the author

Ghyslaine Carla Bert Schopen was born on November 22 1979 in Heerlen, the Netherlands. She finished her HAVO at the Broekland College in Hoensbroek in 1997. In 2000, she finished her VWO at the Bernadinus College in Heerlen. In September 2000, she started her study Animal Sciences at Wageningen University. During her study, her major specialization was in Animal Health, Welfare and Care, and her minor specialization was in Animal Breeding and Genetics. Five years later, she graduated and received her Master degree in Animal Sciences. In December 2005, she started her PhD study at the Animal Breeding and Genomics Centre of Wageningen University. This PhD study was part of the Milk Genomics Initiative and the results are described in this thesis. Since February 2010, she is working as researcher quantitative genetics at Research & Development of the Genetic Products department at CRV, the main cooperative cattle improvement organization in the Netherlands.

Ghyslaine Carla Bert Schopen werd geboren op 22 november 1979 te Heerlen. In 1997 behaalde zij haar HAVO diploma bij het Broekland College te Hoensbroek. Vervolgens ging zij naar het VWO waar zij in 2000 haar VWO diploma behaalde bij het Bernadinus College te Heerlen. In september 2000 begon Ghyslaine aan haar studie dierwetenschappen aan de Wageningen Universiteit. Tijdens haar studie specialiseerde zij zich in diergezondheid, welzijn en verzorging en haar tweede specialisatie was in fokkerij en genetica. Na vijf jaar behaalde zij haar master diploma in dierwetenschappen. In december 2005 begon Ghyslaine aan haar aio project bij de leerstoelgroep fokkerij en genetica. Dit aio project was onderdeel van het Milk Genomics Initiative en de resultaten van het aio project staan beschreven in dit proefschrift. Sinds februari 2010 werkt Ghyslaine als onderzoeker kwantitatieve genetica bij Research & Development van de afdeling genetische producten bij CRV, de grootste coöperatie rundveeverbetering van Nederland.

List of publications

Papers in refereed journal

Schopen, G.C.B, H. Bovenhuis, M.H.P.W. Visker, and J.A.M. van Arendonk. 2008. Comparison of information content for microsatellites and SNPs in Poultry and Cattle. Short communication. Animal Genetics 39: 451 – 453.

Schopen G.C.B., J.M.L. Heck, H. Bovenhuis, M.H.P.W. Visker, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. Journal of Dairy Science 92: 1182-1191.

Schopen G.C.B., P.D. Koks, J.A.M. van Arendonk, H. Bovenhuis, and M.H.P.W. Visker. 2009. Whole genome scan to detect quantitative trait loci for bovine milk protein composition. Animal Genetics 40: 524–537.

Demeter, R. M., G.C.B. Schopen, A.G.J.M. Oude Lansink, M.P.M. Meuwissen, and J.A.M. van Arendonk. 2009. Effects of milk fat composition, DGAT1, and SCD1 on fertility traits in Dutch Holstein cattle. Journal of Dairy Science 92: 5720-5729.

Bouwman, A.C., G.C.B. Schopen, H. Bovenhuis, M.H.P.W. Visker, and J.A.M. van Arendonk. 2010. Genome-wide scan to detect quantitative trait loci for milk urea nitrogen in Dutch Holstein Friesian cows. Journal of Dairy Science. In Press.

Papers submitted or in preparation

Schopen G.C.B., M.H.P.W. Visker, P.D. Koks, E. Mullaart, J.A.M. van Arendonk, and H. Bovenhuis. 2010. Whole genome association study for milk protein composition in dairy cattle. Submitted.

Schopen, G.C.B., M.P.L. Calus, M.H.P.W. Visker, J.A.M. van Arendonk, and H. Bovenhuis. Single and multiple SNP genome wide association analysis in dairy cattle. In preparation.

Conference abstracts

Schopen, G.C.B., H. Bovenhuis, M.H.P.W. Visker, and J.A.M. van Arendonk. 2007. Comparison of Single Nucleotide Polymorphism

and microsatellite polymorphism for QTL mapping. Proceedings of the 58th annual meeting of the European association for animal production (EAAP), Dublin, Ireland, August 26-29, book of abstracts No. 13, p. 168.

Schopen, G.C.B., P.D. Koks, J.A.M. van Arendonk, H. Bovenhuis, and M.H.P.W. Visker. 2008. QTL detection for milk protein composition of bovine milk. XXXI Conference of the International Society for Animal Genetics (ISAG), Amsterdam, the Netherlands, July 20-24, book of abstract, session 5000, poster nr. 5017.

Schopen, G.C.B., J.M.L. Heck, H. Bovenhuis, M.H.P.W. Visker, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2008. Genetic parameters for milk protein composition of dairy cows. Proceedings of the 59th annual meeting of the European association for animal production (EAAP), Vilnius, Lithuania, August 24-27, book of abstracts No. 14, p. 115.

Schopen, G.C.B., P.D. Koks, J.A.M. van Arendonk, H. Bovenhuis, and M.H.P.W. Visker. 2009. Quantitative trait loci detection for milk protein composition in Dutch Holstein-Friesian cows. Proceedings of the 60th annual meeting of the European association for animal production (EAAP), Barcelona, Spain, August 24-27, book of abstracts No. 15, p. 183.

