The genetic background of bovine milk infrared spectra

Qiuyu Wang

Thesis committee

Promotor
Prof. Dr H. Bovenhuis
Personal chair at Animal Breeding and Genomics
Wageningen University & Research

Other members
Prof. Dr H. Hogeveen, Wageningen University & Research
Prof. Dr M. De Marchi, University of Padova, Italy
Prof. Dr H. Soyeurt, University of Liège, Belgium
Dr T. Huppertz, FrieslandCampina, the Netherlands

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

The genetic background of bovine milk infrared spectra

Qiuyu Wang

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday May 21, 2019
at 4 p.m. in the Aula.

Wang, Q.
The genetic background of bovine milk infrared spectra,
160 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2019)
With references, with summaries in English

ISBN: 978-94-6343-902-2
DOI: https://doi.org/10.18174/472245

Abstract

Wang, Q. (2019). The genetic background of bovine milk infrared spectra.
PhD thesis, Wageningen University, Wageningen, the Netherlands

Milk infrared (IR) spectroscopy is a cheap, quick and high-throughput technique that has been widely used to determine milk components. It has been used as the standard method for routine quantification of fat, protein and lactose content of milk, and it is a promising technique to obtain information about milk composition. The aim of this thesis was to explore the genetic background of bovine milk IR spectra, identify the environmental factors affecting milk IR spectra, and combined use genotypic information and milk IR spectra in predicting dairy cattle phenotypes. Two studies were conducted to explore the genetic background of milk IR spectra of Holstein Friesian dairy cows in the Netherlands. Studies were focused on individual IR wavenumbers, and results showed that for many of them 20 to 60% of variation can be attributed to genetic factors. Polymorphisms of individual gene diacylglycerol O-acyltransferase 1 (DGAT1), κ-casein (CSN3) and β-lactoglobulin (LGB), as well as lactation stage of dairy cows and the different dates of IR analysis have significant effect on the values of milk IR spectra. Genome wide association study (GWAS) identified the associated genomic regions. In addition to the regions that related to milk fat, protein and lactose content, this thesis detected 3 new regions related to phosphorus, orotic acid and citric acid content in milk. Knowledge of the genetic background of milk IR spectra could enhance the prediction for dairy cattle phenotypes. This thesis investigated if combined use of genotypes of dairy cows can improve the prediction for milk fat composition. Results suggest that prediction accuracy of unsaturated fatty acids can be considerably improved by adding stearoyl-CoA desaturase (SCD1) genotypes of dairy cows. Predicting methane (CH_4) emission based on milk IR spectra is of great interest due the environmental impact of dairy production. This thesis showed the importance of validation strategy in interpreting the results of predicting CH_4 emission. This result has general value in milk IR prediction for dairy cattle phenotypes that a block cross validation with farms as block could reflect the true predicative ability for independent observations. This thesis also suggested to predict based on IR wavenumbers from water absorption regions of the spectra as a negative control, to detect potential problem due to dependency structure in the data.

To my father who never leaves me walking alone
Finally I become the man you wanted me to be

Contents

1

General introduction

1.1 Milk composition

Bovine milk serves as an important human food source and is used to manufacture many dairy products such as cheese, butter and yogurt. During the last few decades, dairy cattle breeding has been a significant factor in improving milk yield and in this way contributes to meet the increasing demand for milk products. Nowadays, although the global milk consumption is still growing, milk consumption has been declining in developed countries such US and some countries in the EU (Canadian Dairy Information Centre, 2018). This might be due to a decline in population growth and because the saturation level for dairy consumption has been reached. In addition, dairy production has been facing concerns on its environmental sustainability and animal health. The changes in dairy production circumstances suggest that the dairy industry should put more emphasis on sustainable dairy production. This change requires new phenotypes which can serve as indicators for health and energy status, environmental impact, and milk quality, to support management or selective breeding decisions. Acquiring these indicators on a large scale and in a cost effective way is challenging. Milk composition, however, might provide indicators for these traits on a large scale and in a cost effective way.

1.1.1 Milk composition and human health

The milk components fat and protein are currently used in many countries to determine the value of raw milk. Milk fat consists of many different fatty acids and other important nutrients like the fat soluble vitamins (German and Dillard, 2006; Jelen, 2007). Saturated fatty acids account for approximately 70% of the total fatty acids in milk and lead to increased cholesterol levels and an increased risk of cardiovascular disease, whereas the relatively small fraction of polyunsaturated fatty acids in milk have been identified as beneficial to human health (Jensen, 2002; German and Dillard, 2006). Milk protein consists of approximately 80% casein (α_{S1}-casein, α_{S2}-casein, β-casein and κ-casein) and 20% whey (α-lactabumin and β-lactoglobulin) proteins. Milk protein provides many essential amino acids, e.g. lysine, leucine, methionine and threonine (FOX, 1998). Milk protein also includes multifunctional proteins such as lactoferrin that has antimicrobial activity and is part of the innate immune defence (Farnaud and Evans, 2003). Bovine milk also contains several important minerals such as calcium, magnesium, phosphorus, potassium,

selenium, and zinc, which are especially important for pregnant women and children (Black et al., 2002).

1.1.2 Milk composition and manufacturing properties of dairy products

Milk composition is also related to manufacturing properties of milk. Milk fat composition has an influence on melting properties of fat and therefore on the texture of dairy products such as butter, ice cream, yogurt (e.g., Chen et al., 2004). A softer texture is related to a lower melting point, which is associated with an increased unsaturated fatty acid content (e.g., Couvreur et al., 2006). Milk protein composition plays an important role in cheese production. Casein proteins are positively correlated with cheese yield (Wedholm et al., 2006; Heck et al., 2009). Milk protein composition affects milk coagulation properties which is a crucial parameter in cheese making (Auldist et al., 2004; Bonfatti et al., 2010; Penasa et al., 2010).

1.1.3 Milk composition and farm management

Milk composition can serve as an indicator which can be used by farmers to assist in dairy farm management (Hamann and Krömker, 1997). Major milk components e.g. milk fat and protein, are routinely recorded and their changes may indicate changes in the farm management or health status of dairy cows. Bastin et al. (2012) suggested the use of milk fat composition as an indicator for body reserve mobilization and fertility of dairy cows. Schukken et al. (2003) suggested the use of somatic cell counts in milk to monitor milk quality and udder health of dairy cows, since a high somatic cell count is associated with risk of mastitis. Clinical mastitis has also been related to lactoferrin content in milk (Kawai et al., 1999). Changes in β-hydroxybutyrate, milk fatty acids, and fat to protein ratio are related to energy balance of dairy cows which is related to metabolic disorders like ketosis (Van Knegsel et al., 2007; Friggens et al., 2007; Van Haelst et al., 2008; Stoop et al., 2009). Milk urea content, combined with milk protein content, can serve as an indicator for nitrogen efficiency and ammonia emission of dairy cows (Frank and Swensson, 2002). Excreted nitrogen and phosphorus in manure may lead to water pollution and therefore will affect the environmental impact of dairy production. Another topic of concern for the dairy sector is methane (CH_4) emission. Methane is a major greenhouse gas contributing to global warming. Methane production in dairy cows is related to the volatile fatty acids acetate, butyrate and propionate

which are produced during fermentation in the rumen. These volatile fatty acids are also involved in the fatty acids synthesis in the mammary gland. Therefore, it has been suggested that milk composition can be used as an indicator for methane emission of dairy cows (Chilliard et al., 2009; Dijkstra et al., 2011).

Most dairy cattle breeding programs, driven by the milk pricing system, focus on milk production traits such as milk, fat and protein yield (Miglior et al., 2005). In order to meet future demands of modern dairy consumers, dairy cattle breeding programs have to put more emphasis on detailed milk composition, which can be modified by selective breeding (Bovenhuis et al., 2013). In addition, dairy cattle breeding goals will be extended with novel traits that are related to the environmental impact of dairy production, e.g. methane emission. Monitoring farm management requires routine measurements of milk composition which can serve as an indicator for several traits. Therefore the need to record information on milk components is rapidly growing.

1.2 Milk infrared spectroscopy

1.2.1 Principles of milk infrared spectroscopy

Milk composition can be quantified using different techniques. Milk fat composition can be quantified by gas chromatography (GC; de Jong and Badings, 1990; Collomb et al., 2002). Milk protein composition can be quantified by high-performance liquid chromatography (HPLC; Bobe et al., 1998; Bordin et al., 2001), and capillary zone electrophoresis (CZE; de Jong et al., 1993; Heck et al., 2008). However these techniques are costly, time consuming and at present not suited for large scale routine recording of milk components. Alternatively, milk infrared (IR) spectroscopy is a rapid, inexpensive and high-throughput technique for recording milk composition. Nowadays, milk IR spectroscopy is the method of choice for routine milk recording in many countries and is used to quantify milk fat-, protein-, lactose- and urea content (ICAR, 2012). Therefore it is of interest to investigate the potential of milk IR spectroscopy to predict additional milk components (e.g., fat and protein composition) or other phenotypes relevant for dairy production.

IR spectroscopy is one of the vibrational spectroscopy techniques that is based on the interaction between analyzed matter and electromagnetic waves. Electromagnetic radiation is characterized by wavelength (in the unit of nanometer, 10^{-9} meter) and wavenumber (also referred to as frequency; in the unit of reciprocal wavelength in centimeter, cm^{-1}). The relationship between wavelength and wavenumber is

$$v = \frac{10^7}{\lambda(nm)} \qquad \text{Equation 1.1}$$

where v represents the wavenumber, λ is wavelength and 10^7 is a constant.

For convenience, the milk IR spectra are reported in wavenumbers (cm^{-1}) in this thesis. The spectra consist of measurements at a range of individual wavenumbers with a resolution of 3.8558 cm^{-1}.

The electromagnetic radiation used in milk IR spectroscopy consists of two spectral regions: the near-infrared region (NIR, wavenumber 12,500-4,000 cm^{-1}) and the mid-infrared region (MIR, wavenumber 4,000-400 cm^{-1}). The NIR region is the first spectral region exhibiting absorption bands related to molecule vibrations. It is characterized by harmonics and combination bands. The MIR region is the main region of vibrations in molecules. This region contains information allowing the identification and characterization of structure, as well as the conformation of organic structures such as polysaccharides, proteins, and lipids. The ranges of infrared, ultraviolet (UV) and visible (VIS) spectral regions are shown in Figure 1.1. Energy level of the electromagnetic radiation is directly proportional to wavenumbers.

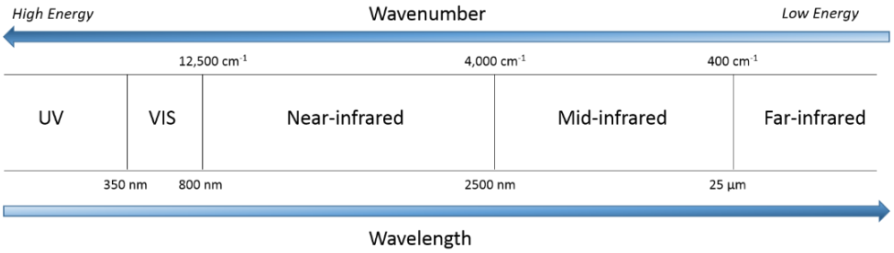


Figure 1.1 The electromagnetic spectrum. Modified from Dufour. (2009)

1.2.2 Molecular vibration and absorption band

For the use of milk IR spectroscopy, bovine milk samples are crossed by electromagnetic radiation, which induces vibrations of chemical bonds within a molecule and thus absorptions of energy from the incoming electromagnetic radiation. There are two types of vibration movements: stretching and bending. Stretching vibration changes the bond length. In symmetric stretching the bonds vibrate in and out simultaneously, while in asymmetric stretching the bonds vibrate in opposite directions. Bending vibration changes the angle between the bonds and atom. The two types of bending vibrations are in plane bending, when atoms stay within the same plane, and out of plane bending, when atoms move outside the original plane. Typical stretching and bending vibrations are shown in Figure 1.2 using the methylene group ($–CH_2$) as an example.

A chemical bond can be considered as a spring that needs a force or energy to compress or extend, responding to Hooke's law. The position of an absorption band by a chemical bond depends on the strength of the chemical bond and the molecule weights of the two atoms. The wavenumber of absorption band can be estimated by Equation 1.2.

$$v_0 = \frac{1}{2\pi c}\sqrt{\frac{K}{\mu}} \qquad \text{Equation 1.2}$$

where v_0 represents wavenumber, c is the speed of light, K is the strength of chemical bond, and μ is the reduced mass of two attached atoms.

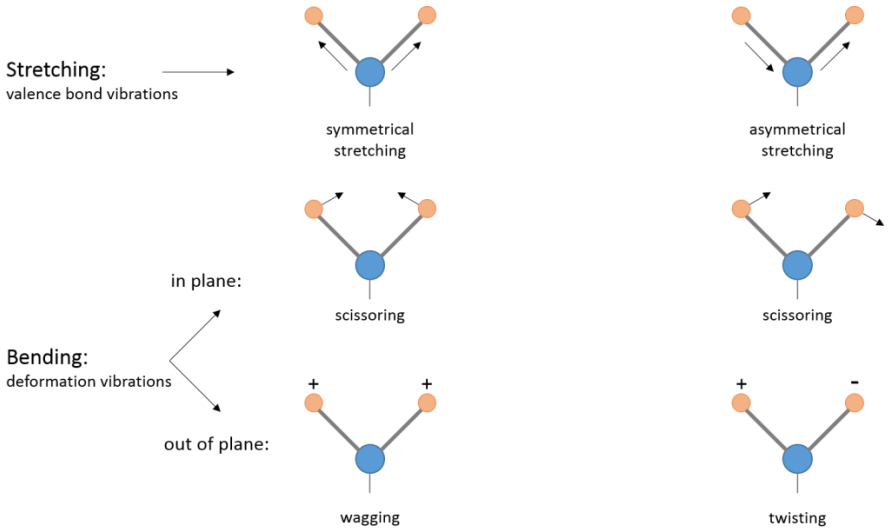


Figure 1.2 Types of stretching and bending vibrations, illustrated on a methylene group ($–CH_2$). Modified from Eck. (2014)

In general, a stronger chemical bond is stiffer, and therefore is associated with a higher K value and therefore vibrates at higher wavenumbers. The chemical bond between two light atoms has a smaller μ and therefore also vibrates at higher wavenumbers. The types of vibrations also influence the absorption position for the same chemical bonds. Bending vibrations are less energetic than stretching and thus vibrate at lower wavenumbers.

The intensity of the absorption band depends on the difference between the two atoms involved in the chemical bond. A larger difference will result in a stronger absorption. For example, the C=O bond formed by different and highly polarized atoms, show a stronger absorption band than a C=C bond.

1.2.3 Assignment of chemical bonds and milk components to spectral regions

Bovine milk contains various organic components e.g. fat, protein, and carbohydrates. The molecules contain various chemical bonds that induce vibrations due to absorption of electromagnetic radiation at different wavenumbers. The absorption at adjacent wavenumbers can be induced by chemical bonds that are abundant in molecules of a certain milk component.

Some absorption bands in the spectral regions can be assigned to specific milk components.

Spectral regions related to fat, protein and lactose in milk are shown in Figure 1.3. In general, wavenumbers of the triacylglycerol ester linkage C–O symmetric stretching (approx. 1,175 cm^{-1}), C=O stretching (approx. 1,750 cm^{-1}), and acyl chain C–H symmetric and asymmetric stretching (2,800-3,000 cm^{-1}) are commonly used to determine milk fat content. The bending vibrations of acyl chain C–H can be found at low wavenumbers, e.g. scissoring at 1,463 cm^{-1}, wagging at 1,123 cm^{-1}. The amide I, II and III bands (1,200-1,700 cm^{-1}) can be used to determine milk protein content. The Amide I band is due to C=O stretching in the polypeptide and is shown at 1,600-1,700 cm^{-1}. Amide II band is due to N–H in plane bending and C–N stretching vibrations and is shown at 1,500-1,600 cm^{-1}. The relatively weak amide III band is due to a combination of N–H wagging, C–C stretching, C–N stretching and C–O wagging and is shown at 1,200-1,400 cm^{-1}. The bond between carbon atom and hydroxyl group, C–OH (approx. 1,080 cm^{-1}) can be used to determine carbohydrates like lactose (Diem, 2015).

In the IR analyses of milk samples, there is always an issue due to the main component of milk: water. Water molecules are very polar and strong infrared absorbers with absorption bands at wavenumbers between 3,000-3,600 cm^{-1} (–OH stretching) and between 1,600-1,700 cm^{-1} (–OH bending) (Safar et al., 1994). The absorption of water is intense and masks the absorption bands of other chemical bonds, for example C–H stretching in carbohydrates around 3,200 cm^{-1}, amide I band in protein at 1,600-1,700 cm^{-1}, and C=C stretching at 1,640-1,666 cm^{-1}. Therefore in practice, these wavenumbers are regarded as noise and assumed not to contain valuable information on milk composition.

The application of milk IR spectroscopy has become an important topic in dairy cattle breeding, since milk IR spectroscopy has been proposed to be able to predict various dairy cattle phenotypes. The milk IR spectra can be regarded as a comprehensive reflection of milk composition. The IR profile might contain information on more than total fat, protein, and lactose content, however, it is not clear what information on milk composition is actually captured by the milk IR spectra. Due to the complexity of components in milk,

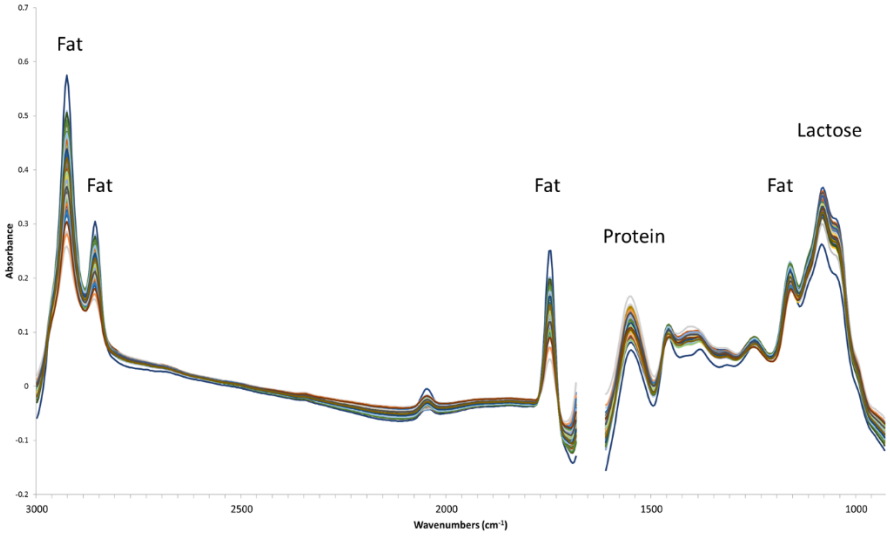


Figure 1.3 The milk infrared (IR) regions representing major milk components

it is difficult to relate individual wavenumbers to detailed milk composition, especially the components in low concentrations.

1.2.4 Terminology in this thesis

In this thesis, the milk IR spectrum was determined by a Fourier-transform infrared spectrometer, in which a mathematical Fourier transform was applied to convert the raw data expressed in a time-domain into actual spectra expressed in a frequency-domain. Various terms and abbreviations, e.g. Fourier transform infrared spectroscopy (FTIR, e.g. Rutten et al., 2011), Mid-infrared spectroscopy (MIR, e.g. Soyeurt et al., 2011) have been used in others' scientific publications on the same topic. In this thesis, the term milk infrared (IR) spectroscopy will be used.

1.3 Prediction of dairy cattle phenotypes using milk infrared spectroscopy

Prediction of dairy cattle phenotypes, including milk composition based on milk IR spectroscopy, has been a topic of many studies over the past ten years (reviewed by De Marchi et al., 2014). However, several issues remain unresolved.

1.3.1 IR prediction of milk fat composition

Several studies pointed at the potential of milk IR spectroscopy to predict milk fat composition (Soyeurt et al., 2006,2011; Rutten et al., 2009; De Marchi et al., 2011; Ferrand et al., 2011). When fatty acids were expressed as a percentage of total fat, Soyeurt et al. (2006) showed that the validation coefficient of determination (R^2) in prediction, was 0.67 for C14:0, 0.50 for C16:0, and 0.53 for C18:1 (based on 49 milk samples). Rutten et al. (2009) used 3,622 milk samples collected in winter and summer and reported R^2 of 0.73 for C14:0, 0.71 for C16:0 and 0.84 for C18:1. Both studies reported higher prediction accuracies when fatty acids were expressed per unit of milk (milk basis) as compared to per unit of fat (fat basis). However, from the perspective of the dairy industry there is more interest in changing fat composition, for example the proportion of unsaturated fatty acids (fat basis), than changing the total amount of unsaturated fatty acids per unit of milk (milk basis). The later can also be achieved by changing fat content of milk. In addition, it was shown that more accurate IR predictions of fat composition were obtained for major fatty acids than for fatty acids in low concentrations. The relation between concentration of fatty acids and prediction accuracy has been discussed by Rutten et al. (2009) and De Marchi et al. (2011).

1.3.2 IR prediction of milk protein composition

Several studies showed moderate IR prediction accuracy for milk protein composition (De Marchi et al., 2009a; Rutten et al., 2011; Bonfatti et al., 2011). Bonfatti et al. (2011) used 1,517 milk samples and showed R^2 of 0.66 for α_{S1}-casein, 0.49 for α_{S2}-casein, 0.53 for β-casein and 0.63 for κ-casein, 0.31 for α-lactabumin and 0.64 for β-lactoglobulin. These prediction accuracies were higher than those reported by De Marchi et al. (2009a) and Rutten et al. (2011). The average R^2 for lactoferrin was 0.71 (Lopez-Villalobos et al., 2009; Soyeurt

et al., 2007; Soyeurt et al., 2012). In general these moderate prediction accuracies for milk protein composition indicate that milk IR spectroscopy is not suited for setting up a milk payment system based on milk protein composition.

1.3.3 IR prediction of other phenotypes

Soyeurt et al. (2009) used 87 milk samples and investigated the possibility of IR spectroscopy for predicting milk mineral composition. Results showed cross-validation R^2 of 0.87 for calcium, 0.85 for phosphorus, 0.36 for potassium and 0.65 for sodium and magnesium content in milk. Toffanin et al. (2015) showed that calcium and phosphorus content in milk can be predicted by milk IR spectroscopy with R^2 of 0.56 and 0.72 respectively. Milk coagulation properties can be predicted by milk IR spectroscopy (Dal Zotto et al., 2008; De Marchi et al., 2009b; De Marchi et al., 2013) with R^2 ranging from 0.62 to 0.76 for rennet coagulation time, and ranging from 0.37 to 0.70 for curd firmness. De Marchi et al. (2009b) showed that acidity of milk can be predicted with R^2 of 0.59 for pH and of 0.66 for titratable acidity, while Toffanin et al. (2015) showed R^2 of 0.74 for titratable acidity.

Moreover, milk IR spectroscopy has been proposed for prediction of traits related to health, energy status, and environmental impact of dairy cattle. Acetone content in milk, as an indicator for ketosis, can be predicted by milk IR spectroscopy with R^2 of 0.81 (Hansen, 1999). More recent studies confirmed the feasibility of milk IR spectroscopy as a screening tool for subclinical ketosis (Heuer et al., 2001; De Roos et al., 2007; Van Knegsel et al., 2010). McParland et al. (2011) used 268 dairy cows with multiple lactations and showed R^2 of 0.45 to 0.52 for energy balance. In a follow up study, the dataset was extended with dairy cows from different countries (McParland et al., 2012). Dehareng et al. (2012) conducted 2 experiments and 3 dietary treatments on 11 Holstein dairy cows to predict methane production (g CH_4/day) and methane intensity (g CH_4/kg of milk). High R^2 ranging from 0.68 to 0.79 were found for these traits in different scenarios.

1.4 Genetic background of milk infrared spectroscopy

The genetic background of milk composition such as heritability, genetic correlations, effects of genes, has been intensively studied. Genomic regions associated with milk production traits such as milk yield, fat and protein content (e.g., Daetwyler et al., 2008; Pryce et al., 2010; Cole et al., 2011), milk fat composition and milk protein composition (Bouwman et al., 2011; Schopen et al., 2011) have been identified using genome-wide association studies. Several studies showed that milk composition is significantly affected by some gene polymorphisms, e.g. Diacylglycerol O-acyltransferase 1 (DGAT1) (e.g., Grisart et al., 2002), Stearoyl-CoA Desaturase (SCD1) (e.g., Schennink et al., 2008), κ-casein (CSN3) and β-lactoglobulin (LGB) (e.g., Heck et al., 2009). Genomic regions or gene polymorphisms that have been shown to significantly affect milk composition are expected to affect wavenumbers in the IR spectra as well, provided that changes in the corresponding milk component are reflected in the IR spectra. Therefore studying the genetic background of milk IR spectra will provide information on which milk components are actually captured by the milk IR spectra.

Milk composition is also affected by environmental factors like herd, lactation stage and age at first calving (Schutz et al., 1990; Stoop et al., 2009; Walker et al., 2004). It is of interest to investigate to which extent environmental factors affect milk IR spectra. In addition, individual wavenumbers may be affected by genetic differences between animals.

The accuracy of IR spectroscopy to predict milk composition might be improved, especially for components in low concentrations. Both genotypes and milk IR spectra contain information on milk composition and combining both information sources may improve prediction accuracy. The improved prediction accuracy might contribute to improved tools for farm management.

1.5 Aim and outline of this thesis

The aim of this thesis was to explore the genetic background of milk IR spectra of dairy cows, and to investigate the feasibility to predict methane emission and detailed milk fat composition using milk IR spectroscopy. In **chapter 2**, we quantified the effects of four genes (DGAT1, SCD1, CSN3 and LGB) and

systematic environmental factors (e.g., lactation stage and date of IR analyses) on individual milk IR wavenumbers. We estimated the heritability and the variation due to differences between herds for individual milk IR wavenumbers. In **chapter 3**, we performed genome-wide association studies on a selected set of IR wavenumbers and identified genomic regions affecting these wavenumbers. In **chapter 4** we predicted methane emission of individual dairy cows based on milk IR spectroscopy using different validation strategies. The methane emission was measured using a sensor that was installed in an Automatic Milking System. In **chapter 5** we investigated if combining milk IR spectroscopy with genotypic information of dairy cows could improve prediction of milk fat composition. The general discussion (**chapter 6**) focused on three main topics: 1. Between-season differences in the genetic background of milk IR spectra; 2. Prediction of DGAT1 genotypes based on milk IR spectra; 3. Ways to extract more information on milk composition based on milk IR analyses.

1.6 References

Auldist, M. J., K. A. Johnston, N. J. White, W. P. Fitzsimons, and M. J. Boland. 2004. A comparison of the composition, coagulation characteristics and cheese making capacity of milk from Friesian and Jersey dairy cows. J. Dairy Res. 71: 51-57.

Bastin, C., D. P. Berry, H. Soyeurt, and N. Gengler. 2012. Genetic correlations of days open with production traits and contents in milk of major fatty acids predicted by mid-infrared spectrometry. J. Dairy Sci. 95: 6113-6121.

Black, R. E., S. M. Williams, I. E., Jones, and A. Goulding. 2002. Children who avoid drinking cow milk have low dietary calcium intakes and poor bone health. Am. J. Clin. Nutr. 76: 675-680.

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.

Bonfatti, V., G. Di Martino, A. Cecchinato, L. Degano, and P. Carnier. 2010. Effects of β-κ-casein (CSN2-CSN3) haplotypes, β-lactoglobulin (BLG) genotypes, and detailed protein composition on coagulation properties of individual milk of Simmental cows. J. Dairy Sci. 93: 3809-3817.

Bonfatti, V., G. Di Martino, and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and

contents of protein genetic variants of individual milk of Simmental cows. J. Dairy Sci. 94: 5776-5785.

Bordin, G., F. C. Raposo, B. De la Calle, and A. R. Rodriguez. 2001. Identification and quantification of major bovine milk proteins by liquid chromatography. J. Chromatogr. A. 928: 63-76.

Bouwman, A. C., H. Bovenhuis, M. H. Visker, and J. M. van Arendonk. 2011. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC genetics. 12: 1.

Bovenhuis, H., M. H. P. W. Visker, and A. Lundén. 2013. Selection for milk fat and milk protein composition. Adv. Anim. Biosci. 4: 612-617.

Canadian Dairy Information Centre. 2018. http://aimis-simia-cdic-ccil.agr.gc.ca/rp/index-eng.cfm?action=pR&r=264&pdctc=.

Chen, S., G. Bobe, S. Zimmerman, E. G. Hammond, C. M. Luhman, T. D. Boylston, A. E. Freeman, and D. C. Beitz. 2004. Physical and sensory properties of dairy products from cows with various milk fatty acid compositions. J. Agric. Food Chem. 52: 3422-3428.

Chilliard, Y., C. Martin, J. Rouel, and M. Doreau. 2009. Milk fatty acids in dairy cows fed whole crude linseed, extruded linseed, or linseed oil, and their relationship with methane output. J. Dairy Sci. 92: 5199-5211.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, B. A. Crooker, C. P. van Tassell, J. Yang, S. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. BMC genomics. 12: 1.

Collomb, M., U. Bütikofer, R. Sieber, B. Jeangros, and J. O. Bosset. 2002. Composition of fatty acids in cow's milk fat produced in the lowlands, mountains and highlands of Switzerland using high-resolution gas chromatography. Int. Dairy J. 12: 649-659.

Couvreur, S., C. Hurtaud, C. Lopez, L. Delaby, and J. L. Peyraud. 2006. The linear relationship between the proportion of fresh grass in the cow diet, milk fatty acid composition, and butter properties. J. Dairy Sci. 89: 1956-1969.

Daetwyler, H. D., F. S. Schenkel, M. Sargolzaei, and J. A. B. 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.

Dal Zotto, R., M. De Marchi, A. Cecchinato, M. Penasa, M. Cassandro, P. Carnier, L. Gallo, and G. Bittante. 2008. Reproducibility and repeatability

of measures of milk coagulation properties and predictive ability of mid-infrared reflectance spectroscopy. J. Dairy Sci. 91: 4103-4112.

Dehareng, F., C. Delfosse, E. Froidmont, H. Soyeurt, C. Martin, N. Gengler, N., A. Vanlierde, and P. Dardenne. 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Anim. 6: 1694-1701.

De Jong, C., and H. T. Badings. 1990. Determination of free fatty acids in milk and cheese procedures for extraction, clean up, and capillary gas chromatographic analysis. J. High. Resolut. Chromatogr. 13: 94-98.

De Jong, N., S. Visser, and C. Olieman. 1993. Determination of milk proteins by capillary electrophoresis. J. Chromatogr. A. 652: 207-213.

De Marchi, M., C. C. Fagan, C. P. O'Donnell, A. Cecchinato, R. Dal Zotto, M. Cassandro, M. Penasa, and G. Bittante. 2009b. Prediction of coagulation properties, titratable acidity, and pH of bovine milk using mid-infrared spectroscopy. J. Dairy Sci. 92: 423-432.

De Marchi, M., M. Penasa, A. Cecchinato, M. Mele, P. I. E. R. Secchiari, and G. Bittante. 2011. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Anim. 5: 1653-1658.

De Marchi, M., V. Bonfatti, A. Cecchinato, G. Di Martino, and P. Carnier. 2009a. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Ital. J. Anim. Sci. 8: 399-401.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2013. Prediction of coagulating and noncoagulating milk samples using mid-infrared spectroscopy. J. Dairy Sci. 96: 4707-4715.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits1. J.Dairy Sci. 97: 1171-1186.

De Roos, A. P. W., H. J. C. M. van Den Bijgaart, J. Hørlyk, and G. De Jong. 2007. Screening for subclinical ketosis in dairy cattle by Fourier transform infrared spectrometry. J. Dairy Sci. 90: 1761-1766.

Diem, M. 2015. Modern vibrational spectroscopy and micro-spectroscopy: theory, instrumentation and biomedical applications. John Wiley & Sons.

Dijkstra, J., S. M. van Zijderveld, J. A. Apajalahti, A. Bannink, W. J. J. Gerrits, J. R. Newbold, H. B. Perdok, and H. Berends. 2011. Relationships between methane production and milk fatty acid profiles in dairy cattle. Anim. Feed Sci. Tech. 166: 590-595.

Dufour, E. 2009. Principles of infrared spectroscopy. Pages 1-27 in Infrared Spectroscopy for Food Quality Analysis and Control. D. W. Sun. ed. Acad. Press, San Diego, CA.

Eck, M. 2014. Performance Enhancement of Hybrid Nanocrystal-polymer Bulk Heterojunction Solar Cells: Aspects of Device Efficiency, Reproducibility, and Stability (Doctoral dissertation, Universität).

Farnaud, S., and R. W. Evans. 2003. Lactoferrin—a multifunctional protein with antimicrobial properties. Mol. Immunol. 40: 395-405.

Ferrand, M., B. Huquet, S. Barbey, F. Barillet, F. Faucon, H. Larroque, O. Leray, J.M. Trommenschlager, and M. Brochard. 2011. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. Chemom. Intell. Lab. Syst. 106: 183-189.

Fox, P. F., P. L. McSweeney, and L. Paul. 1998. Dairy chemistry and biochemistry (No. 637 F6.). London: Blackie Academic & Professional.

Frank, B., and C. Swensson. 2002. Relationship between content of crude protein in rations for dairy cows and milk yield, concentration of urea in milk and ammonia emissions. J. Dairy Sci. 85: 1829-1838.

Friggens, N. C., C. Ridder, and P. Løvendahl. 2007. On the use of milk composition measures to predict the energy balance of dairy cows. J. Dairy Sci. 90: 5453-5467.

German, J. B., and C. J. Dillard. 2006. Composition, structure and absorption of milk lipids: a source of energy, fat-soluble nutrients and bioactive molecules. Crit. Rev. Food Sci. Nutr. 46: 57-92.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in Dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12: 222-231.

Hamann, J., and V. Krömker. 1997. Potential of specific milk composition variables for cow health management. Livest Prod Sci. 48: 201-208.

Hansen, P. W. 1999. Screening of dairy cows for ketosis by use of infrared spectroscopy and multivariate calibration. J. Dairy Sci. 82: 2005-2010.

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.

