Fine mapping and genomic prediction for detailed milk composition

Grum Gebreyesus Teklewold

Supervisors committee in Aarhus University

Main supervisors
Dr Albert Johannes Buitenhuis
Associate professor, Center for Quantitative Genetics and Genomics
Aarhus University, Tjele, Denmark

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

Co-supervisor
Prof. Dr Mogens Sandø Lund
Professor, Center for Quantitative Genetics and Genomics
Aarhus University, Tjele, Denmark

This research was conducted under the joint auspices of the Graduate School of Science and Technology (GSST), Aarhus University, Denmark and the Graduate School of Wageningen Institute of Animal Sciences (WIAS), the Netherlands as part of the Erasmus Joint Doctorate Program "EGSABG"

Thesis committee Wageningen University

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

Co-promotors
Dr Albert Johannes Buitenhuis
Associate professor, Center for Quantitative Genetics and Genomics
Aarhus University, Tjele, Denmark

Prof. Dr Mogens Sandø Lund
Professor of Center for Quantitative Genetics and Genomics
Aarhus University, Tjele, Denmark

Other members
Prof. Dr Martino Cassandro, University of Padova, Italy
Prof. Dr Walter Gerrits, Wageningen University & Research, Netherlands
Dr Freddy Fikse, Växa Sverige, Sweden
Dr Søren Borg, Aarhus University, Denmark

This research was conducted under the joint auspices of the Graduate School of Science and Technology (GSST), Aarhus University, Denmark and the Graduate School of Wageningen Institute of Animal Sciences (WIAS), the Netherlands as part of the Erasmus Joint Doctorate Program "EGSABG"

Fine mapping and genomic prediction for detailed milk composition

Unraveling the genetics of milk composition towards implementing selective breeding

Grum Gebreyesus Teklewold

Thesis

submitted in fulfillment of the requirements for the joint degree of doctor between
Aarhus University
by the authority of the Head of Graduate School of Science and Technology, and
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 Head of Graduate School of Science at Aarhus
University and by the Academic Board of Wageningen University
to be defended in public
on Friday December 14, 2018
at 12.30 p.m. in Foulum, Aarhus University.

Grum Gebreyesus
Fine mapping and genomic prediction for detailed milk composition.

PhD thesis, Aarhus University, Denmark and Wageningen University, the Netherlands (2018)

With summaries in English, Danish and Dutch

ISBN: 978-87-93643-92-5
DOI: https://doi.org/10.18174/461864

Abstract

Gebreyesus, G. (2018). Fine mapping and genomic prediction for detailed milk composition. PhD thesis, Aarhus University, Denmark and Wageningen University, the Netherlands.

Suitability of milk for processing into high-value products, such as cheese and butter, is affected by its protein and fatty acid (FA) composition. In addition, there are consumer concerns to some specific components of milk, whilst increasing preferences for others, mainly on health grounds. Therefore, economic and consumer pressures are driving interests in altering the detailed protein and FA composition of milk. Among potential strategies to alter detailed milk composition, genetic improvement provides cumulative effects carried over generations for a unit of investment. Selective breeding requires large-scale availability of data for accurate estimation of genetic parameters and prediction of genetic merits. Measurement for detailed milk protein and FA traits is currently limited to experimental scales due to costly and time-consuming analytical techniques. The PhD study aimed at improving accuracy of genetic parameter estimation and prediction of breeding values as well as understanding the genetic backgrounds of detailed milk protein and FA traits using efficient quantitative approaches. It is shown that improved accuracy of parameter estimation and genomic prediction is possible for scarcely recorded traits using multi-trait analyses. Advantages of existing multi-trait models is limited when genetic correlation between analyzed traits is weak. With a novel Bayesian model considering heterogeneous correlation structures over the genome, we show that despite weak genome-wide correlation, there exist genomic-regions explaining strong correlation and that it is possible to utilize such "local" correlations for accurate multi-trait genomic prediction. Combining datasets from different populations of a breed was another strategy investigated and shown to benefit genome-wide association (GWA) and genomic prediction for scarcely recorded traits. It is also demonstrated that existing linear genomic prediction models can be extended to incorporate GWA findings for further gain in prediction accuracy. Post-GWA analyses with multiple data sources including tissue-specific gene expression, ontology and pathway information can help refine GWA findings and provide potent information for genomic prediction models. Novel genomic regions and candidate genes established in the study contribute to the knowledge base on the complex genetic backgrounds of milk FA traits. The findings suggest that genomic selection for detailed milk composition is possible. Novel methods presented in the thesis will be of value for genomic prediction in other scarcely recorded traits of economic importance.

Contents

1

General introduction

1.1. Milk production and dairy processing

According to the FAOSTAT database, milk is produced in all countries of the world. Cow's milk in particular has been an important basis of animal-source protein across different cultures and socio-economic classes globally (FAO, 2017). Milk and dairy products are vital components of human nutrition. Especially early in life, milk provides all the nutrients crucial for vitality and development. Rich in essential nutrients, milk makes substantial contribution to meeting the human body's needs for calcium, magnesium, selenium, riboflavin, vitamin B12 and vitamin B5 (FAO, 2013). Several proteins essential for human nutrition are present in milk. For instance casein, which contains many essential amino acids, makes up approximately 80% of the total proteins in milk (Farrell et al. 2004).

The global production of cow's milk in the year 2016 was estimated at 700 million tons (IDF, 2017). With a staggering share of 24% of the global cow's milk production in 2016, the European Union (EU-28) stands as the single biggest milk producer in the world (Eurostat, 2017). According to the Eurostat database, 96.8 % of the whole milk available to the dairy sector in the European Union member countries in 2016 was processed into different products including 9.6 million tons of cheese and 2.4 million tons of butter and other products. In recent years, the global dairy sector has witnessed market price variability and major shifts in consumer preferences prompted with changes in life style. Future projections indicate an overall decrease in consumption of fluid milk and growing demand for processed products such as cheese. Following the ascendance of the fast food industry, more dairy products will be used as ingredients to meet the rising demand for processed foods such as pizza or pastry (European Commission, 2017). World market price variability, and the projected continuation of such commodity price volatility, has also prompted interests in producing more high-value products like cheese, butter or infant formula. In the EU, despite the lifting of milk quota in 2015, only slight increase was observed in the volume of milk production in 2016, while at the same time the price of farm milk continued to fall (Eurostat, 2017). The projected growth in milk production over the coming decade also stands only at 0.8% per year, while the projections for cheese production alone over the same outlook period is estimated at 1.4% per year (OECD/FAO, 2017). Such global and regional trends place the production of specialized value-added milk products at the center of the future of the dairy industry.

1.2. Detailed milk composition on the spotlight

Suitability of whole milk for processing into high-value products is affected by its protein and FA profile. In addition, the protein and FA composition of milk affect its nutritional value. There are increasing consumer concerns over certain protein and FA components of milk in relation to health. Simultaneously, there are increasing preferences to other components, such as the omega FAs. Such economic and societal pressures have led to mounting interests on the detailed protein and FA composition of milk from farmers, the dairy industry and researchers.

1.2.1. Milk protein composition

Bovine milk is composed of several proteins, the majority of which are the caseins (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN), that collectively make up around 80% of total protein in cow's milk (Farrell et al., 2004), and the whey proteins (α-LA and β-LG). Of the casein proteins, β-CN and α_{S1}-CN were reported as the two most abundant components of the total milk protein across different cattle breeds (e.g. Bevilacqua et al., 2006, Schopen et al., 2009). The casein proteins also exist in several post-translational modified forms such as phosphorylation and glycosylation. Of the casein proteins, α_{S1}-CN is heavily phosphorylated which in some studies is linked to common milk allergic reactions in early childhood (e.g., Høst, 2002). The α_{S1}-CN has two common phosphorylation states, with eight (α_{S1}-CN-8P) and nine (α_{S1}-CN-9P) phosphorylated serine residues. Apart from phosphorylation, some milk proteins also undergo glycosylation: a posttranslational modification occurring through the action of various glycosyltransferases (O'Riordan et al., 2014). Among the caseins, κ-CN is known to exist in abundance in glycosylated forms.

The relative concentration of the different proteins in milk have been shown to affect its technological properties that are essential for the profitability of the dairy processing industry. Milk coagulation, for instance, is affected by its protein composition (Jõudu et al., 2008; Bonfatti et al., 2013). An increase in the proportion of caseins is known to increase cheese yield. Variations in concentrations of the different proteins in the milk also underlie differences in rennet clotting time and curd firmness (Wedholm et al., 2006; Jõudo et al., 2008). Higher concentration in the milk of the glycosylated form of κ-CN has also been shown to reduce rennet coagulation time (e.g. Bonfatti et al., 2014). The concentration of β-LG in the milk is also reported to affect the heat stability of milk; increases in β-LG correspond to increased heat stability (Feagan, 1979). In addition to its relevance for cheese production, protein composition might be of relevance for other aspects of dairy products and dairy processing; for example, production of infant formula (Rutten et al., 2011).

1.2.2. Milk fatty acid composition

The fatty acids (FAs) in milk can be classified on basis of their chain length i.e., short chain (4 - 10 carbons), intermediate (12-16) and long chain FAs (18-22 carbons). FAs are also classified according to degree of saturation. Accordingly, FAs with only single bonds in the carbon chains are saturated FAs, whereas FAs with at least one double bond are termed as unsaturated FAs. Depending on the number of double bonds, the unsaturated FAs can be further classified as monounsaturated (one double bond) (MUFA) or polyunsaturated (more than one double bond) (PUFA). Milk FAs are also further categorized based on isomerization. Approximately 70% of the FAs in milk are believed to be saturated (Givens, 2010). FAs arise in milk via different pathways. The short and intermediate chain FAs (C4-C14), and approximately 50% of C16:0, are *de novo* synthesized in the mammary gland (Chilliard et al., 2000). Acetate, from fermentation of carbohydrates in the rumen, and hydroxybutyrate, formed from absorbed butyrate in the rumen epithelium, are believed to be major sources of carbon in the *de novo* synthesis of FAs in ruminants (Bauman and Griinari, 2003). In addition, there are also evidences that some of the FAs are synthesized *de novo* from propionate in ruminant tissues, including the mammary gland (Massart-Leën and Massart, 1983; Vlaeminck et al., 2006). There are also indications that some of the short-chain FAs might arise from the feed. For instance, the study of Heck et al., (2012) indicated that C12:0 is not completely synthesized *de novo* but rather also partly derived from the feed. Most short chain FAs are not present in plant-based feed for ruminants. The findings of Heck et al. (2012) however suggest that if cows were fed with short chain FAs of other sources, they would probably end up in the milk.

The long chain FAs (≥C18), and the rest of C16:0, are suggested to be derived from blood lipids (Chilliard et al., 2000). Majority of these circulating FAs are of dietary and microbial origin absorbed from the intestine which are taken up from circulation by the mammary epithelial cells (Bauman and Griinari, 2001). The remaining circulating FAs taken up by the mammary epithelial cells originate from mobilization of body fat reserves (Bauman and Griinari, 2003). FAs assumed to be mainly deposited in body fat reserves include C16:0 and C18:0 (Clarke, 1993). Proportion of milk FAs arising from mobilized body fat reserves is estimated to be between 4% and 8% in cows at positive energy balance (Pennington and Davis, 1975; Bauman and Griinari, 2003). However, the importance of FA synthesis from body fat reserves increases when cows are in negative energy balance (Balmain et al., 1952).

The FA composition in the milk affects its physical properties and processing abilities such as melting point and hardness of butter, crystallization and fractionation of milk fat (Chilliard, 2000). Milk FA composition has also attracted mounting interests in relation to the effects on its nutritional quality. In human nutrition, the FA composition of milk is a determinant of its risk factor status for several diseases. Saturated FAs in the diet, specially the C14:0 and C16:0 (German and Dillard, 2006), which is believed to be the most abundant FA in milk (Givens, 2010), have been linked to raised plasma cholesterol concentrations. In contrast, some group of FAs are considered beneficial in brain development and reducing cardiovascular disease risks. Specifically, the poly-unsaturated omega3 FA has for long been a target of high research interest in relation to perceived beneficiary effects on human health, particularly in relation to coronary heart disease risks (e.g. Ascherio et al., 1996). A recent meta-analysis study involving ten trials with over 70,000 sampled individuals (Aung et al., 2018) however found no evidence to support the claims of beneficial association of omega3 FA supplement in diets with cardiovascular disease risks. The omega3 FAs arise in milk mainly from grass-based feeding. Studies have shown that dairy production systems with relatively more grazing and fresh pasture based systems allow production of milk with more omega3 concentrations. Organic milk, for instance, was shown to have significantly higher proportions of omega3s than conventional milk (Benbrook et al., 2013; Hein et al., 2018).

1.3. Prospects to include detailed milk composition in breeding goals

Milk yield, recorded throughout lactation, forms the first category (Boichard and Brochard., 2012) of the 30 – 40 traits commonly recorded (Banos, 2010) in present dairy cattle selection schemes across several developed countries. Recordings for milk composition traits are currently limited to protein, fat and lactose contents as well as somatic cell counts analyzed using mid-infrared (MIR) spectrometry.

Economic and societal pressures stipulate implementation of selective breeding to alter milk protein and FA composition towards increasing yield of high-value products and addressing consumer concerns over specific components. The importance of including these traits have long been indicated (Boichard and Brochard, 2012). The rationale behind such calls back then are even more concrete today. Moreover, new technologies are now increasingly available to set and support new breeding goals (Merks et al., 2012). However, inclusion of new phenotypes in existing breeding goals comes with its own challenge and requires consideration of several other criteria apart from the economic values of the potential trait(s). Primarily, the expected genetic gain from selection for the candidate trait should outweigh the costs required for routine evaluations, among other things. The yearly genetic is defined as:

$$\Delta \boldsymbol{G} = \frac{\boldsymbol{i} * \boldsymbol{r} * \boldsymbol{\sigma_a}}{\boldsymbol{L}},$$

Such that *i* is the selection intensity, *r* is the prediction accuracy, $\boldsymbol{\sigma_a}$ the additive genetic standard deviation, while *L* is generation interval. Therefore, genetic gain is directly proportional to prediction accuracy, section intensity and genetic standard deviation and inversely related to the generation interval. Different factors affect prediction accuracy including sources and amount of information used and the prediction strategy applied. The amount of information available for prediction of genetic merits depends on availability of rapid and cheaper analytical techniques of measuring the phenotypes of interest.

1.3.1. Analytical methods for determining milk composition traits

Over the past decades, different methods have been introduced for separation and quantification of detailed protein contents in foods. Some among these methods include polyacrylamide gel electrophoresis with urea (urea – PAGE) (Farkye et al., 1991), capillary zone electrophoresis (Recio et al., 2001), mass spectrometry (MS) methods (Mann et al., 2001) and reversed–phase high performance liquid chromatography (RP-HPLC) (Veloso et al., 2002). Currently, the electrospray ionization mass spectrometry (ESI-MS) methods are most commonly used to quantify detailed milk protein composition. These methods combine physical separation techniques of liquid chromatography (LC) with the mass analysis capabilities of mass spectrometry (MS) (Fenn et al., 1989).

The reversed–phase high performance liquid chromatography (RP-HPLC) has also been applied for separation of FAs in the milk (e.g. Gresti et al., 1993). However, gas chromatography (GC) method is most commonly used due to its efficiency and is considered as the golden standard in quantifying milk FA composition. Nevertheless, its requirement for preparation of an esterified compound has been a disadvantage in terms of the time consuming process and requires specialized skills (Rodriguez et al., 2014).

Both the LC/ESI-MS (for milk proteins) and the GC (milk fats) are currently considered the golden standards in quantifying detailed milk protein and fat composition traits highly accurately. However, the methods are costly and time taking limiting the feasibility of large-scale phenotyping for the detailed milk protein and fat composition traits. Samples determined with such methods are currently limited to small and experimental scales. Recently, mid-infrared (MIR) and Fourier transform infrared (FTIR) spectrometry methods for predicting milk protein (e.g. De Marchi et al., 2009; Bonfatti et al., 2011a; Ferrand et al., 2012) and FA composition traits (e.g. Soyeurt et al., 2006; De Marchi et al., 2011) have attracted interests as alternatives to otherwise expensive LC/ESI-MS and/or GC quantified phenotypes. Studies continue to look into the concordance between golden standard measured phenotypes and IR predicted measurements (e.g. Poulsen et al., 2014; Rodriguez et al., 2014).

1.3.2. Genetic parameters

Genetic parameters, including heritability of the trait and the correlations with other traits in the breeding goal, are crucial information in planning for selective breeding. Genetic parameters are essential to compare the significance of genetic factors vis-a-vise non-genetic factors across traits and in different environments. Particularly, heritability estimates enable prediction of the response to selection and are key inputs in designing breeding schemes. Genetic parameters in populations are not constant (Visscher et al., 2008) and can change due, for instance, to inbreeding (Wade et al., 1996) or selection (Beniwal et al., 1992). Hence, estimation of genetic parameters for quantitative traits, in relation to designing and/or evaluation of breeding schemes, is not a one-time activity, and rather requires periodic updating.

Few studies reported genetic parameters estimates for detailed milk protein composition measured using LC/EMS (e.g. Schopen et al., 2009; Bonfatti et al., 2011b; Buitenhuis et al., 2016). These studies show medium to high heritability for most of the milk protein traits. Similarly, scanty information is available on the genetic parameters of milk FA traits quantified using the GC technique (e.g. Stoop et al., 2008; Krag et al., 2013; Bilal et al., 2014). Across these different studies, moderate to high heritability estimates have been reported for most *de novo* synthesized FAs while generally lower heritability estimates were reported for the intermediate and long chain FAs.

Studies reporting genetic parameters on LC/EMS quantified milk protein and GC measured milk fat composition traits are based on limited samples due to associated expenses. Sampling error in estimation of heritability values is a function the sample size and relatedness structure, as well as bias due to confounding (Visscher et al., 2008). Therefore, thousands of observations might be needed to attain very precise estimates (Visscher et al., 2008).

Genetic parameter estimation for scarcely recorded traits might benefit from the use of genetic markers (Visscher et al., 2008) by estimating the realized relationship as opposed to expected relationships from pedigree information. Especially with the increasing affordability of high density (HD) genotyping, it is now possible to capture most of the variants with small to large effects and hence allowing more accuracy in estimating genetic parameters. At the phenotypic level, availability of information on multiple related traits and simultaneous estimation of heritability estimates might improve accuracy of estimates. Some studies based on real (Eaglen et al., 2012) and simulated data (e.g. Mathew et al., 2016) indicate that different variance components can be estimated with higher accuracies and/or lower standard errors of prediction using multi-traits analysis. By allowing incorporation of information on genetic correlations into analysis, multi-trait models significantly improve the accuracy of the genetic parameter estimates. In chapter 2, we implement multi-trait analysis for estimation of genetic parameters in the detailed milk protein composition traits based on relationship matrix computed using genotype data imputed to full sequence.

1.3.3. Genomic selection

For the past several decades, dairy cattle breeding schemes involved estimation of breeding values from performances recorded on candidates and their relatives using the pedigree relationship. This relied on progeny testing schemes to ensure accuracy of selecting candidates for sex-limited traits such as milk yield, therefore, requiring costly facilities and resulting in longer generation intervals. Following the availability of genetic markers, the emergence of genomic selection has revolutionized the cattle breeding system. Genomic selection refers to the selection of breeding candidates in quantitative models that make use of information from all genetic markers available throughout the genome. The use of genetic-marker information in statistical models for the prediction of genetic merits, also known as marker-assisted selection, has been around for decades. Initially, such approaches were based on genotyping animals for only a few markers to detect QTL with linkage studies and using the information for prediction of breeding values. In contrast, many QTLs often with small effects control quantitative traits and as a result, the proportion of the genetic variance explained by the QTL limits the benefit from marker-assisted selection. Rethinking this approach, Meuwissen et al. (2001) showed the possibility of simultaneously estimating the effects of all available genetic markers covering the whole genome to predict genetic merits and/or future phenotypes. Prediction of breeding merits in such approach requires the estimation of the effects of all available markers in a reference population of individuals that are both phenotyped and genotyped. The estimated effects of the genome-wide markers are then used to predict the genomic estimated breeding values (GEBVs) in a training population that only have genotype data. With phenotypic records no longer necessary for all selection candidates, selection for sex-limited traits at younger ages has resulted in considerable reductions in costs and generation intervals.