Demeter, R.M., G.C.B. Schopen, A.G.J.M. Oude Lansink, M.P.M. Meuwissen, and J.A.M. van Arendonk. 2009. Effects of milk fat composition, DGAT1 and SCD1 on fertility traits in Dutch Holstein cattle. ADSA-CSAS-ASAS Joint Annual Meeting, Montreal, Quebec, Canada, July 12-16, p.123 (35).

Schopen, Ghyslaine, Patrick Koks, Johan van Arendonk, Henk Bovenhuis, and Marleen Visker. 2009. Quantitative trait locus detection for milk protein composition in dairy cattle. 6th International Symposium on Milk Genomics & Human Health, Paris, France, September 28-30.

Awarded Presentations

Schopen, G.C.B., H. Bovenhuis, M.H.P.W. Visker, and J.A.M. van Arendonk. 2007. Comparison of Single Nucleotide Polymorphism

and microsatellite polymorphism for QTL mapping. 58th annual meeting of the European association for animal production (EAAP), Dublin, Ireland, August 26-29. Best poster award of the EAAP 2007 (Rommert Politiek Award).

Schopen, G.C.B., M.H.P.W. Visker, J.A.M. van Arendonk, and H. Bovenhuis. 2008. QTL detection for milk protein composition in dairy cattle. XXXI Conference of the International Society for Animal Genetics (ISAG), Amsterdam, the Netherlands, July 20-24. One of the best poster award 2008.

Schopen, G.C.B., J.M.L. Heck, H. Bovenhuis, M.H.P.W. Visker, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2008. Genetic parameters for milk protein composition of dairy cows. 59th annual meeting of the European association for animal production (EAAP), Vilnius, Lithuania, August 24-27. Best poster award of the 'Animal Genetics' session.

Training and Supervision Plan

The Basic Package (3 credits)

Course on Philosophy of Science and/or Ethics	2006
WIAS Introduction Course	2007

Scientific Exposure (21 credits)

International conferences

8th WCGALP,Belo Horizonte (BR),august 13-18	2006
3rd International Symposium Milk Genomics & Human Health, Brussel (Bel), September 19-21	2006
58th Annual meeting of the EAAP, Dublin (IR), august 26-29	2007
31th Conference of the ISAG, Amsterdam (NL), July 20-24	2008
59th Annual meeting of the EAAP,Vilnius (LT), August 24-27	2008
60th Annual meetin gof the EAAP,Barcelona (SP), August 24-27	2009
6th International Symposium Milk Genomics & Human Health, Paris (Fr), September 28-30	2009

Seminars and workshops

KNAW Special Colloquium, Amsterdam (NL), March 14-17	2006
F&G connection days,Vught (NL), November	2006/08
WIAS Science Day, Wageningen (NL), March	2006-09
Genetics of milk quality, Wageningen (NL), April 29	2009

Presentations

WIAS Science Day (poster)	2007
58th Annual meeting of the EAAP (poster)	2007
59th Annual meeting of the EAAP (poster)	2008
31th Conference of the ISAG (poster)	2008
F&G connection days (oral)	2008
WIAS Science Day (oral)	2009
Genetics of milk quality (oral)	2009
60th Annual meeting of the EAAP (oral)	2009
6th International Symposium Milk Genomics & Human Health (oral)	2009

In-Depth Studies (11 credits)

Bayesian Analysis in Animal Breeding, Göttingen (GER), June 6-10	2006
Fortran 95 for Animal Breeding and Quantitative Genetics, Lelystad (NL), June 11-15	2007
Understanding genotype environment interactions, Wageningen (NL), June 25-29	2007
Linear models in animal breeding, Wageningen (NL), July 2-6	2007
QTL Mapping, MAS, and Genomic Selection, Lelystad (NL), March 10-14	2008
Introduction to R for statistical analysis, Wageningen, 21-22 April	2008
Nutrient Density of Milk, Wageningen (NL), January 26-28	2009
Use of High-density SNP Genotyping for Genetic Improvement of Livestock, Ames (USA), June 1-10	2009

Professional Skills Support Courses (8 credits)

Techniques for Writing and Presenting a Scientific paper	2006
Project- and Time management	2006
Tutorship course	2006
PhD Competence assessment	2006
Personal Efficiency	2006
Groepsgesprekken	2007
Gespreksvaardigheden	2007
Presentation Skills	2007
Course Supervising MSc thesis work	2007

Didactic Skills Training (7 credits)

Assisting Genetic Improvement of Livestock (GIL)	2006
Supervising HBO stagaire	2007
Supervising 2 minor Msc theses	2007/08
Supervising 1 major Mcs thesis	2009

Management Skills Training (11 credits)

Secretary of WIAS associated PhD students (WAPS) council	2006/07
Chairman of WIAS associated PhD students (WAPS) council	2007/08
Organization WIAS Science Day	2008

Total credits: 61

[1] one credit equals a study load of approximately 28 hours

Colophon

The research as described in this thesis is part of the Milk Genomics Initiative, funded by Wageningen University, the Dutch dairy association NZO, cooperative cattle improvement organization CRV, and the Dutch technology foundation STW.

This thesis was printed by Wöhrmann Print Service, Zutphen, the Netherlands.