Heuer, C., H. J. Luinge, E. T. G. Lutz, Y. H. Schukken, J. H. van Der Maas, H. Wilmink, and J. P. T. M. Noordhuizen. 2001. Determination of acetone in cow milk by Fourier transform infrared spectroscopy for the detection of subclinical ketosis. J. Dairy Sci. 84: 575-582.

ICAR (International Committee for Animal Recording). 2012. International agreement of recording practices – Guidelines approved by the General Assembly held in Cork, Ireland on June 2012. ICAR, Rome. Italy.

Jelen, P. 2007. Innovative uses of milk in human nutrition and health. Proceedings of the 35th Biennal Session of ICAR, Kuopio, Finland. EAAP publication 121. EAAP, Rome, Italy.

Jensen, R. G. 2002. The composition of bovine milk lipids: January 1995 to December 2000. J. Dairy Sci. 85: 295-350.

Kawai, K., S. Hagiwara, A. Anri, and H. Nagahata. 1999. Lactoferrin concentration in milk of bovine clinical mastitis. Vet. Res. Commun. 23: 391-398.

Lopez-Villalobos, N., S. R. Davis, E. M. Beattie, J. Melis, S. Berry, S. E. Holroyd, R. J. Spelman, and R. G. Snell. 2009. Breed effects for lactoferrin concentration determined by Fourier transform infrared spectroscopy. In Proceedings of the New Zealand Society of Animal Production (Vol. 69, pp. 60-64). New Zealand Society of Animal Production.

McParland, S., G. Banos, B. McCarthy, E. Lewis, M. P. Coffey, B. O'Neill, M. O'Donovan, E. Wall, D. P. Berry. 2012. Validation of mid-infrared spectrometry in milk for predicting body energy status in Holstein-Friesian cows. J. Dairy Sci. 95: 7225-7235.

McParland, S., G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J. Dairy Sci. 94: 3651-3661.

Miglior, F., B. L. Muir, and B. J. van Doormaal. 2005. Selection indices in Holstein cattle of various countries. J. Dairy Sci. 88: 1255-1263.

Penasa, M., M. Cassandro, D. Pretto, M. De Marchi, A. Comin, S. Chessa, R. Dal Zotto, and G. Bittante. 2010. Influence of composite casein genotypes on additive genetic variation of milk production traits and coagulation properties in Holstein-Friesian cows. J. Dairy Sci. 93: 3346-3349.

Pryce, J. E., S. Bolormaa, A. J. Chamberlain, P. J., Bowman, K. Savin, M. E. Goddard, and B. J. Hayes. 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J. Dairy Sci. 93: 3331-3345.

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. J. Dairy Sci. 94: 5683-5690.

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.

Safar, M., D. Bertrand, P. Robert, M. F. Devaux, and C. Genot. 1994. Characterization of edible oils, butters and margarines by Fourier transform infrared spectroscopy with attenuated total reflectance. J. Am. Oil Chem. Soc. 71: 371-377.

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.

Schopen, G.C.B., M. H. P. W. Visker, P. D. Koks, E. Mullaart, J. A. M. van Arendonk, and H. Bovenhuis. 2011. Whole-genome association study for milk protein composition in dairy cattle. J. Dairy Sci. 94: 3148-3158.

Schukken, Y. H., D. J. Wilson, F. Welcome, L. Garrison-Tikofsky, and R. N. Gonzalez. 2003. Monitoring udder health and milk quality using somatic cell counts. Vet. Res. 34: 579-596.

Schutz, M. M., L. B. Hansen, G. R. Steuernagel, and A. L. Kuck. 1990. Variation of milk, fat, protein, and somatic cells for dairy cattle. J. Dairy Sci. 73: 484-493.

Soyeurt, H., C. Bastin, F. G. Colinet, V. R. Arnould, D. P. Berry, E. Wall, F. Dehareng, H. N. Nguyen, P. Dardenne, J. Schefers, J. Vandenplas, K. Weigel, M. Coffey, L. Théron, J. Detilleux, E. Reding, N. Gengler, and S. McParland. 2012. Mid-infrared prediction of lactoferrin content in bovine milk: potential indicator of mastitis. Anim. 6: 1830-1838.

Soyeurt, H., D. Bruwier, J. M. Romnee, N. Gengler, C. Bertozzi, D. Veselko, and P. Dardenne. 2009. Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. J. Dairy Sci. 92: 2444-2454.

Soyeurt, H., F. G. Colinet, V. R. Arnould, P. Dardenne, C. Bertozzi, R. Renaville, D. Portetelle, and N. Gengler. 2007. Genetic variability of lactoferrin content estimated by mid-infrared spectrometry in bovine milk. J. Dairy Sci. 90: 4443-4450.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty

acids across multiple breeds, production systems, and countries. J.Dairy Sci. 94: 1657-1667.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J.Dairy Sci. 89: 3690-3695.

Stoop, W. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2009. Effect of lactation stage and energy status on milk fat composition of Holstein-Friesian cows. J. Dairy Sci. 92: 1469-1478.

Toffanin, V., M. De Marchi, N. Lopez-Villalobos, and M. Cassandro. 2015. Effectiveness of mid-infrared spectroscopy for prediction of the contents of calcium and phosphorus, and titratable acidity of milk and their relationship with milk quality and coagulation properties. Int. Dairy J. 41: 68-73.

Van Haelst, Y. N. T., A. Beeckman, A. T. M. van Knegsel, and V. Fievez. 2008. Short Communication: Elevated concentrations of oleic acid and long-chain fatty acids in milk fat of multiparous subclinical ketotic cows. J. Dairy Sci. 91: 4683-4686.

Van Knegsel, A. T. M., H. van den Brand, J. Dijkstra, W. M. van Straalen, R. Jorritsma, S. Tamminga, and B. Kemp. 2007. Effect of glucogenic vs. lipogenic diets on energy balance, blood metabolites, and reproduction in primiparous and multiparous dairy cows in early lactation. J. Dairy Sci. 90: 3397-3409.

Van Knegsel, A. T. M., S. G. A. van der Drift, M. Horneman, A. P. W. De Roos, B. Kemp, and E. A. M. Graat. 2010. Ketone body concentration in milk determined by Fourier transform infrared spectroscopy: Value for the detection of hyperketonemia in dairy cows. J. Dairy Sci. 93: 3065-3069.

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. Crop Pasture Sci. 55: 1009-1028.

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.

2

Genetic and environmental variation in bovine milk infrared spectra

Qiuyu Wang, Alex Hulzebosch, Henk Bovenhuis

Animal Breeding and Genomics Centre, Wageningen University & Research,

PO Box 338, 6700AH, Wageningen, the Netherlands

Journal of Dairy Science (2016) 98(8): 6793-6803

Abstract

Milk infrared (IR) spectroscopy is widely used to determine milk composition. In this study 1,060 milk IR wavenumbers ranging from 925 to 5,008 cm^{-1} of 1,748 Holstein Friesian cows on 371 herds in the Netherlands were available. The extent to which IR wavenumbers are affected by genetic and environmental factors was investigated. Inter-herd heritabilities of 1,060 milk IR wavenumbers ranged from 0 to 0.63 indicating that the genetic background of IR wavenumbers differs considerably. The majority of the wavenumbers have moderate to high inter-herd heritabilities ranging from 0.20 to 0.60. The diacylglycerol O-acyltransferase 1 (DGAT1), stearoyl-CoA desaturase (SCD1), κ-casein (CSN3) and β-lactoglobulin (LGB) polymorphisms are known to have large effect on milk composition and therefore we studied the effects of these polymorphisms on individual milk IR wavenumbers. The DGAT1 polymorphism had highly significant effects on many wavenumbers. In contrast, the SCD1 polymorphism did not significantly affect any of the wavenumbers. The SCD1 is known to have a strong effect on the content of C10:1, C12:1, C14:1, and C16:1 fatty acids. Therefore, these results suggest that milk IR spectra contain little direct information on these mono unsaturated fatty acids. The CSN3 and LGB polymorphisms had significant effects on a few wavenumbers. Differences between herds explained 10 to 25% of the total variance for most wavenumbers. This suggests that the wavenumbers of milk IR spectra are indicative for differences in feeding and management between herds. The wavenumbers between 1,619 and 1,674 cm^{-1} and between 3,073 and 3,667 cm^{-1} are strongly influenced by water absorption and usually excluded when setting up prediction equations. However, we found that some of the wavenumbers in the water absorption region are affected by the DGAT1 polymorphism and lactation stage. This suggests that these wavenumbers contain useful information regarding milk composition.

Key words:

bovine milk, IR wavenumbers, heritability, herd, genetic polymorphisms

2.1 Introduction

Milk produced by dairy cows is a complex product consisting of many components (Jelen, 2007). However, at present only fat and protein content are routinely recorded and considered in most dairy cattle breeding programs (Miglior et al., 2005). Despite this, both from a nutritional and a manufacturing perspective, detailed fat and protein composition are of interest. Several studies suggested protein composition is related to milk coagulation and cheese yield (e.g., Wedholm et al., 2006), since a greater casein content is preferable for cheese making. Due to its relatively high concentration of saturated fatty acids, the consumption of bovine milk fat has been associated with negative effects on human health (e.g., German and Dillard, 2006) and therefore changing milk fat composition by means of selective breeding might be of interest. Moreover, milk composition can serve as an indicator for the cow's health status (e.g., Vlaeminck et al., 2006; Van Haelst et al., 2008) and methane emission (e.g., Chilliard et al., 2009). For breeding and management purposes, large scale routinely collected measurements are needed and therefore traits should be easy to measure at relatively low costs. Analytical methods like gas chromatography to quantify milk fat composition, or high-performance liquid chromatography and capillary zone electrophoresis to quantify milk protein composition, are expensive and time-consuming. Therefore these methods are less suited for large scale routine measurements.

Fourier transform infrared (IR) spectroscopy is a fast and cost effective method widely used to determine milk composition. It is the standard method for routine quantification of fat, protein and lactose content of milk (ICAR, 2012). Several studies showed that IR spectra also can be used to determine milk fat composition (e.g., Soyeurt et al., 2006; Rutten et al., 2009). Other studies investigated possibilities to predict milk protein composition based on IR spectra (Bonfatti et al., 2011; Rutten et al., 2011). Furthermore, recent research reported the ability of milk IR spectra to predict traits such as milk coagulation, ketone bodies and energy status of dairy cows (De Marchi et al., 2014).

Some studies analysed the genetic background of milk IR wavenumbers (Soyeurt et al., 2010; Bittante and Cecchinato, 2013). Soyeurt et al. (2010) analysed milk IR spectra of 1,594 first parity Holstein cows and found

substantial differences in heritability between wavenumbers and concluded that not all IR wavenumbers are of genetic interest. Bittante and Cecchinato (2013) studied the IR spectra of 1,064 Brown Swiss cows and reported that approximately 30% of the wavenumbers have heritability larger than 0.10.

Besides quantifying the combined action of all genes on IR wavenumbers, it might be of interest to study effects of some individual genes with known and large effect on milk composition. Polymorphisms in diacylglycerol O-acyltransferase 1 (DGAT1), stearoyl-CoA desaturase 1 (SCD1), κ-casein (CSN3) and β-lactoglobulin (LGB) have been shown to have important effects on milk composition (e.g., Schennink et al., 2008; Heck et al., 2009). Their effects on individual IR wavenumbers can provide insight in the information that is captured by the whole IR spectra. Furthermore, it has been shown that IR spectra can be used to predict LGB genotypes (Rutten et al., 2011) and CSN1S1 haplotypes (Berget et al., 2010). Quantifying the effects of DGAT1, SCD1 and CSN3 polymorphisms on individual IR wavenumbers can give insight in the possibilities of predicting genotypes for these polymorphisms based on IR spectra.

It is well known that milk composition is also affected by feed and management strategies, e.g. feed composition influences milk fat content and fat composition (e.g., Chilliard et al., 2007) and dietary energy intake influences milk protein content (Emery, 1978). There is an increasing interest of consumers in the authenticity of milk as they purchase biological and organic products at higher price. Milk IR spectroscopy might be one of the methods that enable discriminating milk samples produced by cows fed different diets (Valenti et al., 2013). Quantifying herd effects will give insight in the extent to which feed and management differences are reflected by individual wavenumbers. Herd effects for individual IR wavenumbers of bovine milk have not been quantified before.

The aim of this study was to quantify the contribution of genetic and environmental effects to the variation in milk IR wavenumbers. Furthermore, we aimed at quantifying the effects of polymorphisms in DGAT1, SCD1, CSN3 and LGB on milk IR wavenumbers.

2.2 Materials and methods

2.2.1 Data

In this study, one morning milk samples from 1,748 first parity Holstein Friesians cows located on 371 herds have been collected for analysis. The data was collected from February till March of 2005. All cows have at least 87.5% Holstein Friesian genes. The population consisted of 5 large paternal half-sib families from proven sires (98-196 daughters per sire), and 50 small paternal half-sib families from test sires (8-23 daughters per sire), as well as 168 cows descending from 44 other proven sires (1-25 daughters per sire) to assure at least 3 cows per herd. The pedigree of the cows was provided by CRV (Cooperative cattle improvement organization, Arnhem, the Netherlands).

Milk samples were conserved using sodium azide (0.03% wt/wt) at 4°C all times. Subsequently, IR spectra were recorded in a 10 mL milk sample using the MilkoScan FT 6000 equipment (FOSS, Denmark) at the certified laboratory of the Milk Control Station (Zutphen, The Netherlands). All milk samples used in this study were analysed on the same MilkoScan FT 6000. The IR spectra consisted of the transmittance values measured at 1,060 wavenumbers ranging from 925 to 5,008 cm^{-1}.

2.2.2 Genotypes

Blood samples were collected for DNA isolation. The genotyping procedure for DGAT1 K232A and SCD1 A239V polymorphisms were described by Schennink et al. (2008). Genotypes of CSN3 were determined as described by Heck et al. (2009). The polymorphisms associated with the known protein variants for LGB were genotyped using a SNaPshot assay as described by Visker et al. (2011).

Among the 1,748 cows with milk IR data, 1,625 cows had DGAT1 genotypes, 1,579 cows had SCD1 genotypes, 1,534 cows had CSN3 genotypes and 1,542 cows had LGB genotypes. For some cows the genotypes were missing because either no DNA sample was available or the sample could not be genotyped unambiguously. The allele frequencies were 60.0% for A allele and 40.0% for K allele of DGAT1, 73.0% for A allele and 27.0% for V allele of

SCD1, 58.3% for A allele and 41.7% for B allele of LGB and 60.4%, 30.1%, 9.5% for CSN3 A, B and E allele respectively.

2.2.3 Statistical analysis

A series of analyses were performed to quantify the effects of several factors on the 1,060 milk IR wavenumbers. The following model was used:

$$y_{ijklm}=\mu+\beta_1*lactst_{ijklm}+\beta_2*afc_{ijklm}+season_i+sirecode_j+$$
$$date_k+herd_l+a_m+e_{ijklm}, \qquad \text{Equation 2.1}$$

where y_{ijklm} is the transmittance value of the IR wavenumber; μ is the general mean; $lactst_{ijklm}$ is a covariate for the effect of lactation stage (in days) with regression coefficient β_1 ; afc_{ijklm} is a covariate for the effect of age at first calving with regression coefficient β_2 ; $season_i$ is the fixed effect season of calving (June-Aug 2004, Sept-Nov 2004 or Dec 2004-Jan 2005); $sirecode_j$ is the fixed effect accounting for possible differences in genetic level between the groups of proven bull daughters and young bull daughters; $date_k$ is the fixed effect accounting for the effect of 17 days at which IR analyses of milk samples took place; $herd_l$ is a random effect of herd *l*, distributed as $N(\mathbf{0}, \mathbf{I}\sigma_h^2)$, with identity matrix **I** and herd variance σ_h^2; a_m is a random additive genetic effect of animal *m*, distributed as $N(\mathbf{0}, \mathbf{A}\sigma_a^2)$, with additive genetic relationship matrix **A** and the additive genetic variance σ_a^2 . The additive genetic relationship matrix was constructed based on 12,548 animals. e_{ijklm} is a random residual effect, distributed as $N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, with identity matrix **I** and error variance σ_e^2.

The inter-herd heritability for individual wavenumbers was calculated as

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_h^2+\sigma_e^2} \qquad \text{Equation 2.2}$$

ASReml (Gilmour et al., 2009) was used to perform single trait analyses in order to assess the significance (P-values) of the fixed effects and REML estimates of variance components.

The effects of the DGAT1, SCD1, CSN3 and LGB polymorphisms were estimated using equation 2.1 but extended with a fixed effect. The effects of these 4 polymorphisms were determined in separate analyses. Individuals whose genotype was missing were included in the analysis by assigning them to a separate genotype class. Missing genotypes appeared to be randomly distributed across other effects in the model.

Significance tests for systematic environmental factors were performed for each of the 1,060 wavenumbers. To adjust for multiple testing we used a Bonferroni correction. To determine the number of independent traits a Principle Component Analysis (PCA) was performed on the milk IR spectra, indicating that 99% of the variation can be described based on 45 principal components (result not shown). To adjust for multiple testing we therefore assumed 50 independent tests and consequently an effect was considered significant if $-\text{Log}_{10}(P)$ was larger than 3 (i.e., $-\text{Log}_{10}(0.05/50)$), where P represents the significance of the effect.

2.3 Results

The mean, 1st percentile and 99th percentile of the 1,060 milk IR wavenumbers are shown in Figure 2.1. The IR wavenumbers between 1,619 and 1,674 cm^{-1} and between 3,073 and 3,667 cm^{-1} showed larger variation than others. These wavenumbers represent the absorption peaks of water and will be referred to as the water absorption region.

2.3.1 Fixed effects

Age at first calving, season of calving and differences between groups of proven and young bull daughters did not show significant effects on any of the 1,060 wavenumbers. Lactation stage had significant effects on 457 wavenumbers. The highest $-\text{Log}_{10}(P)$ of 15.6 for lactation stage was found for wavenumber 2,495 cm^{-1}. Date of IR analysis had significant effect on 80% of the wavenumbers, especially on the wavenumbers in water absorption region. The $-\text{Log}_{10}(P)$ of lactation stage and date of IR analysis are shown in Figure 2.2.

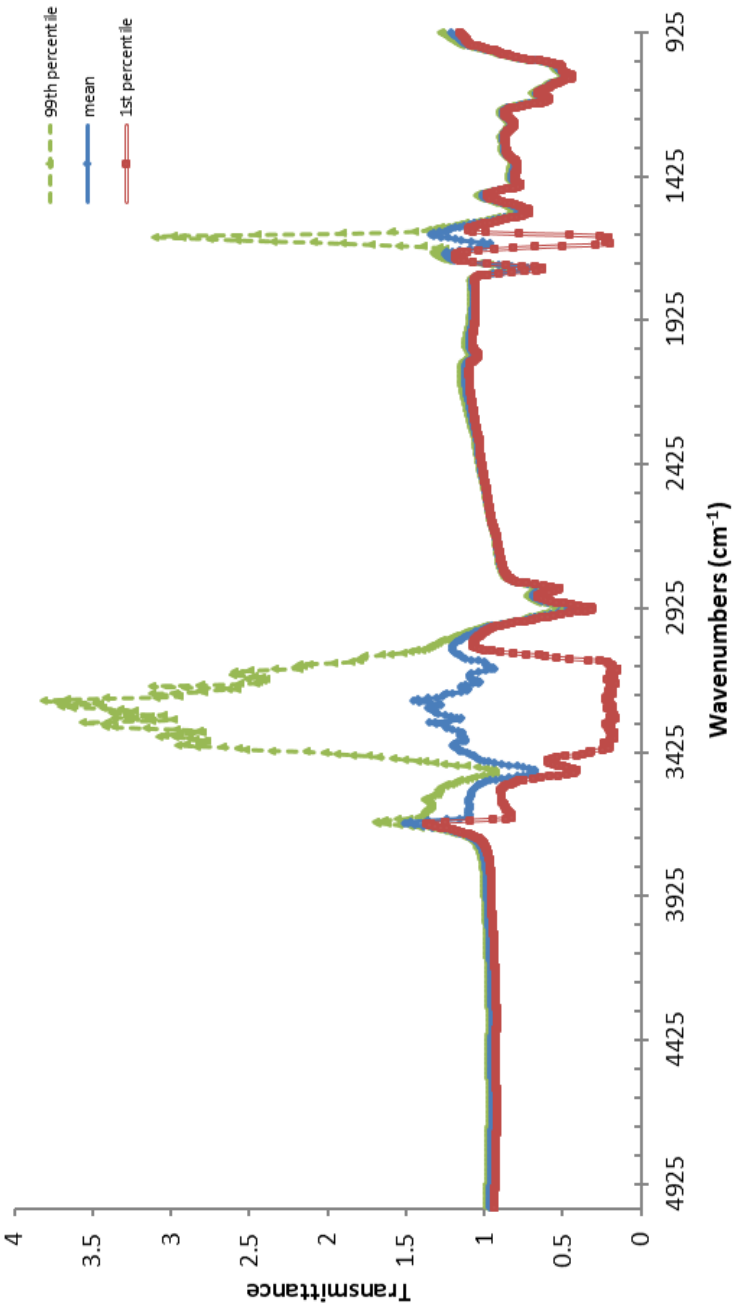


Figure 2.1 The mean transmittance for 1,060 milk IR wavenumbers, and the corresponding 1st percentile and the 99th percentile, based on records of 1,748 dairy cows.

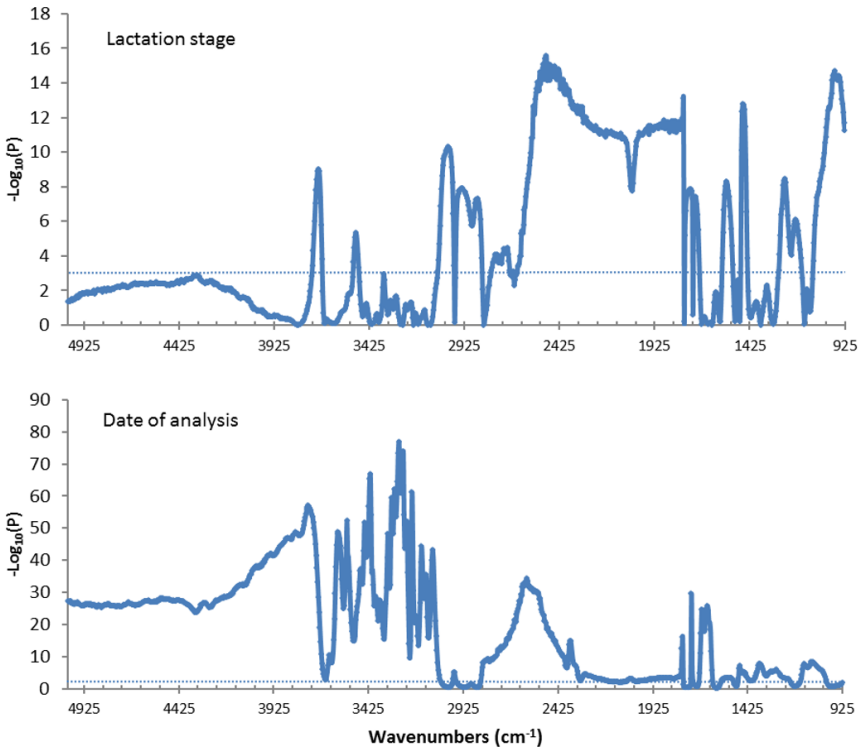


Figure 2.2 The significance of the effect of lactation stage and date of analysis on 1,060 milk IR wavenumbers. The horizontal line indicates a threshold at $-Log_{10}(P)$ of 3.

2.3.2 Variance components

Genetic, herd and residual variances as a fraction of the total phenotypic variance are shown in Figure 2.3. Genetic factors explained more than 40% of the total variation for wavenumbers from 925 to 1,257 cm^{-1}, 1,454 to 1,465 cm^{-1}, 2,811 to 2,973 cm^{-1} and 4,122 to 5,008 cm^{-1}. Furthermore, for most of the wavenumbers from 1,693 to 2,479 cm^{-1} more than 40% of the total variation was due to genetic factors except for 1,724 cm^{-1} (19.6%) and 1,770 cm^{-1} (29.3%). On the other hand, genetic factors explained almost no variation and residual variance accounted for more than 90% of the total phenotypic variance for most of wavenumbers in the water absorption region. An

exception was some wavenumbers near 3,154 cm^{-1} for which genetic factors explained up to 18.7% of the total variation.

Inter-herd heritabilities of the 1,060 wavenumbers were 0.36 on average, ranging from 0 to 0.63. In total there were 197 wavenumbers with heritabilities lower than 0.20 and 291 wavenumbers with heritabilities between 0.20 and 0.40. There were 560 wavenumbers with heritabilities between 0.40 and 0.60, and 12 wavenumbers with heritabilities larger than 0.60.

For 806 wavenumbers, differences between herds contributed more than 10% of the total phenotypic variance. The herd variance accounted for up to 28% of the total phenotypic variance for wavenumber 3,717 cm^{-1}, which was the wavenumber with the largest contribution of herd variance.

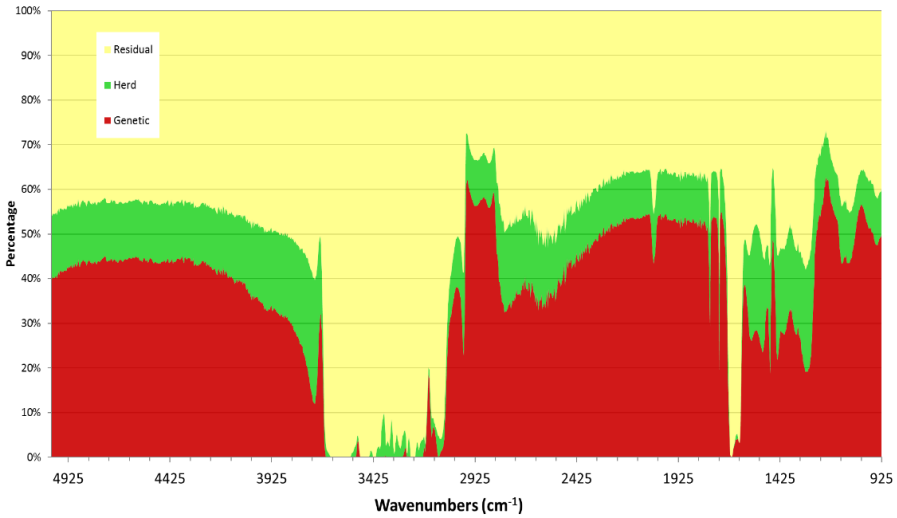


Figure 2.3 Percentage of the total variation of 1,060 milk IR wavenumbers explained by genetic, herd and residual effects.

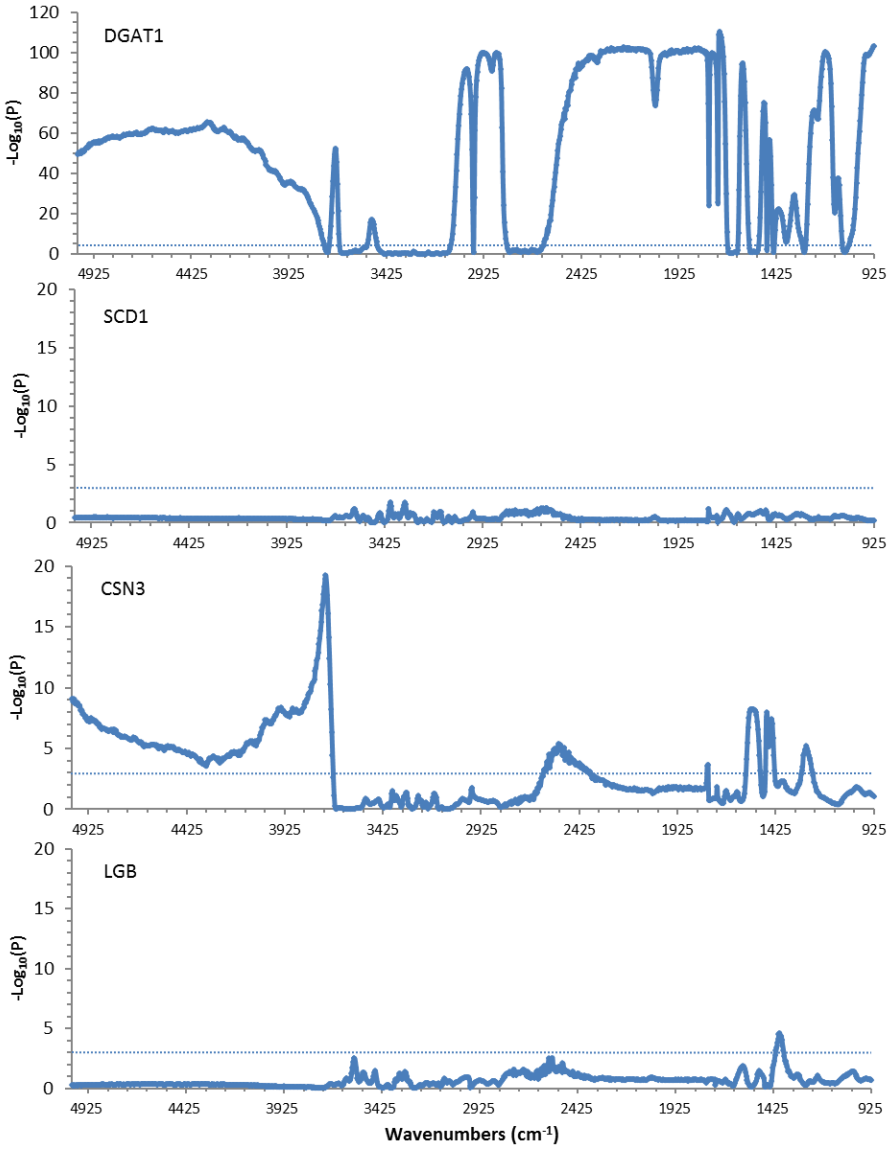


Figure 2.4 The significance of the effect of DGAT1, SCD1, CSN3 and LGB polymorphisms on 1,060 milk IR wavenumbers. The horizontal lines indicate a threshold at $-Log_{10}(P)$ of 3.

2.3.3 Effects of DGAT1, SCD1, CSN3 and LGB

Figure 2.4 shows the -Log_{10}(P) of DGAT1, SCD1, CSN3 and LGB polymorphisms on the 1,060 milk IR wavenumbers. The DGAT1 polymorphism had extremely significant effects on many wavenumbers. For 121 wavenumbers DGAT1 polymorphism had a -Log_{10}(P) larger than 100 and the highest -Log_{10}(P) was 110.4. The DGAT1 had no significant effect on 216 wavenumbers, most of which belongs to the water absorption region. However, wavenumbers from 3,466 to 3,543 cm^{-1} were significantly affected by the DGAT1 polymorphism with the highest -Log_{10}(P) value of 17.0. The SCD1 polymorphism did not significantly affect any of the wavenumbers.

Significant effects of the CSN3 polymorphism were found on 5 regions: wavenumbers from 1,238 to 1,292 cm^{-1}, 1,431 to 1,477 cm^{-1}, 1,504 to 1,573 cm^{-1}, 2,371 to 2,607 cm^{-1}, and 3,682 to 5,008 cm^{-1}. The largest -Log_{10}(P) of 19.2 was found for wavenumber 3,717 cm^{-1}. The LGB polymorphism showed significant effects on wavenumbers between 1,377 and 1,415 cm^{-1}.

2.4 Discussion

In this study we investigated the effects of genetic and environmental factors on milk IR wavenumbers. Besides quantifying the total genetic variance of individual wavenumbers we also estimated the effects of DGAT1, SCD1, CSN3 and LGB polymorphisms. The polymorphisms in DGAT1, CSN3 and LGB significantly affected several wavenumbers whereas the SCD1 polymorphism had no significant effect on any of the wavenumbers. Differences between herds accounted for more than 10% of the variation in many IR wavenumbers. These herd effects might reflect feeding or management differences between farms.

Wavenumbers from the water absorption region showed large variation (Figure 2.1) and a small fraction of the phenotypic variance was explained by genetics (Figure 2.3). These wavenumbers represent the absorption peaks of water and because water is the main component of milk and a very strong infrared absorber, water will mask the effects of other components (e.g., Safar et al., 1994; Karoui et al., 2010). Wavenumbers 3,466 till 3,543 cm^{-1} of the water absorption region were significantly affected by lactation stage (Figure

2.2) and the DGAT1 polymorphism (Figure 2.4). The wavenumbers of the water absorption region are usually excluded when setting up prediction equations (e.g., De Marchi et al., 2009). Our results suggest that some of these wavenumbers contain information on milk composition.

2.4.1 Fixed effects

Lactation stage showed significant effects on many wavenumbers (Figure 2.2). It is well known that fat and protein contents change during lactation (e.g., Schutz et al., 1990). Furthermore, Stoop et al. (2009) showed that milk fat composition is affected by lactation stage. Lactation stage has a smaller effect on milk protein composition (e.g., Walker et al., 2004). These changes in milk composition during lactation are reflected by the milk IR wavenumbers. In our dataset, lactation stage ranged from 63 to 282 days and no early lactation records were available. Since the milk composition changes considerably from early to middle lactation, it is expected that stronger effects of lactation stage will be observed when records covering the complete lactation are included in the analysis.

To our knowledge no other studies specifically quantified the effect of date of analysis on the milk IR spectra. In our study the milk samples were analysed in a short time period of less than two months. Furthermore, the MilkoScan FT 6000 spectrometer was calibrated at regular times according to the manufacturer instructions. However, for many wavenumbers, we found significant effects of date of analysis which suggest instability of the IR spectrometer across dates. Many wavenumbers showed a time trend, which would point at a drift of the spectrometer. The wavenumbers in water absorption region showed highly significant effects of date of analysis. Grelet et al. (2015) standardized milk IR spectra from different laboratories, and reported that standardization coefficients were less adapted in the second study one month later. This also might be due to a time trend or perturbations of the spectrometers.

In the current study milk samples from one herd were analysed on the same day shortly after collection. Therefore the effects of herd and date of analysis were confounded. We found that the variance due to differences between herds increased considerably for some wavenumbers when date of analysis was not accounted for in the model.