1.3.3.1. Genomic prediction models

With the possibility of simultaneously estimating the effects of all available genome-wide markers, genomic prediction, also referred to as the "black box" method, would allow prediction of genetic merits without the need to identify the underlying causal QTLs. This statistical breakthrough was later followed by the first sequencing of the bovine genome (Consortium, 2009), which led to availability of many thousands of variants in the form of single nucleotide polymorphisms (SNPs). Technological advances over the last decades have led to rapid reduction of genotyping cost, resulting in the availability of thousands of genotyped reference animals. With the emergence of next generation sequencing and significant developments in genotype imputations techniques, it is now possible to capture large proportions of the genomic variants. With such increasing availability of hundreds of thousands of genetic markers, often much bigger than the number of genotyped individuals, an issue with the simultaneous estimation of all marker effects has been the dimensionality of the marker data.

Meuwissen et al. (2001) presented different methods of overcoming dimensionality of genotype data often larger than genotyped animals: least square estimation, BLUP-based approaches (SNP-BLUP) and Bayesian methods. The BLUP based approaches assume SNP effects to come from a normal distribution and have the same variance. An extension of this method, GBLUP (VanRaden, 2008) constructs genomic relationship matrix replacing the additive relationship matrix used in traditional BLUP approach. However, for some traits, the assumption of normally distributed QTL effects does not fit well to prior knowledge of QTL with major effects. Hence, various Bayesian approaches, including the initial Bayes A and Bayes B models of Meuwissen at al. (2001), have been suggested with alternative prior assumptions for QTL effects. Bayes B assumes that π proportion of the markers have no effect on the trait, while a prior distribution of inverted chi-square is assumed for the rest (1- π) markers. Bayes A assumes all markers have effects with a prior distribution of an inverted chi square. Thus, Bayes A can be considered as special case of Bayes B where the proportion π of markers with no effect is assumed zero. Series of extensions and variants of these methods in what Gianola, (2013) described as “the Bayesian alphabets” continue to be proposed. The most of known of these is the Bayes C_{π} suggested by Habier et al. (2011). The proportion π of markers with zero effects is treated as known in Bayes A (π =0) and Bayes B (π > 0). Arguing that π should be treated as an unknown and inferred from the data as the shrinkage of SNP effects is affected by π, Habier et al. (2011) suggested Bayes C_{π} where π is assigned a prior and estimated during the analysis. Other Bayesian models suggested include Bayes R (Erbe et al., 2012), in which variants are assigned to one of several normal distributions with different variances. Most of the suggested Bayesian models commonly rely on estimation of locus-specific (co)-variances, leading to estimation of too many parameters as genotype data often contained many thousands of markers. This is especially problematic when available phenotypic data is limited, such as in scarcely recorded traits, where there is little information to estimate thousands of parameters. In such cases, Gianola et al. (2009) suggests to group markers, where markers within a group explain similar variances, to estimate group-specific variances and hence reducing the number of parameters to estimate.

Studies have compared the different Bayesian and GBLUP prediction models for prediction accuracies and computational efficiency. For most traits, Bayesian models showed similar accuracy or small superiority to the GBLUP approaches (Hayes et al., 2009; Daetwyler et al., 2010; Clark et al., 2011). Larger advantages in prediction accuracies were however reported for the Bayesian models for traits controlled by few QTL (Cole et al., 2009; Legarra et al., 2011) and when the genetic relationship between reference populations and selection candidates is weak (Gao et al., 2013; Van den Berg et al., 2015). However, typical Bayesian models are implemented in Markov chain Monte Carlo (MCMC) algorithms that are computationally demanding (Mossel and Vigoda, 2006) and are not straightforward. Therefore, application of the Bayesian models for routine evaluations is so far limited.

1.3.3.2. Computing accuracy of genomic prediction

Cross-validation strategies are commonly used to calculate genomic prediction accuracies while some deterministic methods have also been proposed (e.g. Goddard, 2009; Wientjes et al., 2015a). Cross-validation strategies commonly involve dividing the studied individuals into training and validation sets. The training population will have both phenotypes and genotypes used to estimate marker effects. Phenotypes of the validation population are masked and their genomic breeding values (GEBVs) are predicted based on their genotypes and the marker effects estimated on the training population. The accuracy of genomic prediction in cross-validation is ideally defined as the correlation between the estimated GEBVs and true breeding values and. Since the true breeding values remain unknown in real scenarios, different approximations are followed to assess the reliability of genomic prediction.

Accuracy of genomic prediction is affected by several factors related to the studied population, the traits of interest, marker and QTL properties, and prediction model applied. Factors related to the population affecting prediction accuracy include reference and effective population sizes, which determines LD range as well as genetic relatedness within and between reference and validation population. The LD between available markers and the underlying QTL as well as the minor allele frequencies (MAF) of markers and the underlying QTLs (Wientjes et al., 2015b) also influences prediction accuracy. Genetic architecture of studied traits have also been shown to affect genomic prediction accuracy. Important factors in this regard include heritability of the trait in the studied population and the number of loci affecting the trait and the distribution of their effects (Daetwyler et al., 2010; Hayes et al., 2010; Gianola, 2013).

1.3.3.3. Genomic prediction for scarcely recorded traits

Some immerging traits of economic importance in livestock breeding are expensive and/or difficult to measure at large-scale limiting the available reference population size. Accuracy of genomic prediction for such scarcely recorded traits remain low. Small reference population is associated with low power to estimate marker effects leading to less accurate estimates of genetic merits of individuals. Approaches suggested to improve prediction reliabilities for such traits include multi-traits analyses and combining datasets from different population/breeds (e.g. Calus et al., 2011; Lund et al., 2011).

While most of the initially suggested genomic prediction models were developed on the basis of single-trait scenario, models that allow simultaneous estimation of breeding values for two or more traits have also been proposed (e.g. Calus et al., 2011; Jia and Jannink, 2012; Hayashi and Iwata, 2013). Multi-trait prediction for scarcely recorded traits allows the use information from other large-scale recorded indicator traits and relatives (Henderson and Quaas, 1976). Advantages from simultaneous evaluation for multiple traits have been shown to depend on the genetic correlation between the trait of interest and indicator traits (Calus et al., 2011), genetic architecture of the target traits (Jia and Jannick, 2012) and the prediction models used (Calus et al., 2011; Jia and Jannick, 2012). We show in chapter 3 that multi-trait genomic prediction models might benefit by accounting for heterogeneous correlation structures of the genome.

Other efforts to improve genomic prediction accuracy for scarcely recorded traits or populations of numerically small size have been to combine reference of multiple populations/breeds (de Roos et al., 2009; Lund et al., 2011; Erbe et al., 2012; Calus et al., 2018). The potential benefits of combining multi-population/breed datasets have been shown to largely depend on genetic relatedness between the populations (Habier et al., 2007; Lund et al, 2014), consistency in LD between the markers and QTLs affecting the trait (de Roos et al., 2008; Pryce et al., 2011) and consistency in allele substitution effects between the populations (Wientjes et al., 2015b). In chapter 6, we implement genomic prediction for the milk FA traits using combined reference populations of the Chinese, Danish and Dutch Holstein and investigate gains in prediction reliability compared to genomic prediction using population-specific data.

1.3.3.4. Genetic architecture and genomic prediction accuracy

The genetic architecture of traits, most importantly the number of QTLs affecting the trait and distribution of the effects sizes, affect accuracy of genomic prediction (Daetwyler et al., 2010; Hayes et al., 2010; Gianola, 2013). Larger differences in accuracies have been reported between different genomic prediction models for traits controlled by few QTL with large effects than traits following the infinitesimal genetic model (e.g. Cole et al., 2009; Hayes et al., 2010; Legarra et al., 2011; Gao et al., 2015). Unlike the assumptions behind the GBLUP models where polymorphisms across the genome explain similar variance, regions harboring QTLs of large to intermediate effect might explain larger proportions of in the genome-wide variance for traits controlled by few QTL. Different approaches can be adopted to account for such heterogeneous covariance structures to improve prediction reliabilities. In chapter 3, we develop and implement novel single- and multi-traits Bayesian models and show that improvement in prediction accuracies, compared to the traditional GBLUP and Bayes A models, is dependent on the genetic architecture of studied milk protein traits.

1.3.3.5. Augmenting prediction models with biological information

Apart from using priors to account for the number of QTL and the distribution of effects to improve prediction reliability, genomic prediction models might also benefit from identifying the underlying QTLs and the mechanistic pathways of control. While the black box method, where all markers are used without prior information on the causative QTL, have proved to work and has revolutionized the livestock breeding industry, such models incorporating biological information might be especially important to improve prediction reliability for difficult and/or expensive to measure traits. One effective strategy to detecting QTLs underlying quantitative traits has been the study of genome-wide associations between genetic markers and phenotypes of interest. In chapter 6, we show that genomic prediction using linear models assuming different effects for associated genomic regions improve prediction accuracies compared to the traditional linear models which implicitly assume all markers explaining a common (co)variance genome-wide.

1.3.4. Genome-wide association

Genome-wide association (GWA) refers to quantitative method to detect associations between genetic markers and mutations that affect a trait. Apart from enabling to understand the genetic mechanisms behind quantitative traits, findings of GWA studies can be incorporated into genomic prediction models to enhance prediction accuracies. While genomic prediction uses all available variants across the genome to predict GEBVs, there is enormous potential for GWA analysis to play in improving genomic prediction accuracy by selecting significantly associated variants for models that can put different weight for different group of variates.

Different quantitative approaches have been adopted in GWA experiments. The most common approach, at least in the livestock gene-mapping community, has been the implementation of linear mixed models where the effect of each marker is tested at a time and the relationship matrix, constructed from the pedigree or marker information, is used to account for the population stratification. The consequence of this approach, especially in cattle breeds with long range LD, is that all markers in LD with the causative QTL will highly likely show significant associations thus limiting the mapping precision. As the approach involves testing each SNP at a time, it also comes with the problem of multiple testing and the associated stringent corrections. Multi-marker models to simultaneously test all markers for association with quantitative traits have also been suggested in Bayesian framework. Several studies have accordingly reported genomic regions associated with quantitative traits based on Bayesian models developed for genomic prediction of breeding values (e.g. Guo et al., 2016; Speidel et al., 2018).

Association mapping studies have reported various regions of the bovine genome in connection to the milk protein and milk fat composition. For milk protein composition traits, polymorphisms in the β-casein, κ-casein and β-lactoglobulin genetic variants have been shown to underlie substantial proportion of the genetic variations (e.g. Heck et al., 2009) while several other regions have also been linked through association studies. Schopen et al. (2011), for instance reported that the *DGAT1* polymorphism in BTA 14 is associated with α_{S1}-CN and α_{S2}-CN contents in milk. Buitenhuis et al. (2016) also reported significant associations between the different milk protein composition traits and SNPs on several chromosomes across the bovine genome. Regions on BTA 14, 19 and 26 have been associated to the genetic variations in most of the milk FAs traits (Schennink et al., 2009a; Stoop et al., 2009; Bouwman et al., 2011; 2012; Li et al., 2014). The *DGAT1* gene have been suggested as the most likely candidate for the region on BTA 14 (e.g. Grisart et al., 2002) with the *K* allele of the DGAT1 K232A polymorphism shown to have significant effects on several FA traits (Schennink et al., 2007; Bovenhuis et al., 2016). However, discovery of other mutations in the *DGAT1* region affecting some FA traits (Lehnert et al., 2015) suggest that more candidate polymorphisms could be harbored in the region other than the DGAT1 K232A. On BTA 19, the *FASN* have been suggested as the candidate gene (Schennink et al., 2009b). However, using haplotype analysis Bouwman at al., (2014) showed significant effect of an additional nearby larger gene (*CCDC57*) for some FAs suggesting that there could be more QTLs than just *FASN* in the rather broader region frequently reported on BTA 19. The *SCD1* have been suggested as the candidate gene for the BTA 26 region connected to milk FAs (e.g. Mele et al., 2007). The *SCD* enzyme is shown to be involved specifically in the desaturation process to synthesize MUFAs through introducing a double bond in the delta-9 position (Ntambi and Miyazaki, 2003). However, the genomic regions on BTA14, 19 and 26 combined explain between 3.6 to 50 % of the genetic variations across the milk FA traits (e.g. Bouwman et al., 2011, 2012). Additional regions explaining fractions of the remaining genetic variations are thus expected. Some such additional regions have so far been reported (e.g. Bouwman et al., 2011, 2012; Li et al., 2014; Li et al., 2015; Buitenhuis et al., 2014). The synthesis and metabolism of FA in milk is a complicated process involving many pathways. Hence, with the use of high density markers and larger datasets more incites are to be expected on additional genomic regions explaining smaller to intermediate fractions of the genetic variation in the milk FA traits. In Chapter 5, we undertake a GWA scan for the milk FA traits based on a large multi-population dataset comprising the Chinese, Danish and Dutch Holstein Friesian and report novel, as well as previously reported genomic regions for the FA traits.

GWA studies so far reported for milk composition traits in general are often based on numerically smaller datasets due to the associated expense for quantifying the traits using the golden standard methods. Studies have recently resorted to use of FTIR predicted milk composition phenotypes, as cheaper alternatives to costly GC and/or LC-EMS measured phenotypes, for GWA studies (e.g. Sanchez et al., 2017;

Olsen et al., 2017; Knutsen et al., 2018). However, lack of detections of some of the well-established causal polymorphisms for the FA traits cast doubts over the reliability of such IR predicted phenotypes for GWA studies. For instance, GWA studies of Olsen et al. (2017) and Knutsen et al. (2018) using the FTIR predicted FA phenotypes in the Nordic Red cattle did not detect any significant association near the *DGAT1* region. Lack of segregation of the A variant of the DGAT1 K232A polymorphism have been suggested by the studies as the potential reason for the lack of detections in the *DGAT1* region. However, both GWA studies (Olsen et al., 2017 and Knutsen et al., 2018) did not also detect any significant associations in the *SCD1* region despite the fact that the *SCD1* allele is known to segregate in the Nordic Red cattle (Knutsen et al., 2018). Additionally, Wang et al. (2016) also reported no significant effect of the *SCD1* polymorphism was observed on any of the milk IR wavenumbers in samples from the Dutch Holstein.

The statistical power of GWA experiments to detect associations between markers and a quantitative trait depends on the sample size (Hong and Park, 2012), the distribution of effect sizes and the frequency of causal QTLs (unknown) in the population (Visscher et al., 2017), and the LD between the available markers and the causal QTLs.

Currently, with the increased availability of high density and sequence level genotypes, most GWA studies evaluate hundreds of thousands of SNP markers (Hong and Park, 2012), requiring much larger sample size to attain an adequate statistical power (Klein, 2007; Spencer et al., 2009). This makes it particularly challenging for scarcely recorded traits, such as the detailed milk protein and fat composition traits, to detect QTLs explaining intermediate to small effects.

Combining datasets from different populations and breeds have been implemented as strategy to increase sample size available for GWA studies. Such strategies combined either raw datasets for joint GWA (mega-analysis) (e.g. Veerkamp et al., 2012; Sanchez et al., 2017) or summaries of individual GWA studies for meta-analyses (e.g. Rubio et al., 2015; Bouwman et al., 2018). Advantages of combining multi-population datasets for GWA studies might be affected by the degree of heterogeneity between the samples to be combined. Such heterogeneity could arise due to genetic distance between the populations (Lund et al., 2014). In such cases, combining different populations of the same/related breed is more beneficial compared to combining different breeds. Differences between trait measurements and different environmental exposures can also cause heterogeneity by introducing genotype-by-environment interactions and thus adding more noise. In chapter 4, we use a multi-population dataset comprising the Chinese, Danish and Dutch Holstein population to investigate the advantages and challenges of different data combining strategies for multi-population GWA.

Apart from detection power, precision of detection is equally important for GWA studies to be of practical use in augmenting genomic prediction models. Due to the long range LD in cattle breeds, GWA studies often detect broad genomic regions. Such broad regions often contain several positional candidate QTLs/genes posing difficulty to untangle the true causative QTLs. Significant associations established using GWA between markers and a quantitative trait are not also directly informative of the causative gene or the mechanistic path through which the detected genomic region affects the associated trait. Integration of GWA and functional genomics (gene function annotation and ontology analyses) may help to prioritize positional candidate QTL regions for the true causative signals and take mechanistic insights into their effects (e.g. Hou et al., 2014; Littlejohn et al., 2016; Pegolo et al., 2017). The use of tissue-specific data (e.g. expressions of coding and regulatory (microRNA) genes) is also increasingly popular in the prioritization of GWA signals (e.g. Fang et al., 2018).

Therefore, despite the challenge in large-scale phenotyping, accuracy of predicting genetic merits for the detailed milk protein and FA composition might be possible through utilization of information from correlated traits, multiple related populations and the biology underlying the traits. This thesis explores such possibilities in the case of bovine milk protein and fatty acid composition.

1.4. Aim and outline of the thesis

The aim of this PhD thesis is two folds: 1) explore different statistical models and approaches that allow accurate genetic analyses including genetic parameters estimation, genome-wide association and genomic prediction with limited data; and 2) Applying developed methods to study the genetic backgrounds and implement genomic prediction for the detailed milk composition traits. In chapter 2, we examine the utility of multi-trait analyses in estimating the genetic parameters for detailed milk protein composition using numerically small dataset. In chapter 3, we develop and implement novel Bayesian single- and multi-trait genomic prediction models to improve prediction accuracy for the detailed milk composition traits by accounting for the genetic architecture and disentangling heterogeneous correlation structures with large-scale recorded related traits. Chapter 4 explores different data combining strategies to improve GWA detection power for detailed milk FA traits making use of multi-population datasets including the Chinese, Danish and Dutch Holstein populations. Chapter 5 further characterize genomic region detected for the milk FA traits using multi-population GWA and post-GWA analysis with multiple-data sources. In chapter 6, we use the data combining strategies investigated in chapter 4 to implement genomic prediction for the milk FA traits using multi-population reference in linear models that allow incorporation of the GWA findings from chapter 5. Finally, chapter 7 links findings of the different chapters to the broader context of implementing selective breeding for the milk protein and FA composition traits. General discussion on contributions of the PhD study to the knowledge base, research questions requiring future studies and opinions on current topics pertaining to the milk protein and FA composition, such as infrared prediction of phenotypes, are presented.

References

Ascherio A., Rimm E.B., Giovannucci E.L., Spiegelman D., Stampfer M., Willett W. C. (1996). Dietary fat and risk of coronary heart disease in men: Cohort follow up study in the United States. Br Med J 313 (7049) 84-90.

Aung T., Halsey J., Kromhout D., Gerstein H.C., Marchioli R., Tavazzi L., Geleijnse J.M., Rauch B., Ness A., Galan P., Chew E.Y., Bosch J., Collins R., Lewington S., Armitage J., Clarke R. (2018). Omega-3 Treatment Trialists' Collaboration. Associations of Omega-3 Fatty Acid Supplement Use With Cardiovascular Disease Risks: Meta-analysis of 10 Trials Involving 77 917 Individuals. JAMA Cardiol. 1;3(3):225-234. doi: 10.1001/jamacardio.2017.5205.

Balmain J.H., Folley S.J., Glascock R.F. (1952). Effects of insulin and of glycerol in vitro on the incorporation of [carboxy-14C]acetate into the fatty acids of lactating mammary gland slices with special reference to species differences. Biochem. J. 52:301–6

Banos G. (2010). Past, present and future of international genetic evaluations of dairy bulls. In Proceedings of the 9th World Congress of Genetics Applied to Livestock. Production, Paper 0033. German Society for Animal Science, Leipzig, Germany.

Bauman D.E., Griinari J.M. (2001). Regulation and nutritional manipulation of milk fat: low-fat milk syndrome. Livest Prod Sci 70:15-29.

Bauman D.E., Griinari J.M. (2003). Nutritional regulation of milk fat synthesis. Annu. Rev. Nutr. 23: 203-227.

Benbrook C.M., Butler G., Latif M.A., Leifert C., Davis D.R. (2013). Organic production enhances milk nutritional quality by shifting fatty acid composition: a United States-wide, 18 month study. PloS One 8:e82429.

Beniwal B. K., Hastings L. M., Thompson R., Hill W. G. (1992). Estimation of changes in genetic parameters in selected lines of mice using REML with an animal model.1. Lean mass. Heredity 69, 352–360.