2.4.2 Heritability

In our study most wavenumbers had moderate to high heritabilities ranging from 0.20 to 0.60. The estimated heritabilities for wavenumbers in this study were considerably higher than those reported in other studies. Bittante and Cecchinato (2013) analysed transmittance values at 1,056 milk IR wavenumbers (930 to 5,000 cm^{-1}) of milk from 1,064 Italian Brown Swiss cows. Both studies showed a similar pattern of heritabilities for the different wavenumbers. Bittante and Cecchinato (2013) reported that heritabilities of milk IR wavenumbers ranged from 0 to 0.27. In their study 578 wavenumbers had heritabilities between 0.05 and 0.10 and 261 wavenumbers had heritabilities between 0.10 and 0.20. The lower heritability reported by Bittante and Cecchinato (2013) as compared to the current study might among others be due to different instruments used (FOSS FT 6000 vs. FT 120 spectrometer). However the most likely reason is the difference in breeds studied which is in agreement with previous studies who reported higher heritabilities for milk fat and protein content in Dutch Holstein Friesian than Italian Brown Swiss. Based on largely the same cows as included in the current study, Stoop et al. (2007) estimated heritabilities of 0.52 for fat%, 0.60 for protein% and 0.64 for lactose%. For Italian Brown Swiss, Samoré et al. (2012) reported heritabilities of 0.12 for fat%, 0.28 for protein% and 0.25 for lactose%. These differences are in line with the observed differences in heritability estimates of milk IR wavenumbers between Bittante and Cecchinato (2013) and our study.

Soyeurt et al. (2010) estimated heritabilities for transmittance values of milk IR wavenumbers from 1,594 first parity Holstein cows in the Walloon Region of Belgium. Heritabilities in the study by Soyeurt et al. (2010) ranged from 0.00 to 0.42. These estimates are on average approximately 0.1 lower than estimates from our study. Bastin et al. (2011) analysed milk samples from the Walloon Region and estimated heritabilities of approximately 0.40 for fat% and 0.45 for protein%. These estimates are 0.12 lower for fat% and 0.15 lower for protein% than the estimates reported by Stoop et al. (2007). These differences are in line with the difference in heritability estimates for wavenumbers in the current study and those reported by Soyeurt et al. (2010).

2.4.3 Effects of individual genes

Several studies showed that the DGAT1 K232A polymorphism is especially associated with milk fat content and fat composition (e.g., Grisart et al., 2002; Schennink et al., 2008). The K allele is associated with a higher fraction of C16:0, a higher SFA/UFA ratio and lower fractions of C14:0, unsaturated C18 and CLA (e.g., Schennink et al., 2008). In our study, DGAT1 showed highly significant effects on numerous milk IR wavenumbers, which is in line with its large effect on milk composition. The largest DGAT1 effect was found for wavenumber 1,716 cm^{-1} with $-Log_{10}(P)$ of 110.4. Furthermore, highly significant DGAT1 effects on wavenumbers between 1,735 and 1,762 cm^{-1} were observed. These wavenumbers are associated with carboxylic acid and ester C=O bond stretching (Safar et al., 1994). The DGAT1 also had highly significant effects on wavenumbers between 1,160 and 1,180 cm^{-1}. This region represents the triglyceride ester linkage C–O stretching (Safar et al., 1994). The significant DGAT1 effects on most of the wavenumbers from 2,800 to 2,975 cm^{-1} can be explained as these wavenumbers are associated with alkyl C–H stretching (e.g., Safar et al., 1994; Yang and Irudayaraj, 2000), which is abundant in fat. Figure 2.4 also showed highly significant effects of DGAT1 on wavenumbers around 3,686 cm^{-1}. However it is not clear which chemical bonds are associated with these wavenumbers.

It is known that SCD1 is responsible for the desaturation of fatty acids. Schennink et al. (2008) reported that the SCD1 polymorphism has no significant effect on fat% but a large effect on fat composition. Using mainly the same animals as in the current study, Duchemin et al. (2013) showed that SCD1 has highly significant effects on C10:1, C12:1, C14:1, and C16:1 fatty acids. In the current study we didn't find significant effect of the SCD1 polymorphism on any of the wavenumbers (Figure 2.4). This suggests that there is little direct information in the IR spectra on C10:1, C12:1, C14:1, and C16:1 fatty acids. Milk IR prediction equations for these fatty acids therefore might be based on their correlations with total milk fat content. This would be in agreement with Eskildsen et al. (2014) who suggested that predictions of individual fatty acids by IR measurements in milk are indirect and are based on covariation between the fatty acids and total fat content. Interestingly, SCD1 has no significant effect on the total fraction of unsaturated fatty acids

which can be explained by the negative association of the SCD1 A239V polymorphism V allele with medium chain unsaturated fatty acids (e.g., C10:1, C12:1, C14:1) and the positive association with longer chain unsaturated fatty acids (e.g., C16:1) (Duchemin et al., 2013). Therefore, these results do not provide evidence that the IR spectra contain little direct information on the total fraction of unsaturated fatty acids.

The CSN3 polymorphism has been shown to be associated with protein content (e.g., Bovenhuis et al., 1992) and the relative concentrations of the 6 main milk proteins (e.g., Heck et al., 2009). The B allele of CSN3 is associated with a higher protein% and a higher relative concentration of κ-casein and α_{S2}-casein, as well as a lower relative concentration of α-lactabumin and α_{S1}-casein in milk (Heck et al., 2009). Casein is expected to have absorption peaks around wavenumbers 1,250 cm^{-1}, 1,550 cm^{-1} and 1,650 cm^{-1} due to amide III, amide II and amide I bands, respectively (Osborne and Fearn, 1986), while in this study we found significant effects of CSN3 polymorphism on wavenumbers around 1,269 cm^{-1} and 1,550 cm^{-1} (Figure 2.4). The effects on wavenumbers around 1,269 cm^{-1} might be due to amide III or phosphate bands (Hewavitharana and van Brakel, 1997). Furthermore, the significant effects of CSN3 on wavenumbers between 1,504 and 1,573 cm^{-1} coincide with amide II band. This is mainly due to N–H bending and C–N stretching (Garidel and Schott, 2006). We did not detect significant CSN3 effects on wavenumbers around 1,650 cm^{-1} due to amide I bands which might be because this is in the water absorption region.

We also found significant effects of CSN3 polymorphism on other wavenumbers (Figure 2.4). A spectral peak close to 1,469 cm^{-1} was also observed by De Marchi et al. (2010) and this region might be associated with proteins. The significant effect of CSN3 around wavenumber 2,529 cm^{-1} might be explained by its relation with S–H stretching (Hewavitharana and van Brakel, 1997), which commonly binds to whey protein. Therefore, this effect might be explained by the CSN3 effect on the relative concentration of whey protein. We found a highly significant effect of CSN3 for wavenumber 3,717 cm^{-1}, but this wavenumber is not known to be associated with any specific chemical bond.

Several studies showed significant associations between the LGB polymorphism and milk protein composition (e.g., Lunden et al., 1997; Heck et al., 2009). Cows with the LGB BB genotype have a higher casein and lower β-lactoglobulin content than cows with the LGB AA genotype and therefore the LGB B allele is preferred for cheese production (e.g., Van den Berg et al., 1992; Boland and Hill, 2001; Wedholm et al., 2006). In this study, LGB polymorphism had significant effects on wavenumbers between 1,377 and 1,415 cm^{-1}. The highest $-Log_{10}(P)$ in this region was approximately 4.6. This significant effect of LGB polymorphism might be due to the association with C–N stretching at 1,414 cm^{-1} (Dufour, 2009). Notably, we did not find any wavenumbers which were significantly affected by both the CSN3 and LGB polymorphisms.

The wavenumbers from 3,700 to 5,008 cm^{-1} are difficult to interpret because the spectra are complex and combined by overlapping peaks and variations. Wavenumbers 4,033 to 4,350 cm^{-1} can be attributed to combination bands of C–H, which is abundant in fatty acids. Wavenumbers 4,500 to 5,000 cm^{-1} can be attributed to vibrations of N–H and C=O group of proteins (Subramanian and Rodrigucz-Saona, 2009). These might explain the significant effects of DGAT1 and CSN3 polymorphism on these wavenumbers.

It has been shown that it is possible to predict genotypes of polymorphism known to be associated with milk composition based on milk IR spectra. Rutten et al. (2011) showed that LGB genotypes can be predicted based on IR spectra. Our study showed that DGAT1 and CSN3 genotypes have larger effects than LGB on the IR wavenumbers. This suggests that IR spectra might be used to predict DGAT1 and CSN3 genotypes. The K allele of DGAT1 K232A polymorphism is associated with higher fat%, protein%, and fat yield, but lower milk yield and protein yield (Bovenhuis et al., 2015).The CSN3 B allele is associated with a higher protein% (Heck et al., 2009). As these genotypes have distinct effects on milk composition, knowledge of these genotypes might be of interest. The accuracy of predicted genotypes might be increased by combining IR information with pedigree information. Conversely, genotypes of individual cows and their genotypic effects on milk composition might be combined with IR spectra to improve prediction of milk composition.

2.4.4 Herd

Our milk samples were collected from numerous farms throughout the Netherlands, which is a good representation of herds in the Netherlands. The herd variance quantifies the relative importance of herd effect which reflects differences due to feeding, hygiene, and husbandry. There have been many studies showing the impact of feed on milk composition (e.g., Grummer, 1991; Palmquist et al., 1993; Slots et al., 2009). Valenti et al. (2013) demonstrated that based on IR data it is possible to distinguish milk from hay- and pasture-based systems and those from maize silage- and pasture-based systems.

Herd variation for wavenumbers might also reflect differences between herds in the cow's health status and body conditions. Some metabolic diseases such as ketosis may affect milk composition. Several studies showed that milk IR spectra can be used to screen cows for subclinical ketosis (e.g., Hansen, 1999; Heuer et al., 2001; De Roos et al., 2007). Furthermore, in addition to cell count measurements, milk IR spectra might provide information regarding mastitis (Batavani et al., 2007). McParland et al. (2011) indicated that energy status of dairy cows can be predicted based on IR spectra. On the basis of routine prediction based on IR spectra, the predicted energy status could provide information about dairy farm management or body conditions of individual cows.

2.5 Conclusions

This study showed that genetic differences between cows explain a large part of the variation in milk IR wavenumbers. Furthermore we showed that the DGAT1 polymorphism significantly affected many IR wavenumbers. The polymorphisms of CSN3 and LGB also significantly affected some of the wavenumbers but no significant effect of SCD1 on any of the wavenumbers was found. Differences between herds accounted for a considerable part of phenotypic variance of individual wavenumbers and these wavenumbers might be of interest to discriminate milk from farms with different feeding or management regimes. Some wavenumbers are strongly influenced by water absorption and usually excluded when setting up prediction equations. However, we found that some of the wavenumbers in the water absorption region are significantly affected by DGAT1 polymorphism and lactation stage.

This suggested that these wavenumbers contain information on milk composition.

2.6 Acknowledgements

The China Scholarship Council is acknowledged for funding the PhD project of Qiuyu Wang. Cooperative Cattle Improvement Organization (CRV) is acknowledged for the sampling cows providing and imputation of genotypes. Milk Control Station (Zutphen, The Netherlands) is acknowledged for infrared spectra data. This study is part of the Dutch Milk Genomics Initiative and the project “Melk op Maat”, funded by Wageningen University (Wageningen, the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), CRV, the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands).

2.7 References

Bastin, C., N. Gengler, and H. Soyeurt. 2011. Phenotypic and genetic variability of production traits and milk fatty acid contents across days in milk for Walloon Holstein first-parity cows. J. Dairy Sci. 94: 4152-4163.

Batavani, R. A., S. Asri, and H. Naebzadeh. 2007. The effect of subclinical mastitis on milk composition in dairy cows. Iran. J. Vet. Res. 8: 205-211.

Berget, I., H. Martens, A. Kohler, S. K. Sjurseth, N. K. Afseth, B. Narum, and S. Lien. 2010. Caprine CSN1S1 haplotype effect on gene expression and milk composition measured by Fourier transform infrared spectroscopy. J. Dairy Sci. 93: 4340-4350.

Bittante, G., and A. Cecchinato. 2013. Genetic analysis of the Fourier-transform infrared spectra of bovine milk with emphasis on individual wavenumbers related to specific chemical bonds. J. Dairy Sci. 96: 5991-6006.

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

Bonfatti, V., G. Di Martino, and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and

contents of protein genetic variants of individual milk of Simmental cows. J. Dairy Sci. 94: 5776-5785.

Bovenhuis, H., J. A. M. van Arendonk, and S. Korver. 1992. Associations between milk protein polymorphisms and milk production traits. J. Dairy Sci. 75: 2549-2559.

Bovenhuis, H., M. H. P. W. Visker, H. J. F. van Valenberg, A. J. Buitenhuis, and J. A. M. van Arendonk. 2015. Effects of the DGAT1 polymorphism on test-day milk production traits throughout lactation. J. Dairy Sci. 98: 6572-6582.

Chilliard, Y., F. Glasser, A. Ferlay, L. Bernard, J. Rouel, and M. Doreau. 2007. Diet, rumen biohydrogenation and nutritional quality of cow and goat milk fat. Eur. J. Lipid Sci. Technol. 109: 828-855.

Chilliard, Y., C. Martin, J. Rouel, and M. Doreau. 2009. Milk fatty acids in dairy cows fed whole crude linseed, extruded linseed, or linseed oil, and their relationship with methane output. J. Dairy Sci. 92: 5199-5211.

De Marchi, M., V. Bonfatti, A. Cecchinato, G. Di Martino, and P. Carnier. 2010. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Ital. J. Anim. Sci. 8: 399-401.

De Marchi, M., C. C. Fagan, C. P. O'Donnell, A. Cecchinato, R. Dal Zotto, M. Cassandro, M. Penasa, and G. Bittante. 2009. Prediction of coagulation properties, titratable acidity, and pH of bovine milk using mid-infrared spectroscopy. J. Dairy Sci. 92: 423-432.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97: 1171-1186.

De Roos, A. P. W., H. J. C. M. van Den Bijgaart, J. Hørlyk, and G. De Jong. 2007. Screening for subclinical ketosis in dairy cattle by Fourier transform infrared spectrometry. J. Dairy Sci. 90: 1761-1766.

Duchemin, S., H. Bovenhuis, W. M. Stoop, A. C. Bouwman, J. A. M. van Arendonk, and M. H. P. W. Visker. 2013. Genetic correlation between composition of bovine milk fat in winter and summer, and DGAT1 and SCD1 by season interactions. . J. Dairy Sci. 96: 592-604.

Dufour, E. 2009. Principles of infrared spectroscopy. Pages 1-27 in Infrared Spectroscopy for Food Quality Analysis and Control. D. W. Sun, ed. Acad. Press, San Diego, CA.

Emery, R. S. 1978. Feeding for increased milk protein. J. Dairy Sci. 61: 825-828.

Eskildsen, C. E., M. A. Rasmussen, S. B. Engelsen, L. B. Larsen, N. A. Poulsen, and T. Skov. 2014. Quantification of individual fatty acids in

bovine milk by infrared spectroscopy and chemometrics: Understanding predictions of highly collinear reference variables. J. Dairy Sci. 97: 7940-7951.

Garidel, P., and H. Schott. 2006. Fourier-transform midinfrared spectroscopy for analysis and screening of liquid protein formulations part. 2: detailed analysis and applications. BioProcess Int. 4: 48-55.

German, J. B., and C. J. Dillard. 2006. Composition, structure and absorption of milk lipids: a source of energy, fat-soluble nutrients and bioactive molecules. Crit. Rev. Food Sci. Nutr. 46: 57-92.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in Dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12: 222-231.

Grelet, C., J. F. Pierna, P. Dardenne, V. Baeten, F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. J. Dairy Sci. 98: 2150-2160.

Grummer, R. R. 1991. Effect of feed on the composition of milk fat. J. Dairy Sci. 74: 3244-3257.

Hansen, P. W. 1999. Screening of dairy cows for ketosis by use of infrared spectroscopy and multivariate calibration. J. Dairy Sci. 82: 2005-2010.

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.

Heuer, C., H. J. Luinge, E. T. G. Lutz, Y. H. Schukken, J. H. van der Maas, H. Wilmink, and J. P. T. M. Noordhuizen. 2001. Determination of acetone in cow milk by Fourier transform infrared spectroscopy for the detection of subclinical ketosis. J. Dairy Sci. 84: 575-582.

Hewavitharana, A. K., and B. van Brakel. 1997. Fourier transform infrared spectrometric method for the rapid determination of casein in raw milk. Analyst. 122: 701-704.

ICAR (International Committee for Animal Recording). 2012. International agreement of recording practices – Guidelines approved by the General Assembly held in Cork, Ireland on June 2012. ICAR, Rome. Italy.

Jelen, P. 2007. Innovative uses of milk in human nutrition and health. Proceedings of the 35th Biennal Session of ICAR, Kuopio, Finland. EAAP publication 121. EAAP, Rome, Italy.

Karoui, R., G. Downey, and C. Blecker. 2010. Mid-infrared spectroscopy coupled with chemometrics: A tool for the analysis of intact food systems and the exploration of their molecular structure− Quality relationships− A review. Chem. Rev. 110: 6144-6168.

Lunden, A., M. Nilsson, and L. Janson. 1997. Marked effect of β-lactoglobulin polymorphism on the ratio of casein to total protein in milk. J. Dairy Sci. 80: 2996-3005.

McParland, S., G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J. Dairy Sci. 94:3651–3661.

Miglior, F., B. L. Muir, and B. J. van Doormaal. 2005. Selection indices in Holstein cattle of various countries. J. Dairy Sci. 88: 1255-1263.

Osborne, B. G., and T. Fearn. 1986. Near Infrared Spectroscopy in Food Analysis. Longman, Harlow, UK.

Palmquist, D. L., A. D. Beaulieu, and D. M. Barbano. 1993. Feed and animal factors influencing milk fat composition. J. Dairy Sci. 76: 1753-1771.

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. J. Dairy Sci. 94: 5683-5690.

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Prediction of β-lactoglobulin genotypes based on milk Fourier transform infrared spectra. J. Dairy Sci. 94: 4183-4188.

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.

Safar, M., D. Bertrand, P. Robert, M. F. Devaux, and C. Genot. 1994. Characterization of edible oils, butters and margarines by Fourier transform infrared spectroscopy with attenuated total reflectance. J. Am. Oil Chem. Soc. 71: 371-377.

Samorè, A. B., F. Canavesi, A. Rossoni, and A. Bagnato. 2012. Genetics of casein content in Brown Swiss and Italian Holstein dairy cattle breeds. Ital. J. Anim. Sci. 11: 196-202.

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.

Schutz, M. M., L. B. Hansen, G. R. Steuernagel, and A. L. Kuck. 1990. Variation of milk, fat, protein, and somatic cells for dairy cattle. J. Dairy Sci. 73: 484-493.

Slots, T., G. Butler, C. Leifert, T. Kristensen, L. H. Skibsted, and J. H. Nielsen. 2009. Potentials to differentiate milk composition by different feeding strategies. J. Dairy Sci. 92: 2057-2066.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89: 3690-3695.

Soyeurt, H., I. Misztal, and N. Gengler. 2010. Genetic variability of milk components based on mid-infrared spectral data. J. Dairy Sci. 93: 1722-1728.

Stoop, W. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2009. Effect of lactation stage and energy status on milk fat composition of Holstein-Friesian cows. J. Dairy Sci. 92: 1469-1478.

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.

Subramanian, A., and L. Rodrigucz-Saona. 2009. Fourier transform infrared (FTIR) spectroscopy. Pages 145-178 in Infrared Spectroscopy for Food Quality Analysis and Control. D. W. Sun, ed. Acad. Press, San Diego, CA.

Valenti, B., B. Martin, D. Andueza, C. Leroux, C. Labonne, F. Lahalle, H. Larroqued, P. Brunschwige, C. Lecomtef, M. Brochardg, and A. Ferlay. 2013. Infrared spectroscopic methods for the discrimination of cows' milk according to the feeding system, cow breed and altitude of the dairy farm. Int. Dairy J. 32: 26-32.

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. Nederlands melk en Zuiveltijdschrift. 46: 145-168.

Van Haelst, Y. N. T., A. Beeckman, A. T. M. van Knegsel, and V. Fievez. 2008. Short Communication: Elevated Concentrations of Oleic Acid and Long-Chain Fatty Acids in Milk Fat of Multiparous Subclinical Ketotic Cows. J. Dairy Sci. 91: 4683-4686.

Visker, M. H. P. W., B. W. Dibbits, S. M. Kinders, H. J. F. van Valenberg, J. A. M. van Arendonk, and H. Bovenhuis. 2011. Association of bovine β-casein

protein variant I with milk production and milk protein composition. Anim. Genet. 42: 212–218.

Vlaeminck, B., V. Fievez, S. Tamminga, R. J. Dewhurst, A. van Vuuren, D. de Brabander, and D. Demeyer. 2006. Milk odd-and branched-chain fatty acids in relation to the rumen fermentation pattern. J. Dairy Sci. 89: 3954-3964.

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. Crop Pasture Sci. 55: 1009-1028.

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.

Yang, H., and J. Irudayaraj. 2000. Characterization of semisolid fats and edible oils by Fourier transform infrared photoacoustic spectroscopy. J. Am. Oil Chem. Soc. 77: 291-295.

3

Genome wide association study for milk infrared wavenumbers

Qiuyu Wang and Henk Bovenhuis

Animal Breeding and Genomics Centre, Wageningen University & Research,

PO Box 338, 6700AH, Wageningen, the Netherlands

Journal of Dairy Science (2018) 101(3): 2260-2272

Abstract

Individual wavenumbers of the infrared (IR) spectra of bovine milk have been shown to be moderately to highly heritable. The objective of this study was to identify genomic regions associated with individual milk IR wavenumbers. This is expected to provide information about the genetic background of milk composition and give insight in the relation between IR wavenumbers and milk components. For this purpose a genome wide association study (GWAS) was performed for a selected set of 50 individual IR wavenumbers measured on 1,748 Dutch Holstein cows. Significant associations were detected for 28 of the 50 wavenumbers. In total 24 genomic regions distributed over 16 bovine chromosomes were identified. Major genomic regions associated with milk IR wavenumbers were identified on chromosomes 1, 5, 6, 14, 19 and 20. Most of these regions also showed significant associations with fat%, protein% or lactose%. However, we also identified some new regions which were not associated with any one of these routinely collected milk composition traits. On chromosome 1 two new genomic regions were identified and we hypothesise that they are related to variation in milk phosphorus content and orotic acid, respectively. On chromosome 20 a new genomic region was identified which seem to be related to citric acid. Identification of genomic regions associated with milk phosphorus content, orotic acid and citric acid suggest that the milk infrared spectra contain direct information on these milk components. Consequently milk infrared analyses probably can be used to predict these milk components, which have low concentrations in milk. This can lead to novel applications of milk IR spectroscopy for dairy cattle breeding and herd management.

Key words:

bovine milk, IR wavenumbers, genome wide association study

3.1 Introduction

Infrared (IR) spectroscopy is a fast and relatively cheap method to determine milk composition. It is the standard method for routine quantification of fat, protein and lactose content of milk (ICAR, 2012). Several studies showed that IR can also be used to determine milk fat composition (e.g., Soyeurt et al., 2006; Rutten et al., 2009). More recently it has been suggested that milk IR spectra also can be used to predict other characteristics like e.g. negative energy balance and methane emission of dairy cows (McParland et al., 2011; Dehareng et al., 2012). The IR spectra are caused by the absorption of electromagnetic radiation at wavenumbers that are correlated to the vibrations (stretching and bending) of specific chemical bonds within a molecule (Sun, 2009). The analysis of milk IR applies mainly the region 400-4,000 cm^{-1} or 2,500-25,000 nm, due to active vibrations of various chemical bonds in different small regions. By assigning absorption bands of chemical bonds in IR to milk components, we can identify milk IR wavenumbers associated with some common milk components such as milk fat, milk protein and lactose. The assignment is based on the fact that the major chemical bonds with known vibration frequencies are abundant in these common milk components. However, due to the complexity of milk, it is sometimes difficult to relate wavenumbers to specific components. Milk IR spectra result from a comprehensive scan of milk and may contain information on milk components that are currently not quantified.

Genome wide association study (GWAS) has been performed for routinely recorded milk production traits like milk yield, fat and protein content (e.g., Daetwyler et al., 2008; Pryce et al., 2010; Cole et al., 2011). More recently some studies performed GWAS for detailed milk fat and protein composition (Bouwman et al., 2011; Schopen et al., 2011). Buitenhuis et al. (2013) performed a GWAS for milk components based on nuclear magnetic resonance (NMR) spectroscopy. More recently Sanchez et al. (2016) performed a GWAS based on IR-predicted milk protein composition. The bovine milk IR spectrum might provide unique information about the genetic background of milk composition and to our knowledge, no GWAS based on bovine milk IR wavenumbers has been reported previously.

Genetic analyses of bovine milk IR spectra showed that most IR wavenumbers are heritable (e.g., Soyeurt et al., 2010; Bittante and Cecchinato, 2013). Wang et al. (2016) showed that the majority of the wavenumbers have moderate to high inter-herd heritabilities ranging from 0.20 to 0.60. Performing GWAS for milk IR wavenumbers will provide new possibilities to identify genomic regions responsible for differences in milk composition and enhance our understanding of the background of bovine milk IR spectra.

The objective of this study was to perform GWAS for a representative set of milk IR wavenumbers in order to unravel the genetic background of milk IR spectra and understand the relationship between milk components and IR wavenumbers.

3.2 Materials and methods

3.2.1 Data

Data used for this study was based on one morning milk sample from 1,748 first parity Holstein Friesians cows. Milk samples were collected from February till March 2005. All cows have at least 87.5% Holstein Friesian genes. The population consisted of 5 large paternal half-sib families from proven sires (98 to 196 daughters per sire), and 50 small paternal half-sib families from test sires (8 to 23 daughters per sire), as well as 168 cows descending from 44 other proven sires (1 to 25 daughters per sire) to assure at least 3 cows per herd. The pedigree of the cows was supplied by CRV (Cooperative cattle improvement organization, Arnhem, the Netherlands).

Milk IR spectra were recorded in a 10 mL milk sample using the same MilkoScan FT 6000 equipment (FOSS, Denmark) at the certified laboratory of the Dutch Milk Control Station QLIP (Zutphen, the Netherlands). The spectra consist of the transmittance values measured at 1,060 individual wavenumbers between 925 and 5,008 cm^{-1}. Transmittance quantifies the amount of light passing through the milk sample.

Phenotypic correlations among individual wavenumbers were analysed using PROC CLUSTER in SAS 9.3 (SAS Institute, 2001). More than 95% of the phenotypic variance in the 1,060 wavenumbers could be explained based on

50 clusters. From each of these 50 clusters 1 wavenumber was selected for the GWAS. If a cluster contained multiple consecutive wavenumbers, the middle wavenumber was selected. We performed GWAS for fat%, protein% and lactose% based on the same milk samples. Fat%, protein% and lactose% of the same milk samples were predicted based on milk IR analyses by the Dutch Milk Control Station QLIP (Zutphen, the Netherlands).

3.2.2 Genotypes

Blood samples were collected from the cows for DNA isolation. A custom 50K SNP Infinium Array (Illumina, San Diego, CA, USA) designed by CRV was used for genotyping. The cows were genotyped for 50,688 SNPs. The SNPs were mapped using the bovine genome assembly BTAU_4.0 (Liu et al., 2009). Among these SNPs, 778 were not mapped to any of the *Bos taurus* (BTA) chromosomes and therefore were assigned to BTA 0. In total 591 SNPs were assigned to the X chromosome.

3.2.3 Statistical analysis

The association study for each IR wavenumber was performed based on the following model:

$$y_{ijklmn}=\mu+\beta_1 * lactst_{ijklmn}+\beta_2 * afc_{ijklmn} + season_i + sirecode_j + date_k + SNP_l + herd_m + a_n + e_{ijklmn}$$

Equation 3.1

where y_{ijklmn} is the transmittance value of the IR wavenumber, or predicted milk fat%, protein%, and lactose%; μ is the general mean; $lactst_{ijklmn}$ is a covariate for the effect of lactation stage (63 to 282 days) with regression coefficient β_1; afc_{ijklmn} is a covariate for the effect of age at first calving with regression coefficient β_2; $season_i$ is the fixed effect accounting for season of calving (June-Aug 2004, Sept-Nov 2004 or Dec 2004-Jan 2005); $sirecode_j$ is the fixed effect accounting for possible differences in genetic level between the groups of proven bull daughters and young bull daughters; $date_k$ is the fixed effect accounting for the effect of day at which IR analyses of milk samples took place (k ranges from 1 to 17); SNP_l is the fixed effect of SNP genotype; $herd_m$ is a random effect of herd m, distributed as $N(\mathbf{0}, \mathbf{I}\sigma_h^2)$, with identity matrix $\mathbf{I}$ and

herd variance σ_h^2; a_n is a random additive genetic effect of animal *n*, distributed as *N* (**0**, $\mathbf{A}\sigma_a^2$), with additive genetic relationship matrix **A** and the additive genetic variance σ_a^2. The additive genetic relationship matrix was constructed based on 12,548 animals. e_{ijklmn} is a random residual effect, distributed as *N* (**0**, $\mathbf{I}\sigma_e^2$), with identity matrix **I** and error variance σ_e^2.

In the GWAS variance components were fixed at values estimated and presented by Wang et al. (2016) using the same data (see Table 3.1). Analyses were performed using ASReml (Gilmour et al., 2009).

The genome wide false discovery rate (FDR) was used to determine significance. FDR were determined for each trait separately using the R (R Core Team, 2013) package 'qvalue'. Associations with a FDR < 0.01 were considered significant. A sensitivity analysis was performed in case the lead SNP had a genotype class with at least 1 but less than 5 observations. In that case the records for the minor genotype class were excluded from the analysis and the analysis was repeated based on the remaining genotype classes.

In the current study a genomic region containing at least two SNPs located within 10 Mbp with a significant effect on a trait is defined as a significant region for that trait. The region extends until the last significant SNP was not followed by another significant SNP within 10 Mbp.

3.3 Results and discussion

Table 3.1 shows the IR wavenumbers that were selected for the current study, and IR-predicted fat%, protein%, and lactose%. Table 3.1 shows the proportion of the variance due to herd and genetics for each of the wavenumbers and information regarding chemical bonds with specific absorption bands for the selected wavenumbers (Sun, 2009; Diem, 2015). Several of the selected wavenumbers were located in the water absorption region (between 1,619 and 1,674 cm^{-1} and between 3,073 and 3,667 cm^{-1}). These wavenumbers were included in the GWAS as a result of the applied wavenumber selection procedure, i.e. these wavenumbers tend to have low correlations with other wavenumbers. Some of the selected wavenumbers are important for the quantification of milk fat%, protein% and lactose% due to the absorptions of abundant chemical bonds in these molecules.

Table 3.1 Descriptive summary of selected milk IR wavenumbers.

Trait	Wavenumber (cm^{-1})	Herd[1]	Heritability[2]	σ_p	Specific bond[3]	Regions
IR wavenumber						
WN20	999	0.10	0.53	0.0109	C-H	13, 14_a, 19, 20_a
WN34	1053	0.09	0.51	0.0106	C-O asymmetric stretch, C-N stretch	6_a, 14_a, 19, 20_b, 28
WN50	1114	0.13	0.44	0.0090	C-O, C-C	6_a, 10, 14_a, 19, 20_b
WN72	1199	0.10	0.63	0.0072	C-O, C-N	5_a, 5_b, 10, 14_a
WN80	1230	0.14	0.54	0.0075	C-O	5_a, 5_b, 10, 14_a, 14_b, 20_a
WN106	1330	0.19	0.27	0.0056	C-H symmetric bend	6_a, 14_a, 19, 28
WN126	1407	0.19	0.28	0.0084	C-N stretch	11, 14_a, 20_b
WN142	1469	0.23	0.32	0.0061	C-H asymmetric bend	6_b, 10, 14_a, 14_b, 15, 20_a
WN149	1496	0.16	0.30	0.0121	C-H	5_a, 13, 14_a
WN156	1523	0.24	0.26	0.0183	N-H bending, C-N stretch	6_b, 15, 29
WN176	1600	0.10	0.39	0.0342	C=O Stretch, C-N stretch	5_a, 5_b, 6_a, 13, 14_a, 20_b
WN185	1635	0.01	0.04	0.5836	O-H bend, C=O stretch	
WN191	1658	0.00	0.01	0.3563	O-H bend, C=C stretch	
WN194	1670	0.00	0.01	0.1487	O-H bend, C=C stretch	
WN208	1724	0.19	0.20	0.0036	O-H bend	1_a, 14_a, 20, 28
WN220	1770	0.20	0.29	0.0022	C=O Stretch	6_b, 14_a
WN279	1997	0.10	0.54	0.0089	C=O Stretch	5_a, 5_b, 10, 14_a, 20_a
WN291	2044	0.11	0.44	0.0113	-	
WN414	2518	0.12	0.39	0.0020	P=O	1_b, 6_b, 14_a
WN432	2587	0.14	0.34	0.0016	P=O	1_b, 14_a, 19, 28
WN452	2664	0.16	0.37	0.0021	S-H stretch	1_b, 6_a, 10, 19
WN470	2734	0.17	0.35	0.0028	-	1_b, 6_a, 19, 28
WN542	3011	0.11	0.38	0.0254	-	13, 14_a
WN561	3085	0.03	0.02	0.1631	C-H stretch	14_a
WN567	3108	0.05	0.00	0.3152	v =C-H	
WN572	3127	0.01	0.06	0.3999	O-H stretch	
WN576	3142	0.04	0.07	0.4350	O-H stretch	
WN580	3158	0.01	0.15	0.4933	O-H stretch	
WN584	3173	0.02	0.02	0.4702	O-H stretch	
WN587	3185	0.04	0.00	0.4931	O-H stretch	
WN594	3212	0.03	0.00	0.5310	O-H stretch	
WN600	3235	0.00	0.00	0.6113	O-H stretch	
WN607	3262	0.01	0.00	0.6212	O-H stretch	
WN613	3285	0.04	0.00	0.6136	O-H stretch	
WN617	3301	0.03	0.00	0.5564	O-H stretch	
WN622	3320	0.01	0.00	0.6359	O-H stretch	
WN626	3335	0.09	0.00	0.6233	O-H stretch	
WN637	3378	0.09	0.00	0.5675	O-H stretch	
WN641	3393	0.02	0.00	0.5447	O-H stretch	
WN644	3405	0.02	0.00	0.5293	O-H stretch	
WN653	3439	0.01	0.00	0.2371	O-H stretch	
WN659	3463	0.00	0.00	0.1431	O-H stretch	
WN668	3497	0.00	0.03	0.0996	O-H stretch	14_a
WN679	3540	0.00	0.00	0.0751	O-H stretch	
WN689	3578	0.00	0.00	0.0844	O-H stretch	
WN707	3648	0.01	0.00	0.1216	O-H stretch	
WN717	3686	0.17	0.32	0.0250	O-H stretch	5_a, 5_b, 6_b, 10, 14_a, 20_a
WN728	3729	0.27	0.14	0.0089	O-H stretch	6_b
WN764	3867	0.18	0.32	0.0076	O-H stretch	6_b, 13, 14_a
WN886	4338	0.13	0.45	0.0091	O-H stretch	5_a, 6_a, 13, 14_a
IR-predicted milk components						
Fat%		0.10	0.52	0.6901		5_a, 5_b, 6_a, 13, 14_a
Protein%		0.18	0.46	0.2898		5_b, 6_b, 14_a,b, 15, 20_a, 26, 29
Lactose%		0.07	0.59	0.1417		6_a, 19, 28

[1] It describes the proportion of total variance explained by the differences between herds. Herd = $\frac{\sigma_h^2}{\sigma_a^2+\sigma_h^2+\sigma_e^2}$, where σ_a^2 is additive genetic variance, σ_h^2 is the variance due to differences between herds, σ_e^2 is the residual variance.