Bevilacqua C., Helbling J.C., Miranda G., and Martin P. (2006). Translational efficiency of casein transcripts in the mammary tissue of lactating ruminants. Reproduction, nutrition, development 46(5):567-578.

Bilal G., Cue R.I., Mustafa A.F., Hayes J.F. (2014). Short communication: Genetic parameters of individual fatty acids in milk of Canadian Holsteins. J Dairy Sci.97(2):1150-6. doi: 10.3168/jds.2012-6508

Boichard D., Brochard M. (2012). New phenotypes for new breeding goals in dairy cattle. Animal. 6(4):544-50. doi: 10.1017/S1751731112000018.

Bonfatti V., Di Martino G., Carnier P. (2011a). 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(12):5776-85. doi: 10.3168/jds.2011-4401.

Bonfatti V., Cecchinato A., Gallo L., Blasco A., Carnier P. (2011b). Genetic analysis of detailed milk protein composition and coagulation properties in Simmental cattle. J Dairy Sci. 94(10):5183-93. doi: 10.3168/jds.2011-4297.

Bonfatti V., Gervaso M., Rostellato R., Coletta A., Carnier P. (2013). Protein composition affects variation in coagulation properties of buffalo milk. J Dairy Sci. 96(7):4182-90. doi: 10.3168/jds.2012-6333.

Bonfatti V., Chiarot G., Carnier P. (2014). Glycosylation of κ-casein: genetic and nongenetic variation and effects on rennet coagulation properties of milk. J. Dairy Sci. 97: 1961- 1969 doi: 10.3168/jds.2013-7418

Bouwman A.C., Bovenhuis H., Visker M.H., van Arendonk J.A. (2011). Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet. 11;12:43.

Bouwman A.C., Visker M.H., van Arendonk J.A., Bovenhuis H.(2012). Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. BMC Genet. 29;13:93. doi: 10.1186/1471-2156-13-93.

Bouwman A.C., Daetwyler H.D., Chamberlain A.J., Ponce C.H., Sargolzaei M., Schenkel F.S. et al. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 50(3):362-367. doi: 10.1038/s41588-018-0056-5

Bovenhuis H., Visker M.H.P.W., Poulsen N.A., Sehested J., van Valenberg H.J.F., van Arendonk J.A.M., Larsen L.B., Buitenhuis A.J. (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(4):3113-3123.

Bovine Genome Sequencing and Analysis Consortium. (2009). The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science. 324: 522-528. 10.1126/science.1169588.

Buitenhuis B., Poulsen N.A., Gebreyesus G., Larsen L.B. (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 Genet. 2;17:114.

Calus M.P., Veerkamp R.F. (2011). Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol. 43:26.

Calus M.P., Goddard M.E., Wientjes Y.C.J., Bowman P.J., Hayes B.J. (2018). Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. J Dairy Sci. 101(5):4279-4294. doi:10.3168/jds.2017-13366.

Chilliard Y., Ferlay A., Mansbridge R.M., Doreau M. (2000). Ruminant milk fat plasticity: nutritional control of saturated, polyunsaturated, trans and conjugated FA. Ann. Zootech. 49: 181-205.

Clark S.A., Hickey J.M., van der Werf J.H. (2011). Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol. 17;43:18.

Clarke S.D. (1993). Regulation of fatty acid synthase gene expression: An approach for reducing fat accumulation. J. Anim. Sci. 71:1957–1965.

Cole J.B., VanRaden P.M., O'Connell J.R., Van Tassell C.P., Sonstegard T.S., Schnabel R.D., Taylor J.F., Wiggans G.R. (2009). Distribution and location of genetic effects for dairy traits. J Dairy Sci. 92(6):2931-46. doi: 10.3168/jds.2008-1762.

Daetwyler H.D., Pong-Wong R., Villanueva B., Woolliams J.A. (2010). The impact of genetic architecture on genome-wide evaluation methods. Genetics. 185(3):1021-31. doi: 10.1534/genetics.110.116855.

De Marchi M., Bonfatti V., Cecchinato A., Di Martino G., Carnier P. (2009). Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Italian Journal of Animal Science 8, 399–401.

De Marchi M., Penasa M., Cecchinato A., Mele M., Secchiari P., Bittante G. (2011). Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal. 5(10):1653-8.

de Roos A.P., Hayes B.J., Spelman R.J., Goddard M.E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 179:1503–12.

de Roos A.P., Hayes B.J., Goddard M.E. (2009). Reliability of genomic predictions across multiple populations. Genetics. 183(4):1545-53.

Eaglen S. A., Coffey M., Woolliams J., Wall E. (2012). Evaluating alternate models to estimate genetic parameters of calving traits in United Kingdom Holstein-Friesian dairy cattle. Genetics Selection Evolution 44(1):23.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A., Goddard M.E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95, 4114 – 4129.

Eurostat. (2017). Milk and milk product statistics. Online publication: Data extracted in October 2017. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php/Milk_and_milk_product_statistics

European Commission. (2017). EU Agricultural outlook for the agricultural markets and income 2017-2030. Available online: https://ec.europa.eu/agriculture/sites/agriculture/files/markets-and-prices/medium-term-outlook/2017/2017-fullrep_en.pdf

Fang L., Sørensen P., Sahana G., Panitz F., Su G., Zhang S., Yu Y., Li B., Ma L., Liu G., Lund M.S., Thomsen B. (2018). MicroRNA-guided prioritization of genome-wide association signals reveals the importance of microRNA-target gene networks for complex traits in cattle. Sci Rep. 19;8(1):9345.

FAO, (2013). Milk and dairy products in human nutrition. Food and Agriculture Organization of the United Nations. Rome, 2013. E-ISBN 978-92-5-107864-8

FAO, (2017). The future of food and agriculture: Trends and challenges. Food and Agriculture Organization of the United Nations. Rome, 2017. ISSN 2522-722X

Farkye N.Y., Kiely L.J., Allshouse R.D., Kindstedt P.S. (1991). Proteolysis in mozzarella cheese during refrigerated Storage. J. Dairy Sci. 74 1433.

Farrell Jr., Jimenez-Flores H.M., Bleck G.T., Brown E.M., Butler J.E., Creamer L.K., Hicks C.L., Hollar C.M., Ng-Kwai-Hang K.F., and Swaisgood H.E. (2004). Nomenclature of the Proteins of Cows' Milk—Sixth Revision. Journal of dairy science 87(6):1641-1674.

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

Fenn J.B., Mann M., Meng C.K., Wong S.F., Whitehouse C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules.1989. Science 246:64–71

Ferrand M., Miranda G., Larroquet H., Guisnel S., Leray O., Lahalle F., Brochard M., Martin P. (2012). Determination of protein composition in milk by mid-infrared spectrometry. Presented at ICAR Meeting, 28 May-1 June, Cork, Ireland. Online: https://www.icar.org/wp-content/uploads/2015/09/Ferrand_PPT.pdf

Gao H., Su G., Janss L., Zhang Y., Lund M.S. (2013). Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. J. Dairy Sci., pp. 4678-4687

Gao N., Li J., He J., Xiao G., Luo Y., Zhang H., Chen Z., Zhang Z. 2015. Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model. BMC Genet. 14;16:120. doi: 10.1186/s12863-015-0278-9.

German J.B., Dillard C.J. (2006). Composition, Structure and Absorption of milk lipids: A source of energy, fat-soluble nutrients and bioactive molecules. Crit. Rev. in Food Sci. and Nutr. 46: 57-92.

Gianola D., de los Campos G., Hill W.G., Manfredi E., Fernando R. (2009). Additive genetic variability and the Bayesian alphabet. Genetics. 183:347–63.

Gianola D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. Genetics. 194(3):573-96. doi: 10.1534/genetics.113.151753.

Givens D.I. (2010). Milk and meat in our diet: good or bad for health? Animal 4, 1941-1952.

Goddard M., (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, (136)245–257.

Gresti J., Bugaut M., Maniogui G., Bezard J. (1993). Composition of molecular species of triacylglycerols in bovine milk fat. J. Dairy Sci. 76:1850–1869.

Grisart B., Coppieters W., Farnir F., Karim L., Ford C., Berzi P., Cambisano N., Mni M., Reid S., Simon P., Spelman R., Georges M., Snell R. (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.

Guo X., Su G., Christensen O.F., Janss L., Lund M.S. (2016). Genome-wide association analyses using a Bayesian approach for litter size and piglet mortality in Danish Landrace and Yorkshire pigs. BMC Genomics. 18;17:468.

Habier D., Fernando R.L., Dekkers J.C. (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics. 177(4):2389-97.

Habier D., Fernando R.L., Kizilkaya K., Garrick D.J. (2011). Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 23;12:186.

Hayashi T., Iwata H. (2013). A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics. 31;14:34. doi: 10.1186/1471-2105-14-34.

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

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J., Goddard M.E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6:e1001139.

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

Heck J.M., van Valenberg H.J., Bovenhuis H., Dijkstra J., van Hooijdonk T.C. (2012). Characterization of milk fatty acids based on genetic and herd parameters. J Dairy Res. 79(1):39-46. doi: 10.1017/S0022029911000641.

Hein L., Sørensen L.P., Kargo M., Buitenhuis A.J. (2018). Genetic analysis of predicted fatty acid profiles of milk from Danish Holstein and Danish Jersey cattle populations. J Dairy Sci. 101(3):2148-2157. doi: 10.3168/jds.2017-13225.

Henderson C.R., Quaas R.L.(1976). Multiple trait evaluation using relatives records. J Anim Sci. 1976;43:1188–97

Hong E.P., Park J.W. (2012). Sample size and statistical power calculation in genetic association studies. Genomics Inform. 10(2):117-22.

Hou L., Ma T., Zhao H. (2014). Incorporating functional annotation information in prioritizing disease associated SNPs from genome wide association studies. Sci China Life Sci. 57(11):1072-9. doi: 10.1007/s11427-014-4754-7.

Høst A. 2002. Frequency of cow's milk allergy in childhood, Ann. Allergy Asthma Immunology. 89: 33–37

IDF (International Dairy Federation). (2017). The world dairy situation 2017: Bulletin of the International Dairy Federation 489/2017. International dairy federation/Iederation Internationale du Lait. Brussels, Belgium.

Jia Y., Jannink J.L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 192(4):1513-22.

Jõudu I., Henno M., Kaart T., Püssa T., Kärt O. (2008). The effect of milk protein contents on the rennet coagulation properties of milk from individual dairy cows. Int. Dairy J. 18:964–967.

Klein R.J. (2007). Power analysis for genome-wide association studies. BMC Genet. 8:58.

Knutsen T.M., Olsen H.G., Tafintseva V., Svendsen M., Kohler A., Kent M.P., Lien S. (2018). Unravelling genetic variation underlying de novo-synthesis of bovine milk fatty acids. Sci Rep. 1;8(1):2179. doi: 10.1038/s41598-018-20476-0.

Krag K., Poulsen N.A., Larsen M.K., Larsen L.B., Janns L. and Buitenhuis B. (2013). Genetic parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a Bayesian approach. BMC Genet 14:79.

Legarra A., Robert-Granié C., Croiseau P., Guillaume F., Fritz S. (2011). Improved Lasso for genomic selection. Genet Res (Camb). 2011;93:77–87.

Lehnert K., Ward H., Berry S.D., Ankersmit-Udy A., Burrett A., Beattie E.M., et al. (2015). Phenotypic population screen identifies a new mutation in bovine DGAT1 responsible for unsaturated milk fat. Sci Rep. 2015 Feb 26;5:8484.

Li C., Sun D., Zhang S., Wang S., Wu X., Zhang Q., Liu L., Li Y., Qiao L. (2014). Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. PLoS One. May 23;9(5):e96186.

Li X., Buitenhuis A.J., Lund M.S., Li C., Sun D., Zhang Q., Poulsen N.A., Su G. (2015). Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. J Dairy Sci. 98(11):8152-63.

Littlejohn M.D., Tiplady K., Fink T.A., Lehnert K., Lopdell T., Johnson T., Couldrey C., Keehan M., Sherlock R.G., Harland C., Scott A., Snell R.G., Davis S.R., Spelman R.J. (2016). Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. Sci Rep. 5;6:25376.

Lund M.S., Roos A.P., Vries A.G., Druet T., Ducrocq V., Fritz S., Guillaume F., Guldbrandtsen B., Liu Z., Reents R., Schrooten C., Seefried F., Su G. (2011). A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol. 12;43:43.

Lund M. S., Su G., Janss L., Guldbrandtsen B., and Brøndum R. F. (2014). Genomic evaluation of cattle in a multi-breed context. Liv. Sci. 166: 101-110.

Mann M., Hendrickson R., Pandey A. (2001). Analysis of proteins and proteomes by mass spectrometry. Annu. Rev. Biochem. 70, 437–473

Massart-Leën A.M., Massart D.L. 1981. The use of clustering techniques in the elucidation or confirmation of metabolic pathways. Biochemical Journal 196 611–618

Mathew B., Holand A.M., Koistinen P., Léon J., Sillanpää M.J. (2016). Reparametrization-based estimation of genetic parameters in multi-trait animal model using Integrated Nested Laplace Approximation. Theor Appl Genet. 129(2):215-25.

Mele M., Conte G., Castiglioni B., Chessa S., Macciotta N.P., Serra A., Buccioni A., Pagnacco G., Secchiari P. (2007). Stearoyl-coenzyme A desaturase gene polymorphism and milk fatty acid composition in Italian Holsteins. J Dairy Sci.

Merks J.W., Mathur P.K., Knol E.F. (2012). New phenotypes for new breeding goals in pigs. Animal. 6(4):535-43. Review.

Meuwissen T.H., Hayes B.J., Goddard M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29

Mossel E., Vigoda E. (2006). Limitations of Markov Chain Monte Carlo algorithms for Bayesian inference of phylogeny. The Annals of Applied Probability. 16(4): 2215–2234. DOI: 10.1214/105051600000000538

Ntambi J.M. and Miyazaki M. (2003). Recent insights into stearoyl-CoA desaturase-1. Curr Opin Lipidol. 14:255–61

OECD/FAO 2017. OECD-FAO Agricultural Outlook 2017-2026: Special focus: Southeast Asia. OECD Publishing, Paris. http://dx.doi.org/10.1787/agr_outlook-2017-en

Olsen H.G., Knutsen T.M., Kohler A., Svendsen M., Gidskehaug L., Grove H., Nome T., Sodeland M., Sundsaasen K.K., Kent M.P., Martens H., Lien S. (2018). Genome-wide association mapping for milk fat composition and fine mapping of a QTL for de novo synthesis of milk fatty acids on bovine chromosome 13. Genet Sel Evol. 13;49(1):20. doi: 10.1186/s12711-017-0294-5.

O'Riordan N., Kane M., Joshi L., Hickey R.M. (2014). Structural and functional characteristics of bovine milk protein glycosylation. Glycobiology. 24(3):220-36.

Pegolo S., Dadousis C., Mach N., Ramayo-Caldas Y., Mele M., Conte G., Schiavon S., Bittante G., Cecchinato A. (2017). SNP co-association and network analyses identify E2F3, KDM5A and BACH2 as key regulators of the bovine milk fatty acid profile. Sci Rep. 11;7(1):17317. doi: 10.1038/s41598-017-17434-7.

Pennington J.A., Davis C.L. (1975). Effects of intraruminal and intra-abomasal additions of cod-liver oil on milk fat production in the cow. J. Dairy Sci. 58:49–55

Poulsen N.A., Eskildsen C.E.A., Skov T., Larsen L.B., Buitenhuis A.J. (2014). Comparison of genetic parameters estimation of fatty acids from gas chromatography and FT-IR in Holsteins. In proceedings: 10th World Congress of Genetics Applied to Livestock Production, August 17 - 22, 2014. Vancouver, BC Canada.

Pryce J.E., Gredler B., Bolormaa S., Bowman P.J., Egger-Danner C., Fuerst C., Emmerling R., Sölkner J., Goddard M.E., Hayes BJ. (2011). Short communication: Genomic selection using a multi-breed, across-country reference population. J Dairy Sci. 94(5):2625-30. doi: 10.3168/jds.2010-3719.

Recio I., Ramos M., López-Fandiño R. (2001). Capillary electrophoresis for the analysis of food proteins of animal origin. Electrophoresis, 22, 1489-1502.

Rodriguez M.A., Petrini J., Ferreira E.M., Mourão L.R., Salvian M., Cassoli L.D., Pires A.V., Machado P.F., Mourão G.B. (2014). Concordance analysis between estimation methods of milk fatty acid content. Food Chem. 1;156:170-5.

Rubio Bernal Y.L., Gualdrón Duarte J.L., Bates R.O., Ernst C.W., Nonneman D., Rohrer G.A., King D.A., Shackelford S.D., Wheeler T.L., Cantet R.J. and Steibel J.P. (2015). Implementing meta-analysis from genome-wide association studies for pork quality traits. J Anim Sci. 93(12):5607-17. doi: 10.2527/jas.2015-9502

Rutten M. J., Bovenhuis H., Heck J. M., and van Arendonk J. A. (2011). Predicting bovine milk protein composition based on Fourier transform infrared spectra. Journal of dairy science 94(11):5683-5690.

Sanchez M.P., Govignon-Gion A., Croiseau P., Fritz S., Hozé C., Miranda G., Martin P., Barbat-Leterrier A., Letaïef R., Rocha D., Brochard M., Boussaha M., Boichard D. (2017). Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet Sel Evol. 18;49(1):68.

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

Schennink A., Stoop W.M., Visker M.H.P.W., van der Poel J.J., Bovenhuis H., van Arendonk J.A.M. (2009a). Short communication: Genome-wide scan for bovine milk-fat composition. II. Quantitative trait loci for long-chain fatty acids. J. Dairy Sci. 92:4676–4682.

Schennink A., Bovenhuis H., Léon-Kloosterziel K.M., Van Arendonk J.A.M., Visker M. H. P. W. (2009b). Effect of polymorphisms in the FASN, OLR1, PPARGC1A, PRL and STAT5A genes on bovine milk-fat composition. Anim. Genet. 40:909–916.

Schopen G.C., Heck J.M., Bovenhuis H., Visker M.H., van Valenberg H.J., van Arendonk J.A. (2009). Genetic parameters for major milk proteins in Dutch Holstein-Friesians. Journal of dairy science 92(3):1182-1191.

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

Speidel S.E., Buckley B.A., Boldt R.J., Enns R.M., Lee J., Spangler M.L., Thomas M.G. (2018). Genome-wide association study of Stayability and Heifer Pregnancy in Red Angus cattle. J Anim Sci. 3;96(3):846-853

Soyeurt H., Dardenne P., Dehareng F., Lognay G., Veselko D., Marlier M., Bertozzi C., Mayeres P., Gengler N. (2006). Estimating fatty acid content in cow milk using mid-infrared spectrometry. J Dairy Sci. 89(9):3690-5.

Spencer C.C., Su Z., Donnelly P., Marchini J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 2009;5:e1000477.

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

Stoop W.M., Schennink A., Visker M.H., Mullaart E., van Arendonk J.A., Bovenhuis H. (2009). Genome-wide scan for bovine milk-fat composition. I. Quantitative trait loci for short- and medium-chain fatty acids. J Dairy Sci. 92(9):4664-75.

van den Berg S., Calus M.P., Meuwissen T.H., Wientjes Y.C. (2015). Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. BMC Genet. 23;16:146. doi: 10.1186/s12863-015-0305-x.

VanRaden P.M. (2008). Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414–23

Veerkamp R.F., Coffey M., Berry D., de Haas Y., Strandberg E., Bovenhuis H., Calus M. and Wall E. (2012). Genome-wide associations for feed utilisation complex in primiparous Holstein-Friesian dairy cows from experimental research herds in four European countries. Animal. 6(11):1738-49.

Veloso A., Teixeira N., Ferreira I. (2002). Separation and quantification of the major casein fractions by RPHPLC and urea-polyacrylamide gel electrophoresis – detection of milk adulterations, J. of Chromatography A,967, 209-218

Visscher P. M., Hill W. G., Wray N. R. (2008). Heritability in the genomics era--concepts and misconceptions. Nature reviews. Genetics 9(4):255-266.

Visscher P.M., Wray N.R., Zhang Q., Sklar P., McCarthy M.I., Brown M.A., Yang J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 6;101(1):5-22. doi: 10.1016/j.ajhg.2017.06.005.

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

Wade M. J., Shuster S. M., Stevens L. (1996). Inbreeding: its effect on response to selection for pupal weight and the heritable variance in fitness in the flour beetle, Tribolium castaneum. Evolution 50, 723–733.

Wang Q., Hulzebosch A., Bovenhuis H. (2016). Genetic and environmental variation in bovine milk infrared spectra. JDS 99 (8):6793-803.

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

Wientjes Y.C., Veerkamp R.F., Bijma P., Bovenhuis H., Schrooten C., Calus M.P.L. (2015a). Empirical and deterministic accuracies of across population genomic prediction. Genet Sel Evol. 2015;47:5.

Wientjes Y.C., Calus M.P., Goddard M.E., Hayes B.J. (2015b). Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genet Sel Evol. 8;47:42. doi: 10.1186/s12711-015-0124-6.

2

Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein

G. Gebreyesus[1,2], M. S. Lund[1], L. Janss[1], N. A. Poulsen[3], L. B. Larsen[3], H. Bovenhuis[2], and A. J. Buitenhuis[1]

[1]Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark; [2]Animal Breeding and Genomics Centre, Wageningen University, PO Box 338, 6700 AH Wageningen, the Netherlands; [3] Department of Food Science, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark

[Journal of Dairy Science] (2016) 99:2863–2866

Abstract

Genetic parameters were estimated for the major milk proteins using bivariate and multi-trait models based on genomic relationships between animals. The analyses included, apart from total protein percentage, α_{S1}-casein (CN), α_{S2}-CN, β-CN, κ-CN, α-lactalbumin, and β-lactoglobulin, as well as the posttranslational sub-forms of glycosylated κ-CN and α_{S1}-CN-8P (phosphorylated). Standard errors of the estimates were used to compare the models. In total, 650 Danish Holstein cows across four parities and days in milk ranging from 9 to 481 d were selected from 21 herds. The multi-trait model generally resulted in lower standard errors of heritability estimates, suggesting that genetic parameters can be estimated with high accuracy using multi-trait analyses with genomic relationships for scarcely recorded traits. The heritability estimates from the multi-trait model ranged from low (0.05 for β-CN) to high (0.78 for κ-CN). Genetic correlations between the milk proteins and the total milk protein percentage were generally low, suggesting the possibility to alter protein composition through selective breeding with little effect on total milk protein percentage.

Key words: genetic parameter, milk protein, multi-trait model, genomic relationship

2.1 Introduction

Milk protein composition plays an important role in the technological properties of milk (Ikonen et al., 1999; Bittante et al., 2012). Changes in relative concentrations of individual milk proteins have a major effect on milk coagulation properties (Bonfatti et al., 2011) and coagulation ability of milk is essential in cheese making (Cassandro et al., 2008). The major milk proteins include α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG. In addition, several posttranslational modifications of these proteins exist in milk.

Previous studies have shown that considerable genetic variation exists in the composition of milk protein (Bobe et al., 1999; Schopen et al., 2009), presenting the opportunity to alter milk protein composition through selective breeding. Reliable estimates of genetic parameters, including heritability and genetic covariance, are crucial to evaluate the potential for breeding. Quantifying specific milk proteins requires specialized and costly equipment, making it difficult and expensive to measure the traits. As a result, sufficient phenotypic data are not available for reliable estimation of genetic parameters. One effective strategy to deal with such scarcely recorded traits could be implementation of multi-trait models that take advantage of information from correlated traits (Calus and Veerkamp, 2011).

Generally, only a few studies have previously estimated genetic parameters for specific milk proteins (Schopen et al., 2009; Bonfatti et al., 2011) and their posttranslational sub-forms (Bijl et al., 2014). More importantly, none of the previous studies has estimated genetic parameters for milk protein profile using multi-trait analyses.

In this study, we estimated genetic parameters for the major milk proteins (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA and β-LG), the posttranslational sub-forms (glycosylated κ-CN and α_{S1}-CN-8P, where P = phosphorylated serine), as well as protein percentage using bivariate and multi-trait models with genomic relationships between animals and compared standard errors of the estimated genetic parameters.

2.2 Material and Methods

Morning milk samples were obtained from 650 cows from 21 herds in Denmark. The cows were in different stages of lactation (d 9 to 481 in milk) and parity 1 to 4. The liquid chromatography/electrospray ionization-mass spectrometry (LC/ESI-MS) methods were used to profile the milk proteins. Details on screening of samples and quantification of milk proteins were previously described by Jensen et al. (2012).

Of the total cows, 372 were genotyped using the BovineHD Illumina BeadChip. The remaining 278 cows were genotyped with the BovineSNP50 beadchip. Genomic DNA was extracted from ear tissue. Genotypes were subsequently imputed to full sequence in a 2-step procedure. The 278 cows genotyped with the BovineSNP50 chip were first imputed to the BovineHD (777k) level using a multi-breed reference of 3,383 animals including the 372 HD genotyped cows used in this study. The true and imputed HD data for the 2 cow groups were then merged and imputed to the whole-genome sequence level using a multi-breed reference of 1,228 animals from the "1000 bull genomes" project (http://www.1000bullgenomes.com/) and data from Aarhus University using IMPUTE2 v2.3.1 (Howie et al., 2011).

The genomic relationship matrix was calculated as described by the first method presented in VanRaden, (2008). In total, 3.7 million SNP markers spread over BTA1 to BTA29 were included to calculate the G matrix.

The REML approach in DMU was used to estimate genetic parameters and variance components (Madsen and Jensen, 2010). Bivariate and multi-trait analyses were performed and compared using standard errors for the estimated heritability.

The general model used was:

$$y_{ijkl} = \mu + parity_i + herd_j + b_1 \times DIM_k + b_2 + \exp^{-0.05xDIM_k} + animal_l + e_{ijkl}, \qquad [1]$$

where y_{ijkl} was the observation of animal *l*, in parity *i* and herd *j*; μ was the fixed mean effect; b_1 was the regression coefficient for DIM_k; and DIM_k was a covariate describing the effect of days *k* in milk. Wilmink adjustment ($\exp^{-0.05\times DIM}$) was used for DIM, b_2 was the regression coefficient for the Wilmink adjustment; $animal_l$ was the random additive genetic effect based on **G** of animal *l* with distribution $N(0, \mathbf{G}\sigma_a^2)$, and e_{ijkl} was the random residual effect, which was assumed to be normally distributed with $\boldsymbol{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where **G** is the genomic relationship matrix, **I** was the identity matrix, σ_a^2 was the genetic variation, σ_e^2 was the residual variation.

The bivariate analyses were run for each milk protein analyzed in combination with protein percentage. For the multi-trait analysis, all nine traits were fitted simultaneously. Correlations between traits were based on the multi-trait analyses.

2.3 Results and Discussion

Table 2.1 summarizes the descriptive statistics for the milk protein profile and the total milk protein percentage. Mean protein content in the sampled milk was 3.38%. The major proteins (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG) made up 83% of the total milk protein fraction. The caseins constituted 72.3% of the total protein, with β-CN and α_{S1}-CN alone contributing to 34.1 and 26.8% of the total milk protein, respectively. The whey proteins constituted 10.8% of the total protein.

The α_{S1}-CN-8P accounted for 19.2% of the total milk protein and 71.6% of the α_{S1}-CN fraction of the total protein percentage. This was comparable to previous findings, in which α_{S1}-CN-8P accounted for 21.3% of the total protein (Bijl et al., 2014) and 74% of the α_{S1}-CN (Heck et al., 2008) in the Dutch Holstein population.

Table 2.1. Descriptive statistics[1] of milk protein profile and the total milk protein percentage

Protein or fraction[2]	Mean (%)	CV (%)	5% quantile	95% quantile
α_{S1}-CN	26.8	9	25.5	28.1
α_{S1}-CN-8P	19.2	11	17.7	20.7
α_{S2}-CN	5.3	20	4.5	5.9
β-CN	34.1	10	31.5	36.7
κ-CN	6.1	18	5.3	6.9
Glycosylated κ-CN	1.7	47	1.2	2.0
α-LA	3.3	19	2.9	3.6
β-LG	7.5	21	6.5	8.4
Total protein (%)	3.38	9	3.2	3.55

[1]Mean = phenotypic mean of the trait.

[2]Protein composition was expressed as percentage fractions of the total milk protein percentage (wt/wt); total protein was expressed as percentage (%) of the total milk yield; individual proteins comprise only the peaks identified as intact proteins and isoforms; that is, α_{S1}-CN (comprises α_{S1}-CN 8P + 9P), α_{S2}-CN (comprises α_{S2}-CN 11P + 12P), β-CN (comprises β-CN 4P + 5P), and κ-CN (comprises κ-CN G + 1P), where P = phosphorylated serine group.

Heritability values and standard errors of estimation from the bivariate and multi-trait models are given in Table 2.2. Generally, the heritability estimates for the milk proteins were moderate to high except for β-CN, which had the lowest estimates (0.01–0.05). Glycosylated κ-CN (0.44) and α_{S2}-CN (0.36) had moderate heritability values estimated using the two models; κ-CN had the highest heritability estimates in both models (0.78–0.79). Generally, standard errors were lower for heritability estimates using the multi-trait model for all traits except β-CN. The standard errors of estimation from the multi-trait model in the current study (0.08–0.10) were also lower compared with previous studies, including that of Schopen et al. (2009; 0.08–0.12), Buitenhuis et al. (2014; 0.12–0.21), and Bobe et al. (1999; 0.07–0.12). Given the observed medium to high genetic correlations between the proteins, the multi-trait analyses might have benefitted from use of information from correlated traits.

Heritability values estimated with the multi-trait model were comparable to estimates from previous studies. Intraherd heritability of 0.66 for total protein and 0.64 for κ-CN estimated by Schopen et al. (2009) were comparable to the current estimates of 0.59 and 0.78, respectively. The estimates for α_{S2}-CN (0.36) in the current study were also comparable to the earlier reported value of 0.30 (Ikonen et al., 1997) but lower than estimates of Schopen et al. (2009; 0.73). β-casein had the lowest heritability estimate (0.05) in this study. This was in agreement with previous estimates for β-CN by Buitenhuis et al. (2014; 0.05), but considerably lower than estimates by Ikonen et al. (1997; 0.33 to 0.40) and Schopen et al. (2009; 0.25).

Table 2.2 Heritability values and standard errors of heritability estimates

Trait	Bivariate model		Multi-trait model	
	h^2	SE	h^2	SE
α_{S1}-CN	0.13	0.10	0.15	0.09
α_{S1}-CN-8P[1]	0.14	0.10	0.14	0.09
α_{S2}-CN	0.36	0.10	0.30	0.10
β-CN	0.01	0.07	0.05	0.08
κ-CN	0.44	0.10	0.44	0.09
Glycosylated κ-CN	0.79	0.09	0.78	0.08
α-LA	0.22	0.11	0.25	0.10
β-LG	0.56	0.11	0.54	0.10
Total protein (%)	0.53	0.11	0.59	0.10

[1]Where P = phosphorylated serine.

Genetic and phenotypic correlations estimated using multi-trait analysis are given in Table 2.3. Generally, the milk protein compositions had low to medium genetic correlation with total protein (−0.02–0.32). Higher genetic (0.75) as well as phenotypic (0.61) correlations were also observed between the two whey proteins (β-LG and α-LA).

The genetic correlation between α_{S1}-CN-8P and β-LG was low (−0.01). Bijl et al. (2014) have previously reported a distinct effect of β-LG protein variants and β-LG concentration on α_{S1}-CN-8P concentration. Nonetheless, according to Bijl et al. (2014), the mechanism behind the established association between β-LG protein variants, as well as β-LG concentration and α_{S1}-CN-8P concentration remains unclear, calling for further study. Genetic correlations between the milk proteins and the total milk protein percentage were generally low, except for α_{S1}-CN-8P (0.38) and β-LG (0.27).

Table 2.3. Genetic (above diagonal) and phenotypic (below diagonal) correlations[1]

Trait[2]	α_{S1}-CN-8P	α_{S1}-CN	α_{S2}-CN	β-CN	Glyc κ-CN	κ-CN	α-LA	β-LG	Total protein(%)
α_{S1}-CN-8P		0.78	-0.11	-0.38	-0.42	-0.23	-0.59	-0.07	0.32
α_{S1}-CN	0.86		-0.51	-0.02	-0.33	-0.36	-0.46	-0.15	0.15
α_{S2}-CN	0.29	0.12		-0.78	-0.08	-0.21	0.18	0.09	0.12
β-CN	0.41	0.47	0.04		0.37	0.49	0.13	0.03	0.11
Glyc κ-CN	-0.04	0.03	0.02	0.05		0.81	0.16	-0.04	-0.02
κ-CN	0.20	0.16	0.05	0.21	0.68		-0.005	-0.02	0.12
α-LA	0.10	0.08	0.09	0.09	-0.02	0.06		0.60	-0.03
β-LG	0.006	0.01	0.04	-0.001	-0.07	-0.02	0.77		0.24
Total protein(%)	0.18	0.03	0.12	-0.07	0.10	0.18	-0.04	0.13	

[1] SE (0.06–0.31)

[2] P = phosphorylated serine; Glyc = glycosylated

2.4 Conclusions

Our results suggest that genetic parameters can be estimated with high accuracy for scarcely recorded traits using multi-trait analysis with genomic relationships between animals. Lower genetic correlations between the milk proteins with total protein percentage reported in this study also suggest that altering milk protein compositions through selective breeding might have little or no effect on the total protein percentage.

Acknowledgements

The authors acknowledge help from Rasmus Froberg Brøndum during the imputations of genotype data. This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research). The study used data from Danish-Swedish Milk Genomics Initiative (www.milkgenomics.dk), as well as the Milk Levy Fund (Denmark) projects “Phenotypic and genetic markers for specific milk quality parameters” and “The importance of the metagenome for milk composition and quality.” A. J. B., N. A.P., and L.B.L. were supported by the project “Breeding high value milk: BIG MILK” of the Milk Levy Fund, Denmark (2015-2018). G.G. is enrolled in the Erasmus-Mundus joint doctorate European Graduate School in Animal Breeding and Genetics.

References

Bijl E., van Valenberg H. J. F., Huppertz T., van Hooijdonk A. C. M., Bovenhuis H. (2014). Phosphorylation of αS1-casein is regulated by different genes. Journal of dairy science 97(11):7240-7246.

Bittante G., Penasa M., Cecchinato A. (2012). Invited review: Genetics and modeling of milk coagulation properties. Journal of dairy science 95(12):6843-6870.

Bobe G., Beitz D. C., Freeman A. E., and Lindberg G. L. (1999). Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. Journal of dairy science 82(12):2797-2804.

Bonfatti V., Cecchinato A., Gallo L., Blasco A., and Carnier P. (2011). Genetic analysis of detailed milk protein composition and coagulation properties in Simmental cattle. Journal of dairy science 94(10):5183-5193.

Buitenhuis A. J., Poulsen N. A., Larsen L. B. (2014). Estimation of genetic parameters for the protein profile in Danish Holstein milk. Proceedings, 10th World congress of genetics applied to livestock production. https://asas.org/docs/default-source/wcgalp-posters/614_paper_9289_manuscript_543_0.pdf?sfvrsn=2

Calus M. P. L., Veerkamp R. F. (2011). Accuracy of multi-trait genomic selection using different methods. Genetics selection evolution (Paris) 43(1):26.

Cassandro M., Comin A., Ojala M., Dal Zotto R., De Marchi M., Gallo L., Carnier P., Bittante G. (2008). Genetic parameters of milk coagulation properties and their relationships with milk yield and quality traits in Italian Holstein cows. Journal of Dairy Science, 91: 371–376

Heck J. M. L., Olieman C., Schennink A., van Valenberg H. J. F., and Visker M. H. P. W. (2008). Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis. International dairy journal 18(5):548-555.

Howie B., Marchini J., and Stephens M. (2011). Genotype imputation with thousands of genomes. G3 (Bethesda) 1:457–470. http:// dx.doi.org/http://dx.doi.org/10.1534/g3.111.001198.

Ikonen T., Ojala M., and Syväoja E.L. (1997). Effects of composite casein and beta-lactoglobulin genotypes on renneting properties and composition of bovine milk by assuming an animal model. 1997:12.

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

Jensen H. B., Poulsen N. A., Andersen K. K., Hammershoj M., Poulsen. H. D. (2012). Distinct composition of bovine milk from Jersey and Holstein-Friesian cows with good, poor, or noncoagulation properties as reflected in protein genetic variants and isoforms. Journal of dairy science 95(12):6905-6917.

Madsen P., Jensen J. (2010). A User's Guide to DMU. Version 6, Release 5.0. University of Aarhus, Faculty Agricultural Sciences (DJF), Department of Genetics and Biotechnology, Research CentreFoulum, Tjele, Denmark.

Schopen G. C., J. M. Heck, H. Bovenhuis, M. H. Visker, H. J. van Valenberg, and J. A. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. Journal of dairy science 92(3):1182-1191.

VanRaden P. M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91(11):4414-4423.

3

Modeling heterogeneous (co)variances from adjacent-SNP groups improves genomic prediction for milk protein composition traits

Grum Gebreyesus[1,2], Mogens Sandø Lund[1], Bart Buitenhuis[1], Henk Bovenhuis[2], Nina A Poulsen[3] and Luc G Janss[1]

[1]Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark; [2]Animal Breeding and Genomics Centre, Wageningen University, PO Box 338, 6700 AH Wageningen, the Netherlands; [3]Department of Food Science, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark

[GSE] (2017) 49:89

Abstract

Accurate genomic prediction requires a large reference population, which is problematic for traits that are expensive to measure. Traits related to milk protein composition are not routinely recorded due to costly procedures and are considered to be controlled by a few quantitative trait loci (QTL) of large effect. The amount of variation explained may vary between regions leading to heterogeneous (co)variance patterns across the genome. Genomic prediction models that can efficiently take such heterogeneity of (co)variances into account can result in improved prediction reliability. In this study, we developed and implemented novel univariate and bivariate Bayesian prediction models, based on estimates of heterogeneous (co)variances for genome segments (BayesAS). Available data consisted of milk protein composition traits measured on cows and de-regressed proofs (DRP) of total protein yield derived for bulls. Single-nucleotide polymorphisms (SNPs), from 50K SNP arrays, were grouped into non-overlapping genome segments. A segment was defined as one SNP, or a group of 50, 100, or 200 adjacent SNPs, or one chromosome, or the whole genome. Traditional univariate and bivariate genomic best linear unbiased prediction (GBLUP) models were also run for comparison. Reliabilities were calculated through resampling strategy and using deterministic formula.

BayesAS models improved prediction reliability for most of the traits compared to GBLUP models and this gain depended on segment size and genetic architecture of the traits. The gain in prediction reliability was especially marked for the protein composition traits β-CN, κ-CN and β-LG, for which prediction reliabilities were improved by 49 percentage points on average using the MT-BayesAS model with 100-SNP segment size compared to the bivariate GBLUP. Prediction reliabilities were highest with the BayesAS model that uses 100-SNP segment size. The bivariate versions of our BayesAS models resulted in extra gains of up to 6% in prediction reliability compared to the univariate versions.

Substantial improvement in prediction reliability was possible for most of the traits related to milk protein composition using our novel BayesAS models. Grouping adjacent SNPs into segments provided enhanced information to estimate parameters and allowing the segments to have different (co)variances helped disentangle heterogeneous (co)variances across the genome.

Key words: Milk protein composition, genomic prediction, heterogeneous (co)variance, BayesAS, SNP grouping

3.1 Introduction

The protein composition of milk determines its technological characteristics such as the cheese-making properties. Major proteins in milk include the caseins (αS1-, αS2-, β- and κ-CN) and whey proteins (α-lactalbumin, and β-lactoglobulin). The heritability of the relative proportion of these proteins in bovine milk is moderate to high (Bobe et al., 1999; Schopen et al., 2009; Gebreyesus et al., 2016), which provides the opportunity to alter the protein composition of milk through selective breeding. Prediction of genetic merit for traits related to milk protein composition has never been reported and one reason for this is that measurements of the detailed protein composition of milk is currently limited to experimental samples due to costly and time-consuming analytical techniques.