[2] heritability = $\frac{\sigma_a^2}{\sigma_a^2+\sigma_h^2+\sigma_e^2}$, where σ_a^2 is additive genetic variance, σ_h^2 is the variance due to differences between herds, σ_e^2 is the residual variance.

[3] This is according to Sun. 2009 and Diem. 2015.

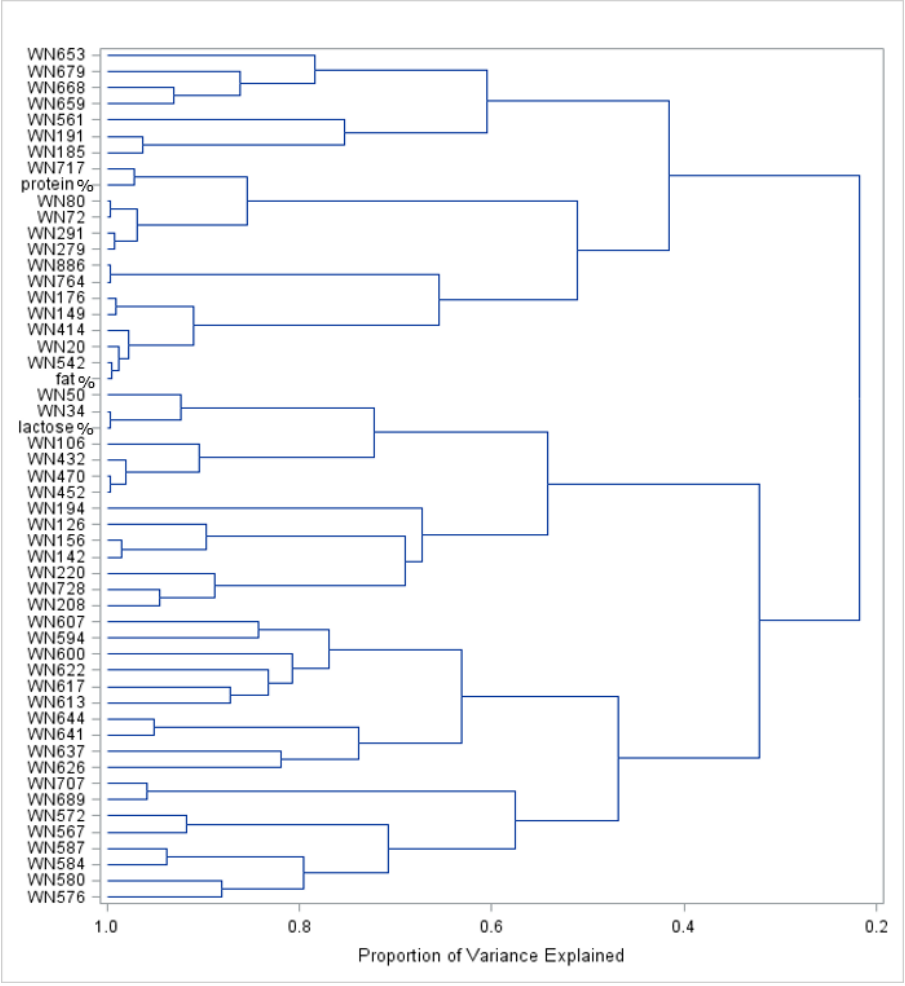


Figure 3.1 Cluster tree based on principal component analysis of phenotypic correlations among the IR wavenumbers and fat%, protein%, lactose% analysed in this study.

Correlations among wavenumbers analysed in this study, including IR-predicted fat%, protein%, and lactose% are graphically illustrated using a cluster tree (Figure 3.1). Wavenumbers from the water absorption region (e.g., WN587, WN626, and WN641) show weak correlations with other groups of wavenumbers. Many of the selected wavenumbers are related with IR-predicted fat%, protein% or lactose%. Some of them show especially strong correlations with fat% (WN542), protein% (WN717), or lactose% (WN34).

The significant genomic regions detected in the GWAS are summarized in Table 3.1 and genome wide association plots for fat%, protein% and lactose% are shown in Figure 3.2. For the IR wavenumbers not all genome wide association plots are shown; Figure 3.3 shows plots for those wavenumbers that show a strong signal in some of the chromosomal regions. The y-axis in Figure 3.2 and 3.3 are cut of at a $-Log_{10}(P)$ value of 20.

The GWAS results indicated that 28 out of the 50 wavenumbers studied are significantly associated with at least one genomic region. Significant associations were detected for in total 24 genomic regions distributed over 16 bovine chromosomes. No genomic regions were associated with all selected wavenumbers. Chromosomes BTA 1, BTA 5, BTA 6, BTA 14, BTA 19 and BTA 20 showed significant associations with multiple wavenumbers or multiple regions on those chromosomes showed significant associations. Results from these chromosomes will be discussed in more details.

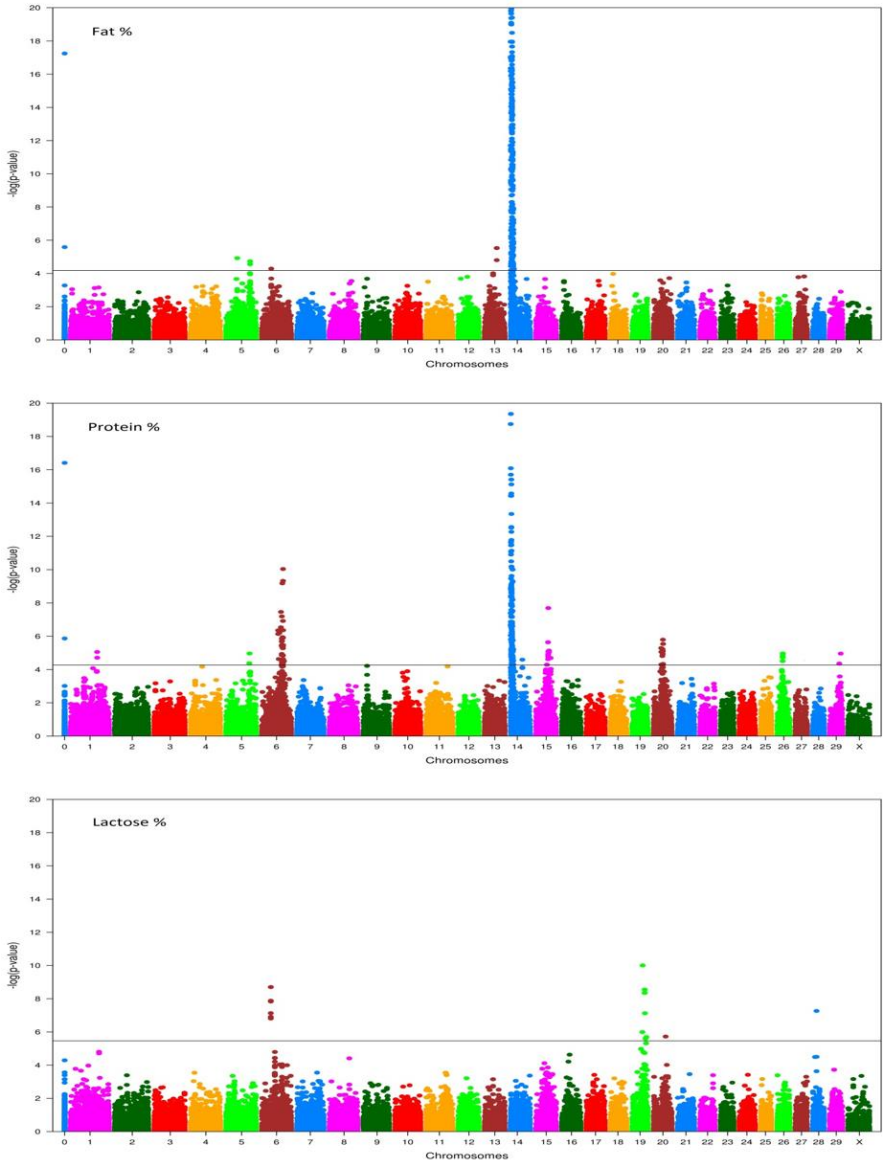
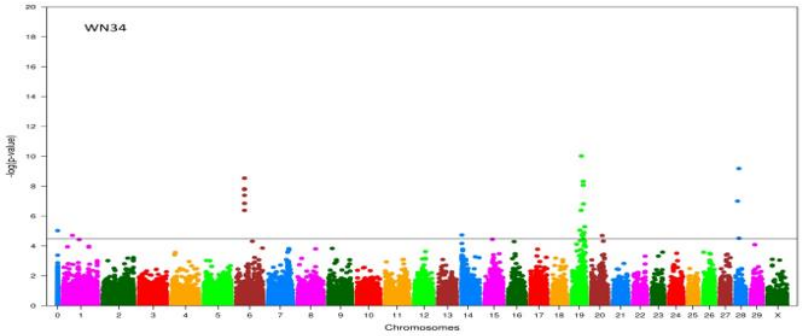
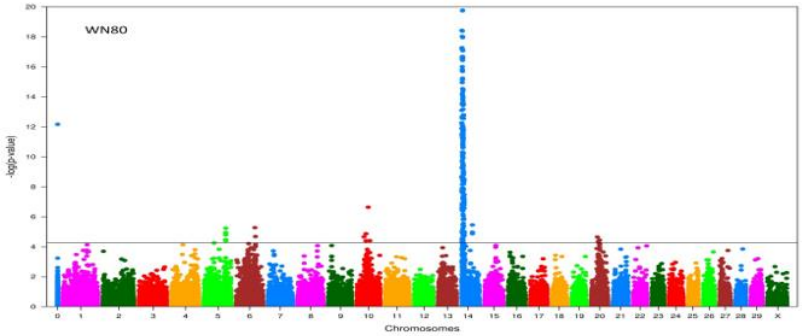
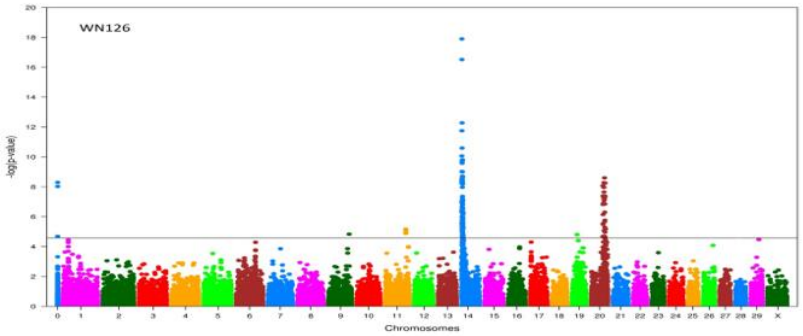
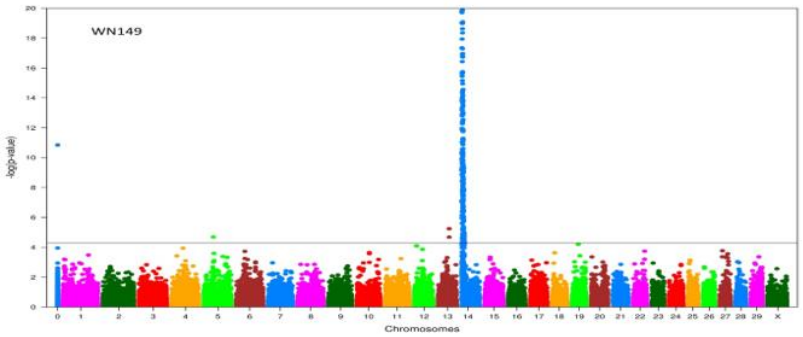


Figure 3.2 Manhattan plots of genome wide association studies on IR-predicted fat%, protein% and lactose%. The genomic position is represented along the x-axis and chromosome numbers are given. The $-Log_{10}(P)$ values of SNPs are given on y-axis and cut off at 20. The horizontal lines represent the 0.01 false discovery rate thresholds.


WN34
-log(p-value)
Chromosomes


WN80
-log(p-value)
Chromosomes


WN126
-log(p-value)
Chromosomes


WN149
-log(p-value)
Chromosomes

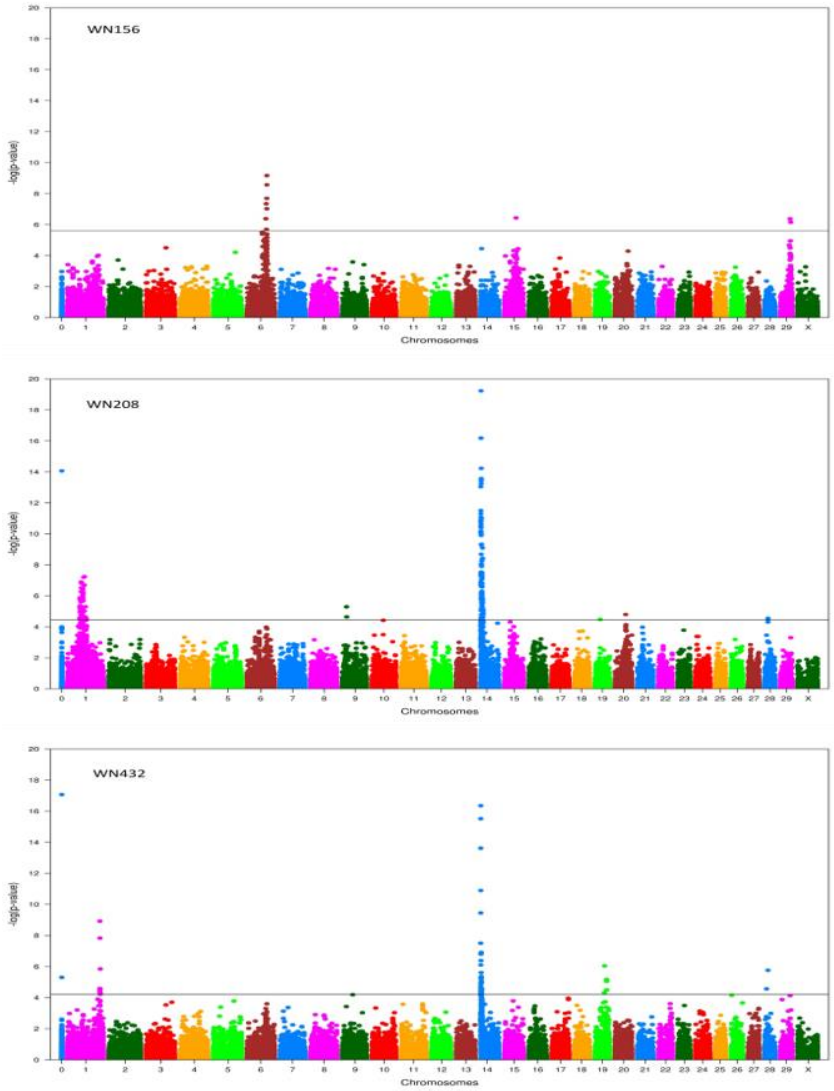


Figure 3.3 Manhattan plots of genome wide association studies on representative IR wavenumbers, showing all the significant genomic regions. The genomic position is represented along the x-axis and chromosome numbers are given. The $-Log_{10}(P)$ values of SNPs are given on y-axis and cut off at 20. The horizontal lines represent the 0.01 false discovery rate thresholds.

3.3.1 BTA 1

We detected 2 significant regions on BTA 1; region 1_a, from 56.9 to 80.8 Mbp (see Figure 3.3, WN208) and region 1_b, from 145.4 to 147.3 Mbp (see Figure 3.3, WN432). These two regions did not show significant associations in our study with the routinely recorded milk composition traits fat%, protein% and lactose% (Figure 3.2), and to our knowledge these regions also have not been reported to be associated with routinely recorded milk production traits in other studies. Region 1_a is associated with WN208 and the lead SNP is rs29022932 with a $-Log_{10}(P)$ of 7.2 is located at 76.6 Mbp. Region 1_b showed significant associations for WN414, WN432, WN452 and WN470 with $-Log_{10}(P)$ ranging from 6.2 to 8.9. SNP rs29019625 showed the strongest association for WN414, and an adjacent SNP rs43281569 showed the strongest association for WN432, WN452, and WN470.

Region 1_a, associated with WN208, harbours the gene uridine monophosphate synthase (UMPS), which catalyses the last two steps of de novo pyrimidine synthesis converting orotic acid to 5'-monophosphate. This gene is known for the lethal genetic defect DUMPS (Deficiency of Uridine Mono Phosphate Synthase). Female carriers of this defect are known to have elevated levels of orotic acid in their milk and urine (Robinson et al. 1984). Buitenhuis et al. (2013) reported a QTL for orotic acid in the same region as the current study detected. Buitenhuis et al. (2013) quantified orotic acid using NMR Spectroscopy. Based on these results we hypothesise that milk IR spectra can be used to predict orotic acid in milk.

In region 1_b located near 146 Mbp, we found several SNPs (e.g., rs29019625) significantly associated with WN414, WN432, WN452, and WN470. Some of these wavenumbers are related to P=O chemical bonds (Table 3.1). Using Inductively Coupled Plasma Emission Spectroscopy Buitenhuis et al. (2015) and Kemper et al. (2016) identified a QTL at 144.4 Mbp for milk phosphorus concentration. These studies suggested SLC37A1 (solute carrier family 37 member 1) as the most likely candidate gene, explaining approximately 10% phenotypic variance of phosphorus concentration. SLC37A1 functions as a phosphorus antiporter (Chou et al., 2013), transporting glucose-6-phosphate and phosphorus in opposite directions.

It has been shown that milk phosphorous content can be predicted based on milk IR spectra (Soyeurt et al., 2009; Toffanin et al., 2015, Bonfatti et al., 2016). This prediction might be based on the relation between milk protein content and milk phosphorus content. However, region 1_b does not show an association with protein% in our study (Figure 3.2) and therefore these results suggest that the milk IR spectra contain direct information on milk phosphorus content. The current study provides additional evidence that milk phosphorus content can be predicted by milk IR spectra. Large scale quantification of milk phosphorus content using infrared analyses might result in management strategies to better feed cows according their phosphorus requirement and in more efficient use of phosphorus by the dairy sector.

3.3.2 BTA 5

We detected 2 significant regions on BTA 5; region 5_a, from 43.0 to 47.0 Mbp (see Figure 3.2 for fat% and Figure 3.3 for WN149) and region 5_b, from 97.4 to 101.1 Mbp (see Figure 3.2 for fat% and protein% and Figure 3.3 for WN80). Region 5_a contained 2 SNPs (rs41616530 and rs29014575) showing significant effects on multiple IR wavenumbers (e.g., WN72, WN80, WN149, WN176, WN717) with $-Log_{10}(P)$ ranging from 4.2 to 5.1, and on fat% with $-Log_{10}(P)$ of 4.9. Figure 3.1 shows that wavenumbers associated with region 5_a are phenotypically correlated with protein% and fat%, however protein% didn't show any association in this region. This region has not been associated with milk production traits in other studies.

In region 5_b, SNPs were associated with WN72, WN80, WN176, WN279, WN717 ($-Log_{10}(P)$ up to 5.5) and fat% ($-Log_{10}(P)$ up to 4.7). The lead SNP rs29016908 was located at 101.1 Mbp. Region 5_b also contained several SNP (lead SNP rs41569048) significantly associated with protein% ($-Log_{10}(P)$ up to 4.9). The same region was previously reported by Schopen et al. (2011). Based on sequence-based imputation and expression QTL mapping, Littlejohn et al. (2016) identified microsomal glutathione S-transferase 1 (MGST1) as the most likely causal gene for the observed associations with fat% on BTA 5. The MGST1 is located at 93.5 Mbp and Littlejohn et al. (2016) found this chromosomal region to be strongly associated with milk fat%, protein% and milk yield.

3.3.3 BTA 6

We detected 2 significant regions on BTA 6; region 6_a, near 37.1 Mbp (see Figure 3.2 for lactose% and Figure 3.3 for WN34) and region 6_b, from 68.0 to 93.5 Mbp (see Figure 3.2 for protein% and Figure 3.3 for WN156). Region 6_a was associated with WN34, WN50, WN106, WN452, WN470, WN728 and lactose%. For these wavenumbers the lead SNP rs81154100 showed $-Log_{10}(P)$ values ranging from 4.5 to 8.5. The association detected for region 6_a with lactose% is in line with Kemper et al. (2016). As shown in Figure 3.1, several of the wavenumbers associated with region 6_a are strongly correlated with lactose%. Significant associations for this region have also been reported for somatic cell score (Daetwyler et al., 2008) which is known to be negatively correlated with lactose% (e.g., Welper et al. 1992).

Region 6_a harbours the gene ATP binding cassette, subfamily G, member 2 (ABCG2*,* located at 37.4 Mbp). The ABCG2 has been suggested to be the causative gene underlying the QTL (Cohen-Zinder et al., 2005; Olsen et al., 2007). The expression of ABCG2 in the mammary gland is induced during late pregnancy and lactation period, contributing to the secretion of nutrients into the milk (Jonker et al., 2005), for instance, a major role for ABCG2 in the secretion of antimicrobials and riboflavin into the milk of ruminants has been reported (Real et al., 2011a; Otero et al., 2016).

In region 6_b, many SNPs were very significantly associated with WN142, WN156, WN717, WN728, and WN764. For these wavenumbers the lead SNP rs43703016 showed $-Log_{10}(P)$ values ranging from 7.9 to 17.6 and for protein% the $-Log_{10}(P)$ value was 10.0. As shown in Figure 3.1, WN142, WN156, and WN728 were phenotypically strongly correlated, and WN717 was strongly correlated with protein%.

This region contains the 4 casein coding genes. The lead SNP rs43703016 located at 88.5 Mbp was the most significant SNP associated with these traits in the region. This SNP is one of the two SNPs that are causal for protein variants A and B of κ-casein gene (CSN3). Schopen et al. (2011) showed that this SNP is strongly associated with protein%, β-lactoglobulin content, casein index and especially κ-casein content. Previously we detected significant effects of CSN3 milk protein variants on WN142, WN156, WN717, WN728, and WN764 (Wang et al., 2016).

3.3.4 BTA 14

We detected 2 significant regions on BTA14: region 14_a, from 0 to 18.5 Mbp (see Figure 3.2 for fat%, and protein% and Figure 3.3 for WN80, WN126, WN149, WN208 and WN432) and region 14_b, near 49.3 Mbp (see Figure 3.3 for WN80). Region 14_a was strongly associated with 21 of the 50 selected IR wavenumbers, as well as protein% and fat%. WN279 showed the highest $-Log_{10}(P)$ of 103.4. Region 14_a contains the diacylglycerol O-acyltransferase 1 (DGAT1) gene. Previously Wang et al. (2016) showed that the DGAT1 K232A polymorphism has significant effects on many milk IR wavenumbers. It has been shown in several studies that the DGAT1 K232A polymorphism has major effects on milk yield, milk fat%, protein% (Grisart et al., 2002). Furthermore, Bovenhuis et al. (2016) showed that the DGAT1 K232A polymorphism has significant effects on milk fatty acid, protein and mineral composition. We also found the significant association of region 14_a for WN668, which is part of the water absorption region. This agrees with our previous finding that the DGAT1 K232A polymorphisms are significantly associated with wavenumbers in the water absorption region (Wang et al., 2016). In this previous study we found that wavenumbers from 3,466 to 3,543 cm^{-1} (including WN668) were significantly affected by the DGAT1 K232A polymorphism with the highest $-Log_{10}(P)$ value of 17.0. The signals for wavenumbers in the so called water absorption region are dominated by the effect of water in milk, but not completely. This suggests that these wavenumbers contain useful information regarding milk composition.

Remarkably, we detected another region on BTA 14 (14_b) affecting milk composition. Region 14_b was significantly associated with WN80 and WN142. Multiple SNPs in this region (e.g., lead SNP rs41668861), which are located at approximately 49.3 Mbp were significantly associated with WN80 ($-Log_{10}(P)$ of 5.4; Figure 3.3) and WN142 ($-Log_{10}(P)$ of 4.8). The SNPs in this region were also associated with protein% ($-Log_{10}(P)$ of 4.6; Figure 3.2). The SNPs in this region remained significant after correcting for the DGAT1 K232A polymorphisms suggesting that this is the effect of a different QTL. Bennewitz et al. (2004) suggested that besides the DGAT1 K232A polymorphism there are other QTL on BTA 14 affecting milk production traits. Other studies reported there is a QTL in the region between 55 and 79 Mbp affecting milk production traits (Ashwell et al., 2004; Kolbehdari et al., 2009).

3.3.5 BTA 19

We detected significant associations for the region from 42.2 to 59.0 Mbp on BTA 19. The associations for this region are illustrated for lactose% in Figure 3.2 and for WN34 and WN432 in Figure 3.3. The lead SNP rs29020588, was significantly associated with WN20, WN34, WN50, WN106, WN432, WN452, and WN470 with $-Log_{10}(P)$ ranging from 4.6 to 10.0 and lactose% ($-Log_{10}(P)$ of 10.0). Another highly significant SNP was rs109400579 located at 52.0 Mbp and showed in general slightly lower $-Log_{10}(P)$ for the wavenumbers as compared to SNP rs29020588. Cecchinato et al. (2014) found a significant association for lactose% in Brown Swiss at 48.8 Mbp and suggested that the growth hormone 1 gene (GH1) is involved.

As shown in Figure 3.1, all the wavenumbers associated with this region on BTA 19, except WN20, are strongly correlated**.** Bouwman et al. (2011) reported that BTA 19 is strongly associated with short and medium chain saturated fatty acids (e.g., C14:0) and with long chain unsaturated fatty acids (e.g., C18:1). Bouwman et al. (2014) fine mapped the region on BTA 19 associated with fatty acids and showed that the most significant SNPs were located in an linkage disequilibrium block that contained the genes fatty acid synthase (FASN) and coiled-coil domain containing 57 (CCDC57). The FASN is involved in de novo fatty acids synthesis and has been associated with fat% and medium and long chain fatty acid content of milk (Roy et al., 2006; Morris et al., 2007). Medrano et al. (2010) showed that CCDC57 is expressed in the mammary gland but this gene has not been related to milk fat composition. Another gene ATP citrate lyase (ACLY) located at 42.7 Mbp is a critical enzyme linking glucose catabolism to lipogenesis by providing acetyl-CoA from mitochondrial citrate for fatty acid and cholesterol biosynthesis. Some other genes located in this region are signal transducer and activator of transcription 5A (STAT5A), sterol regulatory element binding transcription factor 1 (SREBF1). These genes are possibly related to fat composition or lactose%. This study didn't detect association for fat% in this region, which indicates that the QTL is responsible for relative amount of fatty acids but not total milk fat content.

3.3.6 BTA 20

We detected 2 significant regions on BTA 20: region 20_a, from 27.3 to 39.1 Mbp (see Figure 3.2 for protein% and Figure 3.3 for WN20, WN80, WN142, WN279, WN717) and region 20_b, from 52.8 to70.7 Mbp (see Figure 3.3 for WN126). The lead SNP (rs41257066) for region 20_a showed -Log_{10}(P) ranging from 4.9 to 5.8. This region harbours the growth hormone receptor (GHR) gene which has been reported to be associated with milk yield and composition (e.g., Blott et al. 2003, Kadri et al, 2015). Region 20_b is associated with WN106, WN176 and especially WN126 (-Log_{10}(P) of 8.6). Another SNP in this region was also associated with lactose% (Figure 3.2), however, this SNP was not associated with WN126. Therefore, the highly significant association detected for WN126 was not detected for the routinely recorded milk composition traits fat%, protein% and lactose%. For the same genomic region on BTA 20, Buitenhuis et al. (2013) reported a significant association for citric acid in Danish Holstein and therefore WN126 might be related to citric acid. Recently, Grelet et al. (2016) showed that citrate in milk can be predicted with good accuracy based on milk IR data.

3.3.7 Additional genomic regions

BTA 10. A SNP located at 51.6 Mbp was significantly associated with WN50, WN72, WN80, WN142, WN279, WN452 and WN717 with -Log_{10}(P) ranging from 4.8 to 6.6. The association of this region is shown for WN80 in Figure 3.3. In Danish studies this region on BTA 10 has been associated with milk fat composition and glycosylated κ-casein content (Buitenhuis et al. 2014; 2016). However, based on largely the same animals as used in the current study, Bouwman et al. (2012) did not detect significant associations of this region with milk fat composition.

BTA 11. The SNP located at approximately 95 Mbp (e.g., lead SNP rs29014608) were significantly associated with WN126 (-Log_{10}(P) up to 5.2). The association for this region is shown in Figure 3.3 for WN126. Wang et al. (2016) showed that the β-lactoglobulin polymorphism (LGB) has a significant effect on WN126. Schopen et al. (2011) reported that the region from 84.3 to 110.2 Mbp is associated with milk protein composition, and SNP rs41255679 located at 107.2 Mbp is associated with β-lactoglobulin content and the casein index. This SNP is located in the promoter region of the LGB gene and is

known to be in linkage disequilibrium with variants A and B of LGB (Ganai et al., 2009). Bedere and Bovenhuis (2017) showed that the tail part of BTA 11 contains more than one mutation with an effect on β-lactoglobulin content.

BTA 13. Several SNPs on BTA 13 (e.g., lead SNP rs41658332) located from 50.6 to 50.8 Mbp were significantly associated with WN20, WN149, WN176, WN542, WN764, WN886 ($-Log_{10}(P)$ ranging from 4.2 to 5.8) and fat% ($-Log_{10}(P)$ up to 5.5). Figure 3.1 shows that these wavenumbers are correlated with each other and with fat%. The associations detected for this region are illustrated in Figure 3.2 for fat% and in Figure 3.3 for WN149. Bouwman et al. (2011) found a SNP located at 64.8 Mbp, which is in the gene acyl-CoA synthetase short-chain family member 2 (ACSS2). This is a candidate gene for C6:0, C8:0 and C10:0. Cole et al. (2011) reported several SNPs located between 50 and 60 Mbp are related to fat yield and fat%.

BTA 15. In the region around 51.9 Mbp, several SNPs were significantly associated with WN142, WN156, and protein% with $-Log_{10}(P)$ ranging from 5.5 to 7.7. SNP rs110249976 showed the most significant associations with these traits. This SNP was also reported by Schopen et al. (2011) to be significantly associated with protein%. The associations of this region were shown in Figure 3.2 for protein% and in Figure 3.3 for WN156.

BTA 28. The SNP rs29016491 at 14.9 Mbp was significantly associated with WN34, WN50, WN106, WN208, WN432, WN470 ($-Log_{10}(P)$ ranging from 4.7 to 9.2), and lactose% ($-Log_{10}(P)$ up to 7.3). Another SNP at 8.7 Mbp also showed significant associations with these wavenumbers. The associations of this region were shown in Figure 3.2 for lactose% and in Figure 3.3 for WN34. Most of these wavenumbers were associated with absorption bands of C–O and C–H bonds (Table 3.1), which are abundant in lactose.

BTA 29. The significant associations were detected at 45.4 Mbp for WN156 and protein%. SNP rs29026584 was the most significant SNP affecting WN156 ($-Log_{10}(P)=6.2$) and protein% ($-Log_{10}(P)=5.0$). This SNP was also reported by Schopen et al. (2011) to be significantly associated with protein%. Pryce et al. (2010) also found this region to be associated with protein%. WN156 is related to the band of N–H stretching vibration. Therefore this genomic region might contain information on milk protein content. The

associations for this region are illustrated in Figure 3.2 for protein% and in Figure 3.3 for WN156.

In the current study we presented GWAS results for milk IR wavenumbers. Most of the identified genomic regions have been previously associated with fat%, protein% or lactose%. Interestingly also a number of regions were identified which could not be related to these routinely recorded traits. For instance, genomic regions were identified which in other studies have been associated with phosphorus, orotic acid and citric acid. This suggests that there is genetic variation in these milk components and that the milk IR spectra contain information about these milk components. Phosphorus, orotic acid and citric acid are components with low concentration in milk. Detection of genomic regions based on IR wavenumbers for these low concentration components suggest that it might be of interest to calibrate milk IR spectra for other low concentration components. Large-scale recording of milk phosphorus content based on milk IR prediction equations can contribute to improved phosphorus efficiency of the dairy sector whereas there is currently no clear application for large scale recording of orotic or citric acid.

3.4 Conclusions

A GWAS was performed for 50 individual bovine milk IR wavenumbers, resulting in 24 significant genomic regions on 16 bovine chromosomes. Out of the 50 individual milk IR wavenumbers, 28 wavenumbers showed significant associations with at least one genomic region. Genomic regions, on chromosome 5, 6, 14, 19, and 20, showed significant associations with milk IR wavenumbers as well as milk composition traits, such as fat%, protein%, and lactose%. Remarkably, chromosome 1 contains 2 regions which are associated with milk IR wavenumbers but did not show an association with routinely recorded milk composition traits fat%, protein%, or lactose%. One of these regions has been shown to be associated with orotic acid and the other with phosphorus content of milk. This suggests that milk IR spectra contain information about phosphorus and orotic acid content of milk. The current study shows that GWAS based on milk IR wavenumbers revealed new genomic regions affecting milk composition that were not identified based on

routinely recorded milk composition traits. Furthermore, this study shows the GWAS provides further insight in the information that is captured by the milk IR spectra. This can lead to novel applications of milk IR spectroscopy for dairy cattle breeding and herd management.

3.5 Acknowledgements

The China Scholarship Council is acknowledged for funding the PhD project of Qiuyu Wang. Cooperative Cattle Improvement Organization (CRV) is acknowledged for the sampling cows providing and imputation of genotypes. Milk Control Station (Zutphen, The Netherlands) is acknowledged for infrared spectra data. This study is part of the Dutch Milk Genomics Initiative and the project “Melk op Maat”, funded by Wageningen University (Wageningen, the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), CRV, the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands).

3.6 References

Ashwell, M. S., D. W. Heyen, T. S. Sonstegard, C. P. van Tassell, Y. Da, P. M. VanRaden, M. Ron, J. I. Weller, and H. A. Lewin. 2004. Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. J. Dairy Sci. 87: 468-475.

Bedere, N and H. Bovenhuis. 2017. Characterizing a region on BTA11 affecting β-lactoglobulin content of milk using high-density genotyping and haplotype grouping. BMC genetics 18:17

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

Bittante, G., and A. Cecchinato. 2013. Genetic analysis of the Fourier-transform infrared spectra of bovine milk with emphasis on individual

wavenumbers related to specific chemical bonds. J. Dairy Sci. 96: 5991-6006.

Blott, S., J. J. Kim, S. Moisio, A. Schmidt-Kuntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, and D. Johnson. 2003. Molecular dissection of a quantitative trait locus: A phenylalanineto-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.

Bonfatti, V., L. Degano, A. Menegoz, and P. Carnier. 2016. Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. J. Dairy Sci. 99: 8216-8221.

Bouwman, A. C., H. Bovenhuis, M. H. Visker, and J. M. van Arendonk. 2011. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC genetics. 12: 43.

Bouwman, A. C, M. H. Visker, J. M. van Arendonk and H. Bovenhuis. 2012. Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. BMC genetics. 13: 93.

Bouwman, A. C., M. H. Visker, J. M. van Arendonk, and H. Bovenhuis. 2014. Fine mapping of a quantitative trait locus for bovine milk fat composition on Bos taurus autosome 19. J. Dairy Sci. 97: 1139-1149.

Bovenhuis, H., M. H. P. W. Visker, N. A. Poulsen, J. Sehested, H. J. F. van Valenberg, J. A. M. van Arendonk, L. B. Larsen, and A. J. Buitenhuis. 2016. Effects of the diacylglycerol o-acyltransferase 1 (DGAT1) K232A polymorphism on fatty acid, protein, and mineral composition of dairy cattle milk. J. Dairy Sci. 99: 3113-3123.

Buitenhuis, A. J., U. K. Sundekilde, N. A. Poulsen, H. C. Bertram, L. B. Larsen, and P. Sørensen. 2013. Estimation of genetic parameters and detection of quantitative trait loci for metabolites in Danish Holstein milk. J Dairy Sci. 96: 3285-3295.

Buitenhuis, B., L. L. Janss, N. A. Poulsen, L. B. Larsen, M. K. Larsen, P. Sørensen. 2014. Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. BMC genomics. 15:1112.

Buitenhuis, B., N. A. Poulsen, G. Gebreyesus, L. B. Larsen. 2016. Estimation of genetic parameters and detection of chromosomal regions affecting the major milk proteins and their post translational modifications in Danish Holstein and Danish Jersey cattle. BMC genetics. 17:114.

Buitenhuis, B., N. A. Poulsen, L. B. Larsen, and J. Sehested. 2015. Estimation of genetic parameters and detection of quantitative trait loci for minerals in Danish Holstein and Danish Jersey milk. BMC genetics. 16: 52.

Cecchinato, A., C. Ribeca, S. Chessa, C. Cipolat-Gotet, F. Maretto, J. Casellas, and G. Bittante. 2014. Candidate gene association analysis for milk yield, composition, urea nitrogen and somatic cell scores in Brown Swiss cows. Animal. 8: 1062-1070.

Chou, J. Y., H. S. Jun, and B. C. Mansfield. 2013. The SLC37 family of phosphate-linked sugar phosphate antiporters. Mol. Aspects Med. 34: 601-611.

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, A. Everts-van der Wind, J. H. Lee, J. K. Drackley, M. R. Band, A. G. Hernandez, M. Shani, and H. A. Lewin. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 15: 936-944.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, B. A. Crooker, C. P. van Tassell, J. Yang, S. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. BMC genomics. 12: 1.

Daetwyler, H. D., F. S. Schenkel, M. Sargolzaei, and J. A. B. 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.

Dehareng, F., C. Delfosse, E. Froidmont, H. Soyeurt, C. Martin, N. Gengler, A. Vanlierde, and P. Dardenne. 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Animal. 6: 1694-1701.

Diem, M. 2015. Modern Vibrational Spectroscopy and Micro-Spectroscopy: Theory, Instrumentation and Biomedical Applications. John Wiley & Sons, Hoboken, NJ.

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, and R. Thompson. 2009. ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Grelet, C., Bastin, C., Gelé, M., Davière, J.-B., Johan, M., Werner, A., Reding, R., Fernandez Pierna, J.A., Colinet, F.G., Dardenne, P., Gengler, N.,

Soyeurt, H., Dehareng, F. 2016 Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. J. Dairy Sci. 99: 4816-4825.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in Dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12: 222-231.

ICAR (International Committee for Animal Recording). 2012. International agreement of recording practices – Guidelines approved by the General Assembly held in Cork, Ireland on June 2012. ICAR, Rome. Italy.

Jonker, J. W., G. Merino, S. Musters, A. E. van Herwaarden, E. Bolscher, E. Wagenaar, E. Mesman, T. C. Dale, and A. H. Schinkel. 2005. The breast cancer resistance protein BCRP (ABCG2) concentrates drugs and carcinogenic xenotoxins into milk. Nat. Med. 11: 129.

Kadri, N. K., B. Guldbrandtsen, S. Mogens Lund, and G. Sahana. 2015. Genetic dissection of milk yield traits and mastitis resistance quantitative trait loci on chromosome 20 in dairy cattle. J. Dairy Sci. 98: 9015–9025.

Kemper, K. E., M. D. Littlejohn, T. Lopdell, B. J. Hayes, L. E. Bennett, R. P. Williams, X. Q. Xu, P. M. Visscher, M. J. Carrick, and M. E. Goddard. 2016. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. BMC genomics. 17: 858.

Kolbehdari, D., Z. Wang, J. R. Grant, B. Murdoch, A. Prasad, Z. Xiu, E. Marques, P. Stothard, S.S. Moore. 2009. A whole genome scan to map QTL for milk production traits and somatic cell score in Canadian Holstein bulls. J. Anim. Breed. Genet. 126: 216-227.

Littlejohn, M. D., K. Tiplady, T. A. Fink, K. Lehnert, T. Lopdell, T. Johnson, C. Couldrey, M. Keehan, R. G. Sherlock, C. Harland, A. Scott, R. G. Snell, S. R. Davis, and R. J. Spelman. 2016. Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. Sci. Rep. 6: 25376.

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.

McParland, S., G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J. Dairy Sci. 94: 3651–3661.

Medrano, J., G. Rincon, and A. Islas-Trejo. 2010. Comparative analysis of bovine milk and mammary gland transcriptome using RNASeq. Page 125 in Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany. German Society for Animal Science, Neustadt, Germany.

Morris, C.A., N. G. Cullen, B. C. Glass, D. L. Hyndman, T. R. Manley, S. M. Hickey, J. C. McEwan, W. S. Pitchford, C. D. Bottema, and M. A. Lee. 2007. Fatty acid synthase effects on bovine adipose fat and milk fat. Mamm. Genome, 18: 64-74.

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

Otero, J. A., V. Miguel, L. González-Lobato, R. García-Villalba, J. C. Espín, J. G. Prieto, G. Merino, A. I. Álvarez. 2016. Effect of bovine ABCG2 polymorphism Y581S SNP on secretion into milk of enterolactone, riboflavin and uric acid. Animal. 10: 238-247.

Pryce, J. E., S. Bolormaa, A. J. Chamberlain, P. J., Bowman, K. Savin, M. E. Goddard, and B. J. Hayes. 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J. Dairy Sci. 93: 3331-3345.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Real, R., L. González-Lobato, M. F. Baro, S. Valbuena, A. de la Fuente, J. G. Prieto, A. I. Álvarez, M. M. Marques, and G. Merino. 2011. Analysis of the effect of the bovine adenosine triphosphate-binding cassette transporter G2 single nucleotide polymorphism Y581S on transcellular transport of veterinary drugs using new cell culture models. J. Anim. Sci. 89: 4325-4338.

Robinson, J. L., D. B. Dombrowski, G. W. Harpestad, and R. D. Shanks. 1984. Detection and prevalence of UMP synthase deficiency among dairy cattle. J. Hered. 75: 277-280.

Roy, R., L. Ordovas, P. Zaragoza, A. Romero, C. Moreno, J. Altarriba, and C. Rodellar. 2006. Association of polymorphisms in the bovine FASN gene with milk-fat content. Anim. Genet. 37: 215-218.

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.

Sanchez, M. P., A. Govignon-Gion, M. Ferrand, M. Gelé, D. Pourchet, Y. Amigues, S. Fritz, M. Boussaha, A. Capitan, D. Rocha, G. Miranda, P. Martin, M. Brochard, D. Boichard. 2016. Whole-genome scan to detect quantitative trait loci associated with milk protein composition in 3 French dairy cattle breeds. J. Dairy Sci. 99: 8203-8215.

SAS Institute. 2011. SAS/STAT User's Guide: Release 9.3. SAS Inst., Cary, NC.

Schopen, G.C.B., M. H. P. W. Visker, P. D. Koks, E. Mullaart, J. A. M. van Arendonk, and H. Bovenhuis. 2011. Whole-genome association study for milk protein composition in dairy cattle. J. Dairy Sci. 94: 3148-3158.

Soyeurt, H., D. Bruwier, J. M. Romnee, N. Gengler, C. Bertozzi, D. Veselko, and P. Dardenne. 2009. Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. J. Dairy Sci. 92: 2444-2454.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89: 3690-3695.

Soyeurt, H., I. Misztal, and N. Gengler. 2010. Genetic variability of milk components based on mid-infrared spectral data. J. Dairy Sci. 93: 1722-1728.

Sun, D. W., ed. 2009. Page 3-27 in Infrared Spectroscopy for Food Quality Analysis and Control. Academic Press/Elsever, London, UK.

Toffanin, V., M. De Marchi, N. Lopez-Villalobos, and M. Cassandro. 2015. Effectiveness of mid-infrared spectroscopy for prediction of the contents of calcium and phosphorus, and titratable acidity of milk and their relationship with milk quality and coagulation properties. Int. Dairy J. 41: 68-73.

Wang, Q., A. Hulzebosch, and H. Bovenhuis. 2016. Genetic and environmental variation in bovine milk infrared spectra. J. Dairy Sci. 99:6793–6803.

Welper, R. D., and A. E. Freeman. 1992. Genetic Parameters for Yield Traits of Holsteins, Including Lactose and Somatic Cell Score1. J. Dairy Sci. 75: 1342-1348.

4

Validation strategy can result in overoptimistic view on the ability of milk infrared spectra to predict methane emission of dairy cattle

Qiuyu Wang and Henk Bovenhuis

Animal Breeding and Genomics Centre, Wageningen University & Research,

PO Box 338, 6700AH, Wageningen, the Netherlands

Accepted by Journal of Dairy Science

Abstract

Due to its environmental impact, it is of great interest to reduce methane (CH_4) emission of dairy cattle and selective breeding might contribute to this. However, this requires a rapid and cheap measurement technique which can be used to quantify CH_4 emission for a large number of individual dairy cows. Milk infrared (IR) spectroscopy has been proposed as a predictor for CH_4 emission. This study investigated the feasibility of milk IR spectra to predict breath sensor measured CH_4 of 801 dairy cows on 10 commercial farms. To evaluate the prediction equation we used random and block cross validation. Using random cross validation we found a R^2val of 0.49 which suggests that milk IR spectra are informative in predicting CH_4 emission. However based on block cross validation, with farms as blocks, a negligible R^2val of 0.01 was obtained, indicating that milk IR spectra cannot be used to predict CH_4 emission. Random cross validation thus results in an overoptimistic view on the ability of milk IR spectra to predict CH_4 emission of dairy cows. The difference between both validation strategies can be explained by the confounding effects of farm and date of milk IR analysis, which introduces a correlation between batch effects on the IR analyses and the farm average CH_4. Breath sensor measured CH_4 is strongly influenced by farm specific conditions which magnifies the problem. Based on random cross validation, also milk IR wavenumbers from water absorption regions showed moderate accuracy (R^2val=0.25) but not based on block cross validation (R^2val=0.03). These results therefore indicate that in the current study random cross validation results in an overoptimistic view on the ability of milk IR spectra to predict CH_4 emission. We suggest prediction based on wavenumbers from water absorption regions as a negative control to identify potential dependence structures in the data.

Key words:

validation strategy, CH_4 emission, milk infrared spectroscopy, prediction

4.1 Introduction

Methane (CH_4) is emitted by dairy cattle during anaerobic fermentation of feed in the rumen (Gerber et al., 2013). CH_4 is a potent greenhouse gas and in this way dairy production contributes to global warming. In addition, the emitted CH_4 accounts for a loss of 2 to 12% gross energy intake of the cow (Johnson and Johnson, 1995), which negatively affects feed efficiency. Therefore there is a need to reduce CH_4 emission from dairy cattle and selective breeding might contribute to this (Wall et al., 2010). This would require the measurements of CH_4 or CH_4 indicators on a large number of individual dairy cows.

Several methods have been suggested to quantify CH_4 emission from individual dairy cows (Hammond et al., 2016; Negussie et al., 2017). Some studies investigated the possibility of predicting CH_4 emission based on milk infrared (IR) spectroscopy (Dehareng et al., 2012; Vanlierde et al.,2015; Shetty et al., 2017; Van Gastelen et al., 2018; Vanlierde et al., 2018). Milk IR spectroscopy is a method to analyse milk composition and widely used to routinely quantify milk fat-, protein- and lactose content. The hypothesized relationship between CH_4 emission and milk composition is based on the observation that hydrogen and volatile fatty acids produced during microbial fermentation in the rumen are both involved in the synthesis pathways of CH_4 and de novo milk fatty acids (short and medium-chain fatty acids). Therefore milk IR spectra might be used to predict CH_4 and has the advantage as compared to alternative indicators for CH_4 emission that it can be easily applied on a large scale without substantial cost. Results from studies predicting CH_4 emission based on milk IR spectra are not consistent; Vanlierde et al. (2015) reported a R^2val of 0.77 for CH_4 production whereas Shetty et al. (2017) reported a R^2val of 0.13. These differences warrant further research into the potential to quantify CH_4 emission of individual cows using milk IR spectra.

A common strategy to evaluate the performance of prediction equations is random cross validation. The basic idea behind random cross validation is to randomly split the data in a part that is used for building the prediction equation (calibration set) and a part used for evaluating the performance of the prediction equation (validation set). When the validation set is independent

from the calibration set, cross validation should provide a good estimate of prediction error of the equation when applied to new data. However, recently a number of studies in different fields of science pointed at situations in which random cross validation resulted in overoptimistic estimation of predicative ability (Qin et al. 2016; Roberts et al. 2017). Problems with random cross validation might occur when there are dependence structures in the data and in these situations block cross validation, where data is not split randomly but strategically and accounting for the dependence structures, can address these problems. The issue of dependence structures might also be relevant for milk IR spectra prediction of traits relevant for dairy cows, e.g. when milk samples on farms are collected and analysed in batches. Therefore in this study we compared results from random cross validation with block cross validation using farms as blocks.

The objective of this study was to investigate the feasibility of milk IR spectra to predict breath sensor measured CH_4 emission of individual dairy cattle on commercial farms using different validation strategies.

4.2 Materials and methods

4.2.1 Methane

CH_4 data was collected for in total 1,508 lactating dairy cows located on 11 commercial dairy farms in the Netherlands between November 2013 and March 2016. The number of cows per farm ranged from 62 to 224 and more than 85% of the cows were of the Holstein Friesian breed.

CH_4 was measured in the breath of cows during milking in Automatic Milking Systems (Lely Astronaut A4, Lely Industries NV, Maassluis, the Netherlands) with non-dispersive infrared sensors (Acreo Swedish ICT, Kista, Sweden). This method is commonly referred to as the sniffer method. Sensors were located in the front gate of the Automatic Milking System, above the feeding trough at the same level as the cow's nose. In total 4 different sensors were used during data collection and the sensors were moved from one farm to the other. The sensors measured CH_4 concentrations (expressed in ppm) continuously twice per second and these measurements were averaged over the period a cow visited the Automatic Milking System. The average CH_4

concentrations were log_{10}-transformed to make the trait resemble a normal distribution and the transformed trait will be simply referred to as CH_4 in the remaining part of this article. More details about data collection and definition of the CH_4 phenotype is described by Van Engelen et al. (2018).

4.2.2 Milk infrared spectra

Milk samples were collected during routine milk production recording for all lactating cows on the 11 commercial dairy farms involved in the current study. Milk recording was on farms with Automatic Milking Systems and milk samples were collected during a cows' visit to the milking unit. Sample collection took place during routine milk recording as implemented by CRV (Cooperative cattle improvement organization, Arnhem, the Netherlands). Milk IR spectra were obtained by analysing the milk samples using FOSS FT6000 instruments by Qlip B.V. (Leusden, the Netherlands). Milk IR predicted fat%, protein%, and lactose% were based on the same milk samples and provided by CRV.

4.2.3 Data editing

The CH_4 and milk IR spectra data were combined to build the dataset for IR prediction of CH_4. All CH_4 measurements of a cow for a period ranging from 2 days before and 2 days after the date of milk recording (a 5-day period) were averaged to represent CH_4 emission of a cow. Average CH_4 is based on 10 to 12 CH_4 measurements. Using a repeatability of 0.27 (Van Engelen et al., 2018) it was calculated that CH_4 emission of a cow based on the average of 11 repeated measurements can be estimated with 90% accuracy. Cows with less than 5 CH_4 measurements during the 5-day period were eliminated. Due to this restriction several cows were eliminated from the data set, among others all records from 1 particular farm. The final dataset consisted of 801 cows from 10 commercial farms.

4.2.4 Methane prediction

The partial least squares regression (PLSR) (Wold et al., 1983) procedure in SAS 9.3 (SAS Institute, 2001) was used to develop CH_4 prediction equations. Milk IR spectra consist of 1,060 individual wavenumbers ranging from 925 to 5,008 cm^{-1}. The spectra were converted from transmittance to absorbance values. In total 275 individual wavenumbers from 3 spectral regions of milk IR spectra were used to develop prediction equations: 925 to 1,584 cm^{-1}, 1,719

to 1,784 cm^{-1} and 2,652 to 2,976 cm^{-1}. Wavenumbers in these regions are known to contain most of the information on milk composition (e.g., Capuano et al., 2014) and they will be referred to as “informative IR wavenumbers”.

Wavenumbers from the water absorption regions are generally considered to contain mainly noisy information and not considered informative for predicting milk composition or any other trait. Therefore we expect that prediction of CH_4 based on these wavenumbers will result in negligible prediction accuracy and we considered this prediction as a negative control. Based on results from Wang et al. (2016) we selected wavenumbers from the water absorption regions which contain negligible information on milk composition. In total 114 wavenumbers from 1,623 to 1,670 cm^{-1}, 3,166 to 3,254 cm^{-1}, 3,285 to 3,463 cm^{-1}, and 3,547 to 3,659 cm^{-1} were selected. These wavenumbers were selected based on the following criteria: they should not be significantly affected by systematic environmental effects (except for date of IR analysis), they should not be significantly affected by DGAT1 genotypes and more than 90% of the variation should be unexplained (residual variance).

The optimal number of latent variables for the PLS prediction equation was determined based on the lowest root mean squared error of a 10-fold cross validation procedure using CVTEST option in PROC PLS in SAS 9.3 (SAS Institute, 2001). The optimal number of latent variables was 15 for the 275 informative IR wavenumbers and 4 for the 114 wavenumbers from the water absorption regions and these values were used as input for building the CH_4 prediction equations.

Fat%, protein%, and lactose% are milk composition traits which are routinely collected and predicted based on milk IR measurements and might contain information on CH_4. As interest is in the additional information that might be available in the full milk IR spectra for predicting CH_4, we also predicted CH_4 based on these routinely collected milk composition traits.

4.2.5 Validation strategies

For the described CH_4 prediction we applied 2 different validation strategies: random cross validation and block cross validation with farms as blocks. The commonly used random cross validation refers to a strategy in which the samples are randomly assigned to a calibration or a validation set. This procedure is repeated and results are averaged. In the current study 640 cows

(approx. 80%) were randomly assigned to the calibration set and 161 cows to the validation set (approx. 20%) and the procedure was repeated 50 times. The averaged coefficient of determination was calculated based on the calibration samples (R^2cal) along with root mean squared error of calibration (RMSEC). The validation samples were used to calculate the validation coefficient of determination (R^2val) and the validation mean squared error of prediction (RMSEP).

In the block cross validation strategy samples were assigned to a calibration or a validation set based on the farm from which the samples were taken. In the current study the calibration set consisted of 9 farms and the validation set consisted of samples from the remaining farm. This procedure was repeated 10 times such that samples from each farm were validated based on the prediction equation that was calibrated based on data from the other 9 farms.

4.2.6 Repeated observations

Several studies that predicted CH_4 emission based on milk composition used data that included repeated observations on the same cow (e.g., Dehareng et al., 2012). Consequently, when using random cross validation different samples from the same cow might be present both in the calibration and the validation set. This might affect the evaluation of the prediction equation as it creates dependencies between calibration and validation sets. Therefore we also constructed CH_4 prediction equations based on cows with repeated observations. A second milk recording was available on 4 out of the 10 farms. In total 234 cows had repeated CH_4 and milk IR observations available and these 468 observations were used in this analysis. The repeated observations had the same data editing procedure as described above. The number of cows on each of the 4 farms ranged from 39 to 92. The time interval between 2 observations (2 routine milk recordings) of a cow was approximately 1 month. Observations that were collected first were assigned to the calibration set and the records that were collected later were assigned to the validation set.

4.3 Results

4.3.1 Descriptive statistics

The average log_{10}-transformed CH_4 based on all 801 records from the first sampling period was 2.250 with a standard deviation of 0.267. The number of records per farm ranged from 34 to 116. There were substantial differences between farms in average log_{10}-transformed CH_4 (ranging from 2.026 till 2.657) and the within farm standard deviation (ranging from 0.072 till 0.188) was substantially smaller than the standard deviation based on all records (0.267). The Pearson correlation coefficients between CH_4 and routinely recorded milk production traits were generally low: 0.03 for fat%, 0.07 for protein%, and -0.14 for lactose%. Pearson correlations between CH_4 and each of the 275 informative IR wavenumbers ranged from -0.17 to 0.25. Pearson correlations between CH_4 and each of the 114 wavenumbers from the water absorption regions ranged from -0.23 to 0.29.

4.3.2 Prediction of methane

Performance of the different CH_4 prediction equations evaluated using random cross validation and block cross validation are in Table 4.1. Based on random cross validation, prediction of CH_4 based on routinely recorded milk production traits fat%, protein% and lactose% was poor (R^2val=0.02). Prediction based on 275 informative IR wavenumbers and evaluation using random cross validation indicated a moderate prediction accuracy (R^2val=0.49). A surprisingly high prediction accuracy was found based on wavenumbers from the water absorption regions (R^2val=0.25).

Block cross validation gave a completely different view on the ability to predict CH_4 based on milk IR spectra. Prediction based on 275 informative IR wavenumbers showed an R^2val averaged over 10 replicates of 0.01 indicating that the predictive power of milk IR spectra for CH_4 is negligible. Results for each of the 10 replicates are shown in Table 4.2. In each replicate samples from one farm were validated based on the prediction equation that was calibrated based on samples from the other 9 farms. The R^2val ranged from 0.00 to 0.03 while R^2cal ranged from 0.53 to 0.67. These results show that the small R^2val is not due to specific conditions on one or a few farms.

Table 4.1 Evaluation of log_{10}-transformed CH_4 prediction using random cross validation and block cross validation[1]

	Random cross validation[2]				Block cross validation[3]			
Predicators	R^2cal	R^2val	RMSEC	RMSEP	R^2cal	R^2val	RMSEC	RMSEP
Fat%+Prot%+Lact%	0.02	0.02	0.265	0.266	0.02	0.04	0.263	0.266
IR informative	0.55	0.49	0.180	0.192	0.58	0.01	0.171	0.408
IR water absorption	0.32	0.25	0.221	0.233	0.33	0.03	0.217	0.309

[1] R^2cal = coefficient of determination for calibration; R^2val = coefficient of determination for validation; RMSEC = root mean squared error of calibration; RMSEP = root mean squared error of validation.

[2] Results are averaged from 50 replicates.

[3] Results are averaged from 10 replicates and in each replicate samples from 1 farm were validated by samples from the other 9 farms.

Table 4.2 Descriptive statistics and evaluation of log_{10}-transformed CH_4 prediction based on informative IR wavenumbers and block cross validation. Results are shown for each of the 10 replicates where records from one of the farms are used for validation[1]

Calibration					Validated	Validation				
No. Cow	Mean	SD	R^2cal	RMSEC	Farm	No. Cow	Mean	SD	R^2val	RMSEP
733	2.264	0.273	0.57	0.178	A	68	2.095	0.115	0.00	0.298
767	2.254	0.272	0.56	0.181	B	34	2.160	0.072	0.01	0.222
702	2.209	0.258	0.54	0.175	C	99	2.542	0.105	0.00	0.363
733	2.253	0.275	0.67	0.159	D	68	2.216	0.156	0.01	0.516
729	2.262	0.272	0.57	0.178	E	72	2.128	0.174	0.03	0.266
721	2.205	0.235	0.60	0.149	F	80	2.657	0.188	0.00	0.671
698	2.281	0.269	0.62	0.166	G	103	2.037	0.126	0.03	0.467
749	2.222	0.253	0.62	0.155	H	52	2.653	0.080	0.03	0.714
685	2.288	0.269	0.53	0.183	I	116	2.026	0.094	0.00	0.438
692	2.264	0.283	0.57	0.184	J	109	2.160	0.093	0.00	0.128

[1] Mean = averaged log_{10}-transformed CH_4; R^2cal = coefficient of determination for calibration; R^2val = coefficient of determination for validation; RMSEC = root mean squared error of calibration; RMSEP = root mean squared error of validation.

Block cross validation also resulted in low prediction accuracy based on wavenumbers from the water absorption regions (Table 4.1, R^2val=0.03). Validation coefficient of determination based on block cross validation for routinely recorded milk production traits fat%, protein% and lactose% was negligible (R^2val=0.04) and in line with results obtained using random cross validation.

4.3.3 Repeated observations

Results for validation of CH_4 prediction based on repeated observations of the same cow are shown in Table 4.3. The predictions were based on 275 informative IR wavenumbers and validation was based on 1 replicate. Results are presented for each farm separately as well as based on all 234 cows. For each prediction, we found a large discrepancy between the R^2val and corresponding R^2cal. For individual farms, the R^2val for CH_4 ranged from 0.10 to 0.21. Based on data from all 234 cows the R^2val for CH_4 was 0.07.

Table 4.3 Prediction of repeated observations on the farms with 2 measurements of log-transformed CH_4 and milk IR spectra[1, 2]

Farm	Number of cows	R^2cal	R^2val	RMSEC	RMSEP
A	58	0.77	0.10	0.057	0.526
C	45	0.72	0.10	0.047	0.277
H	39	0.78	0.21	0.037	0.850
J	92	0.63	0.19	0.066	0.094
Total	234	0.81	0.07	0.109	0.364

[1] R^2cal = coefficient of determination for calibration; R^2val = coefficient of determination for validation; RMSEC = root mean squared error of calibration; RMSEP = root mean squared error of validation.

[2] Results are based on 1 replicate. Prediction equations were developed on the first collected samples and validated on the later collected samples.

4.4 Discussion

In this study we investigated the feasibility of milk IR spectra for predicting breath sensor measured CH_4 emission of individual dairy cows on commercial farms. The CH_4 prediction was evaluated using different validation strategies. Results from random cross validation suggested a moderate prediction accuracy (R^2val=0.49) whereas block cross validation with farms as blocks indicated that milk IR spectra cannot be used to predict CH_4 emission (R^2val=0.01). Wavenumbers from water absorption regions are commonly assumed to contain no information on milk composition however these wavenumbers resulted in a R^2val of 0.25 for CH_4 based on random cross validation whereas a R^2val of 0.03 was obtained based on block cross validation. These results indicate that random cross validation leads to an overoptimistic view on the ability of milk IR spectra to predict CH_4 emission of individual dairy cows.

4.4.1 Validation strategy

The current study showed large differences in results between random cross validation and block cross validation. Random cross validation has been commonly applied in studies that use milk IR spectra to predict CH_4 emission (e.g., Dehareng et al., 2012; Shetty et al., 2017) or other traits of dairy cows (e.g., Soyeurt et al., 2006; Rutten et al., 2009). However, in the current study we show that random cross validation can lead to misleading conclusions. Qin et al. (2016) indicate that random cross validation underestimates the error rate of the prediction equation when predictors are analysed in batches and there are systematic differences between batches. In the current study CH_4 measurements were taken by installing breath sensors consecutively in different farms. Milk samples were collected on these farms during routine milk production recording. Consequently milk IR analyses of all milk samples collected from a farm were performed on the same day, and probably using the same spectrometer, whereas milk samples from different farms were analysed on different days, and possibly different spectrometers. As a consequence systematic differences between the date of milk IR analysis, which might include differences between spectrometers, are confounded with farm. Grelet et al. (2015) showed the importance of differences between spectrometers. Furthermore, Wang et al. (2016) showed significant effects for

date of milk IR analysis, and the time trend in milk IR wavenumbers indicates instability of the spectrometer in time. Confounding of farm with the date of milk IR analysis causes that errors associated with the date of milk IR analysis become correlated with average CH_4 emission of a farm. The problem is illustrated by prediction of CH_4 based on wavenumbers from the water absorption regions which resulted in a R^2val of 0.25. These wavenumbers show highly significant effects for the date of milk IR analysis (Wang et al., 2016) and contain negligible information on the actual milk composition. Due to this confounding wavenumbers from the water absorption regions explained part of the between herd variation in CH_4 emission. Correlations due to confounding of farm and date of milk IR analysis will not be transmitted to new data points and are spurious. These correlations are broken down in block cross validation with farms as blocks, however, they will remain undetected when using random cross validation and therefore random cross validation will lead to wrong conclusions. Similar discrepancies between validation strategies were reported in other fields of science (Gasch et al., 2015; Robert et al., 2017; Meyer et al., 2018). Burman et al. (1994) conducted simulations on dependent observations and showed that classical leave-one-out cross validation can be misleading. They also demonstrated that blocking adapts cross validation to dependency by allowing near independence between calibration and validation sets.

Based on largely the same data as used in the current study, Van Engelen et al. (2018) quantified that 56% of the total variation in CH_4 can be explained by the interaction of day of CH_4 measurement and automatic milking system which reflects mainly variation between farms. The sniffer method used in the current study to quantify CH_4, is known to be affected by specific farm conditions like airflow patterns and barn management (Wu et al., 2016). Differences in R^2val between both validation strategies are directly related to the magnitude of the between farm variation in CH_4 emission.

Prediction of CH_4 based on fat%, protein%, and lactose% suggests no information of breath CH_4 emission is captured by routinely recorded milk composition traits. Interestingly, there is no discrepancy between both validation strategies as both R^2val were negligible. This indicates batch effects on IR analyses do not affect IR predicted fat%, protein%, and lactose%; the

signal to “date of milk IR analysis”-noise ratio does not affect all IR wavenumbers to the same extend.

4.4.2 Negative control

There might be situations where confounding effects of batch analyses are not immediately obvious. To identify potential problems we therefore suggest prediction based on wavenumbers from the water absorption regions as a negative control. This assumes that wavenumbers from the water absorption regions contain negligible information on milk composition. Previously Wang et al. (2016) showed that some of the wavenumbers from the water absorption regions are significantly affected by the DGAT1 polymorphism and lactation stage suggesting that these wavenumbers do contain information regarding milk composition. When we included these wavenumbers in the prediction we obtained a random cross validation R^2val of 0.34, which is considerably higher than the value of 0.25 (Table 4.1). This underlines the importance of carefully selecting wavenumbers from the water absorption regions when used as a negative control.

A commonly used negative control is permutation of the data. We additionally performed 2 different permutation analyses. Firstly, we randomly assigned the milk IR spectra to cows in the dataset. Based on random cross validation this resulted in R^2val<0.01, both for prediction equations based on informative IR wavenumbers and wavenumbers from the water absorption regions. Secondly, the data was permutated by randomly assigning milk IR spectra to another cow within the same farm. Based on random cross validation this resulted in a R^2val of 0.44 for a prediction equation based on informative IR wavenumbers and a R^2val of 0.23 based on wavenumbers from the water absorption regions. When using block cross validation R^2val was negligible, both for prediction based on informative IR wavenumbers and for wavenumbers from the water absorption regions. The second permutation strategy did not affect herd CH_4 averages and therefore these results confirm that errors associated with date of milk IR analysis are correlated with average farm CH_4 emission; batch effects on the IR analyses explain differences in CH_4 between farms. The permutation analysis provides further insight in the problem but will not necessarily be able to identify the problem, i.e.

confounding. Including prediction based on wavenumbers from the water absorption regions is a way to detect potential problems.

4.4.3 Literature

Several studies investigated the feasibility of predicting CH_4 emission for individual cows based on milk IR spectra. The reported random cross validation R^2val ranged from 0.13 (Shetty et al., 2017) to 0.77 (Vanlierde et al., 2015). These large differences between studies can be partly explained by differences in methods used to quantify CH_4 emission: some studies used climate respiration chambers (Van Gastelen et al., 2018; Vanlierde et al., 2018) whereas others used sulfur hexafluoride (SF_6) tracer (Dehareng et al., 2012; Vanlierde et al., 2015) or the sniffer method (Shetty et al., 2017). Similar like the current study Shetty et al. (2017) used the sniffer method and reported a R^2val based on random cross validation of 0.13. This is considerably lower than the value of 0.49 we found using the same validation strategy. Therefore the spurious results obtained based on random cross validation in the current study seem less severe in the study by Shetty et al. (2017). This might be because Shetty et al. (2017) averaged the IR spectra over multiple milk samples. It is expected that averaging milk IR spectra, when analysed in different batches, will reduce date of IR analysis effects. In addition, differences in CH_4 emission between farms might be smaller in Shetty et al. (2017) as compared to the current study. Shetty et al. (2017) used data from 3 farms and also performed an additional calibration based on data from an experimental farm while validating based on commercial farm data, which is identical to the block cross validation as we propose. This block validation strategy showed negligible R^2val and it was concluded that variation in commercial farm data is not included in the experimental data (Shetty et al., 2017).