In livestock breeding, genomic selection has become a successful approach, especially for sex-limited traits, because it speeds up the selection process by reducing generation interval and enables to select new selection candidates at early ages. Accuracy of genomic prediction hinges on a number of factors including size of reference population, heritability of the trait, effective population size, marker density, and the genetic architecture of the trait, in particular, the number of loci that affect the trait and the distribution of their effects (Daetwyler et al., 2008; Goddard, 2009; Meuwissen, 2009). Therefore, prediction accuracy for traits with limited records is still low. However, if the methodology used exploits information about the distribution of the loci that underlie a trait, traits that are controlled by a few quantitative trait loci (QTL) with major effects can be predicted with better accuracy than traits that have a more polygenic nature (Hayes et al, 2010). Several statistical models have been developed for genomic prediction using genome-wide single-nucleotide polymorphisms (SNPs), which include the Bayesian models (e.g. BayesA and BayesB) of Meuwissen et al. (2001), the genomic best linear unbiased prediction (GBLUP) model (VanRaden, 2008) and several extensions of these models. Compared to GBLUP, the Bayesian variable selection models improve considerably prediction reliability for traits that are controlled by a few QTL with major effects (Cole et al., 2009; Legarra et al., 2011). This is mainly due to the assumption that, in the GBLUP model, the variance does not vary across the genome, i.e. it does not take heterogeneity over segments into account. Unlike GBLUP, Bayesian variable selection models allow the variance of SNP effects to differ among loci (VanRaden, 2008). Genome-wide association studies have indicated that a few QTL regions underlie substantial proportions of the genetic variation in traits related to milk protein composition (Schopen et al., 2011). Hence, it is expected that, for traits related to milk protein composition, a model assuming

SNP-specific variances in genomic prediction can result in higher prediction reliability than the GBLUP approach. However when the available dataset is small, as is the case for expensive-to-measure traits, reliable estimation of SNP-specific variances with the Bayesian approach becomes problematic since there are too many parameters to estimate relative to the information available. In such situations, Gianola et al. (2009) suggested to group SNPs according to their common variance. Grouping adjacent SNPs might be advantageous for estimating variances reliably by enhancing the amount of information and reducing the number of parameters to estimate. Adjacent SNPs are very likely to be in linkage disequilibrium (LD) with the same QTL and to have the same variance, which allow us to account for heterogeneity between SNP groups. In this context, SNPs must be properly ordered and grouped such that they are realistically in LD with the same QTL while ensuring that their group size is optimum for the reliable estimation of variances.

Another option that is widely used to deal with traits of limited records is to implement multi-trait models, which simultaneously use information from related traits and individuals (Henderson and Quaas, 1976). In multi-trait analysis, correlation structures between the traits is central to gaining any advantage in prediction reliability over single-trait predictions (Calus and Veerkamp, 2011). Milk protein traits have a low to moderate genetic correlation with routinely recorded traits such as total protein yield (Schopen et al., 2009). However, while the genome-wide correlation is generally low, specific genomic segments may display high genetic correlations between SNP effects for different traits. Therefore, modeling such heterogeneous covariance patterns may result in improved prediction reliability, when using multi-trait models.

In this study, we report genomic prediction reliabilities for traits related to milk protein composition using a relatively small set of cow data by developing novel Bayesian hierarchical models that account for heterogeneous variance structures across regions over the genome. Furthermore, we extend our novel Bayesian models to bivariate scenarios that model heterogeneous covariance structures between milk protein composition traits measured on cows and a large set of bull data with highly accurate de-regressed proofs (DRP) for total protein yield.

3.2 Methods

3.2.1. Animals and phenotypes

Available data comprised two datasets: a relatively small set of cow data with information on traits related to milk protein composition and a large set of bull data with highly accurate total protein yields from regular milk recordings on daughters. Individuals in the two datasets were genetically related i.e. all the cows had their sires in the bull dataset.

Single morning milk samples were collected once from 650 Danish Holstein cows in 21 herds. Cows were sampled at different stages of lactation (days 9 to 481 in milk) and parity (1 to 4). Liquid chromatography/electrospray ionization-mass spectrometry (LC/ESI-MS) methods were used for profiling milk proteins. Details on the identification and relative quantification of milk proteins are in Jensen et al. (2012). We used these methods to quantify milk proteins, including α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG, posttranslational modifications of G-κ-CN and α_{S1}-CN-8P, as well as total protein percentage. In later analyses, β-CN was excluded from the genetic analysis due to very low estimates of its heritability across models (0.01 to 0.05), which made meaningful predictions difficult to obtain given the small sample size.

DRP for milk protein yield were obtained from 5326 progeny-tested Danish Holstein bulls. Estimated breeding values from the Nordic genetic evaluation in January 2013 were used to derive DRP following the methodology described by Schaeffer (2001).

3.2.2. Genotypes

Genotyping was performed using the BovineHD Illumina Beadchip for 372 cows or the BovineSNP50 Beadchip for the remaining 278 cows and all the bulls. SNPs that were overlapping in these two genotyping arrays were combined and subjected to quality control. Quality parameters used to select SNPs were: (1) minimum call rates of 90% for individuals and 95% for loci and (2) exclusion of SNPs with a minor allele frequency (MAF) lower than 5%. Finally, 36,000 SNPs across the 29 bovine autosomes were available for the analyses.

3.2.3. Models

Hierarchical Bayesian models based on genome segments of different sizes (hereafter collectively called BayesAS models) were developed to predict genomic breeding values (GBV). Univariate and bivariate GBLUP models were used to compare performances of the novel Bayesian models.

3.3.3.1. GBLUP models

Univariate (based on cow data only) and bivariate (based on combined cow data and bull DRP) GBLUP models were implemented using DMU (Madsen and Jensen, 2007). The general model used for the univariate analysis (ST-GBLUP) was:

$$y_{ijkl} = \mu_i + parity_{ij} + herd_{ik} + b_{i1}\, DIM_l$$

$$+ b_{i2} * exp^{-0.05*DIM_l} + g_{il} + e_{1ijkl}\,, \qquad (1)$$

where y_{ijkl} are the observations on trait i from cow l, in parity j, and herd k; μ_i is the fixed mean effect for trait i; b_{i1} is the regression coefficient for DIM_l in trait i, which is a covariate describing the effect of days in milk for each cow l; b_{i2} is the regression coefficient for the Wilmink adjustment ($exp^{-0.05*DIM_l}$) of days in milk for trait i; e_{1ijkl} is a random residual effect that is assumed to be normally distributed with $\mathbf{e}_1 \sim N\left(0, \mathbf{I}_1\ \sigma^2_{e_1}\right)$, where $\mathbf{I}_1$ is an identity matrix with dimensions 650 by 650. The effect of g_{il} is a random additive genetic effect for trait i of cow l with distribution $N(0, \mathbf{G}\sigma^2_a)$, where $\mathbf{G}$ is the genomic relationship matrix between cows with dimension 650 by 650 and σ^2_a is the genetic variation in trait i.

To run a bivariate analysis (MT-GBLUP) of DRP on protein yield and each protein composition trait, DRP were modelled as:

$$y_{DRP_l} = \mu_{DRP} + g_{2l} + e_{2l}, \qquad (2)$$

where, y_{DRP_l} is the DRP for bull l; and μ_{DRP} is the corresponding fixed mean effect. g_{2l} is the random additive genetic effect for animal *l* for protein yield with distribution $N(0, \mathbf{G}_2\sigma^2_a)$, where $\mathbf{G}_2$ is the genomic relationship matrix for combined cow and bull population with dimension 5976 by 5976. Distribution of the vectors of the two animal effects in the bivariate models are as follows:

$$\begin{pmatrix}\mathbf{g_1}\\ \mathbf{g_2}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\otimes\mathbf{G_2}\right),$$

$$\text{with } \Sigma = \begin{pmatrix}\sigma_1^2 & \sigma_{1,2}\\ \sigma_{1,2} & \sigma_2^2\end{pmatrix},$$

where, in this case, $\mathbf{g}_1$ is a vector of breeding values for all animals for one of the cow traits based on the covariance matrix $\mathbf{G}_2$ unlike in Model (1); σ_1^2 is the genetic variance for each cow trait and σ_2^2 is the genetic variance for the bull DRP.

The random residual effect e_{2l}, in Model (2), is assumed to be normally distributed with, $\mathbf{e_2} \sim N(0, \mathbf{I}_2\ \sigma_{e_2}^2)$, where $\mathbf{I}_2$ is an identity matrix with dimension 5326 by 5326 and $\sigma_{e_2}^2$ is the residual variation for bull DRP. In the bivariate analysis, the residual covariance for the pair of bivariate traits was set to zero because the observations came from different individuals. The distribution of the vectors of the two residual effects in the bivariate analyses can be described as:

$$\begin{pmatrix} \mathbf{e_1} \\ \mathbf{e_2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{I}_1\sigma_{e_1}^2 & 0 \\ 0 & \mathbf{I}_2\sigma_{e_2}^2 \end{pmatrix}\right).$$

The genomic relationship matrix used in the GBLUP models was calculated as described in the first method presented by VanRaden (2008).

3.2.3.2. BayesAS models

Models that were proposed initially by Janss (2014) were implemented in the MCMC Bayesian framework of the Bayz program (www.bayz.biz). Adjacent SNPs were grouped into non-overlapping genomic segments and the (co)variance for each segment was estimated. Accordingly, six models were implemented, in which a genome segment was defined as: single SNPs or groups of 50, 100, or 200 adjacent SNPs, a complete chromosome or all the SNPs in the genome. The model that considers the whole genome as a segment can be considered basically as a GBLUP model implemented in a Bayesian manner.

Univariate (ST) and bivariate (MT) versions of the BayesAS models were implemented. For the ST-BayesAS models, each protein composition trait (y_{ijkl}) from the cow dataset was run as in the model described below:

$$y_{ijkl} = \mu_i + parity_{ij} + herd_{ik}$$

$$+ b_{i1}\, DIM_l + b_{i2} * exp^{-0.05*DIM_l} + \mathbf{z}_l\, \mathbf{a}_i + e_{1ijkl}. \qquad (3)$$

Model components for fixed effects, covariates and random residual effects are in Model (1). $\mathbf{Z}$ is a matrix with SNP covariates (centered) with dimensions of the number of individuals (n = 650) by the number of loci (m = 36,000) and $\mathbf{z}_l$ is a row of genotypes for animal l, $\mathbf{a}_i$ is a vector of SNP effects for trait i, with length m and with elements $\mathbf{a}_i = \{a_{ijk}\}$, such that a_{ijk} is the effect of SNP k in SNP group j for trait i.

For the MT-BayesAS models, an additional model component was included to run the DRP on total protein yield from bulls (y_{DRP_l}) simultanously with each protein composition trait from cows. The following model was added to run the bivariate MT-BayesAS analyses:

$$y_{DRP_l} = \mu_{DRP} + \mathbf{z}_l\,\mathbf{a}_i + e_{2l}, \qquad (4)$$

$\mathbf{Z}$ in the MT-BayesAS is a matrix with SNP covariates (centered) with dimensions of the number of individuals (n = 5976) by the number of loci (m = 36,000) and $\mathbf{z}_l$ is a row of genotypes for animal l, $\mathbf{a}_i$ is a vector of SNP effects for trait i, with length m and with elements $\mathbf{a}_i = \{a_{ijk}\}$ and the residual term (e_{2l}), is as in Model (2). The index "*i*" here refers to both cow trait and bull DRP run in each bivariate analysis, for sake of simplicity in describing the models. SNP effects between each of the two traits in the bivariate analyses are correlated using latent variables by the following hierarchical model:

$$a_{ijk} = r_{0i} * \mathbf{s}_0 + r_{1ij} * \mathbf{s}_1 + a^*_{ijk}, \qquad (5)$$

where $\mathbf{s}_0 = \{s_{0jk}\}$ and $\mathbf{s}_1 = \{s_{1jk}\}$ are vectors of latent variables with length m, to model average covariance across SNP groups ($\mathbf{s}_0$) and deviations within SNP groups ($\mathbf{s}_1$) using nested regression; r_{0i} is a regression coefficient of $\mathbf{s}_0$ for all SNPs and r_{1ij} is a regression coefficient of $\mathbf{s}_1$ for each SNP group j; and a^*_{ijk} is the residual SNP effect, which is uncorrelated across traits. The latent variables in $\mathbf{s}_0$ and $\mathbf{s}_1$ are assumed to be normally distributed with a variance of 1:

$\mathbf{s}_0 \sim N(0, \mathbf{I})$ and $\mathbf{s}_1 \sim N(0, \mathbf{I})$,

where, $\mathbf{I}$ is an identity matrix with dimensions of number of loci (m = 36,000). Distributional prior assumptions for the regression coefficients of $\mathbf{s}_0$ and $\mathbf{s}_1$ are:

$$r_{0i} \sim U(-\infty, \infty),$$

$$r_{1ij} \sim N(0, \sigma^2_{r_{1i}}),$$

$$\sigma^2_{r_{1i}} \sim U(0, \infty),$$

where $U()$ stands for a uniform distribution across the given interval.

The residual SNP effect a^*_{ijk} is assumed to be normally distributed with a mean of 0 and SNP-group specific variance ($\sigma^2_{a^*_{ij}}$) for which an inverse-chi-square distribution was set with scale SC^2_i and degrees of freedom df_i for all SNP effects in group j:

$$a^*_{ijk} \sim N(0,\ \sigma^2_{a^*_{ij}}),$$

$$\sigma^2_{a^*_{ij}} \sim \chi^{-2}(SC^2_i, df_i).$$

The scale parameter SC^2_i is assumed to have a uniform distribution. The parameter df_i is set so that the spread of the variances of individual SNP-groups around the common scale is controlled (here, a value of 5 was used).

Samples of the posterior distributions of the model parameters are obtained using MCMC techniques, i.e. sampling from conditional distributions. The conditional distributions for all parameters in Eqs (3), (4) and (5) are normal and for variances are scale-inverse chi-squared. For the parameters $\mathbf{s}_0$ and $\mathbf{s}_1$, which are present in the expectation for multiple SNP-effects, the bayz software automatically combines all parts of the likelihoods and combines them with the prior distribution to form the conditional posterior.

$\mathbf{Za}_i$ from Models (3) and (4) computes the genomic values ($\mathbf{g}_i$) at each MCMC cycle. The total explained genomic variance for trait i is computed as the variance of the genomic values from every MCMC cycle:

$$\sigma^2_i = var(\mathbf{Za}_i) = var(\mathbf{g}_i). \qquad (6)$$

The genomic covariance between the cow and bull traits can then be calculated as:

$$\sigma_{cow,bull} = cov(\mathbf{g}_{cow}, \mathbf{g}_{bull}), \qquad (7)$$

where $\mathbf{g}_{cow}$ is a vector of genomic values for all individuals for each cow-trait and $\mathbf{g}_{bull}$ is a vector of genomic values of all individuals for total protein yield. Similarly, genetic values for the individuals at SNP group j ($\mathbf{g}_{ij}$) was calculated at each MCMC sample based on the genotypes and estimated effects of SNPs in group j as:

$$\mathbf{g}_{ij} = \mathbf{Z}_j \, \mathbf{a}_{ij}, \qquad (8)$$

where $\mathbf{Z}_j$ is a matrix of covariates for SNPs within group j, with size of number of individuals by number of SNPs at group j, and $\mathbf{a}_{ij}$ is a vector of effects of SNPs at group j for trait i. Genomic variance for trait i at SNP group j was then calculated from these MCMC samples of individual genetic values as:

$$\sigma_{ij}^2 = var(\mathbf{g}_{ij}). \qquad (9)$$

The proportion of the genomic variance explained by segments was computed for each trait i as: $\frac{\sigma_{ij}^2}{\sigma_i^2}$. The genomic covariance for each cow and bull trait at each SNP group j was then calculated as:

$$\sigma_{cowj,bullj} = cov(\mathbf{g}_{cowj}, \mathbf{g}_{bullj}). \qquad (10)$$

Inferences were based on 500,000 Gibbs samples. The first 50,000 samples were discarded as burn-in, and every 500th sample was saved for post-Gibbs analyses. The mean of the variance and covariance terms, which are calculated in each MCMC iteration, is used later. Convergence was assessed using the R package CODA (Plummer et al., 2006).

The BayesAS models presented in this study can be considered as extensions of the Bayes A model of Meuwissen et al. (2001) mainly different in that estimates of variances are per SNP groups (segments) instead of single SNP. In this case, taking one SNP as a segment might be considered as an approximation to the BayesA approach.

However, there still is a difference in that the scale parameter of the $\chi^{-2}(SC_i^2, df_i)$ prior for $\sigma_{a_{ij}^*}^2$ is treated as unknown instead of being fixed. Moreover, the bivariate versions of BayesAS uniquely use latent variables to model covariances between traits.

3.2.4. Comparison of the predictive ability between models

A resampling strategy using cows in five test sets was implemented to compare models for prediction reliability. Our aim was to avoid sibling relationships between each test set and between the training and test sets. Hence, 197 cows, which had no siblings in the dataset, were selected. In each of the resampled analyses, 100 of the 197 cows were randomly taken for the test set, while the remaining 97 cows from each random sampling were included in the reference population of 550 cows. For all models, prediction reliability for cows was computed as the squared correlation between estimated GBV and the phenotype corrected for fixed effects as in Model (1), divided by heritability estimates (Su et al., 2012) from the complete dataset of 650 cows using Model (1). Since the major practical implication of genomic prediction studies is to assess the predictive ability of models for young bulls with no phenotypic record, reliabilities of models in the MT-BayesAS analyses were computed for bulls using standard errors of predicted GBV using the following formula, as described by Mrode (2014):

$$1 - \frac{{SEP_l}^2}{\sigma_i^2}, \tag{11}$$

where SEP_l is the standard error of prediction (posterior standard deviations from MCMC samples) of GBV for each bull based on its Gibbs samples for each protein composition trait; and σ_i^2 is the total genomic variance calculated as in Eq. (6), which, as an approximation, was taken as the additive genetic variance. Model reliabilities were computed for all bulls, and the average was taken as the model reliability for the respective trait.

Further analyses were conducted using the Gibbs samples from the 100-SNP segment size MT-BayesAS model to assess prediction reliability when varying the proportion of segments, based on ranking of explained genomic variance, were used for prediction. Prediction reliabilities were, accordingly, computed using the top 2% (8), 7% (26), 15% (56), 25% (93), 50% (186), or 75% (279) of all 372 genomic segments included in the analyses. Segments were ranked on estimated variance based on evaluation on the training data with all segments included. Reliabilities were computed for the test sets similarly as in the other BayesAS models and were used to compare the different scenarios.

3.3 Results

3.3.1. Heritability estimates for milk protein composition traits and genomic correlations with total protein yield

Table 3.1 presents heritability estimates for traits related to milk protein composition obtained with the ST-GBLUP model, their genome-wide correlations and covariances with total milk protein yield obtained with the MT-GBLUP model. Heritability estimates were high for κ-CN, G-κ-CN, β-LG, and protein percentage. Heritability estimates were moderate for α_{S2}-CN, but lower for α_{S1}-CN, α_{S1}-CN-8P, and α-LA. Milk protein composition traits showed very low (-0.16 to 0.15) genomic correlations with total milk protein yield. Genome-wide correlations with protein yield were negative for α_{S2}-CN, α_{S1}-CN-8P, and protein percentage. Standard errors of the correlations were higher than the correlation estimates for all traits except for α_{S2}-CN and protein percentage.

Table 3.1 Heritability estimates and genome-wide correlations and covariances with total milk protein yield

[a]Trait	h^2	SE	Covariance	SE	Correlation	SE
α_{S1}-CN	0.14	0.07	0.01	0.05	0.04	0.16
α_{S1}-CN-8P	0.14	0.09	-0.02	0.05	-0.07	0.16
α_{S2}-CN	0.33	0.09	-0.08	0.06	-0.16	0.12
κ-CN	0.69	0.09	0.06	0.05	0.09	0.07
G-κ-CN	0.41	0.09	0.0008	0.04	0.0006	0.10
α-LA	0.15	0.09	0.05	0.05	0.15	0.16
β-LG	0.52	0.10	0.04	0.05	0.07	0.09
Protein%	0.54	0.09	-0.08	0.06	-0.14	0.10

[a]Protein composition expressed as a fraction of the total milk protein percentage by weight wt (wt/wt), protein% expressed as percentage of the total milk yield; individual proteins comprise only the peaks identified as intact proteins and isoforms; that is, α_{S1}-CN (comprises α_{S1}-CN 8P + 9P), α_{S2}-CN (comprises α_{S2}-CN 11P + 12P), κ-CN (comprises κ-CN G 1P + unglycosylated κ-CN 1P), where P = phosphorylated serine group. G-κ-CN = glycosylated-κ-CN; α_{S1}-CN-8P= α_{S1}-CN with 8 phosphorylated serine groups
Heritability (h^2) estimates were from the univariate GBLUP analysis; covariances and correlations are from the bivariate GBLUP model

3.3.2. Prediction reliability of the GBLUP models

Prediction reliabilities were low for all traits (0.03 to 0.21) when using the ST- and MT-GBLUP models (Table 3.2). Compared to the other protein composition traits, β-LG (0.21) and κ-CN (0.16) had the highest prediction reliabilities, whereas α_{S2}-CN and α_{S1}-CN-8P had the lowest (0.03) when using univariate analysis. There was a slight gain in prediction reliability for α_{S2}-CN and protein percentage when bivariate analysis was used. There was no improvement in prediction reliability for κ-CN, G-κ-CN, β-LG, or α_{S1}-CN-8P compared to ST-GBLUP predictions. Prediction reliability was a little lower with the MT-GBLUP model than with univariate prediction for α_{S1}-CN and α-LA.