Most studies used random cross validation to evaluate the prediction of CH_4 emission based on milk IR spectra (Dehareng et al. 2012; Vanlierde et al., 2015; Vanlierde et al., 2018; Van Gastelen et al. 2018). Results from the current study indicate that random cross validation results overoptimistic view on the ability of milk IR spectra to predict CH_4 emission and therefore results presented in these studies might be too optimistic. However, this will only be the case under specific conditions. The description of the data in the

aforementioned studies do not allow drawing conclusions about whether, and if so to what extent, the results reported are affected by the validation strategy.

Several studies showed that milk IR spectra can be used to predict a wide variety of traits like milk fat composition (e.g., Soyeurt et al., 2006; Rutten et al., 2009), milk protein composition (Bonfatti et al., 2011; Rutten et al., 2011), milk coagulation, ketone bodies and energy status of dairy cows (McParland et al., 2011; De Marchi et al., 2014). All of these studies based their conclusion on random cross validation, and as shown in the current study, under certain conditions this might lead to overoptimistic conclusions. Studies by Rutten et al. (2009, 2011) involved almost 400 farms and milk samples from approximately 20 farms were analysed on the same day. Therefore, 20 farms are confounded with date of milk IR analysis. Random cross validation might result in overoptimistic R^2val in case the averaged milk fat or milk protein composition for these farms differs between dates of milk IR analysis. This is unlikely and was confirmed by analysis of the negative control (results not shown).

4.4.4 Predicting on repeated observations

The differences between random and block cross validations indicate that an overoptimistic result was obtained when both calibration and validation datasets contained samples from the same farms. In order to investigate if the presence of the same cows in both calibration and validation sets has an impact on the evaluation of the prediction equation, we performed additional analyses using the repeated observations. As shown in Table 4.3, the R^2val were between 0.10 and 0.21 for individual farms, and 0.07 when based on all cows with repeated observations. These values were rather low but higher than the block cross validation R^2val of 0.04. Although these results are based on a small number of observations and only one replicate, they suggest that having repeated observations on the same cow is a smaller problem than the issue related to confounding of farm and date of milk IR analysis. However when multiple observations per cow are available, the validation strategy should take this into account (Shetty et al., 2017; Vanlierde et al., 2018).

4.5 Conclusions

In this study, we investigated the feasibility of milk IR spectra to predict breath sensor measured CH_4 emission of dairy cows on commercial farms in the Netherlands using different validation strategies. We showed that random cross validation can result in an overoptimistic view on the ability of milk IR spectra to predict CH_4 emission. This is due to confounding of farm and date of milk IR analysis. Whether adjusting for date of milk IR analysis can avoid these issues requires further investigations. In order to identify dependence structures in the data we recommend prediction based on wavenumbers from the water absorption regions as a negative control. The negligible prediction accuracy of CH_4 emission based on block cross validation with farms as blocks indicates that milk IR spectra cannot be used to predict breath sensor measured CH_4 of dairy cows.

4.6 Acknowledgements

The authors acknowledge the China Scholarship Council for funding the PhD project of Qiuyu Wang, Qlip B.V. (Leusden, the Netherlands) for the milk IR spectra, CRV (Cooperative cattle improvement organization, Arnhem, the Netherlands) for providing data of dairy cows, and the 11 farmers for their collaboration. Data used in this project was obtained from the TI Food and Nutrition program 'Reduced methane emissions from dairy cows: towards sustainable dairy cattle production by increased understanding of genetic variation and rumen functioning'.

4.7 References

Bonfatti, V., G. Di Martino, and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. J. Dairy Sci. 94: 5776-5785.

Burman, P., E. Chow, and D. Nolan. 1994. A cross-validatory method for dependent data. Biometrika, 81: 351-358.

Capuano, E., J. Rademaker, H. van den Bijgaart, and S. M. van Ruth. 2014. Verification of fresh grass feeding, pasture grazing and organic farming by FTIR spectroscopy analyses of bovine milk. Food Res. Int. 60: 59-65.

Dehareng, F., C. Delfosse, E. Froidmont, H. Soyeurt, C. Martin, N. Gengler, A. Vanlierde, and P. Dardenne. 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Animal. 6: 1694-1701.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97: 1171-1186.

Gasch, C. K., T. Hengl, B. Gräler, H. Meyer, T. S. Magney, and D. J. Brown. 2015. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D+ T: The Cook Agronomy Farm data set. Spat. Stat. 14: 70-90.

Gerber, P. J., H. Steinfeld, B. Henderson, A. Mottet, C. Opio, J. Dijkman, A. Falcucci, and G. Tempio. 2013. Trackling Climate Change Through Livestock – A Global Assessment of Emissions and Mitilgation Opportunities. Food and Agriculture Organization of the United Nations, Rome, Italy.

Grelet, C., J. F. Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. J. Dairy Sci. 98: 2150-2160.

Hammond, K. J., L. A. Crompton, A. Bannink, J. Dijkstra, D. R. Yáñez-Ruiz, P. O'Kiely, E. Kebreab, M. A. Eugène, Z.Yu, K.J.Shingfield, A. Schwarm, A. N. Hristov, C. K. Reynolds. 2016. Review of current in vivo measurement techniques for quantifying enteric methane emission from ruminants. Anim. Feed Sci. Technol. 219: 13-30.

Johnson, K. A., and D. E. Johnson. 1995. Methane emissions from cattle. J. Anim. Sci. 73: 2483-2492.

McParland, S., G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J. Dairy Sci. 94: 3651–3661.

Meyer, H., C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss. 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Modell Softw. 101: 1-9.

Negussie, E., Y. de Haas, F. Dehareng, R. J. Dewhurst, J. Dijkstra, N. Gengler, D. P. Morgavi, H. Soyeurt, S. van Gastelen, T. Yan and F. Biscarini. 2017. Invited review: Large-scale indirect measurements for enteric methane

emissions in dairy cattle: A review of proxies and their potential for use in management and breeding decisions. J. Dairy Sci. 100: 2433-2453.

Qin, L. X., H. C. Huang, and C. B. Begg. 2016. Cautionary note on using cross-validation for molecular classification. J. Clin. Oncol. 34: 3931-3938.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40: 913-929.

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.

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. J. Dairy Sci. 94: 5683-5690.

SAS Institute. 2011. SAS/STAT User's Guide: Release 9.3. SAS Inst., Cary, NC.

Shetty, N., G. Difford, J. Lassen, P. Løvendahl, and A. J. Buitenhuis. 2017. Predicting methane emissions of lactating Danish Holstein cows using Fourier transform mid-infrared spectroscopy of milk. J. Dairy Sci. 100: 9052-9060.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89: 3690-3695.

Van Engelen, S., H. Bovenhuis, P. P. J. van der Tol, and M. H. P. W. Visker. 2018. Genetic background of methane emission by Dutch Holstein Friesian cows measured with infrared sensors in automatic milking systems. J. Dairy Sci. 101: 2226-2234.

Van Gastelen, S., H. Mollenhorst, E. C. Antunes-Fernandes, K. A. Hettinga, G. G. van Burgsteden, J. Dijkstra, and J. L. W. Rademaker. 2018. Predicting enteric methane emission of dairy cows with milk Fourier-transform infrared spectra and gas chromatography–based milk fatty acid profiles. J. Dairy Sci. 101: 5582-5598.

Vanlierde, A., H. Soyeurt, N. Gengler, F. G. Colinet, E. Froidmont, M. Kreuzer, F. Grandl, M. Bell, P. Lund, D. W. Olijhoek, M. Eugene, C. Martin, B. Kuhla, and F. Dehareng. 2018. Development of an equation for estimating methane emissions of dairy cows from milk Fourier transform mid-infrared

spectra by using reference data obtained exclusively from respiration chambers. J. Dairy Sci. 101: 7618-7624.

Vanlierde, A., M. L. Vanrobays, F. Dehareng, E. Froidmont, H. Soyeurt, S. McParland, E. Lewis, M. H. Deighton, F. Grandl, M. Kreuzer, B. Gredler, P. Dardenne, and N. Gengler. 2015. Hot topic: Innovative lactation-stage-dependent prediction of methane emissions from milk mid-infrared spectra. J. Dairy Sci. 98: 5740-5747.

Wall, E., G. Simm, and D. Moran. 2010. Developing breeding schemes to assist mitigation of greenhouse gas emissions. Animal. 4: 366-376.

Wang, Q., A. Hulzebosch, and H. Bovenhuis. 2016. Genetic and environmental variation in bovine milk infrared spectra. J Dairy Sci. 99: 6793-6803.

Wold, S., H. Martens, and H. Wold. 1983. The multivariate calibration problem in chemistry solved by the PLS method. Page 286-293 in Proc. Conf. on Matrix Pencils, Lecture Notes in Mathematics. A. Ruhe and B. Kagstrom, ed. Springer, Heidelberg, Germany.

Wu, L., P. W. G. Koerkamp, and N. W. Ogink. 2016. Temporal and spatial variation of methane concentrations around lying cubicles in dairy barns. Biosyst. Eng. 151: 464-478.

5

Combined use of milk infrared spectra and genotypes improves prediction of milk fat composition

Qiuyu Wang and Henk Bovenhuis

Animal Breeding and Genomics Centre, Wageningen University & Research,

PO Box 338, 6700AH, Wageningen, the Netherlands

Submitted to Journal of Dairy Science

Abstract

It has been shown that milk infrared (IR) spectroscopy can be used to predict detailed milk fat composition. In addition polymorphisms have been identified with substantial effects on milk fat composition. In this study we investigated the combined use of milk IR spectroscopy and genotypes of dairy cows on the accuracy of predicting milk fat composition. Milk fat composition based on gas chromatography and milk IR spectra were available for 1,456 Dutch Holstein Friesian cows. In addition genotypes for the diacylglycerol acyltransferase1 (DGAT1) K232A, the stearoyl-CoA desaturase1 (SCD1) A293V polymorphisms and a Single Nucleotide Polymorphism (SNP) located in an intron of the fatty acid synthase (FASN) gene were available. Adding SCD1 genotypes to the milk IR spectra resulted in a considerable improvement of the prediction accuracy for the unsaturated fatty acids C10:1, C12:1, C14:1 cis-9, C16:1 cis-9 and their corresponding unsaturation indices. Adding DGAT1 genotypes to the milk IR spectra resulted in an improvement of the prediction accuracy for C16:1 cis-9 and C16 index. Adding genotypes of the FASN SNP to the IR spectra did not improve prediction of milk fat composition. This study demonstrated the potential of combining milk IR spectra with genotypic information from 3 polymorphisms to predict milk fat composition. Prediction accuracy can be further improved by combining milk IR spectra with genomic breeding values.

Key words:

milk fat composition, milk infrared spectroscopy, genotypes, prediction

5.1 Introduction

Bovine milk composition contains many valuable components such as fatty acids, caseins, β-lactoglobulin, α-lactalbumin, lactose, and minerals. Milk fat consists of fatty acids that differ in length and degree of saturation. Saturated fatty acids (SFA) make up approximately 70% of total fat in milk, and some of them have been related to increased risks of cardiovascular disease. On the other hand, unsaturated fatty acids (UFA) such as oleic acid (C18:1 cis-9), conjugated linoleic acid (CLA) and other polyunsaturated fatty acids have been suggested to have beneficial effects on human health (Haug et al., 2007). Moreover, milk fat composition has been associated with energy status and health of dairy cows (Vlaeminck et al., 2006a,b; Van Haelst et al., 2008), and therefore milk fat composition might contain valuable information for dairy farm management (Hamann and Krömker, 1997).

Detailed milk fat composition is currently not quantified during routine milk recording on commercial dairy farms. This is mainly due to the expensive and time consuming quantification of fatty acids using gas chromatography (GC). Milk infrared (IR) spectroscopy is the routine method for quantifying milk fat-, protein- and lactose content (ICAR, 2012). Milk IR spectroscopy has also been suggested as a method to estimate milk fat composition (e.g., Soyeurt et al., 2006; Rutten et al., 2009; Ferrand et al., 2011) The results of these studies show that especially milk fatty acids in low concentration like UFA cannot be predicted accurately based on IR analyses.

Polymorphisms in diacylglycerol O-acyltransferase 1 (DGAT1) and stearoyl-CoA desaturase 1 (SCD1) genes have been shown to have highly significant effects on milk fat composition (e.g., Schennink et al., 2007, 2008; Bovenhuis et al. 2016). In addition, the fatty acid synthase (FASN) gene has been proposed as a candidate gene responsible for fatty acids in milk (Morris et al., 2007). Based on a genome wide association study, Bouwman et al. (2014) identified a SNP located in an intron of the FASN gene with a highly significant effect on C14:0.

We hypothesize that the genotypic information explains part of the variation in milk fat composition that is not captured by the milk IR spectra and therefore combining milk IR spectroscopy and genotypic information can improve prediction of milk fat composition. The aim of this study is to determine the

accuracy of predicting milk fat composition based on a combination of milk IR spectra and genotypic information.

5.2 Materials and methods

5.2.1 Milk IR spectra data

One morning milk sample of first parity Holstein Friesians cows which were between 63 and 263 days in lactation was collected for analyses. The data was collected from February till March of 2005. All cows in this study have at least 87.5% Holstein Friesian genes. The population consisted of 5 large paternal half-sib families from proven sires (98-196 daughters per sire), and 50 small paternal half-sib families from test sires (8-23 daughters per sire), as well as 168 cows descending from 44 other proven sires (1-25 daughters per sire).

Milk samples were conserved using sodium azide (0.03% wt/wt) and stored at 4°C. Milk IR spectra were recorded in a 10 mL milk sample using the MilkoScan FT 6000 equipment (FOSS, Denmark) at the certified laboratory of the Milk Control Station (Zutphen, The Netherlands). All milk samples used in this study were analysed using the same MilkoScan FT 6000. The milk IR spectra consisted of measurements at 1,060 wavenumbers ranging from 925 to 5,008 cm^{-1}.

5.2.2 Fat composition data

Milk fatty acids were measured using gas chromatography at the COKZ laboratory (Netherlands Controlling Authority for Milk and Milk Products, Leusden, the Netherlands). Details on the method used to quantify milk fatty acids can be found in Schennink et al. (2007). Individual fatty acids were expressed on a fat basis (g/100 g of fat) rather than on a milk basis (g/100 g of milk). Although studies showed better prediction can be achieved when fatty acids are expressed on a milk basis (e.g., Soyeurt et al., 2006; Rutten et al., 2009), main interest of the dairy industry is to detect changes in fatty acids independent of milk fat content.

In this study we predicted individual fatty acids as well as groups of fatty acids and unsaturation indices. Groups of fatty acids were C6-12 which is the sum

of C6:0, C8:0, C10:0 and C12:0, C14-16 which is the sum of C14:0 and C16:0, C18u which is the sum of C18:1 trans-4-8, C18:1 trans-9, C18:1 trans-11, C18:1 cis-9, C18:1 cis-11, C18:2 cis-9,12 and C18:3 cis-9,12,15, C18 trans which is the sum of sum of C18:1 trans-4-8, C18:1 trans-9, C18:1 trans-11. In addition total SFA (C4:0, C5:0, C6:0, C7:0, C8:0, C9:0, C10:0, C11:0, C12:0, C13:0, C14:0, C15:0, C16:0, C17:0, C18:0), total UFA (C10:1, C12:1, C14:1, C16:1, C18u, CLA). The ratio between SFA and UFA were defined and included in the category “groups of fatty acids”. Milk fat unsaturation indices were calculated by expressing an unsaturated fatty acid as the proportion of the unsaturated fatty acid plus its saturated substrate, and multiplied by 100, e.g. C14 index = 100*C14:1 cis9/(C14:1 cis9 + C14:0). Unsaturation indices were calculated for the following product and substrate pairs: C10:1 and C10:0 (C10 index), C12:1 and C12:0 (C12 index), C14:1 cis-9 and C14:0 (C14 index), C16:1 cis-9 and C16:0 (C16 index), C18:1 cis-9 and C18:0 (C18 index), CLA and C18:1 trans-11 (CLA index). Records from 17 dairy cows were eliminated because they had at least one milk fatty acid observation that deviated more than five standard deviations from the mean.

5.2.3 Genotypes

Previous genome-wide association studies identified 3 genomic regions with major effects on milk fat composition, located on chromosomes 14, 19 and 26 (Bouwman et al., 2011). On chromosome 14 the DGAT1 K232A polymorphism and on chromosome 26 the SCD1 A239V polymorphism are responsible for the observed effects. Genotypes of these 2 polymorphisms were obtained as described by Schennink et al. (2008). The causal mutation for the association detected on chromosome 19 has not been identified. The Coiled-coil domain-containing protein 57 (CCDC57) and the FASN gene have been suggested as possible candidate genes (Bouwman et al. 2014). Genotypes from the lead SNP on BTA19, rs137372738 (or BovineHD1900014372), located in an intron of the FASN gene, were included in the current study (Bouwman et al. 2014). In the remaining part of this article we will refer to this SNP as “FASN” .

In the study population frequencies for the DGAT1 A allele was 59.6% and 40.4% for the K allele; frequencies for the SCD1 A allele was 72.9% and 27.1% for the V allele; and frequencies for the FASN A allele was 34.0% and 66.0% for the G allele. Records from dairy cows with no genotypic information were

removed. After combining milk IR spectra, fatty acid composition and genotypes the final dataset consisted of 1,456 dairy cows from 355 farms.

5.2.4 Fatty acid prediction

In the current study 1,200 cows were randomly assigned to the calibration set and 256 cows to the validation set, and the procedure was repeated 50 times. The averaged coefficient of determination was calculated based on the validation samples (R^2val). The partial least squares regression (PLSR) (Wold et al., 1983) procedure in SAS 9.3 (SAS Institute, 2001) was used to develop prediction equations. The milk IR spectra were converted from transmittance to absorbance values. In total 275 individual wavenumbers from 3 spectral regions of milk IR spectra were used to develop prediction equations: 925 to 1,584 cm^{-1}, 1,719 to 1,784 cm^{-1} and 2,652 to 2,976 cm^{-1}. Wavenumbers in these regions are associated with vibrations of chemical bonds e.g. triacylglycerol ester linkage C–O symmetric stretching, C=O stretching, and acyl chain C–H symmetric and asymmetric stretching, and these vibrations are informative for quantifying milk fat (Dufour et al., 2009). These wavenumbers will be referred to as “informative IR wavenumbers”.

At first fatty acids were predicted based on genotypes for each of the 3 polymorphisms separately. The 3 polymorphisms were treated as class variables in the predictions. Secondly milk fat composition was predicted based on informative IR wavenumbers. Subsequently the informative IR wavenumbers were combined with information from the 3 described polymorphisms: 1. Prediction based on informative IR wavenumbers and DGAT1 genotypes (referred to as IR+DGAT1); 2. Prediction based on informative IR wavenumbers and SCD1 genotypes (referred to as IR+SCD1); 3. Prediction based on informative IR wavenumbers and FASN genotypes (referred to as IR+FASN); 4. Prediction based on informative IR wavenumbers and all these 3 polymorphisms (referred to as IR+3 polym.).

We used wavenumbers from the water absorption regions as a negative control for predicting milk fat composition (Wang and Bovenhuis, submitted). Wavenumbers from the water absorption regions are expected to contain negligible information on milk fat composition and therefore we expect prediction accuracies based on these wavenumbers to be close to zero. Prediction accuracy different from zero is an indication for problems with the

validation strategy e.g. due to dependencies in the data (Wang and Bovenhuis, submitted). We selected those wavenumbers from the water absorption region which contain negligible information on milk composition; more than 95% of the variation in the selected wavenumbers was unexplained (residual variance) and the selected wavenumbers were not significantly affected by systematic environmental effects (except for date of IR analysis) or DGAT1 genotypes. In total 114 wavenumbers from 1,623 to 1,669 cm^{-1}, 3,164 to 3,253 cm^{-1}, 3,284 to 3,462 cm^{-1}, and 3,547 to 3,658 cm^{-1} were selected (Wang and Bovenhuis, submitted).

5.3 Results and discussion

5.3.1 Descriptive statistics

Descriptive statistics of the gas chromatography measured milk fatty acids, groups of fatty acids and unsaturation indices are shown in Table 5.1. The concentrations of fatty acids are similar to Rutten et al. (2009) who used largely the same dairy cows. Individual fatty acids C16:0, C18:1 cis-9, and C14:0 contributed most to the total fat content, and they are the only fatty acids which account for more than 10% of the total fat. The short to medium chain UFA (i.e. C10:1, C12:1, C14:1 cis-9, C16:1 cis-9) and conjugated linoleic acid (CLA) were in low concentrations. C16:1 cis-9 was the most abundant of them and accounted for less than 1.5% of total milk fat. For the individual fatty acids, coefficients of variation ranged from 6.9% for C6:0 to 26.9% for CLA. The fatty acids in this study showed comparable means but smaller coefficients of variation as compared to Soyeurt et al. (2011) and Fleming et al. (2017). This might be due to the inclusion of multiple dairy cattle breeds in the studies by Soyeurt et al. (2011) and Fleming et al. (2017).

The SFA accounted for approximately 69% of total milk fat and UFA for approximately 25%. The most important UFA, C18:1 cis-9, contributed 18% to the total fat content. The unsaturation indices were relatively low (2.7 to 10.9%) for short and medium chain fatty acids (C10 to C16 index) and much higher for the C18 index (67%). C18:1 cis-9 derived from the diet contributes to the relatively high C18 index, whereas short to medium UFA can only be synthesized in the mammary gland by SCD1. The coefficients of variation were large for C10 to C16 indices and small for C18 index.

Table 5.1 Descriptive statistics of GC-measured individual fatty acids content, groups of fatty acids and unsaturation indices on a fat basis (g/100 g of fat, n=1,456).

Trait	Mean	SD[1]	CV(%)[2]	Min.	Max.
Individual fatty acids					
C4:0	3.52	0.27	7.6	2.70	4.57
C6:0	2.23	0.15	6.9	1.72	2.76
C8:0	1.37	0.13	9.5	0.86	1.85
C10:0	3.03	0.41	13.6	1.51	4.62
C10:1	0.37	0.06	16.5	0.12	0.59
C12:0	4.12	0.69	16.7	1.97	6.85
C12:1	0.12	0.03	23.7	0.04	0.22
C14:0	11.64	0.88	7.6	8.61	14.94
C14:1 cis-9	1.35	0.26	19.1	0.47	2.40
C16:0	32.70	2.71	8.3	22.86	42.95
C16:1 cis-9	1.45	0.31	21.6	0.74	2.92
C18:0	8.72	1.35	15.4	3.91	14.06
C18:1 cis-9	17.99	1.98	11.0	12.34	25.14
C18:2 cis-9,12	1.18	0.26	22.1	0.56	2.54
C18:3 cis-9,12,15	0.41	0.11	25.5	0.15	0.90
CLA	0.39	0.11	26.9	0.19	0.92
Groups of fatty acids					
C6-12	10.76	1.18	11.0	6.61	15.50
C14-16	44.34	2.64	5.9	33.86	53.54
C18 u	21.11	2.22	10.5	15.31	29.54
C18 trans	1.12	0.23	20.2	0.55	2.12
SFA	69.25	2.56	3.7	60.41	76.93
UFA	24.83	2.30	9.3	18.61	32.93
Ratio SFA:UFA	2.82	0.36	12.6	1.85	4.13
Unsaturation indices					
C10 index	10.93	1.78	16.3	3.45	17.92
C12 index	2.74	0.51	18.8	0.75	4.65
C14 index	10.38	1.85	17.9	3.86	17.13
C16 index	4.23	0.79	18.6	2.16	7.48
C18 index	67.40	3.66	5.4	52.00	79.93
CLA index	33.75	3.71	11.0	19.53	50.74
Total index	26.22	2.57	9.8	18.05	35.07

[1] SD = Standard deviation

[2] CV = Coefficient of variation (in %)

5.3.2 Negative control

Table 5.2 shows the R^2val, averaged over 50 replicates for predicting milk fatty acids based on different sources of information. In this study 114 wavenumbers from the water absorption regions were used to predict milk fat composition. The results in Table 5.2 show that prediction based on these wavenumbers showed a negligible R^2val of 0.01 or lower. Wang and Bovenhuis. (submitted) reported that batch effects of milk IR analyses can cause overoptimistic results when using random cross validation, which can be detected by a negative control. Results from the negative control indicate that batch effects do not affect R^2val in the current study.

5.3.3 Milk IR spectra

Prediction based on IR showed R^2val ranging from 0.19 (C18:3 cis-9,12,15) to 0.72 (C6:0) for individual fatty acids, from 0.33 (C18 trans) to 0.78 (SFA) for groups of fatty acids, and from 0.22 (C14 index) to 0.75 (total index) for unsaturation indices. The current study shows higher prediction accuracies as those presented by Rutten et al. (2009), who based their results on largely the same winter milk samples (WW scenario) as the current study. Rutten et al. (2009) calibrated based on 909 samples and validated on the remaining 909 samples. In the current study we calibrated on 1,200 samples and validated on 256 samples and using a larger calibration set increases the prediction accuracy. The current study showed lower prediction accuracies as compared to Rutten et al. (2009) when winter and summer milk samples were combined in a larger calibration set. The current study also showed in general higher prediction accuracy as reported by Soyeurt et al. (2006), which can be explained by the smaller dataset (49 samples). We found higher prediction accuracies for short chain fatty acids but lower prediction accuracies for long chain fatty acids as compared to Fleming et al. (2017) which used a dataset of 2,023 samples.

To our knowledge prediction of milk unsaturation indices based on milk IR spectroscopy have not been reported before. The prediction accuracy was moderate for most of the indices: R^2val ranging from 0.22 for C14 index to 0.57 for C18 index. A remarkably high prediction accuracy was obtained for the total unsaturation index (R^2val of 0.75). The total unsaturation index is strongly determined by C16:0 and C18:1 cis-9, which are the 2 most abundant

Table 5.2 Validation coefficients of determination (R^2val) for predicting milk fat composition based on wavenumbers from water absorptions as a negative control (NC), genotypes, milk IR spectra and using combinations of milk IR spectra.[1]

Trait	NC	DGAT1	SCD1	FASN	IR	IR +DGAT1	IR +SCD1	IR +FASN	IR +3 polym.
Individual fatty acids									
C4:0	0.00	0.01	0.01	0.02	0.67	0.66	0.66	0.67	0.66
C6:0	0.00	0.03	0.01	0.02	0.72	0.73	0.72	0.72	0.71
C8:0	0.00	0.02	0.01	0.06	0.71	0.71	0.70	0.72	0.71
C10:0	0.00	0.01	0.02	0.08	0.71	0.72	0.72	0.71	0.73
C10:1	0.00	0.01	0.15	0.01	0.36	0.37	0.55	0.36	0.55
C12:0	0.00	0.01	0.01	0.04	0.56	0.56	0.56	0.56	0.55
C12:1	0.01	0.01	0.09	0.01	0.37	0.35	0.47	0.36	0.47
C14:0	0.00	0.10	0.02	0.12	0.58	0.63	0.59	0.59	0.65
C14:1 cis-9	0.01	0.01	0.24	0.00	0.23	0.23	0.47	0.22	0.48
C16:0	0.01	0.12	0.01	0.01	0.53	0.55	0.53	0.54	0.57
C16:1 cis-9	0.00	0.13	0.09	0.01	0.27	0.35	0.36	0.29	0.46
C18:0	0.00	0.01	0.01	0.00	0.41	0.39	0.39	0.41	0.40
C18:1 cis-9	0.00	0.13	0.01	0.03	0.68	0.70	0.69	0.69	0.69
C18:2 cis-9,12	0.00	0.04	0.00	0.00	0.28	0.27	0.27	0.27	0.28
C18:3 cis-9,12,15	0.00	0.03	0.00	0.01	0.19	0.20	0.18	0.17	0.18
CLA	0.00	0.06	0.01	0.01	0.40	0.40	0.41	0.40	0.40
Groups of fatty acids									
C6-12	0.01	0.01	0.01	0.07	0.77	0.77	0.77	0.77	0.77
C14-16	0.01	0.07	0.00	0.00	0.55	0.56	0.56	0.56	0.56
C18 u	0.00	0.14	0.01	0.02	0.70	0.71	0.70	0.70	0.71
C18 trans	0.00	0.02	0.00	0.01	0.33	0.32	0.33	0.34	0.32
SFA	0.01	0.10	0.00	0.02	0.78	0.78	0.78	0.78	0.78
UFA	0.00	0.10	0.01	0.02	0.76	0.76	0.76	0.76	0.75
Ratio SFA:UFA	0.00	0.10	0.00	0.02	0.77	0.78	0.77	0.77	0.77
Unsaturation indices									
C10 index	0.01	0.01	0.21	0.03	0.29	0.31	0.51	0.27	0.48
C12 index	0.00	0.01	0.18	0.01	0.29	0.31	0.48	0.28	0.49
C14 index	0.01	0.04	0.28	0.02	0.22	0.26	0.52	0.23	0.54
C16 index	0.00	0.06	0.13	0.01	0.34	0.41	0.46	0.35	0.54
C18 index	0.00	0.08	0.02	0.01	0.57	0.57	0.58	0.57	0.56
CLA index	0.00	0.05	0.03	0.01	0.52	0.50	0.53	0.51	0.53
Total index	0.00	0.09	0.00	0.03	0.75	0.76	0.75	0.75	0.74

[1] NC = negative control, i.e. prediction using wavenumbers from the water absorption regions; DGAT1 = prediction using DGAT1 genotypes; SCD1 = prediction using SCD1 genotypes; FASN = prediction using genotypes of SNP rs137372738 which is located in an intron of FASN.

individual SFA and UFA in milk respectively. However, prediction accuracies for C16:0 (0.53) and C18:1 cis-9 (0.68) were lower than for the total unsaturation index.

5.3.4 Milk IR spectra and DGAT1 genotypes

Table 5.2 shows prediction accuracy for fat composition based on DGAT1 genotypes only. The results show that DGAT1 genotypes explain a considerable part of the variation in several milk fat acids. DGAT1 genotypes explained more than 10% of the variation in C14:0, C16:0, C16:1 cis-9, and C18:1 cis-9, and 9% of the variation in the total unsaturation index. When DGAT1 genotypes were combined with informative IR wavenumbers (IR+DGAT1), prediction accuracies showed small to moderate improvements for C14:0 (+0.05), C16:0 (+0.02) and C18:1 cis-9 (+0.02). A stronger improvement in prediction accuracy was observed for C16:1 cis-9 (+0.08). Small to moderate improvements in prediction accuracies were also observed for the C10 index (+0.02), C12 index (+0.02), C14 index (+0.04), and the C16 index (+0.07).

The DGAT1 gene codes the enzyme that catalyzes the last step of triacylglycerol synthesis and has major effect on milk fat percentage and fat composition (e.g., Schennink et al., 2007; Bovenhuis et al., 2016). Results from the current study show that for some of the fatty acids over 10% of the phenotypic variation can be explained by the DGAT1 polymorphism. The results from the current study show that variation in milk fat composition explained by the IR spectra and by DGAT1 genotypes are not simply additive but the information partly overlaps. Prediction of C14:0 based on DGAT1 has a prediction accuracy of 0.10 and based on IR spectra of 0.58. Combining both information sources does not result in a prediction accuracy of 0.68 but 0.63, indicating that the information captured by both sources are not independent. This is in agreement with Wang et al. (2016) who showed that DGAT1 genotypes have highly significant effects on many individual wavenumbers. Therefore these results indicate that the information on milk fat composition of DGAT1 genotypes has to a large extend been captured by the milk IR spectra. Improvement in prediction by combining both information sources is therefore for some traits absent (e.g., SFA, UFA, C18 index). The largest improvements in prediction accuracies were observed for C16:1 cis-9

(+0.08) and C16 index (+0.07).

5.3.5 Milk IR spectra and SCD1 genotypes

Results in Table 5.2 show that SCD1 genotypes explained a substantial part of the variation in the individual fatty acids C10:1 (15%), C12:1 (9%), C14:1 cis-9 (24%), C16:1 cis-9 (9%) and their unsaturation indices. As compared to prediction based on IR only, IR+SCD1 shows considerable improvements in prediction accuracies for C10:1 (+0.19), C12:1 (+0.10), C14:1 cis-9 (+0.24), C16:1 cis-9 (+0.09), and their corresponding unsaturation indices: C10 index (+0.22), C12 index (+0.19), C14 index (+0.30) and C16 index (+0.12). However, negligible improvement in prediction accuracy was observed for C18 index (+0.01), CLA index (+0.01) and no improvement for the total unsaturation index.

Large effects of SCD1 genotypes on UFA and unsaturation index traits have been reported. Schennink et al. (2008) indicated that the SCD1 V allele is associated with lower C10, C12, and C14 indices, and higher C16, C18, and CLA indices. As compared to the other unsaturation indices the effect of SCD1 was smaller on C18 and CLA indices and the total saturation index was not affected by SCD1 genotypes (Schennink et al., 2008). Duchemin et al. (2013) showed that SCD1 has highly significant effects on C10:1, C12:1, C14:1 cis-9, and C16:1 cis-9 fatty acids. These results are in agreement with the prediction accuracies presented in Table 5.2.

The SCD1 gene codes stearoyl-CoA desaturase that is responsible for Δ9-desaturation of fatty acids which explains the effects of this polymorphism on unsaturated fatty acids and the unsaturation indices. Wang et al. (2016) showed no significant effects of SCD1 genotypes on any of the milk IR wavenumbers. These results suggest that SCD1 genotypes and milk IR spectra contain independent information on milk unsaturation. This is confirmed by results presented in Table 5.2, e.g. prediction accuracy of the C10 unsaturation index is 0.29 base on IR, 0.21 based on SCD1 and 0.51 based on IR+SCD1. Prediction accuracy of individual unsaturated fatty acids based on IR is low but can be improved considerable by adding information on SCD1 genotypes.