Table 3.2 Prediction reliability from univariate and bivariate GBLUP models

[a]Trait	ST-GBLUP	MT-GBLUP
α_{S1}-CN	0.11	0.10
α_{S1}-CN-8P	0.03	0.03
α_{S2}-CN	0.03	0.06
κ-CN	0.16	0.16
G-κ-CN	0.14	0.14
α-LA	0.12	0.11
β-LG	0.21	0.21
Protein%	0.10	0.12

[a]Protein composition expressed as a fraction of the total milk protein percentage by weight wt (wt/wt), protein% expressed as percentage of the total milk yield; individual proteins comprise only the peaks identified as intact proteins and isoforms; that is, α_{S1}-CN (comprises α_{S1}-CN 8P + 9P), α_{S2}-CN (comprises α_{S2}-CN 11P + 12P), κ-CN (comprises κ-CN G 1P + unglycosylated κ-CN 1P), where P = phosphorylated serine group. G-κ-CN = glycosylated-κ-CN; α_{S1}-CN-8P = α_{S1}-CN with 8 phosphorylated serine groups.

3.3.3. Genome segment-wise variances for milk protein composition traits and covariance with total protein yield

Figure 3.1 presents the proportion of genomic variance in milk composition traits explained by each chromosome using the ST-BayesAS model. Marked differences were observed in the proportion of genomic variance explained by genome segments across the traits. For some of the protein composition traits, a single chromosome explained up to or more than half of the genomic variance. For instance, Bos taurus (BTA) chromosome 6 explained 76, 63 and 47 % of the genomic variance for κ-CN, G-κ-CN and α_{S2}-CN, respectively. Likewise, 40% of the genomic variance for β-LG was explained by BTA11 alone.

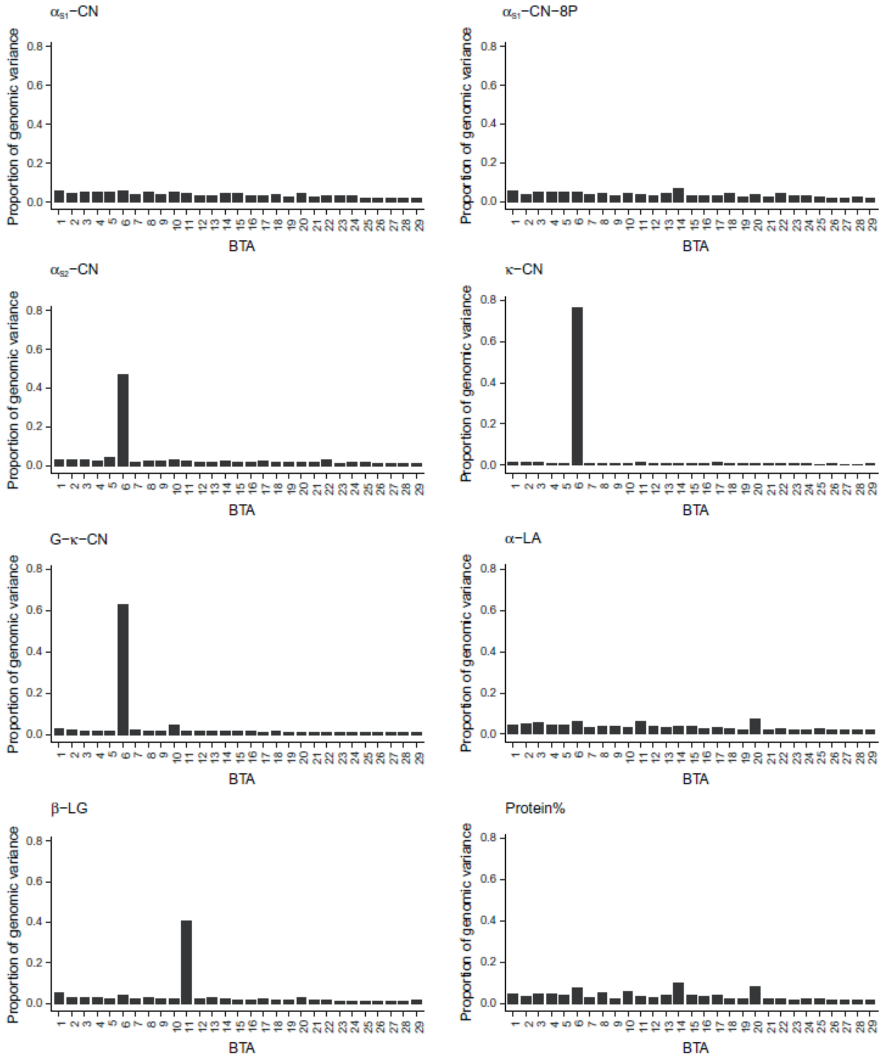

Fig. 3.1 Proportion of genomic variance explained by each chromosome. Proportion of the genomic variance in the milk protein composition traits explained by each chromosome from the ST-BayesAS model taking chromosomes as segments

Figure 3.2 shows the covariances between traits related to milk protein composition and total protein yield explained by genomic segments of 100 SNPs. Across the traits, some segments explained a large part of the covariance, whereas others accounted for nearly no covariance. Covariances between total milk protein yield and a particular trait were positive for some segments and negative for others. For G-κ-CN, κ-CN, β-LG, α_{S2}-CN, and protein percentage, a few segments showed peaks for the explained covariance. Segment 106, corresponding to a group of 100 adjacent SNPs on BTA6, explained a large amount of positive covariance of αS2-CN, κ-CN, and G-κ-CN with total protein yield. Similarly, a sizable proportion of the covariance between β-LG and protein yield was explained by a single segment on BTA11. A segment on BTA14 explained a substantial part of the negative covariance between protein percentage and protein yield. The same segment showed a peak for the covariance between αS1-CN-8P and total milk protein yield compared to the rest of the segments. Although some segments explained relatively more covariance between αS1-CN and total protein yield and between α-LA and total protein yield compared to other segments, the actual covariance values explained by these segments were very low (note the difference in y-axis scales between plots in Figure 3.2).

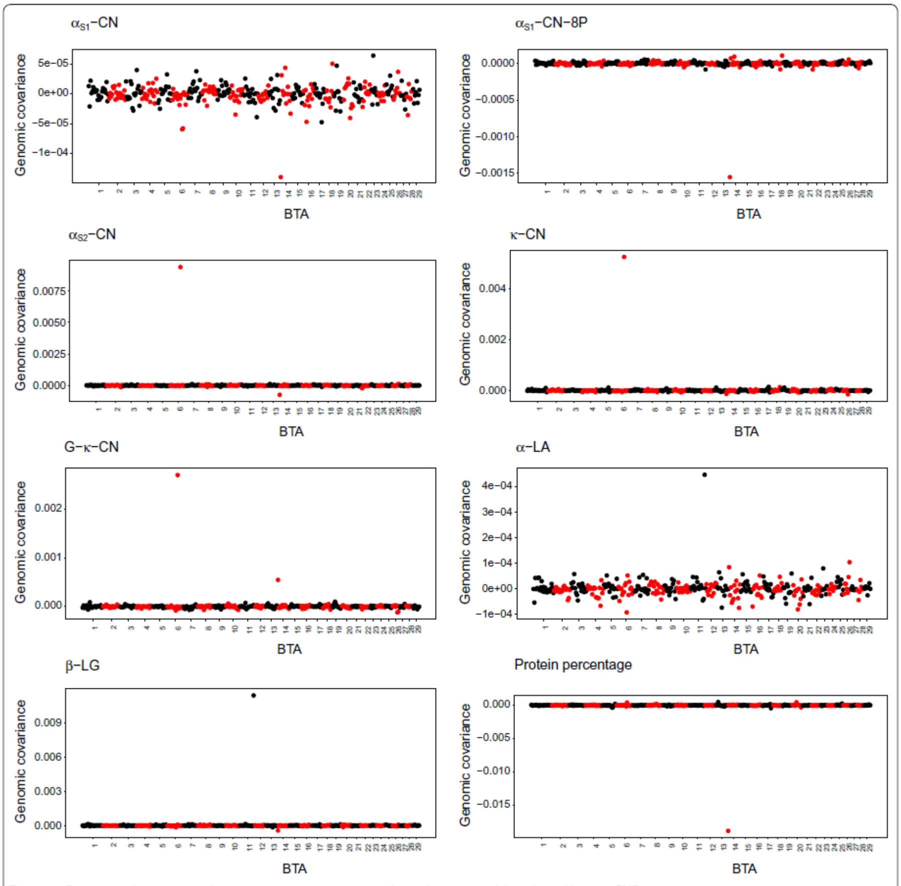

Fig. 3.2 Covariance between each protein composition trait with total protein yield explained by 100-SNP genomic segments

3.3.4. Prediction reliability with BayesAS models

Prediction reliabilities for cows using the BayesAS models were generally high compared to those obtained with the GBLUP models across all traits. Prediction reliabilities using both the MT- (Figure 3.3) and ST-BayesAS models were generally high for most of the highly heritable traits, such as κ-CN, G-κ-CN, and β-LG, using different segment sizes. Using the 100-SNP segment size resulted in the highest prediction reliability for all studied protein composition traits in both univariate and bivariate versions of the BayesAS models. Prediction reliabilities using the 100-SNP segment size with the MT-BayesAS model were 0.76 for G-κ-CN, 0.68 for κ-CN, and 0.52 for β-LG. Expanding the segment size to include all SNPs on a chromosome or the whole genome resulted in the lowest prediction reliabilities with the BayesAS models. The performance of the whole-genome-based model was similar to that of the respective GBLUP models. With the MT-BayesAS model, improvement in prediction reliability reached 63% for G-κ-CN, 52% for κ-CN, 31% for β-LG, and 15% for α_{S2}-CN when using the 100-SNP-based model compared to the whole-genome-based model. Prediction reliabilities were low for α_{S1}-CN, α-LA, and α_{S1}-CN-8P for all BayesAS models and improved minimally by using the 100-SNP-based model compared to the whole-genome approach. The 50- and 100-SNP models performed similarly well for β-LG. However, for the other proteins, the 100-SNP model outperformed both the 50- and 200-SNP based models, which generally showed comparable results. Using the single-SNP segment size resulted in lower performance compared to the 50-, 100-, and 200-SNP-based models for all traits. Prediction reliabilities computed for β-LG and protein percentage using the single-SNP-based MT-BayesAS model were better than when each chromosome (by 13 and 1 percentage points) or the whole genome was used as the segment (by 18 and 2 percentages points), respectively.

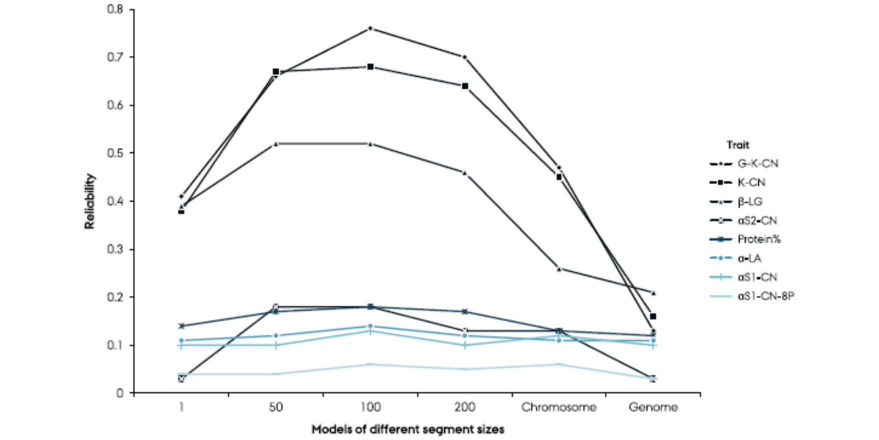

Fig. 3.3 Prediction reliability across MT-BayesAS models. Reliability of models according to segment sizes of 1, 50, 100, and 200 SNPs, chromosome, and whole genome. G-κ-CN = glycosylated-κ-CN; α_{S1}-CN-8P = α_{S1}-CN with eight phosphorylated serine groups

In general, slight additional gains in prediction reliability were achieved using the MT-BayesAS models compared to the univariate BayesAS model (Table 3.3), i.e. 6 and 5 percentage points for G-κ-CN and κ-CN using 100 SNP-segments and the average improvement with this segment size was 3 percentage points. However, improvement in prediction reliability from the MT-BayesAS models declined when the whole genome was taken as segment, which resulted basically in similar performances than the ST version except for β-LG.

Table 3.3 Prediction reliability from univariate and bivariate BayesAS models

[a]Trait	BayesAS-1SNP		BayesAS-100SNP		BayesAS-Genome	
	MT	ST	MT	ST	MT	ST
α_{S1}-CN	0.10	0.09	0.13	0.09	0.10	0.09
α_{S1}-CN-8P	0.04	0.02	0.06	0.03	0.03	0.03
α_{S2}-CN	0.03	0.03	0.18	0.16	0.03	0.03
κ-CN	0.38	0.37	0.68	0.63	0.16	0.16
G-κ-CN	0.41	0.39	0.76	0.70	0.13	0.14
α-LA	0.11	0.09	0.14	0.14	0.11	0.11
β-LG	0.39	0.39	0.52	0.50	0.21	0.19
Protein%	0.14	0.14	0.18	0.17	0.12	0.11

[a]Protein composition expressed as a fraction of the total milk protein percentage

by weight wt (wt/wt), protein% expressed as percentage of the total milk yield; individual proteins comprise only the peaks identified as intact proteins and isoforms; that is, α_{S1}-CN (comprises α_{S1}-CN 8P + 9P), α_{S2}-CN (comprises α_{S2}-CN 11P + 12P), κ-CN (comprises κ-CN G 1P + unglycosylated κ-CN 1P), where P = phosphorylated serine group. G-κ-CN = glycosylated-κ-CN; α_{S1}-CN-8P = α_{S1}-CN with 8 phosphorylated serine groups

3.3.5. Reliabilities of models for bulls

Table 3.4 shows the reliabilities of the MT-BayesAS models for bulls with segments of different sizes. Prediction reliability computed for the cow datasets was higher than that for bulls for G-κ-CN while the reverse was found for α_{S2}-CN. Higher model reliabilities were computed for bulls for α_{S2}-CN, κ-CN, G-κ-CN and β-LG with the 50- and 100-SNP segments compared to the other MT-BayesAS models. On the contrary, prediction reliability did not vary much across models for α_{S1}-CN, α-LA, α_{S1}-CN-8P and protein percentage, which had relatively low reliabilities. Prediction reliabilities obtained from the MT-GBLUP model were similar to those from the genome-based MT-BayesAS model for all traits and hence are not presented in Table 3.4.

Table 3.4 Model reliability for bulls across the MT-BayesAS models

[a]Trait	MT-BayesAS model reliability					
	1	50	100	200	Chromosome	Genome
α_{S1}-CN	0.05	0.06	0.04	0.06	0.05	0.06
α_{S1}-CN-8P	0.06	0.06	0.06	0.06	0.06	0.07
α_{S2}-CN	0.12	0.32	0.32	0.26	0.21	0.14
κ-CN	0.56	0.71	0.71	0.68	0.56	0.21
G-κ-CN	0.42	0.56	0.56	0.54	0.39	0.15
α-LA	0.07	0.07	0.08	0.08	0.08	0.06
β-LG	0.37	0.50	0.51	0.49	0.27	0.19
Protein %	0.23	0.22	0.22	0.21	0.19	0.18

[a]Protein composition expressed as a fraction of the total milk protein percentage by weight wt (wt/wt), protein% expressed as percentage of the total milk yield; individual proteins comprise only the peaks identified as intact proteins and isoforms; that is, α_{S1}-CN (comprises α_{S1}-CN 8P + 9P), α_{S2}-CN (comprises α_{S2}-CN 11P + 12P), κ-CN (comprises κ-CN G 1P + unglycosylated κ-CN 1P), where P = phosphorylated serine group. G-κ-CN = glycosylated-κ-CN; α_{S1}-CN-8P= α_{S1}-CN with

8 phosphorylated serine groups.

3.3.6. Prediction reliabilities with selected genome segments

Figure 3.4 shows prediction reliabilities according to the proportion of selected 100-SNP segments used in the prediction. Using fewer segments that explain large proportions of the variances resulted in higher predictive ability for G-κ-CN, κ-CN, β-LG, α_{S2}-CN, and protein percentage. For these traits, prediction reliability using only 2% (8) of the top-ranked segments resulted in the highest reliability, whereas prediction reliability decreased as more segments were added. In contrast, prediction reliability increased as more segments were added for α-LA, α_{S1}-CN, and α_{S2}-CN-8P, with the highest reliability obtained when all segments were used for prediction.

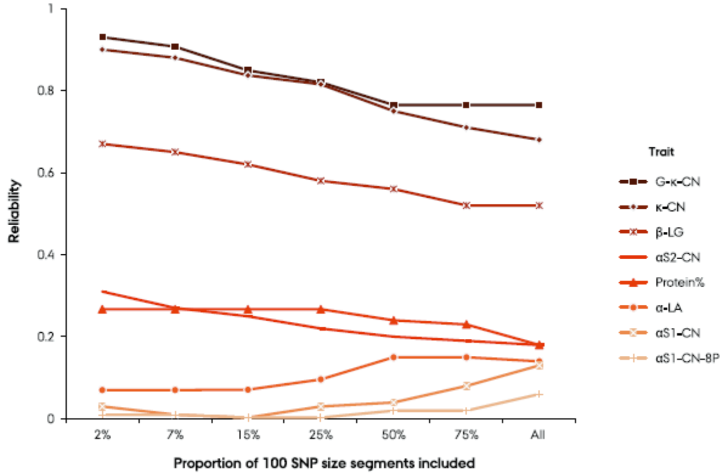

Fig. 3.4 Reliability of prediction using various proportions of genomic segments. Predictions were based on post-Gibbs analyses of samples from the MT-100-BayesA model. Segments were ranked based on explained covariance separately for each training set

3.4 Discussion

3.4.1. ST-GBLUP vs. MT-GBLUP models

Using only the cow dataset with the GBLUP model resulted in low prediction reliability, due to the small size of the training dataset. Reference population size is a key factor that affects reliability of genomic prediction in cattle (Daetwyler et al., 2008; Hayes et al., 2010; VanRaden et al., 2009). Moreover, a small sample size may not sufficiently reflect the genetic variability. For instance, considering a subset of the cow dataset used in this study, Poulsen et al. (2013) showed that the genetic variation of the *CSN1N1* gene was very low in Danish Holstein, with most individuals having the *BB* genotype, which may explain the lower prediction reliability for α_{S1}-CN and its sub-fraction α_{S1}-CN-8P.

Although information on total protein yield from a large number of bulls was added when using the MT-GBLUP model, prediction reliabilities were as poor as, or even worse than, those in the univariate analysis. Thus, addition of information from total milk protein yield was not sufficient to offset the computational burdens of the bivariate analyses, due to the low genome-wide correlation between protein yield and composition traits. Among the milk proteins, the highest genome-wide correlation with total protein yield was measured for α_{S2}-CN (-0.16) and protein percentage (-0.14), for which the MT-GBLUP model resulted in slightly improved prediction reliabilities for cows and bulls. Although α-LA had a correlation of 0.15 with total protein yield, the standard error of the correlation was higher than the correlation estimate (0.16). Although the data used was limited, our findings on genome-wide correlations were comparable to results from previous studies. In the literature, genetic correlations between milk protein percentage and protein yield in different dairy cattle breeds are low, in general (Meyer, 1985; Chauhan and Hayes, 1991; Roman and Wilcox, 2000).