5.3.6 Milk IR spectra and FASN genotypes

Results in Table 5.2 show that the chromosomal region on BTA19 containing the FASN gene explained part of the variation in the short to medium chain SFA: C8:0 (6%), C10:0 (8%), C12:0 (4%), C14:0 (12%). This is in line with previous studies on the effects of this region (Morris et al., 2007; Bouwman et al., 2011). The results of the current study show that prediction based on IR+FASN had a negligible improvement in prediction accuracy as compared to IR (Table 5.2).

The FASN gene codes the fatty acid synthase that catalyzes fatty acid synthesis in the mammary gland. However, multiple candidate genes might be responsible for the effects associated with this chromosomal region on BTA19 (Bouwman et al. 2014). The SNP rs137372738 that was used in the current study has no significant effects on any of the milk IR wavenumbers (results not shown). Therefore, similar as for the SCD1 genotypes, it was expected that the information captured by the genotypes and the milk IR spectra would be relatively independent and prediction accuracies would be additive. However, results did not show an improvement in R^2val based on IR+FASN as compared to prediction based on IR only. Although SNP rs137372738 was not significantly associated with any of the IR wavenumbers, Wang et al. (2016) did show significant associations of other SNP in the same region on BTA19 with several IR wavenumbers and with lactose content. Therefore, this region most likely contains multiple QTL. We performed additional analyses and estimated correlations between the fatty acids and all IR wavenumbers. Correlations were estimated separately for each FASN (rs137372738) genotype class. Results suggest that correlations between e.g. C14:0 and IR wavenumbers for one genotype class differed from correlations estimated for the other two genotype classes. This might be due to the presence of multiple QTL in this genomic region. In this case multiple SNP or haplotype information might need to be combined with IR data rather than information from a single SNP.

5.3.7 Milk IR spectra and 3 genotypes

The predictive ability of milk IR spectra combined with DGAT1, SCD1 and FASN genotypes is shown in Table 5.2. As compared to prediction based on IR only, the joint improvement in R^2val based on genotypic information from 3

polymorphisms was dominated by information contributed by DGAT1 genotypes for predicting C14:0 and C16:0, and by SCD1 genotypes for predicting C10:1, C12:1, C14:1 cis-9 and their unsaturation indices. Both DGAT1 and SCD1 improved prediction of C16:1 cis-9 and C16 index, and effects of both genotypes were additive (IR+3 polym.). The additive effect of DGAT1 and SCD1 genotypes on prediction accuracies of C16:1 cis-9 suggests that both genotypes explain different parts of the genetic variation in C16:1 cis-9. This is in line with results by Schennink et al. (2008).

5.3.8 Perspectives of combined use milk IR spectra and genotypic information

Different approaches have been suggested to improve milk IR prediction of milk fat composition, e.g. pre-treatments of IR wavenumbers (De Marchi et al., 2011; Soyeurt et al., 2011; Ferrand-Calmels et al., 2014), wavenumber pre-selection by genetic algorithms (Ferrand et al., 2011), log-transformation of skewed fatty acid measurements and creating uniform distributed subset samples (Fleming et al., 2017). In the current study we combined milk IR spectra with genotypic information to improve the prediction of milk fat composition. We showed that the combined use of genotypes and milk IR spectra can improve prediction for milk fat composition.

We used information from 3 polymorphisms with major effects on milk fat composition, however, there is no reason to restrict prediction to these 3 polymorphisms. The availability of a reference population consisting of animals with milk fat composition which are genotyped for a large number of SNP can result in the prediction of genomic breeding values. Genomic breeding values estimate the effect of the whole genome on a specific trait and thus provide additional information on top of the information provided by polymorphisms with a major effect. The prediction accuracy based on genomic information is bounded by the square root of the heritability of the trait. Results from the current study show that in some situations genotypic information and information from IR spectra can to a certain extend overlap. In those situations gain in prediction accuracy from including genomic information might be limited.

Combining genotypic information with information from IR spectra might also improve prediction of for example milk protein composition and mineral

composition. Changes in milk composition might be due to feeding strategies but they also might reflect the health status dairy cows. Milk composition has for example been related to negative energy balance (Bastin et al., 2012), udder health (Kawai et al., 1999; Schukken et al., 2003; Ouweltjes et al., 2007), ketosis (Van Knegsel et al., 2010Friggens et al., 2007; Stoop et al., 2009) and nitrogen efficiency (Frank and Swensson, 2002). Therefore, changes in milk composition might be used to monitor efficiency and health of dairy cows. Improved prediction of milk composition based on milk IR spectra combined with genotypic information of dairy cows might therefore contribute to precision livestock farming.

5.4 Conclusions

In this study genotypes of DGAT1, SCD1 and a SNP in the FASN gene were combined with milk IR spectra to investigate if prediction of milk fat composition can be improved. DGAT1 genotypes showed small improvements in prediction accuracy for some of the fatty acids. Adding FASN genotypes did not improve prediction of milk fat composition. The SCD1 genotypes considerably improved prediction for the unsaturated fatty acids and their unsaturation indices. This improved prediction accuracy illustrates the potential of combining milk IR spectroscopy with genomic information to predict milk composition. More accurate prediction of milk composition can result in better farm management indicators.

5.5 Acknowledgements

The China Scholarship Council is acknowledged for funding the PhD project of Qiuyu Wang. Milk Control Station (Zutphen, The Netherlands) is acknowledged for infrared spectra data. This study used data from the Dutch Milk Genomics Initiative and the project “Melk op Maat”, funded by Wageningen University (Wageningen, the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), CRV, the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands).

5.6 References

Bastin, C., D. P. Berry, H. Soyeurt, and N. Gengler. 2012. Genetic correlations of days open with production traits and contents in milk of major fatty acids predicted by mid-infrared spectrometry. J. Dairy Sci. 95: 6113-6121.

Bonfatti, V., G. Di Martino, and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. J. Dairy Sci. 94: 5776-5785.

Bouwman, A. C., H. Bovenhuis, M. H. Visker, and J. A. M. van Arendonk. 2011. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC genetics. 12: 43.

Bouwman, A. C., M. H. Visker, J. A. M. van Arendonk, and H. Bovenhuis. 2014. Fine mapping of a quantitative trait locus for bovine milk fat composition on Bos taurus autosome 19. J. Dairy Sci. 97: 1139-1149.

Bovenhuis, H., M. H. P. W. Visker, N. A. Poulsen, J. Sehested, H. J. F. van Valenberg, J. A. M. van Arendonk, L. B. Larsen, and A. J. Buitenhuis. 2016. Effects of the diacylglycerol o-acyltransferase 1 (DGAT1) K232A polymorphism on fatty acid, protein, and mineral composition of dairy cattle milk. J. Dairy Sci. 99: 3113-3123.

De Marchi, M., M. Penasa, A. Cecchinato, M. Mele, P. I. E. R. Secchiari, and G. Bittante. 2011. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Anim. 5: 1653-1658.

Duchemin, S., H. Bovenhuis, W. M. Stoop, A. C. Bouwman, J. A. M. van Arendonk, and M. H. P. W. Visker. 2013. Genetic correlation between composition of bovine milk fat in winter and summer, and DGAT1 and SCD1 by season interactions. J. Dairy Sci. 96: 592-604.

Dufour, E. 2009. Principles of infrared spectroscopy. Pages 1-27 in Infrared Spectroscopy for Food Quality Analysis and Control. D. W. Sun. ed. Acad. Press, San Diego, CA.

Ferrand, M., B. Huquet, S. Barbey, F. Barillet, F. Faucon, H. Larroque, O. Leray, J. M. Trommenschlager, and M. Brochard. 2011. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. Chemom. Intell. Lab. Syst. 106: 183-189.

Ferrand-Calmels, M., I. Palhière, M. Brochard, O. Leray, J. M. Astruc, M. R. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillat, M. Douguet, F. Faucon-Lahalle, M. Gele, G. Thomas, J. M. Trommenschlager, and H.

Larroque. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 97: 17-35.

Fleming, A., F. S. Schenkel, J. Chen, F. Malchiodi, V. Bonfatti, R. A. Ali, B. Mallard, M. Corredig, and F. Miglior. 2017. Prediction of milk fatty acid content with mid-infrared spectroscopy in Canadian dairy cattle using differently distributed model development sets. J. Dairy Sci. 100: 5073-5081.

Frank, B., and C. Swensson. 2002. Relationship between content of crude protein in rations for dairy cows and milk yield, concentration of urea in milk and ammonia emissions. J. Dairy Sci. 85: 1829-1838.

Friggens, N. C., C. Ridder, and P. Løvendahl. 2007. On the use of milk composition measures to predict the energy balance of dairy cows. J. Dairy Sci. 90: 5453-5467.

Hamann, J., and V. Krömker. 1997. Potential of specific milk composition variables for cow health management. Livest Prod Sci. 48: 201-208.

Haug, A., A. T. Høstmark, and O. M. Harstad. 2007. Bovine milk in human nutrition–a review. Lipids Health Dis. 6: 25.

ICAR (International Committee for Animal Recording). 2012. International agreement of recording practices – Guidelines approved by the General Assembly held in Cork, Ireland on June 2012. ICAR, Rome. Italy.

Kawai, K., S. Hagiwara, A. Anri, and H. Nagahata. 1999. Lactoferrin concentration in milk of bovine clinical mastitis. Vet. Res. Commun. 23: 391-398

Morris, C. A., N. G. Cullen, B. C. Glass, D. L. Hyndman, T. R. Manley, S. M., Hickey, J. C. McEwan, W. S. Pitchford, C. D. K. Bottema, and M. A. Lee. 2007. Fatty acid synthase effects on bovine adipose fat and milk fat. Mamm. Genome. 18: 64-74.

Ouweltjes, W., B. Beerda, J. J. Windig, M. P. L. Calus, and R. F. Veerkamp. 2007. Effects of management and genetics on udder health and milk composition in dairy cows. J. Dairy Sci. 90: 229-238.

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.

SAS Institute. 2011. SAS/STAT User's Guide: Release 9.3. SAS Inst., Cary, NC.

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.

Schennink, A., W. M. Stoop, M. H. Visker, J. M. 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.

Schukken, Y. H., D. J. Wilson, F. Welcome, L. Garrison-Tikofsky, and R. N. Gonzalez. 2003. Monitoring udder health and milk quality using somatic cell counts. Vet. Res. 34: 579-596.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89: 3690-3695.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94: 1657-1667.

Stoop, W. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2009. Effect of lactation stage and energy status on milk fat composition of Holstein-Friesian cows. J. Dairy Sci. 92: 1469-1478

Van Haelst, Y. N. T., A. Beeckman, A. T. M. van Knegsel, and V. Fievez. 2008. Short Communication: Elevated Concentrations of Oleic Acid and Long-Chain Fatty Acids in Milk Fat of Multiparous Subclinical Ketotic Cows. J. Dairy Sci. 91: 4683-4686.

Van Knegsel, A. T. M., S. G. A. van der Drift, M. Horneman, A. P. W. De Roos, B. Kemp, and E. A. M. Graat. 2010. Ketone body concentration in milk determined by Fourier transform infrared spectroscopy: Value for the detection of hyperketonemia in dairy cows. J. Dairy Sci. 93: 3065-3069.

Vlaeminck, B., V. Fievez, A. R. J. Cabrita, A. J. M. Fonseca, and R. J. Dewhurst. 2006a. Factors affecting odd-and branched-chain fatty acids in milk: a review. Anim. Feed Sci. Tech. 131: 389-417.

Vlaeminck, B., V. Fievez, S. Tamminga, R. J. Dewhurst, A. van Vuuren, D. de Brabander, and D. Demeyer. 2006b. Milk odd-and branched-chain fatty acids in relation to the rumen fermentation pattern. J. Dairy Sci. 89: 3954-3964.

Wang, Q., and H. Bovenhuis. Submitted. Validation strategy can result in overoptimistic view on the ability of milk infrared spectra to predict methane emission of dairy cattle. J. Dairy Sci.

Wang, Q., A. Hulzebosch, and H. Bovenhuis. 2016. Genetic and environmental variation in bovine milk infrared spectra. J. Dairy Sci. 99:6793–6803.

Wold, S., H. Martens, and H. Wold. 1983. The multivariate calibration problem in chemistry solved by the PLS method. Page 286-293 in Proc. Conf. on Matrix Pencils, Lecture Notes in Mathematics. A. Ruhe and B. Kagstrom, ed. Springer, Heidelberg, Germany.

6

General discussion

The main aim of this thesis was to investigate the genetic background of bovine milk infrared (IR) spectra of Holstein Friesian dairy cows in the Netherlands, and to investigate if knowledge of the genetic background can improve prediction of dairy cattle phenotypes based on milk IR spectra. In chapter 2, we estimated the heritability of individual milk IR wavenumbers and determined the significance of lactation stage, date of IR analyses, polymorphisms of diacylglycerol O-acyltransferase 1 (DGAT1), κ-casein (CSN3) and β-lactoglobulin (LGB) on individual milk IR wavenumbers. In chapter 3, we detected genomic regions that are significantly associated with selected milk IR wavenumbers. Most of the identified genomic regions were also associated with routinely recorded milk production traits such as fat-, protein-, and lactose content, but also other genomic regions were detected. We found evidence that these regions are related to milk phosphorus content, citric acid and orotic acid. In chapter 4, we predicted methane (CH_4) emission of dairy cattle by milk IR spectroscopy. We used different validation strategies to assess the quality of the predictions. The results showed the importance of using an appropriate validation strategy. We suggest the use of IR wavenumbers from water absorption regions as a negative control. In chapter 5, we showed that combined use of milk IR spectroscopy and genotypes improved the accuracy of predicting milk fat composition. Especially prediction accuracy of unsaturated fatty acids improved when combining milk IR spectroscopy with stearoyl-CoA desaturase-1 (SCD1) genotypes.

In this general discussion, I will first discuss the differences between the genetic background of milk IR spectra collected in winter and summer. Secondly I will discuss the potential of using milk IR spectroscopy to predict DGAT1 genotypes of dairy cattle. Finally I will discuss possible approaches to extract more information from milk IR spectra.

6.1 Seasonal differences in genetic background of milk IR spectra

The very first objective of this thesis was to unravel the genetic background of bovine milk IR spectra. Milk IR spectra consist of IR measurements at more than 1,000 individual wavenumbers. These individual milk IR wavenumbers were treated as dependent variables in a statistical analysis. It is well known that milk composition is affected by genetic factors, e.g. milk fat and protein

composition have substantial heritabilities and individual genes have been detected affecting milk composition (Schennink et al., 2008; Heck et al., 2009a; Bouwman et al., 2014). In addition, milk composition is affected by feed and management strategies, e.g. feed composition influences milk fat content and fat composition (e.g., Chilliard et al., 2007) and dietary energy intake influences milk protein content (Emery, 1978). The effect of differences between farms on milk IR spectra was described in chapter 2: approximately 10 to 25% of total variance in individual IR wavenumbers can be attributed to differences between farms.

All milk samples studied in chapter 2 were collected during a short period in winter (between February and March). Changes in feed may alter milk composition and the values for milk IR wavenumbers. Feed composition of dairy cows differs between seasons. During winter dairy cows in the Netherlands are housed indoor and fed with silage based diets, whereas during summer the majority of the cows graze outdoor. This change in diet most likely will result in changes in milk composition (Palmquist and Bealieu, 1993; Chilliard et al., 2001). In the Netherlands, considerable differences in milk composition exist between seasons. Heck et al. (2009b) showed large seasonal variation in milk fat and protein content. Milk fat content was 4.1% in June and 4.6% in January. Protein content increased from 3.2% in June to 3.4% in December. Seasonal difference also exists in milk fat composition. Short and medium chain saturated fatty acids, are generally increased in winter, whereas long chain unsaturated fatty acids are increased in summer (Duchemin et al., 2013). Capuano et al. (2014) showed that inclusion of fresh grass in the cows' diet can be accurately predicted based on milk IR spectra (more than 98% correctly predicted), which indicates that there are difference between IR spectra of milk collected in different seasons.

This thesis unravelled the genetic background of milk IR spectra of winter milk samples. To have a better understanding of the general applicability of the results presented in this thesis, I performed genetic analyses on another milk IR dataset of milk samples collected during summer, and investigated the correlations between milk IR wavenumbers based on winter and summer milk samples.

6.1.1 Heritability and effects of genes

Additional genetic analyses were performed using the same methods and based on largely the same dairy cows as described in chapter 2. The summer milk samples consisted of 1,757 milk samples from 384 commercial dairy farms collected between May and June in 2005. The effects of polymorphisms in DGAT1, SCD1, CSN3 and LGB and variance components were estimated for individual milk IR wavenumbers. The inter-herd heritability for individual wavenumbers was estimated.

The inter-herd heritabilities of individual IR wavenumbers for both winter and summer samples are shown in Figure 6.1. For the summer IR samples, inter-herd heritabilities of the 1,060 wavenumbers were 0.31 on average (0.36 in winter samples), ranging from 0 to 0.65 (0 to 0.63 in winter samples). In total there were 207 wavenumbers (197 in winter samples) with heritabilities lower than 0.20 and 562 wavenumbers (291 in winter samples) with heritabilities between 0.20 and 0.40. There were 283 wavenumbers (560 in winter samples) with heritabilities between 0.40 and 0.60, and 8 wavenumbers (12 in winter samples) with heritabilities larger than 0.60. Compared to IR wavenumbers in winter samples, the heritability of IR wavenumbers in summer samples had similar pattern but in general estimates were lower. This is in line with Duchemin et al. (2013) that heritability estimates for fat composition in summer samples are slightly lower in general. However in this study given the standard errors of heritability estimates, there are no significant differences between estimates from winter and summer samples. As shown in Figure 6.1, winter and summer samples showed similar heritability in the spectral region related to lactose content (around 1,080 cm^{-1}), which might be related to the constant lactose content in milk across seasons (Heck et al., 2009b). Winter milk samples showed higher heritabilities in the regions related to fat (2,800 to 2,900 cm^{-1}). It is difficult to observe a pattern in differences between winter and summer for IR regions related to protein (around 1,425 cm^{-1}). Analyses of the summer milk samples indicate that substantial variation of many individual milk IR wavenumbers can be attributed to genetics. This confirms results as reported in chapter 2.

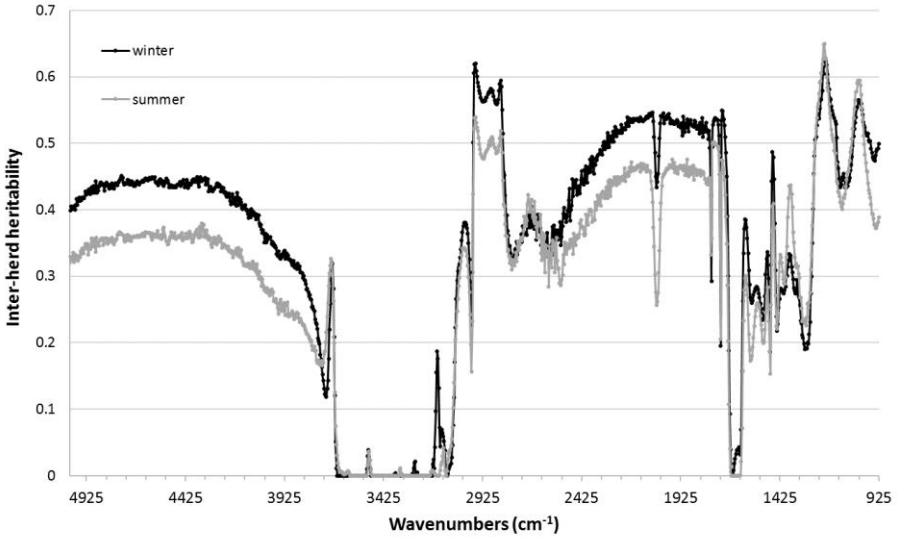


Figure 6.1 Heritability of individual milk IR wavenumbers for winter and summer samples.

The effects of DGAT1 and CSN3 polymorphisms on milk IR wavenumbers from summer samples are shown in Figure 6.2. DGAT1 showed highly significant effects on many IR wavenumbers with a generally lower significance level as compared to winter samples (chapter 2). The wavenumbers in the water absorption regions that affected by DGAT1 were also detected in summer samples. This confirms the findings in chapter 2. Significance of the CSN3 polymorphism showed similar results in summer as in winter samples. Significance levels in the spectral region around 1,250 and 1,450 cm^{-1}, which is related to protein, was higher in summer samples. Polymorphism in LGB showed similar levels of significance in summer and winter samples, and polymorphism in SCD1 showed no significant effects on any wavenumbers (results not shown), which is in line with chapter 2.

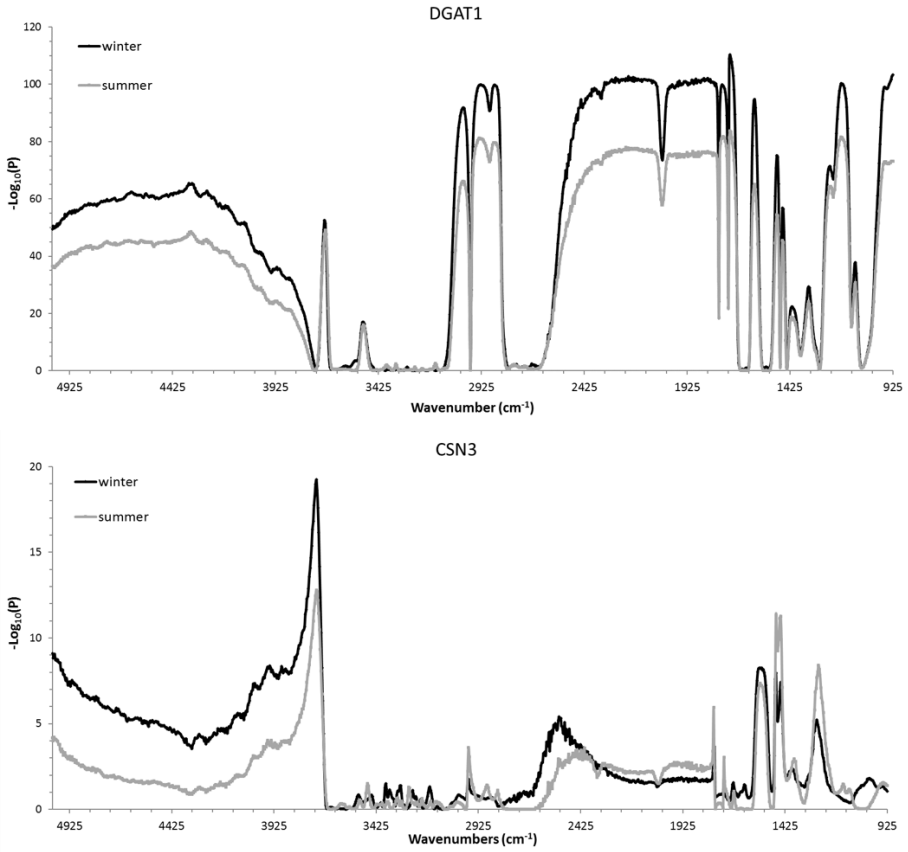


Figure 6.2 Effect of DGAT1 and CSN3 polymorphisms on individual milk IR wavenumbers in winter and summer samples.

6.1.2 Phenotypic and genetic correlations

Furthermore, I calculated correlations between milk IR wavenumbers in winter and summer samples. These correlations were estimated based on 1,572 dairy cows with records in both winter and summer.

Pearson correlations between winter and summer samples based on unadjusted data were estimated for all 1,060 IR wavenumbers. As shown in Figure 6.3, moderate to high phenotypic correlations were observed for the

informative wavenumbers (as defined and used in chapter 4 and 5). The highest correlation (0.78) was found for wavenumber 1,230 cm^{-1}. The wavenumbers in water absorption regions showed low phenotypic correlations. It was remarkable that the highest correlation for wavenumbers in the water absorption region was 0.09 for wavenumber 3,497 cm^{-1} (WN668 as shown in chapter 3). This was the wavenumber in the water absorption region with the highest $-Log_{10}(P)$ value for DGAT1 (Figure 6.3, indicated by a black dot).

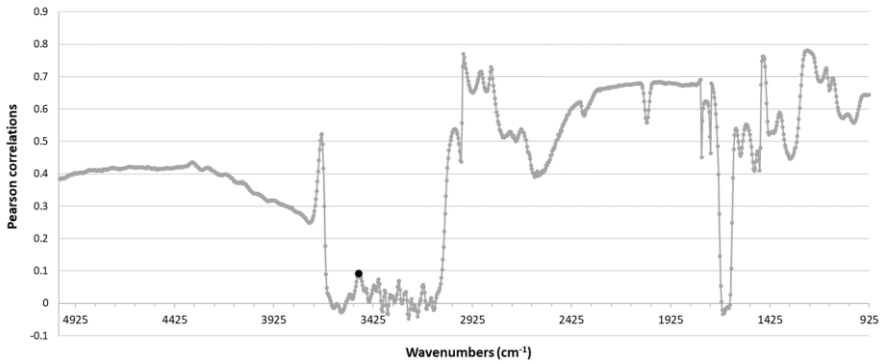


Figure 6.3 Pearson correlations of 1,060 unadjusted milk IR wavenumbers between winter and summer samples. The black dot indicates wavenumber 3,476 cm^{-1} (WN668).

Genetic correlations between winter and summer samples were estimated for 14 individual IR wavenumbers by ASReml 4 (Gilmour et al., 2015). These 14 wavenumbers were part of the wavenumbers selected in chapter 3. Table 6.1 shows the estimated phenotypic, genetic and herd correlations of the selected wavenumbers between winter and summer IR measurements. Most of the wavenumbers have very high genetic correlations between winter and summer and the estimates do not significantly differ from unity. WN208 and WN668 have a lower estimated genetic correlation of 0.66. In chapter 3, WN208 was significantly associated with a genomic region on chromosome 1 and it was hypothesized that this wavenumber could provide information on orotic acid. WN668 is the wavenumber showed the most significant effect of the DGAT1 polymorphism in the water absorption regions. WN668 also

showed the highest Pearson correlation between winter and summer samples (Figure 6.3). The standard errors for the genetic correlations of WN208 and WN668 were larger than for the other wavenumbers, indicating that the current dataset does not contain sufficient information to estimate the genetic correlations for these wavenumbers accurately. For WN668 the heritabilities were very low (<0.06) which explains the large standard error for the estimated genetic correlation. For WN208 heritabilities were 0.25 in winter and 0.29 in summer and lower than the other wavenumbers. The correlations due to herd were moderate, ranging from 0.17 to 0.59 except for WN668 with an estimate that did not differ significantly from 0. This shows that effects of herd management on IR wavenumbers are quite different in summer and in winter.

Table 6.1 Phenotypic correlations, genetic correlations and herd correlations (with standard error in parentheses) between measurements in winter and summer of 14 selected IR wavenumbers.

	Wavenumber (cm^{-1})	Phenotypic correlation	Genetic correlation	Herd correlation
WN20	999	0.56 (0.02)	0.97 (0.05)	0.18 (0.11)
WN34	1,053	0.57 (0.02)	0.99 (0.02)	0.44 (0.10)
WN50	1,114	0.69 (0.02)	0.96 (0.03)	0.59 (0.07)
WN80	1,230	0.78 (0.02)	0.99 (0.01)	0.59 (0.07)
WN126	1,407	0.53 (0.02)	0.99 (0.03)	0.36 (0.09)
WN142	1,469	0.68 (0.02)	0.96 (0.03)	0.56 (0.06)
WN156	1,523	0.50 (0.02)	0.97 (0.05)	0.41 (0.08)
WN208	1,724	0.46 (0.02)	0.66 (0.19)	0.24 (0.07)
WN220	1,770	0.45 (0.02)	0.97 (0.05)	0.30 (0.08)
WN414	2,518	0.47 (0.02)	0.99 (0.05)	0.18 (0.08)
WN432	2,587	0.40 (0.02)	0.97 (0.06)	0.20 (0.07)
WN668	3,497	0.09 (0.03)	0.66 (0.49)	-0.25 (0.34)
WN717	3,686	0.52 (0.02)	0.86 (0.08)	0.27 (0.07)
WN728	3,729	0.26 (0.02)	0.94 (0.10)	0.17 (0.06)

These analyses show that the genetic background of milk IR spectra in winter and summer milk samples is very similar: the high genetic correlations indicate that the milk IR wavenumbers are genetically the same traits in winter and summer. This is in line with Duchemin et al. (2013) which showed strong positive genetic correlations between milk fat composition in winter and summer. Rutten et al. (2009) investigated the prediction of milk fatty acids using both winter and summer samples. It was concluded that the effect of season on the coefficient of determination (R^2) in validation was limited but occasionally a large prediction bias was observed.

These analyses suggest that the genetic background of milk IR spectra is very similar in winter and in summer; significance of the effects of polymorphisms in DGAT1, SCD1, CSN3, and LGB on milk IR wavenumbers are similar; and genetic correlations between milk IR wavenumbers in winter and summer are close to unity. This indicates that the results described in this thesis are not limited to winter milk samples but have general applicability.

6.2 Prediction of DGAT1 genotypes

The diacylglycerol O-acyltransferase 1 (DGAT1) gene, which plays a role in triglyceride synthesis, has been identified as a strong positional and functional candidate gene for the QTL effect on bovine chromosome 14. Associations between DGAT1 polymorphism and multiple milk production traits have been investigated. Many studies presented evidence that the DGAT1 K232A polymorphism has major effects on milk production traits: the lysine residue (K allele) increased milk fat production, fat content and protein content, but decreased milk protein yield and milk production (Grisart et al., 2002; Spelman et al., 2002; Winter et al., 2002; Thaller et al., 2003; Weller et al., 2003; Sanders et al., 2006; Gautier et al., 2007; Schennink et al., 2007; Banos et al., 2008; Berry et al., 2010). Other studies showed that the DGAT1 K232A polymorphism was associated with milk fat composition (Shorten et al., 2004; Schennink et al., 2007; Schennink et al., 2008; Conte et al., 2010; Lu et al., 2015; Bovenhuis et al., 2016). Moreover, Bovenhuis et al. (2016) detected substantial effects of the DGAT1 polymorphism on milk mineral composition. Therefore knowledge about the DGAT1 genotypes of cows might be beneficial for selective breeding and for predicting phenotypes.

Results presented in this thesis showed that DGAT1 genotypes have highly significant effects on milk IR spectra with $-Log_{10}(P)$ values of up to 110.4 (chapter 2). In chapter 3, the genomic region containing the DGAT1 gene was detected in GWAS for many milk IR wavenumbers. In chapter 5, DGAT1 genotypes improved prediction accuracy for some milk fatty acids when combined with milk IR spectra, among others for C16:1 cis-9 and C16 unsaturation index. These results show a strong relation between milk IR spectra and DGAT1 genotypes which suggests that DGAT1 genotypes can be predicted by milk IR spectra. Studies showed that milk IR spectra have some potential to predict LGB genotypes (Rutten et al., 2011), α_{S1}-casein (CSN1S1) genotypes (Berget et al., 2010; Bonfatti et al., 2015) and CSN3 genotypes (Bonfatti et al., 2015). Therefore it is of interest to investigate the ability of milk IR spectra to predict DGAT1 genotypes of dairy cows.

To investigate the potential of milk IR spectra to predict DGAT1 genotypes I used the same methodology as described Rutten et al. (2011). The actual DGAT1 genotypes of individual cows were obtained by the genotyping procedure described by Schennink et al. (2007). DGAT1 genotypes and milk IR spectra were available for 1,636 dairy cows. Of these cows 596 (36.4%) were AA, 771 (47.2%) AK and 269 (16.4%) KK. The total samples were divided into a calibration and a validation set. 1,300 observations (approx. 80%) were randomly selected as calibration data and the remaining 336 observations (approx. 20%) were used as the validation data. The response variable of the prediction model, the DGAT1 genotype, was represented by a dummy variable and coded as 0, 1, 2 for the DGAT1 genotypes AA, AK and KK, respectively. The 275 informative milk IR wavenumbers as described in chapter 4 and 5 were used as predictors. The partial least squares (PLS) procedure in SAS 9.3 (SAS Institute, 2001) was used to perform the analyses. The number of latent variables was 9 and was determined by cross validation.

In the validation process, dummy variables indicating the DGAT1 genotypes, were predicted using the milk IR prediction model. A set of decision rules were used to transform the continuous dummy variables into the three genotypes:

$$\hat{Y} \leq 0.5 \rightarrow \text{DGAT1 genotype} = AA;$$

$$0.5 < \hat{Y} < 1.5 \rightarrow \text{DGAT1 genotype} = AK;$$

$$\hat{Y} \geq 1.5 \rightarrow \text{DGAT1 genotype} = KK.$$

In total 10 replicates of the calibration and validation procedure were performed. The results averaged over the 10 replicates are presented in Table 6.2.

Table 6.2 Observed and predicted DGAT1 genotype combinations.

Percentage [1] (%)		Observed DGAT1 genotype			Correct Prediction[2]
		AA	AK	KK	
Predicted DGAT1 genotype	AA	19.4 ± 1.7	5.5 ± 1.0	0.3 ± 0.3	63.0 ± 2.2
	AK	16.9 ± 1.7	40.0 ± 3.1	12.5 ± 1.4	
	KK	0.1 ± 0.1	1.7 ± 0.5	3.6 ± 0.6	

[1] Averaged proportions over all 336 observations in the validation set and based on 10 replicates. Standard deviations are based on 10 replicates.
[2] The averaged proportion of correctly predicted DGAT1 genotypes, equal to the sum of diagonal elements.

DGAT1 genotypes for 63.0% (±2.2) cows were correctly predicted based on milk IR spectra. The standard deviations for all combinations were small. Rutten et al. (2011) investigated the use of milk IR spectra to predict LGB genotypes and reported that LGB genotypes could be predicted correctly in 74% of the cases. Bonfatti et al. (2015) showed 56% of the CSN1S1 genotypes and 70% of the CSN3 genotypes can be predicted correctly based on milk IR spectra in buffalo. Genotypes of LGB, CSN1S1 and CSN3 have significant effects on milk protein composition whereas DGAT1 has significant effects on many milk components: milk fat content, milk protein content, milk

fatty acid composition and milk mineral composition (e.g., Grisart et al., 2002; Thaller et al., 2003; Schennink et al., 2007; Bovenhuis et al., 2016). The DGAT1 polymorphism showed highly significant effects ($-Log_{10}(P) >100$) on numerous IR wavenumbers, while associations of LGB and CSN3 polymorphisms were significant but much less strong (chapter 2). Therefore I expected that DGAT1 genotypes of individual cows could be predicted accurately based on the milk IR spectra. The fraction of correctly predicted DGAT1 genotypes was 63.0% and lower than expected. Especially when considering that randomly assigning cows to DGAT1 genotypes without any IR information results in correctly assigning 37.5% of the cows to DGAT1 genotypes. This shows that the information in the milk IR spectra to predict DGAT1 genotypes is low and less than expected.