Moreover, all the bivariate analyses in our study involved combination of data on different scales, which may have influenced the computed reliabilities. DRP for milk protein yield were expressed on a lactation basis (305-day production), whereas protein composition traits and percentage were related to one morning milk sample. In our study, prediction reliabilities for the traits related to milk protein composition traits were expected to improve if both traits in the bivariate analyses were on a similar scale.

3.4.2. Predictive ability of BayesAS models

Prediction reliabilities from the resampling showed large improvements with the

ST- and MT-BayesAS models compared to their GBLUP counterparts. The BayesAS models allow for different variances and covariances by SNP groups, which can deviate from the genome-wide (co)variance. This was especially important for some traits for which one or two key segments alone explained a large part of the total variance. Grouping adjacent SNPs seems to have helped to get more reliable estimates from a small dataset while allowing the segments to have different variances disentangled heterogeneous (co)variance patterns and improved prediction reliability. Similarly, a simulation study by Shariati et al. (2012) showed that prediction reliability based on SNP grouping was better than that obtained by SNP-BLUP methods. SNP grouping in the study of Shariati et al. (2012) was based on similar effect sizes. Other grouping options also exist, e.g. depending on LD between SNPs (Dehman et al., 2015). The BayesAS models can also be used to implement such grouping strategies for which segment sizes might vary depending on LD or effect size similarity.

Prediction reliability with the BayesAS models appears to depend highly on the segment sizes considered and the genetic architecture of the traits. Comparison between the BayesAS models with different segment sizes showed that grouping 100 adjacent SNPs resulted in superior performance for all proteins. Grouping 50 SNPs was as predictive as the models based on 100-SNP segments for all traits except G-κ-CN for which prediction reliability improved by 9 percentage points with the 100-SNP segment size model. Taking each SNP as a segment resulted in lower prediction reliability than groups of 50, 100, or 200 adjacent SNPs for most traits. With our BayesAS models, prediction reliabilities decreased as segment size increased beyond 100 SNPs in both the univariate and bivariate analyses. The lowest reliabilities were obtained when considering each chromosome or the whole genome as segments. In other words, the (co)variance between segments was diluted as segment size increased beyond 100 SNPs. Similarly, Brøndum et al. (2012) reported that using a segment size of 100 SNPs resulted in the highest accuracy in an across-breed genomic prediction study for protein, fat, and milk yield using 465,000 SNPs. Defining the optimal segment size, in terms of number of adjacent SNPs, is critical to achieving meaningful gains from the novel models presented here. Optimal segment size should be established for each specific situation, for instance through some resampling strategy, considering the SNP array, species, and LD in the population.

The gain in prediction reliability from using different segment sizes in the BayesAS models also varied across the traits. In both the ST- and MT-BayesAS models, differences in prediction reliability between segment sizes were very large for G-κ-CN, κ-CN, α_{S2}-CN, and β-LG, whereas across all models they were smaller for α_{S1}-

CN, α_{S1}-CN-8P, or α-LA. These results are likely related to the genetic architecture of the protein composition traits investigated. Previous genome-wide association studies found that the proportions of κ-CN, α_{S2}-CN, and β-LG in milk are controlled by major QTL on BTA6 and 11 (Schopen et al., 2011), which carry the casein gene cluster and the gene encoding β-LG (Buitenhuis et al., 2016), respectively. On the one hand, a single chromosome could explain a very large proportion of the variance for some protein composition traits, including G-κ-CN, κ-CN, β-LG, and α_{S2}-CN, which showed the largest improvement in reliability when the heterogeneity of variances across the genome segments was accounted for. On the other hand, the proportion of explained variance by each chromosome was very small for α_{S1}-CN and α-LA, which indicates that many segments contribute small proportions to the average variance. Similarly, Buitenhuis et al. (2016) found no major region that was significantly associated with α_{S1}-CN in the Danish Holstein population, which could be associated to the low genetic variability of the *CSN1N1* gene reported for this population by Poulsen et al. (2013). This result indicates that SNP grouping is more useful for traits that are controlled by QTL with major effects.

Comparison between the univariate and bivariate versions of our BayesAS models showed that for the most informative traits, the MT version resulted in further improvements in prediction reliability of up to 6 percentage points for segment sizes of 100 and 50. While further improvements in prediction reliability of up to 6% from the MT-BayesAS over the univariate versions are still important, it was generally lower than expected. Further investigations are required to understand the impact of genetic architecture of the indicator trait(s) on the potential advantages, over univariate analysis, of our bivariate BayesAS models.

A few segments explained a substantial proportion of the genomic variance for traits related to milk protein composition and their covariance with protein yield. Thus, we investigated the reliability of predictions based on only a few of the best-explaining 100-SNP segments. Predictions based on only 2% (8/372) of the genome segments resulted in the highest prediction reliability for G-κ-CN, κ-CN, β-LG, and α_{S2}-CN. For these proteins, prediction reliability decreased as more segments were added. Inclusion of more segments that explained a smaller proportion of the (co)variance added noise rather than meaningful information. Similarly, in a simulation study based on a GBLUP approach Sarup et al. (2016) demonstrated that including non-causal markers led to dilution of the effect of causal markers and reduced predictive ability. For other protein composition traits, including α_{S1}-CN-8P, α_{S1}-CN, and α-LA, prediction reliability improved as more segments were included, with the highest prediction reliability being obtained when all segments were considered. This result is in agreement with our finding on the proportions of genomic covariance explained by 100-SNP segments, where many segments across the genome contributed small proportions of the average covariance between these traits and total protein yield. In this study, we have used the same dataset to rank the top segments and do the prediction. This could lead to overestimation of reliability and introduce prediction bias. However, such bias is expected to be minimal as the SNP effects in these top segments are re-estimated for prediction with the different proportion of segments.

3.5 Conclusions

A novel BayesAS model, which allows to explore and model heterogeneous variance and covariance patterns across genomic regions, improved prediction reliabilities for milk protein composition traits with small dataset compared to the GBLUP and single-SNP based Bayesian models. The number of adjacent SNPs grouped together affected prediction reliability for the BayesAS models. A segment size of 100 SNPs gave the highest prediction reliability using 36,000 SNPs spread across the genome. For the most informative traits (highest genomic reliability), a further gain in reliability was observed when using the bivariate versions of our BayesAS models compared to univariate counterparts. Our results also show that the gains in prediction reliability achieved by SNP grouping depend on the genetic architecture of the traits. A future study with simulated data would be useful to test our novel BayesAS models with larger datasets.

Authors' contributions

GG planned the study, processed the data, implemented the models and drafted the manuscript. MSL conceived and co-supervised the study and contributed to the discussion of the results. BB supervised the project and contributed to the discussion of the results. HB contributed to the discussion of results. NAP collected the milk samples and contributed to the milk analysis and discussion of the results. LG developed the models and contributed to the discussion of the results. All authors read and approved the final manuscript.

Acknowledgements

GG is enrolled in the Erasmus-Mundus joint doctorate European Graduate School in Animal Breeding and Genetics. This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research. The study used data from the Danish-Swedish Milk Genomics Initiative (www.milkgenomics.dk) and the Milk Levy Fund (Denmark) projects: “Phenotypic and genetic markers for specific milk quality parameters” and “The importance of the metagenome for milk composition and quality". BB and NAP were supported by the project “Breeding high value milk: BIG MILK” of the Milk Levy Fund, Denmark (2015-2018).

References

Bobe G., Beitz D.C., Freeman A.E., Lindberg G.L. (1999). Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. J Dairy Sci. 82:2797-804.

Brøndum R.F., Su G., Lund M.S., Bowman P.J., Goddard M.E., Hayes B.J. (2012). Genome position specific priors for genomic prediction. BMC Genomics. 13:543.

Buitenhuis B., Poulsen N.A., Gebreyesus G., Larsen L.B. (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 Genet. 17:114.

Calus M.P., Veerkamp R.F. (2011). Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol. 43:26.

Chauhan V.P., Hayes J.F. (1991). Genetic parameters for first milk production and composition traits for Holsteins using multivariate restricted maximum likelihood. J Dairy Sci. 74:603-10.

Cole J.B., VanRaden P.M., O'Connell J.R., Van Tassell C.P., Sonstegard T.S., Schnabel

R.D., Taylor J.F., Wiggans G.R. (2009). Distribution and location of genetic effects for dairy traits. J Dairy Sci. 92:2931-46.

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

Dehman A., Ambroise C., Neuvial P. (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. BMC Bioinformatics. 16:148.

Gebreyesus G., Lund M.S., Janss L., Poulsen N.A., Larsen L.B., Bovenhuis H., Buitenhuis A.J. (2016). Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J Dairy Sci. 99:2863-6.

Gianola D., de los Campos G., Hill W.G., Manfredi E., Fernando R. (2009). Additive genetic variability and the Bayesian alphabet. Genetics. 183:347-63.

Goddard M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 136:245-57.

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J., Goddard M.E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 20106:e1001139.

Henderson C.R., Quaas R.L. (1976). Multiple trait evaluation using relatives records. J Anim Sci. 43:1188-97.

Janss L. (2014). Disentangling pleiotropy along the genome using sparse latent variable models. In Proceedings of the 10th World Congress on Genetics Applied to Livestock Production: 17-22 August 2014; Vancouver. 2014.

Jensen H.B., Poulsen N.A., Andersen K.K., Hammershøj M., Poulsen H.D., Larsen L.B. (2012). Distinct composition of bovine milk from Jersey and Holstein-Friesian cows with good, poor, or noncoagulation properties as reflected in protein genetic variants and isoforms. J Dairy Sci. 95:6905-17.

Legarra A., Robert-Granié C., Croiseau P., Guillaume F., Fritz S. (2011). Improved Lasso for genomic selection. Genet Res (Camb). 93:77-87.

Madsen P., Jensen J. (2007). An user’s guide to DMU. A package for analyzing multivariate mixed models. 2007. Version 6, release 4.7. available at http:// dmu.agrsci.dk

Meuwissen T.H. (2009). Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol. 2009;41:35.

Meuwissen T.H., Hayes B.J., Goddard M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-29.

Mrode R.A. (2014). Linear models for the prediction of animal breeding values. 3rd edition. Wallingford: CAB International. 2014.

Meyer K. (1985). Genetic parameters for dairy production of Australian Black and White cows. Livest Prod Sci. 12:205-19.

Plummer M., Best N., Cowles K., Vines K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. R News. 6:7-11.

Poulsen N.A., Bertelsen H.P., Jensen H.B., Gustavsson F., Glantz M., Månsson H.L., Andrén A., Paulsson M., Bendixen C., Buitenhuis A.J., Larsen L.B. (2013). The occurrence of noncoagulating milk and the association of bovine milk coagulation properties with genetic variants of the caseins in 3 Scandinavian dairy breeds. J Dairy Sci.;96:4830-42.

Roman R.M., Wilcox C.J. (2000). Bivariate animal model estimates of genetic, phenotypic, and environmental correlations for production, reproduction, and somatic cells in Jerseys. J Dairy Sci. 83:829-35.

Sarup P., Jensen J., Ostersen T., Henryon M., Sørensen P. (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. BMC Genet. 17:11.

Schaeffer L.R. (2001). Multiple trait international bull comparisons. Livest Prod Sci. 69:145-53.

Schopen G.C., Heck J.M., Bovenhuis H., Visker M.H., van Valenberg H.J., van Arendonk J.A. (2009). Genetic parameters for major milk proteins in Dutch Holstein-Friesians. J Dairy Sci. 2009;92:1182-91.

Schopen G.C., Visker M.H., Koks P.D., Mullaart E., van Arendonk J.A., Bovenhuis H. (2011). Whole-genome association study for milk protein composition in dairy cattle. J Dairy Sci. 94:3148-58.

Shariati M.M., Sørensen P., Janss L. (2012). A two step Bayesian approach for genomic prediction of breeding values. BMC Proc. 6:S12.

Su G., Christensen O.F., Ostersen T., Henryon M., Lund M.S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One.7:e45293.

VanRaden P.M. (2008). Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414-23.

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

4

Combining multi-population datasets for joint genome-wide association and meta-analyses: the case of bovine milk fat composition traits

G. Gebreyesus[1,2], A. J. Buitenhuis[1], N. A. Poulsen[3], M. H. P. W. Visker[2], Q. Zhang[4], H. J. F. van Valenberg[5], D. Sun[4], and H. Bovenhuis[2]

[1]Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark; [2]Animal Breeding and Genomics Centre, Wageningen University, PO Box 338, 6700 AH Wageningen, the Netherlands; [3]Department of Food Science, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark; [4]Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture of China, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China; [5]Dairy Science and Technology Group, Wageningen University and Research, P.O. Box 17, 6700 AA Wageningen, the Netherlands

Submitted [Frontiers in Genetics]

Abstract

In genome-wide association (GWA) studies, sample size is the most important factor affecting statistical power that is under control of the investigator, posing a major challenge in understanding the genetics underlying difficult-to-measure traits. Combining datasets available from different populations for joint or meta-analysis is a promising alternative to increase sample sizes available for GWA studies. Simulation studies indicate statistical advantages from combining raw data or GWA summaries in enhancing quantitative trait loci (QTL) detection power. However, the complexity of genetics underlying most quantitative traits, which itself is not fully understood, is difficult to fully capture in simulated datasets. In this study, population-specific and combined-population GWA as well as different meta-analyses were carried out with the objective of assessing the advantages and challenges of different data combining strategies in enhancing detection power of GWA studies using milk fatty acid (FA) traits as examples. Gas chromatography (GC) quantified milk FA samples and high density (HD) genotypes were available from 1566 Dutch, 614 Danish and 700 Chinese Holstein Friesian cows. Using the joint GWAS, 28 additional genomic regions were detected with significant associations to at least one FA compared to the population-specific analyses. Most of these additional regions were also detected using the different meta-analyses methods employed. Furthermore, using the confirmed regions of diacylglycerol acyltransferase 1 (*DGAT1*) and stearoyl-CoA desaturase (*SCD1*) genes, we show that significant associations were established with more FA traits in the joint GWA than the remaining scenarios. However, there were few regions detected in the population-specific analyses that were not detected using the joint GWA or the meta-analyses. These non-overlapping population-specific detections are shown to be highly likely caused by genotype by feed interactions. Our results show that combining multi-population dataset can be a powerful tool to enhance detection power in GWA studies for scarcely recorded traits. Detection of higher number of regions using the different meta-analyses methods, compared to any of the population-specific analyses, also emphasizes utility of these methods in the absence of raw multi-population datasets to undertake joint GWA.

Key words: multi-population GWAS, fatty acid, milk, meta-analysis, Holstein Friesian, combining datasets

4.1 Introduction

In GWA studies, sample size is the most important factor that affects statistical power and is under control of the investigator. Limited sample size is hence a major hurdle in GWA studies for traits that are difficult or expensive to measure. In the livestock breeding industry, emerging phenotypes of interest for selective breeding are often expensive or difficult to measure. Measurements for such traits are limited to experimental samples from different populations. One option to deal with the limitation of sample size in understanding the genetics underlining such traits could be to combine the available smaller datasets for joint GWA (mega-analysis) or to combine summaries of the individual GWA for meta-analysis.

Combining datasets for large-scale joint GWA has been used as an effective method to increase GWA power in human disease association studies (e.g. Consortium et al., 2013; Sung et al., 2013) and to some extent in livestock studies (e.g. Veerkamp et al., 2012). An alternative approach for the discovery of QTL for common human diseases (Begum et al., 2012) and livestock phenotypes (Rubio et al., 2015; Bouwman et al., 2018) has been the meta-analysis of individual GWA studies. The analysis of data summarized from multiple independent studies is expected to increase power while avoiding the limitations imposed by restrictions on sharing individual-level data (Evangelou and Ioannidis, 2013). Different methods of meta-analysis have been proposed depending on the sources of information used and assumptions regarding SNP effects in the different populations. The most common approaches are to combine p-values/transformed p-values, such as in the weighted z-score method, and the use of SNP effects, as implemented either in fixed or random effects models. Some of these approaches weigh individual studies based on sample sizes and some assume SNP effects are different in the different populations/individual studies. The relative performance of the different meta-analyses approaches depends on existence and extent of heterogeneity between studies and differences in sample sizes.

Theoretical illustrations and simulation studies have indicated statistical advantages from combining datasets and GWA summaries in enhancing QTL detection power (e.g., Costafreda, 2009; Lin and Zeng, 2010). However, the complexity of genetics underlying most quantitative traits, which itself is not fully understood, is difficult to fully capture in simulated datasets. In this context, it is worth investigating the utility of these different approaches of combining datasets using real data and comparing results, for instance, in known QTL regions with confirmed connections to the studied traits.

In this study, we investigated the advantages and challenges pertinent to combining multi-population datasets for joint GWA and meta-analysis of population-specific studies using milk fatty acids (FAs) measured on Dutch, Danish and Chinese Holstein Friesians as example traits. Milk fat composition traits have attracted growing interest, mainly in relation to implication on human health. Better understanding of the genetics underlying these traits could help implement selective breeding for milk with specific fat composition. Detailed milk fatty acid composition is not routinely recorded. Gas chromatography (GC) analysis is currently the method of choice in determination of milk fat composition with high accuracy. However, this method is expensive and time consuming, thus, limiting the measurement of milk fat composition to experimental samples.

Combining multi-population datasets is not straightforward and comes with its own challenges. Heterogeneity of samples from the different populations is a major hurdle. Such heterogeneity might arise, for instance, due to genetic distance between the populations, differences between trait measurements, different environmental exposures, and different genotyping chips (Begum et al., 2012).

With the objective of assessing the advantages and challenges of different data combining strategies in enhancing detection power of GWA studies, this study compared detection of genomic regions for milk FA traits using population-specific GWA studies, joint GWA on combined population dataset and three different approaches of meta-analyses. Detection of significant associations on the previously confirmed regions of *DGAT1* and *SCD1* received due emphasis in comparing results of the different GWA scenarios. Possible sources of heterogeneity in the FA between the sample populations and potential implications of these on the different GWA scenarios are discussed.

4.2 Material and Methods

4.2.1 Animals and phenotypes

Measurements for 13 FA traits including C8:0, C10:0, C12:0, C14:0, C14:1, C15:0, C16:0, C16:1, C18:0, C18:1c9, C18:2n6, C18:3n3 and C18:2 cis-9,trans-11 (CLA) were obtained from test day milk samples of 784 Chinese, 675 Danish and 1566 Dutch Holstein cows. Quantification of the FA traits was based on the GC method as previously presented in details by Li et al. (2014) for Chinese samples, Poulsen et al. (2012) for Danish samples and Stoop et al. (2008) for the Dutch samples. Desaturation indexes were also calculated based on the FA measurements as: C14 index = C14:1/(C14:1+C14:0) x 100; C16 index = C16:1/(C16:1+C16:0) x 100 and C18 index = C18:1c9/ (C18:1c9+C18:0) x 100.

Cows were sampled from 18 herds in China, 22 herds across Denmark and 398 herds in the Netherlands. Stage of lactation in sampled cows ranged between 3 to 700 days in milk in the Chinese population, 9 to 481 days in milk in the Danish population and 60 to 278 days in milk in the Dutch Holstein cows. To standardize the samples from the three countries, only cows at days-in-milk of 60 and above were considered for the association analyses. Thus, 700 Chinese, 614 Danish and 1566 Dutch samples were available for the association analyses. The reason to standardize the dataset by lactation stage is that the genetic determination of milk fat composition traits might be different in the early stages of lactation. There is evidence that effects of genes in early lactation differ from those later in lactation (e.g., Bovenhuis et al., 2015). By excluding early lactation records we eliminate this heterogeneity issue.

4.2.2 Genotypes and Imputation

The BovineSNP50 Beadchip (50K, Illumina) was used to genotype cows in the Chinese dataset. Imputation of the 50K genotypes to HD was then performed with the Fimpute software package (Sargolzaei et al., 2014) using reference population of 96 Chinese Holstein bulls genotyped with BovineHD Beadchip (777K). In the Danish dataset, 278 cows were genotyped using the BovineSNP50 Beadchip. The remaining Danish Holstein cows were genotyped using the BovineHD Beadchip and used as reference to impute the 50K genotypes of the first part of the Danish cows to HD as described in Gebreyesus et al. (2016).