Table 6.2 shows that 25.2% of the cows were predicted to be DGAT1 AA, which is lower than the observed proportion of DGAT1 genotypes (36.4%). Furthermore, 69.4% of dairy cows were predicted to have AK genotype which is higher than the observed proportion (47.2%), and 5.4% of dairy cows were predicted to have KK genotype which is lower than the observed proportion (16.4%). In the dairy cows predicted to have AK genotype, more than 40% of them were actually AA or KK genotypes. The issue is that due to the choice of the cut off values too few animals were predicted to have AA and KK genotypes.

I investigated if it is possible to improve the prediction accuracy. First of all, the decision rule was adjusted so that proportions of predicted genotypes were the same as the observed genotype frequencies: 36.4% for AA, 47.2% for AK and 16.4% for KK. And secondly, after corrected for the systematic environmental effects e.g. lactation stage, season of calving, as well as random effects due to herd and residue, the additive genetic effects of individual milk IR wavenumbers were used as predictors. However neither of these attempts improved the prediction accuracy as compared to PLS based on raw milk IR wavenumbers.

As DGAT1 genotypes have highly significant effects on many milk components, it would be of interest to have an easy, non-invasive and inexpensive method to determine DGAT1 genotypes of dairy cows. The analysis showed a low prediction accuracy of DGAT1 genotypes based on

milk IR spectra. This was surprising as other studies have shown higher prediction accuracies for polymorphisms which showed smaller effects on milk IR wavenumbers. It is worthwhile to investigate if other prediction methods, for example partial least square discriminant analysis (PLS-DA) or machine learning algorithms can improve prediction accuracies.

6.3 Capture more information from milk IR spectra

It is very attractive to use milk IR spectra to capture fine milk composition for breeding and management of dairy cows (Gengler et al., 2016). Milk IR spectroscopy is used e.g. in milk payment systems, prediction of manufacturing properties of dairy products, and prediction of novel phenotypes for monitoring dairy cattle health and dairy farm management. The basis of these applications is that milk IR spectra capture information on milk composition. Major milk components e.g. fat, protein and lactose have direct signals in the milk IR spectra. Signals from major milk components might overwhelm signals of other milk components, e.g. components in low concentrations. In chapter 3, the genome wide association study detected genomic regions that have been related to milk phosphorus content, citric acid and orotic acid. These results suggest that the milk IR spectra contains information on milk components which are currently not routinely quantified. There might be other ways to extract additional information from milk IR spectra.

6.3.1 Extract information from the water absorption regions

Water is the main component of milk. Wavenumbers from water absorption regions (between 1,619 and 1,674 cm^{-1}, and between 3,073 and 3,667 cm^{-1}, as indicated by chapter 2) show much larger phenotypic variance than other wavenumbers, and most of the variance is noise (unexplained residual variance). Therefore the water absorption regions are commonly excluded when building prediction equations. In chapter 2, we identified that some wavenumbers (between 3,466 and 3,543 cm^{-1}) in the water absorption regions are significantly affected by DGAT1 polymorphisms. This was confirmed in the summer milk samples (Figure 6.2). Furthermore, wavenumbers around 3,497 cm^{-1} in the water absorption regions have heritabilities larger than 0, which

was also found in a recent study in Danish Holstein and Jersey cows (Zaalberg et al., 2019).

The water absorption regions are caused by intense and broad absorption bands of water due to hydroxyl group (–OH) stretching and H–O–H bending. Water molecules result in intense absorption and overwhelm signals induced by other milk components in the same region. First of all, the stretching of –OH also exists in carbohydrates for example glucose and in carboxyl group (–COOH) present in fatty acids. Therefore information about these components in the water absorption region might be masked by the overwhelming signal from the water molecules. Secondly, the double carbon bond (C=C) stretching is observed between 1,640 and 1,666 cm^{-1} (Dufour, 2009), which is located in the water absorption. In addition the C=C bond is non-polar and shows weaker absorption as compared to C=O bond. This explains why milk IR spectra capture little information on unsaturated fatty acids. Last but not least, Amide I band due to C=O stretching in the polypeptide shows absorption at 1,600-1,700 cm^{-1}. Amide I band is related to secondary structure of protein, e.g. α-helix, β-sheet (Karoui et al., 2003). Therefore part of information on milk fat, protein and carbohydrates might be masked by water when IR spectroscopy is based on raw milk.

Two approaches can be applied to limit or avoid the effect of water. A obvious approach is to remove water from milk samples. Afseth et al. (2010) applied IR to dried film measurements. This study showed that the dried film approach improved predictions of milk fatty acids as compared to analysing raw milk, especially major saturated fatty acids like C16:0 (R^2 of 0.93 vs. 0.65), C18:0 (R^2 of 0.91 vs. 0.48) and minor fatty acids like CLA (R^2 of 0.86 vs. 0.53), polyunsaturated fatty acids (R^2 of 0.78 vs. 0.52) and total trans-UFA (R^2 of 0.85 vs. 0.54). The concentration of milk samples before IR analyses removed water and increased the content of the component of interest. This study showed a new approach to extract additional information from the milk IR spectra. It is of interest to investigate if prediction of milk composition can be further improved after correcting for the IR signals due to water.

Another approach avoiding water absorption band is to subtract water spectra by using attenuated total reflectance (ATR) combined with IR spectroscopy. ATR is a sampling technique that allows analysing samples directly in solid or

liquid state, for example milk component dissolved in water. When samples are presented in solid or liquid form, the feature of spectral light is determined by the thickness of the solid sample or absorption of the liquid. A larger thickness is associated with a stronger absorption. ATR is ideal for strongly absorbing or thick samples which often produce intense peaks when measured by transmission. The conformation changes of β-lactoglobulin have been studied by Dufour et al. (1994) using ATR technique. Recent studies showed the use of ATR combined with IR spectroscopy in investigating physicochemical and structural changes of protein when processing ultra-heat treatment (UHT) milk (Grewal et al., 2017, 2018), with emphasis on the spectral region between 1,500 and 1,700 cm^{-1}. The available spectra in the region might also be useful to quantify milk protein composition. These type of analyses require combining an ATR unit with IR spectrometer.

6.3.2 Use of processed milk

The study of Afseth et al. (2010) removed water from the milk sample and this way IR absorption signals from other milk components can be revealed. Similarly, removing other milk components than water might improve IR prediction of some components. Removing milk fat and performing analysis of skim milk might for example improve prediction of milk protein composition. Furthermore, removing casein from milk will enhance the ability to quantify whey proteins. Membrane technology has been applied in the dairy industry for separation of milk (Kumar et al. 2013). It is commonly based on ceramic or polymeric membranes with different pore sizes, allowing separation of permeate from retentate. Different types of milk filtrations have been developed, in the order of pore size from large to small, such as micro-filtration, ultra-filtration, nano-filtration and reverse osmosis. The permeate and retentate of each filtration technique contain separated milk component fractions. Applying milk IR spectroscopy to permeate or retentate of different filtration stage might enhance the prediction of certain milk components, and possibly additional milk components. Franzoi et al. (2018) developed very accurate milk IR predictions for protein and sugars in defatted and delactosated milk after ultra- and nano-filtration. This study showed that the use of processed milk has the potential to capture additional information from milk IR spectra. However the procedure of membrane technology might be currently too complicated for routine milk recording scheme.

6.3.3 Use of different prediction methods

The most commonly used method in predicting dairy cattle phenotypes based on milk IR spectra is partial least square (PLS) regression. There are some studies suggesting that Bayesian regression outperforms PLS in predicting dairy cattle phenotypes. Ferragina et al. (2015) compared PLS and Bayesian methods in predicting several milk fatty acids and technological properties. Bayesian models showed a remarkably higher R^2 in validation as compared to PLS; 0.67 for the Bayesian model versus 0.44 based on PLS for C10:0, 0.48 versus 0.30 for C14:1 cis-9, 0.60 versus 0.41 for C16:0, and 0.49 versus 0.26 for C18:0. Increases in R^2 when using Bayesian models were also observed for milk technological property traits e.g. rennet coagulation time, cheese yield and recovery of fat and protein. The authors showed that the Bayesian models with highest prediction accuracy used a smaller number of informative wavenumbers, and therefore prediction was less affected by the uninformative wavenumbers. However Bonfatti et al. (2017) reported small improvements in R^2 for Bayesian methods as compared to PLS: up to 2.3 percentage points for C14:0 and limited improvement for protein composition. Bayesian models even showed lower prediction accuracy for milk mineral composition. Depending on the phenotypes, Bonfatti et al. (2017) concluded that Bayesian models produced statistically significant improvements in prediction accuracy but the absolute amounts of improvement were small or negligible.

Further research could be extended to new approaches like machine learning algorithms to predict dairy cattle phenotypes based on milk IR spectra. Artificial neural networks (ANN) have been suggested for predicting dairy cattle phenotypes, e.g. subclinical ketosis (Ehret et al., 2015) and time of calving (Borchers et al., 2017). Vásquez et al. (2018) indicated that ANN model had slightly better performance than PLS model in predicting hardiness of cheese using spectral imaging data. Dórea et al. (2018) compared the performance of PLS versus ANN in predicting dry matter intake of lactating dairy cows based on milk IR spectra. Both PLS and ANN had similar accuracy when using milk production traits (fat%, protein% and lactose%) as predictors, however when milk production traits were replaced by milk IR wavenumbers, ANN showed better prediction accuracy than PLS (R^2 of 0.67 vs. 0.53) and smaller prediction error. This suggested that additional information from milk

IR spectra can be extracted by ANN which improves prediction of dry matter intake. Comparison between PLS and ANN methods has also been investigated in other fields of science. Perai et al. (2010) showed that ANN outperforms PLS in predicting metabolizable energy level of poultry diets (R^2 of 0.94 vs. 0.36).

It has been demonstrated that ANN can capture nonlinear relationships between the response variable and predictors (milk IR wavenumbers), and therefore it has an advantage over PLS regression where commonly linearity is assumed. The actual advantage of machine learning algorithms over PLS still requires further study. However the aforementioned studies showed promising results.

6.4 Concluding remarks

In this general discussion, I investigated the differences between the genetic background of milk IR wavenumbers in winter and summer samples. The high genetic correlations indicated that the studied individual IR wavenumbers are genetically the same traits. The polymorphism of DGAT1 gene has very significant effect on milk IR wavenumbers, however the ability of milk IR wavenumbers to predict DGAT1 genotypes is low. Some approaches might be helpful to capture more information from milk IR spectra, e.g. prediction using machine learning algorithms.

6.5 References

Afseth, N. K., H. Martens, Å. Randby, L. Gidskehaug, B. Narum, K. Jørgensen, S. Lien, and A. Kohler. 2010. Predicting the fatty acid composition of milk: a comparison of two Fourier transform infrared sampling techniques. Appl. Spectrosc. 64: 700-707.

Banos, J., J. A. Woolliams, B. W. Woodward, A. B. Forbes, and M. P. Coffey. Impact of Single Nucleotide Polymorphisms in Leptin, Leptin Receptor, Growth Hormone Receptor, and Diacylglycerol Acyltransferase (DGAT1) Gene Loci on Milk Production, Feed, and Body Energy Traits of UK Dairy Cows. J. Dairy Sci. 91: 3190-3200.

Berget, I., H. Martens, A. Kohler, S. K. Sjurseth, N. K. Afseth, B. Narum, and S. Lien. 2010. Caprine CSN1S1 haplotype effect on gene expression and milk composition measured by Fourier transform infrared spectroscopy. J. Dairy Sci. 93: 4340-4350.

Berry, D. P., D. Howard, P. O'Boyle, S. Waters, J. F. Kearney, and M. McCabe. 2010. Associations between the K232A polymorphism in the diacylglycerol-O-transferase 1 (DGAT1) gene and performance in Irish Holstein-Friesian dairy cattle. Irish J. Agr. Food. Res. 49: 1-9.

Bonfatti, V., A. Cecchinato, and P. Carnier. 2015. Short communication: Predictive ability of Fourier-transform mid-infrared spectroscopy to assess CSN genotypes and detailed protein composition of buffalo milk. J. Dairy Sci. 98: 6583-6587.

Bonfatti, V., F. Tiezzi, F. Miglior, and P. Carnier. 2017. Comparison of Bayesian regression models and partial least squares regression for the development of infrared prediction equations. J. Dairy Sci. 100: 7306-7319.

Borchers, M. R., Y. M. Chang, K. L. Proudfoot, B. A. Wadsworth, A. E. Stone, and J. M. Bewley. 2017. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. J. Dairy Sci. 100: 5664-5674.

Bouwman, A. C., M. H. Visker, J. A. M. van Arendonk, and H. Bovenhuis. 2014. Fine mapping of a quantitative trait locus for bovine milk fat composition on Bos taurus autosome 19. J. Dairy Sci. 97: 1139-1149.

Bovenhuis, H., M. H. P. W. Visker, N. A. Poulsen, J. Sehested, H. J. F. van Valenberg, J. A. M. van Arendonk, L. B. Larsen, and A. J. Buitenhuis. 2016. Effects of the diacylglycerol o-acyltransferase 1 (DGAT1) K232A polymorphism on fatty acid, protein, and mineral composition of dairy cattle milk. J. Dairy Sci. 99: 3113-3123.

Capuano, E., J. Rademaker, H. van den Bijgaart, and S. M. van Ruth. 2014. Verification of fresh grass feeding, pasture grazing and organic farming by FTIR spectroscopy analyses of bovine milk. Food Res. Int. 60: 59-65.

Chilliard, Y., A. Ferlay, and M. Doreau. 2001. Effect of different types of forages, animal fat or marine oils in cow's diet on milk fat secretion and composition, especially conjugated linoleic acid (CLA) and polyunsaturated fatty acids. Livest. Prod. Sci. 70: 31-48.

Chilliard, Y., F. Glasser, A. Ferlay, L. Bernard, J. Rouel, and M. Doreau. 2007. Diet, rumen biohydrogenation and nutritional quality of cow and goat milk fat. Eur. J. Lipid Sci. Technol. 109: 828-855.

Conte, G., M. Mele, S. Chessa, B. Castiglioni, A. Serra, G. Pagnacco, and P. Secchiari. 2010. Diacylglycerol acyltransferase 1, stearoyl-CoA desaturase 1, and sterol regulatory element binding protein 1 gene polymorphisms and milk fatty acid composition in Italian Brown cattle. J. Dairy Sci. 93: 753-763.

Dórea, J. R. R., G. J. M. Rosa, K. A. Weld, and L. E. Armentano. 2018. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. J. Dairy Sci. 101: 5878-5889.

Duchemin, S., H. Bovenhuis, W. M. Stoop, A. C. Bouwman, J. A. M. van Arendonk and M. H. P. W. Visker. 2013. Genetic correlation between composition of bovine milk fat in winter and summer, and DGAT1 and SCD1 by season interactions. J. Dairy Sci. 96: 592-604.

Dufour, É. Principles of infrared spectroscopy. 2009. Infrared Spectroscopy for Food Quality Analysis and Control; Academic Press: New York, Chapter 1.

Dufour, É., P. Robert, D. Bertrand, and T. Haertlé. 1994. Conformation changes of β-lactoglobulin: An ATR infrared spectroscopic study of the effect of pH and ethanol. J. Protein Chem. 13: 143-149.

Ehret, A., D. Hochstuhl, N. Krattenmacher, J. Tetens, M. S. Klein, W. Gronwald, and G. Thaller. 2015. Use of genomic and metabolic information as well as milk performance records for prediction of subclinical ketosis risk via artificial neural networks. J. Dairy Sci. 98: 322-329.

Emery, R. S. 1978. Feeding for increased milk protein. J. Dairy Sci. 61: 825-828.

Ferragina, A., G. de Los Campos, A. I. Vazquez, A. Cecchinato, and G. Bittante. 2015. Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. J. Dairy Sci. 98: 8133-8151.

Franzoi, M., C. L. Manuelian, L. Rovigatti, E. Donati, and M. De Marchi. 2018. Development of Fourier-transformed mid-infrared spectroscopy prediction models for major constituents of fractions of delactosated, defatted milk obtained through ultra-and nanofiltration. J. Dairy Sci. 101: 6835-6841.

Gautier, M., A. Capitan, S. Fritz, A. Eggen, D. Boichard, and T. Druet. 2007. Characterization of the DGAT1 K232A and variable number of tandem repeat polymorphisms in French dairy cattle. J. Dairy Sci. 90: 2980-2988.

Gengler, N., H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M. L. Vanrobays, A. Lainé, S. Vanderick, C. Grelet, A. Vanlierde, E.

Froidmont, and P. Dardenne. 2016. Capitalizing on fine milk composition for breeding and management of dairy cows[1]. J. Dairy Sci. 99: 4071-4079.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. 2015. ASReml user guide release 4.1. VSN International Ltd, Hemel Hempstead, UK.

Grewal, M. K., J. Chandrapala, O. Donkor, V. Apostolopoulos, L. Stojanovska, and T. Vasiljevic. 2017. Fourier transform infrared spectroscopy analysis of physicochemical changes in UHT milk during accelerated storage. Int. Dairy J. 66: 99-107.

Grewal, M. K., T. Huppertz, and T. Vasiljevic. 2018. FTIR fingerprinting of structural changes of milk proteins induced by heat treatment, deamidation and dephosphorylation. Food Hydrocoll. 80: 160-167.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in Dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12: 222-231.

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. 2009a. Effects of milk protein variants on the protein composition of bovine milk. J. Dairy Sci. 92: 1192-1202.

Heck, J. M. L., H. J. F. van Valenberg, J. Dijkstra, and A. C. M. van Hooijdonk. 2009b. Seasonal variation in the Dutch bovine raw milk composition. J. Dairy Sci. 92: 4745-4755.

Karoui, R., G. Mazerolles, and É. Dufour. 2003. Spectroscopic techniques coupled with chemometric tools for structure and texture determinations in dairy products. Int. Dairy J., 13: 607-620.

Kumar, P., N. Sharma, R. Ranjan, S. Kumar, Z. F. Bhat, and D. K. Jeong. 2013. Perspective of membrane technology in dairy industry: A review. Asian-Australas J. Anim. Sci. 26: 1347.

Lu, J., S. Boeren, T. van Hooijdonk, J. Vervoort, and K. Hettinga. 2015. Effect of the DGAT1 K232A genotype of dairy cows on the milk metabolome and proteome. J. Dairy Sci. 98: 3460-3469.

Palmquist, D. L., A. D. Beaulieu, and D. M. Barbano. 1993. Feed and animal factors influencing milk fat composition. J. Dairy Sci. 76: 1753-1771.

Perai, A. H., H. Nassiri Moghaddam, S. Asadpour, J. Bahrampour, and G. Mansoori. 2010. A comparison of artificial neural networks with other statistical approaches for the prediction of true metabolizable energy of meat and bone meal. Poult. Sci. 89: 1562-1568.

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.

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Prediction of β-lactoglobulin genotypes based on milk Fourier transform infrared spectra. J. Dairy Sci. 94: 4183-4188.

Sanders, K., J. Bennewitz, N. Reinsch, G. Thaller, E.-M. Prinzenberg, C. Kühn, and E. Kalm. 2006. Characterization of the DGAT1 Mutations and the CSN1S1 Promoter in the German Angeln Dairy Cattle Population. J. Dairy Sci. 89: 3164-3174.

SAS Institute. 2011. SAS/STAT User's Guide: Release 9.3. SAS Inst., Cary, NC.

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.

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.

Shorten, P. R., T. B. Pleasants, and G. C. Upreti. 2004. A mathematical model for mammary fatty acid synthesis and triglyceride assembly: the role of stearoyl CoA desaturase (SCD). J. Dairy Res. 71: 385-397.

Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory, and R. G. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85: 3514-3517.

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt, and R. Fries. 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci. 81: 1911-1918.

Vásquez, N., C. Magán, J. Oblitas, T. Chuquizuta, H. Avila-George, and W. Castro. 2018. Comparison between artificial neural network and partial least squares regression models for hardness modeling during the ripening process of Swiss-type cheese using spectral profiles. J. Food Eng. 219: 8-15.

Weller, J. I., M. Golik, E. Seroussi, E. Ezra, and M. Ron. 2003. Population-wide analysis of a QTL affecting milk-fat production in the Israeli Holstein population. J. Dairy Sci. 86: 2219-2227.

Winter, A., W. Krämer, F. A. O. Werner, S. Kollers, S. Kata, G. Durstewitz, J. Buitkamp, J. E. Womack, G. Thaller, and R. Fries. 2002. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA: diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proc. Natl. Acad. Sci. 99: 9300-9305.

Zaalberg, R. M., N. Shetty, L. Janss, and A. J. Buitenhuis. 2019. Genetic analysis of Fourier transform infrared milk spectra in Danish Holstein and Danish Jersey. J. Dairy Sci. 102: 503-510.

Summary

Milk infrared (IR) spectroscopy has been used as the routine method to determine milk production traits such fat, protein, lactose, urea content in milk, and it is a promising technique to obtain detailed milk composition. Several studies suggested that milk IR spectra can be used to predict milk fatty acids, protein composition, mineral composition, milk coagulation properties, acidity, as well as some traits related to health, energy status, and environmental impact of dairy cattle. A great advantage of milk IR spectroscopy is that large number of samples can be analyzed in a cost-effective and a relatively short period. The quantification of milk components is based on the absorption of electromagnetic radiation by chemical bonds in the molecules at different wavenumbers. The absorption at adjacent wavenumbers can be induced by a chemical bond that is abundant in molecules of a certain milk component, and thus information on certain milk components can be captured by milk IR spectra. The genetic background of milk composition has been studied intensively however little is known about the genetic background of milk IR spectra. Therefore it is of interest to understand to what extent milk IR spectra are affected by genetic differences between dairy cows. This thesis aimed at studying the genetic background of milk IR spectra and applying the gained knowledge into prediction of dairy cattle phenotypes such as methane (CH_4) emission and milk fat composition.

Chapter 2 investigates the genetic and environmental variation in individual wavenumbers of milk IR spectra. Inter-herd heritabilities of 1,060 infrared wavenumbers ranged from 0 to 0.63 indicating that the genetic background of infrared wavenumbers differs considerably. The majority of the wavenumbers have moderate to high inter-herd heritabilities ranging from 0.20 to 0.60. Differences between herds explained 10 to 25% of the total variance for most wavenumbers. This suggests that the wavenumbers of milk IR spectra are indicative for differences in feeding and management between herds. This chapter also investigates the effects of systematic environmental and genetic factors on individual wavenumbers of milk IR spectra. Many wavenumbers of milk IR spectra were significantly affected by lactation stage, date of IR analysis, and polymorphisms of gene diacylglycerol O-acyltransferase 1 (DGAT1), κ-casein (CSN3) or β-lactoglobulin (LGB). These genes have major effect on milk composition. In contrast, the stearoyl-CoA desaturase (SCD1) polymorphism did not significantly affect any of the wavenumbers. SCD1 is

known to have a strong effect on the content of C10:1, C12:1, C14:1, and C16:1 fatty acids. Therefore, these results suggest that IR spectra contain little direct information on these mono unsaturated fatty acids. The wavenumbers between 1,619 and 1,674 cm^{-1} and between 3,073 and 3,667 cm^{-1} are strongly influenced by water absorption and usually excluded when setting up prediction equations. However, we found that some of the wavenumbers in the water absorption region are affected by the DGAT1 polymorphism and lactation stage. This suggests that these wavenumbers contain useful information regarding milk composition.

Chapter 3 identifies the genomic regions that are associated with milk IR spectra. For this purpose a genome wide association study (GWAS) was performed for a selected set of 50 individual IR wavenumbers measured on 1,748 Dutch Holstein cows. Significant associations were detected for 28 of the 50 wavenumbers. In total 24 genomic regions distributed over 16 bovine chromosomes were identified. Major genomic regions associated with milk IR wavenumbers were identified on chromosomes 1, 5, 6, 14, 19 and 20. Most of these regions also showed significant associations with fat%, protein% or lactose%. However, we also identified some new regions which were not associated with any one of these routinely collected milk composition traits. On chromosome 1 two new genomic regions were identified and we hypothesise that they are related to variation in milk phosphorus content and orotic acid, respectively. On chromosome 20 a new genomic region was identified which seem to be related to citric acid. Identification of genomic regions associated with milk phosphorus content, orotic acid and citric acid suggest that the milk IR spectra contain direct information on these milk components.

Chapter 4 applies different validation strategies in predicting CH_4 emission of individual dairy cows based on milk IR spectroscopy. Due to its environmental impact it is of great interest to reduce CH_4 emission of dairy cattle and selective breeding might contribute. Milk IR spectroscopy has been proposed as a rapid and cheap measurement technique which can be used to quantify CH_4 emission for a large number of individual dairy cows. This study incorporated breath sensor measured CH_4 of 801 dairy cows on 10 commercial farms. Using random cross validation a coefficient of determination in validation (R^2val) of 0.49 was found which suggest that milk

IR spectra are of interest for predicting CH_4 emission. However based on block cross validation, with farms as blocks, a negligible R^2val of 0.02 was obtained, indicating that milk IR spectra cannot be used to predict CH_4 emission independently. Random cross validation thus results in an overoptimistic view on the ability of milk IR spectra to predict CH_4 emission of dairy cows. The difference between both validation strategies can be explained by confounding of farm and date of milk IR analysis which introduces a correlation between batch effects on the IR analyses and the farm average CH_4. Breath sensor measured CH_4 is strongly influenced by farm specific conditions which magnifies the problem. Based on random cross validation also milk IR wavenumbers from water absorption regions, which contain mainly noisy information, showed moderate accuracy (R^2val=0.25) but not based on block cross validation (R^2val=0.03). These results suggest wavenumbers from water absorption regions as a negative control to identify potential dependence structures in the data.

Chapter 5 investigates the combined use of milk IR spectroscopy and genotypes of dairy cows on the accuracy of predicting milk fat composition. Milk fat composition based on gas chromatography and milk IR spectra were available for 1,456 Dutch Holstein Friesian cows. Genotypes of DGAT1, SCD1 and a single nucleotide polymorphism (SNP) located in an intron of fatty acid synthase (FASN) have significant effects on some milk fatty acids. The genotypes of them were incorporated and combined with milk IR spectra wavenumbers as predicators to milk fat composition. Adding DGAT1 genotypes to the milk IR spectra resulted in an improvement of the prediction accuracy for C16:1 cis-9 and C16 index. Adding SCD1 genotypes to the milk IR spectra resulted in a considerable improvement of the prediction accuracy for the unsaturated fatty acids C10:1, C12:1, C14:1 cis-9, C16:1 cis-9 and their corresponding unsaturation indices. Adding genotypes of the FASN SNP to the IR spectra did not improve prediction of milk fat composition. This study demonstrated the potential of combining milk IR spectra with genotypic information from 3 polymorphisms to predict milk fat composition. Prediction accuracy can be further improved by combining milk IR spectra with genomic breeding values.

Chapter 6 is the general discussion. The first topic addressed the differences in genetic background of milk IR spectra collected in winter and summer. The high genetic correlations of wavenumbers indicated that milk IR spectra collected in different seasons can be regarded as genetically the same trait. The second topic focused on predicting DGAT1 genotypes based on milk IR spectra. DGAT1 genotypes have very significant effects on milk IR wavenumbers, however the ability of milk IR wavenumbers to predict DGAT1 genotypes is low. The third topic was about to extract more information on milk composition captured by milk IR spectra. Some approaches such as analysing processed milk and applying different prediction methods are of interest.

Curriculum Vitae

About the author

Qiuyu Wang was born on 2nd November 1989, in the city of Hefei in China. In 2008, he graduated from Hefei No. 1 High School and wanted to be an engineer of auto vehicles. However he went to Nanjing Agricultural University studying Animal Pharmacy in the college of Veterinary Medicine. Then he entered an intensive training class held by colleges of Veterinary Medicine and Animal Science & Technology. In the 3rd year of his bachelor he decided to choose the direction of animal science, and during the winter holiday he got an opportunity to visit Wageningen University (for a personal purpose, not academic). The experience during this visit made him decide to continue master study in Wageningen after he completed his bachelor with honor in 2012.

Qiuyu was always one of the tallest in school until he arrived at Wageningen. He started master program in Animal Science. He took compulsory courses of both Animal Breeding and Genetics, and Animal Nutrition. He finally chose Animal Breeding and Genetics as his specialization. In 2014 he finished his master and received the scholarship from China Scholarship Council to continue PhD study. The results of his PhD are presented in this thesis entitled "The genetic background of bovine milk infrared spectra".

Qiuyu is (relatively) good at tennis and football. He was one of the top tennis players among all undergraduates in Nanjing. He served as the captain of the Chinese football team in Wageningen, and won the champion of "Chinese in the Netherlands Football Tournament" as the best scorer in 2015. He also has special interest in geography, tourism and photography.

List of publications

Wang, Q., Hulzebosch, A., and Bovenhuis, H. (2016). Genetic and environmental variation in bovine milk infrared spectra. Journal of dairy science, 99(8), 6793-6803.

Wang, Q., and Bovenhuis, H. (2018). Genome-wide association study for milk infrared wavenumbers. Journal of dairy science, 101(3), 2260-2272.

Wang, Q., and Bovenhuis, H. Validation strategy can result in overoptimistic view on the ability of milk infrared spectra to predict methane emission of dairy cattle. (accepted by Journal of Dairy Science).

Wang, Q., and Bovenhuis, H. Combined use of milk infrared spectra and genotypic information improves the prediction of bovine milk fat composition. (submitted to Journal of Dairy Science).

Training and Supervision Plan (TSP)

The Basic Package (3 ECTS)	**Year**
WIAS Introduction Day (mandatory)	2015
Course on philosophy of science and/or ethics (mandatory)	2017
Course on essential skills (Frank Little) (recommended)	2015
Disciplinary Competences (16 ECTS)	
Writing own research proposal	2015
Get started of ASReml	2015
Spectroscopy and Chemometrics training	2015
Genotype by environmental interaction, uniformity and stability	2015
EGSABG Summer Course "Emerging technologies in animal breeding"	2017
Methagene Course "Breeding for complex traits"	2017
Design of Breeding Programs with Genomic Selection	2017
Dairy Protein Biochemistry	2018
Quantitative genetics Discussion Group	2015-2018
Professional Competences (6 ECTS)	
Data Management Planning	2015
Scientific Writing	2015
Scientific Writing	2015
Reviewing a Scientific Paper	2016
Project and Time Management	2016
Brain Training	2018

Presentation Skills (4 ECTS)

Poster presentation in Final OptiMIR Scientific and Expert Meeting	2015
Poster presentation in WIAS Science Day 2016	2016
Oral presentation in EAAP 2016	2016
Oral presentation in EAAP 2017	2017

Teaching competences (6 ECTS)

Supervising practical Genetic Improvements of Livestock	2016-2017
MSc Supervision	2017-2018

Education and Training Total (35 ECTS)*

*One ECTS credit equals a studyload of approximately 28 hours

Acknowledgements

Unfortunately I'm not an emotional person good at profoundly expressing my feelings, but still I have something to say.

This thesis would have been impossible without help and support from many people. I must wholeheartedly acknowledge my promoter and daily supervisor, Henk Bovenhuis. Dear Henk (this appears uncountable times in my emails during recent years), we met in the course of GIL in 2012 when I was a 1st year MSc student, and 2 years later you gave me the opportunity to start my PhD journey. You became my promoter after Johan's dimission. We know the life in this 4 years was not easy. You behaved as the instructor that allow me sitting on the driving seat of my project. At this moment it doesn't matter how many times I complained about your rigor, I really appreciate your patience, profession, creative ideas and critical thinking. I could not have finished my PhD without your help, but what I learned more is the way you think.

I would like to thank the secretaries at ABG, especially Lisette, for your help and support on my project and study life during these years. Thanks Alex for your support on data analysis manuscript. And many thanks all my colleagues in ABG family, e.g. Angela, Biaty, Claudia, Esther, Floor, Gareth, Haibo, Ibrahim, Jeremie, Jovana, Kasper, Langqing and Zhou, Mandy, Maria, Pascal, Sabine, Sanne, Siyuan, Shuwen, Tessa, Tom among others. We shared great moments that will never be eliminated in my memory.

I'd also like to thank my supervisor committee members, Erik Mullaart, Jeroen Heck, Jan Rademaker and Harrie van den Bijgaart, thank you for your time and kind advice.

I am very grateful to all my friends in Wageningen. Xuezhu, thank you for your unreserved support throughout my PhD. Yang Chen, you are a reliable tennis partner that always being active. I hope you keep playing hard and get your driving license as soon as possible. Jun Qiu, we knew each other since our first day in Wageningen and now we both become a PhD, thank you for your many "like" in WeChat moments. Many thanks to my friends, Yu Jiang, Cong Bao, Ruxin He, Zhengwei Wu, Hao Ye, Wei Xu, Yuan He, Sheng Zhang, Wenbiao Shi, Yiru Wang, Yuxi Deng etc. I have special acknowledgement to our Chinese football team, Team Dragon. Playing with

you is one of my best memory these years. Remember, fixed barrack, floating soldiers.

My friends from outside Wageningen have always been a strong support for me as well: Ting Liao, Wenyang Dong, Hao Ren, Lei Wu, Yanan Jiang, Yujie Liu, Caifang Ren, Shizhi Wang. It was great to see some of you in Europe and even travel together. I wish you all the best, no matter in your scientific career or personal life.

During these years I met some of my former teachers in my bachelor university in the Netherlands or other country in Europe. Special thanks to Professors Wen Yao, Feng Wang, Yanli Zhang, Jin Cui and Jie Chen. I highly appreciate your precious advice.

Last but not least, thank my mother and my family. It is you who make me never walk alone. I'm going to use the rest of my life to acknowledge.

Thank you all if you are reading this.

Qiuyu Wang

Colophon

The work of this thesis is part of the Milk Genomics Initiative, funded by Wageningen University (Wageningen, the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), CRV, the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands).

Qiuyu Wang was sponsored by a Chinese Scholarship Council (CSC) fellowship.

Cover design: Ying Wang and Qiuyu Wang

Printed by Digiforce, De Limiet 24, 4131 NR Vianen, the Netherlands