A custom 50K SNP Beadchip, designed by CRV (Arnhem, Netherlands), was used to genotype all cows in the Dutch dataset. A reference population of 1,333 animals from the Dutch Holstein, with HD genotypes was then used to impute to the 50K genotypes to HD as described in Duchemin et al. (2014). SNPs with minor allele frequencies (MAF) less than 0.05 or with a count of one of the genotypes less than 10 in each population were excluded from the association analyses, i.e., population-specific as well as joint GWA. Total of 464,130 SNPs were available in common for the population-specific as well as combined-population analyses. The SNP positions were based on the bovine genome assembly UMD 3.1 (Zimin et al., 2009).

4.2.3 Genome-wide bin-wise linkage disequilibrium and MAF analysis

Genome-wide pair-wise linkage disequilibrium (LD) was calculated between the SNP markers within a 1 Mbp window along the genome using the r^2 as a measure in the Plink program (Purcell et al., 2007). Correlation of MAF in the three populations was assessed for non-overlapping bins of 100 SNPs throughout the genome.

4.2.4 Statistical analysis

4.2.4.1 Test for phenotypic differences

Two tailed t-test was carried out for testing the differences in phenotypic means of the FAs between the three populations using the t.test default function in R (R Core Team, 2017). Similarly, F-test was carried out to test differences in standard deviations pairwise. The test criterion was: $F = \sigma_1/\sigma_2$, where σ_1 is the larger of the two standard deviations and with degrees of freedom *f1* for the larger standard deviation and *f2* for the smaller standard deviation (σ_2).

4.2.4.2 Association analysis

A single-SNP association test was implemented using a mixed linear model in the GCTA program (Yang et al., 2011). Population-specific and combined-population analyses were undertaken using the following statistical model:

$$y_{ijkl} = \mu + parity_i + herd_j + b_1 * DIM_{ijkl} + b_2 * SNP_k + animal_l + e_{ijkl}, \text{ (1)}$$

Where y_{ijkl} is the phenotype of cow l; µ is the fixed effect of mean; parity_i and herd_j are the fixed effects of parity and herd, respectively; b_1 is the regression coefficient for DIM, DIM_{ijkl} is a covariate of days in milk, since only cows with more than 60 days-in-milk were included in the analyses, a linear adjustment for days in milk was sufficient; b*2* is the allele substitution effect for SNP, SNP_k is a covariate indicating the number of copies of a specific allele (0, 1 or 2) of the SNP; and animal is the random additive genetic effect. Animal effects were assumed to be distributed as $\mathrm{N}(0, \mathbf{G}\sigma_a^2)$, where **G** is the genomic relationship matrix constructed excluding the chromosome on which the SNP k is located. Residuals were assumed to be distributed as $\mathrm{N}(0, \mathbf{I}\sigma_e^2)$ where **I** is the identity matrix.

Homogeneity of residuals was assessed by plotting the residuals against predicted phenotypes from the model used to estimate heritability, i.e. model (1) without the SNP effect. For some of the FAs, especially for C18:2n6, C18:3n3 and CLA, residuals tend to increase with the mean indicating heterogeneity of the residual variance (Figure 4.1). Therefore, records for these FAs were log transformed in both the combined as well as the population-specific analyses.

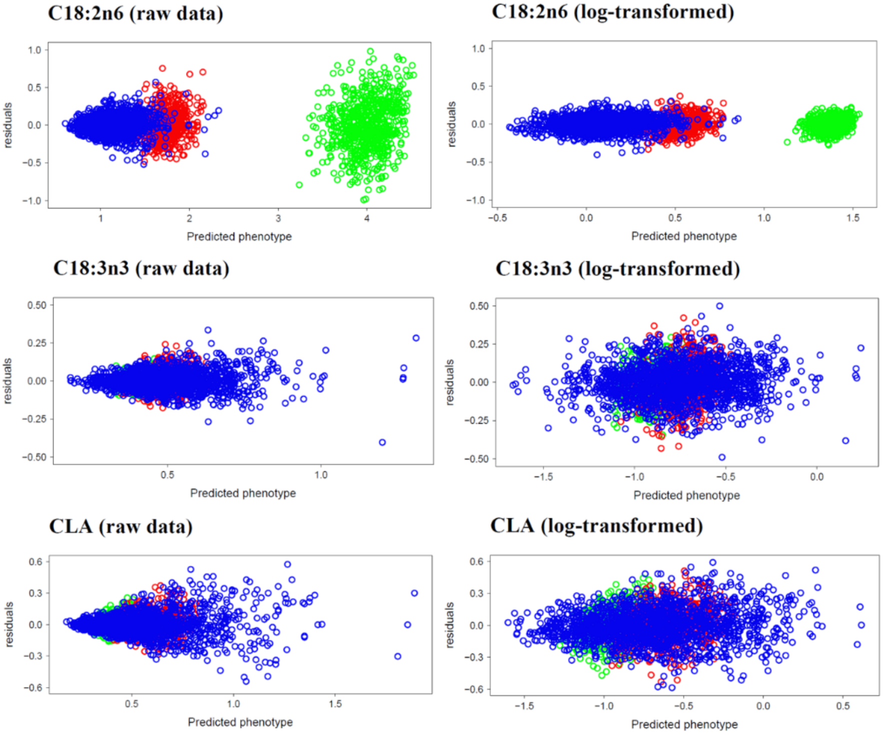

Fig 4.1. Residuals computed from the mixed-linear model with G-matrix constructed from all the SNPs plotted against predicted phenotypes in combined dataset colored according to population (blue dots for Dutch samples, red dots for Danish samples and green dots for Chinese samples) before (left) and after (right) in in C18:2n6, C18:3n3 and CLA

Significance thresholds were determined using a false discovery rate (FDR). Significance thresholds corresponding to FDR of 5% ranged for different FA from –log10 p-value = 3.4 to –log10 p-value = 5. We used a –log10 p-value of 5 as the genome-wide significance threshold for all FA composition traits. To determine if a region harbored one or more QTL, the lead SNP with the highest –log10 p-value on each chromosome was fitted as fixed effect for subsequent association analyses. If a peak around such a “leadSNP” was no more observed in the subsequent analysis, all SNPs around the leadSNP meeting the significance threshold were considered as part of that single region. If one or more peaks remained after the subsequent analysis, a different QTL was assumed and the SNP with the next highest –log10 p-value was taken as the next leadSNP and the procedure repeated until no SNP with significant association (–log10 p-value >5) remained.

Heritability (h^2) estimates were computed for the populations separately as well as the combined dataset as:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}, \qquad (2)$$

Where, σ_a^2 is the additive genetic variance estimated using model 1 but without fitting an effect for SNP and using G constructed from all SNPs, and σ_e^2 is the residual variance.

4.2.4.3 Meta-analysis

We compared the performance of our joint GWA with meta-analysis of summaries from the population-specific studies. Three different methods of meta-analysis were implemented using Plink program (Purcell et al., 2007), i.e., the fixed and random effects models, and the weighted z-score approach.

Fixed-effects model: Fixed effect meta-analysis methods assume SNP effects are similar across studies. Different implementations of the fixed effect meta-analysis model exist. In this study, estimates from each study were combined by weighing them according to the inverse of their standard error following Kavvoura and Ioannidis (2008). In short, a combined effect across population is computed as:

$$B = \frac{\sum_i \beta_i w_i}{\sum_i w_i},$$

where, β_i is the effect estimated for population *i* and w_i is the weight for the corresponding population computed as: $w_i = [Var(\beta_i)]^{-1}$.

Random-effects model: The random effects model assumes that the individual studies estimate different effects for each SNP. In this study, the Cochran's test statistic (Q) was first calculated as: $Q = \sum_i w_i \ (B - \beta_i)$. The between-study variance of heterogeneity, expressed as τ2, is then computed as:

$$\tau^2 = (Q-(N_i-1))/\left(\sum_i w_i - \left(\frac{\sum_i {w_i}^2}{\sum_i w_i}\right)\right),$$

Where, N_i is the number of individual studies, i.e., three in this case. The combined SNP effect is finally calculated as:

$$B^R = \frac{\sum_i \beta_i w_i^R}{\sum_i w_i^R},$$

where w_i^R is the random-effect weight for each population-specific study calculated by incorporating the between-study variance of heterogeneity (τ^2) as: $w_i^R = [\tau^2 + Var\ (\beta_i)]^{-1}$.

With the estimation of heterogeneity test statistic (Q) at SNP level, the random effects model is also able to give an estimate of the degree of heterogeneity between individual studies. However, as a test statistics, the Cochran's test statistic (Q) has been suggested as underpowered when the number of studies used for the meta-analysis is small. Other robust test statistics have been proposed (Higgins and Thompson, 2002) that are suggested to be scale and size invariant (Nakaoka, 2009). In our study, to study the effect of heterogeneity under the different meta-analysis scenarios, one of these heterogeneity statistics (I^2) was computed as:

$$I^2 = 100 * \frac{Q-(N_i-1)}{Q}.$$

Weighted z-scores method: In the weighted z-scores method, p-values are transformed into z-scores taking into account sample sizes (Whitlock, 2005) and direction of SNP effects in individual studies. In this method intermediate z-scores are first computed as: was computed as: $Z_i = \Phi^{-1}\left(1-\frac{p_i}{2}\right) * sign(\Delta_i)$, where Φ is the cumulative distribution function, p_i is the p-value for the i^{th} population and Δ_I is the direction of the SNP effect in population *i*. The overall weighted z-score is then calculated as: $Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$, where, w_i is the square root of sample size of the i^{th} population.

4.2.4.4 Power calculations

To quantify the theoretical expected gain in power as a result of combining the datasets for GWA analysis, we run a power test based on the sample size of each population and the combined dataset for varying scenarios of QTL explained variance assuming similar LD structures using the package ldDesign in R (Ball, 2004). The parameter settings used for the ldDesign package were allele frequencies of p = q = 0.5, assuming same LD between markers and QTL in the different populations with r^2 value of 0.3, and a significance threshold –log10 p-value of 5.

4.3 Results

4.3.1 Descriptive statistics and genetic parameters

Table 4.1 presents phenotypic means and the standard deviations for FA traits in the three populations. Significant differences between the three populations were observed in phenotypic means for several of the FAs. Phenotypic means in the Chinese samples were in general lower for the short and medium chain FAs and higher for most of the long chain FAs. The largest difference in phenotypic means between the three populations was observed for C18:2n6, which was three times higher in the Chinese samples (3.99) compared to the mean values in the Dutch (1.11) and Danish (1.74) samples. Large differences in phenotypic means were also shown for C8:0 and C18:1c9 between the Chinese samples on the one hand and the Dutch and Danish samples on the other. There were significant differences in standard deviations between the populations but not to the same extent as for the means. There were only three FAs where all three populations differ significantly. Standard deviations were generally lower for most FAs in the Chinese sample compared to those for the Dutch and Danish samples.

Table 4.2 presents additive genetic variances and heritability estimates for the studied milk FA traits in the Dutch, Danish and Chinese Holsteins as well as in the combined dataset. Due to relatively small sample sizes estimates of additive genetic variances in general showed large standard errors (some of which were larger than the estimates). For most FAs, additive genetic variances in Dutch and Danish samples were similar but additive genetic variances in the Chinese data differed (lower) from the other two populations. Heritability estimates were higher for most FAs in the Dutch samples compared to the Danish and Chinese samples. Heritability estimates from the combined analysis were moderate to high with the highest value estimated for C14 index (0.53) and the lowest for C18:3n3 (0.21).

Table 4.1. Phenotypic means and standard deviations for the milk FA traits in the different populations and combined dataset

FAs	NL (N=1566)		DK (N=614)		CN (N=700)		Combined (N=2874)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Saturated FAs[1]								
C8:0	1.31^{a}	0.17^{a}	1.47^{b}	0.22^{b}	0.58^{c}	0.22^{b}	1.18	0.38
C10:0	2.87^{a}	0.45^{a}	3.22^{b}	0.56^{b}	2.22^{c}	0.40^{a}	2.80	0.58
C12:0	3.79^{a}	0.72^{a}	3.69^{a}	0.68^{a}	2.94^{b}	0.49^{b}	3.58	0.76
C14:0	11.1^{a}	1.05^{a}	11.6^{b}	1.36^{b}	10.1^{c}	1.14^{ab}	11.0	1.26
C15:0	1.11^{a}	0.19^{a}	1.11^{a}	0.19^{a}	0.99^{b}	0.13^{b}	1.09	0.18
C16:0	29.1^{a}	3.50^{a}	30.1^{b}	3.49^{a}	32.9^{c}	1.84^{b}	30.2	3.53
C18:0	9.84^{a}	1.74^{a}	9.84^{a}	1.91^{a}	12.0^{b}	1.69^{a}	10.3	1.99
Unsaturated FAs[1]								
C14:1	1.38^{a}	0.27^{a}	1.01^{b}	0.28^{a}	0.86^{c}	0.21^{b}	1.19	0.35
C16:1	1.39^{a}	0.29^{a}	1.58^{b}	0.42^{b}	1.64^{b}	0.33^{c}	1.49	0.35
C18:1c9	20.2^{a}	2.78^{a}	19.6^{b}	2.84^{a}	28.3^{c}	2.44^{b}	21.9	4.37
C18:2n6	1.11^{a}	0.25^{a}	1.74^{b}	0.27^{a}	3.99^{c}	0.46^{b}	1.89	1.19
C18:3n3	0.50^{a}	0.16^{a}	0.50^{a}	0.09^{b}	0.42^{b}	0.06^{c}	0.48	0.13
CLA	0.56^{a}	0.26^{a}	0.57^{a}	0.15^{b}	0.41^{b}	0.09^{c}	0.53	0.23
Desaturation indexes [2]								
C14 index	11.0^{a}	1.83^{a}	7.98^{b}	1.89^{a}	7.84^{c}	1.63^{b}	9.71	2.37
C16 index	4.60^{a}	0.91^{a}	4.97^{b}	1.11^{a}	4.74^{c}	0.93^{a}	4.70	0.97
C18 index	67.3^{a}	3.88^{a}	66.6^{b}	3.90^{a}	70.2^{c}	3.27^{b}	67.8	3.98

[a,b,c] phenotypic means and standard deviations on the same row with different superscripts are significantly different at $p<0.001$; NL = Dutch Holstein, DK= Danish Holstein, CN = Chinese Holstein

[1] Expressed in % wt/wt

[2] Desaturation indexes calculated as unsaturated/(unsaturated + saturated) × 100

Table 4.2. Genetic parameters* (+ standard errors) for milk fatty acids in 1566 Dutch, 614 Danish and 700 Chinese Holstein samples

FAs	NL		DK		CN		Combined	
	$\sigma_a^2{}_{(SE)}$	$h^2{}_{(SE)}$	$\sigma_a^2{}_{(SE)}$	$h^2{}_{(SE)}$	$\sigma_a^2{}_{(SE)}$	$h^2{}_{(SE)}$	$\sigma_a^2{}_{(SE)}$	$h^2{}_{(SE)}$
C8:0	$0.01_{(0.04)}$	$0.48_{(0.06)}$	$0.01_{(0.07)}$	$0.33_{(0.10)}$	$0.002_{(0.04)}$	$0.06_{(0.05)}$	$0.008_{(0.04)}$	$0.27_{(0.03)}$
C10:0	$0.09_{(0.11)}$	$0.51_{(0.06)}$	$0.09_{(0.17)}$	$0.36_{(0.10)}$	$0.02_{(0.09)}$	$0.16_{(0.07)}$	$0.07_{(0.09)}$	$0.39_{(0.04)}$
C12:0	$0.12_{(0.14)}$	$0.40_{(0.06)}$	$0.10_{(0.19)}$	$0.30_{(0.10)}$	$0.04_{(0.13)}$	$0.21_{(0.07)}$	$0.09_{(0.11)}$	$0.33_{(0.04)}$
C14:0	$0.27_{(0.21)}$	$0.39_{(0.06)}$	$0.15_{(0.32)}$	$0.14_{(0.10)}$	$0.21_{(0.28)}$	$0.22_{(0.08)}$	$0.21_{(0.17)}$	$0.25_{(0.03)}$
C15:0	$0.006_{(0.04)}$	$0.29_{(0.06)}$	$0.007_{(0.05)}$	$0.27_{(0.10)}$	$0.001_{(0.02)}$	$0.10_{(0.07)}$	$0.004_{(0.02)}$	$0.23_{(0.04)}$
C16:0	$2.79_{(0.66)}$	$0.48_{(0.06)}$	$0.75_{(0.76)}$	$0.12_{(0.09)}$	$0.77_{(0.51)}$	$0.27_{(0.08)}$	$1.80_{(0.48)}$	$0.34_{(0.04)}$
C18:0	$0.78_{(0.39)}$	$0.37_{(0.06)}$	$0.53_{(0.49)}$	$0.23_{(0.10)}$	$0.54_{(0.43)}$	$0.25_{(0.08)}$	$0.52_{(0.29)}$	$0.25_{(0.04)}$
C14:1	$0.03_{(0.07)}$	$0.55_{(0.06)}$	$0.03_{(0.09)}$	$0.49_{(0.10)}$	$0.01_{(0.06)}$	$0.35_{(0.09)}$	$0.03_{(0.05)}$	$0.47_{(0.04)}$
C16:1	$0.05_{(0.08)}$	$0.65_{(0.05)}$	$0.07_{(0.13)}$	$0.42_{(0.10)}$	$0.02_{(0.09)}$	$0.26_{(0.09)}$	$0.05_{(0.07)}$	$0.46_{(0.04)}$
C18:1c9	$1.90_{(0.58)}$	$0.41_{(0.06)}$	$0.37_{(0.66)}$	$0.07_{(0.08)}$	$1.33_{(0.68)}$	$0.24_{(0.08)}$	$1.38_{(0.46)}$	$0.27_{(0.04)}$
C18:2n6	$0.007_{(0.04)}$	$0.27_{(0.06)}$	$0.01_{(0.07)}$	$0.17_{(0.09)}$	$0.03_{(0.12)}$	$0.26_{(0.10)}$	$0.01_{(0.05)}$	$0.18_{(0.03)}$
C18:3n3	$0.002_{(0.02)}$	$0.27_{(0.06)}$	$0.0004_{(0.02)}$	$0.05_{(0.08)}$	$0.0001_{(0.01)}$	$0.05_{(0.06)}$	$0.005_{(0.01)}$	$0.19_{(0.03)}$
CLA	$0.009_{(0.04)}$	$0.32_{(0.06)}$	$0.002_{(0.04)}$	$0.11_{(0.09)}$	$0.001_{(0.02)}$	$0.15_{(0.07)}$	$0.004_{(0.02)}$	$0.21_{(0.04)}$
C14 index	$1.81_{(0.47)}$	$0.62_{(0.05)}$	$2.10_{(0.65)}$	$0.59_{(0.10)}$	$0.89_{(0.51)}$	$0.36_{(0.09)}$	$1.57_{(0.37)}$	$0.53_{(0.03)}$
C16 index	$0.39_{(0.23)}$	$0.55_{(0.06)}$	$0.46_{(0.37)}$	$0.37_{(0.10)}$	$0.17_{(0.25)}$	$0.24_{(0.08)}$	$0.32_{(0.19)}$	$0.38_{(0.04)}$
C18 index	$6.99_{(1.03)}$	$0.49_{(0.06)}$	$3.46_{(1.18)}$	$0.26_{(0.10)}$	$1.90_{(0.83)}$	$0.21_{(0.07)}$	$3.95_{(0.73)}$	$0.31_{(0.04)}$

*parameter estimates were prior to any data transformation; NL = Dutch Holstein, DK= Danish Holstein, CN = Chinese Holstein

4.3.2. Consistency in LD and MAF

The genome-wide LD analysis showed that the three populations have similar LD structures across the genome (Figure 4.2). The maximum mean bin-wise LD was 0.71 for the three populations, while the minimum mean bin-wise LD was 0.07 for the Dutch and Danish populations and 0.06 for the Chinese population. Correlation in MAF between the populations averaged for bins of 100 SNPs throughout the genome was 0.87 between the Danish and Dutch populations and between the Danish and Chinese populations, and 0.81 between the Chinese and Dutch populations.

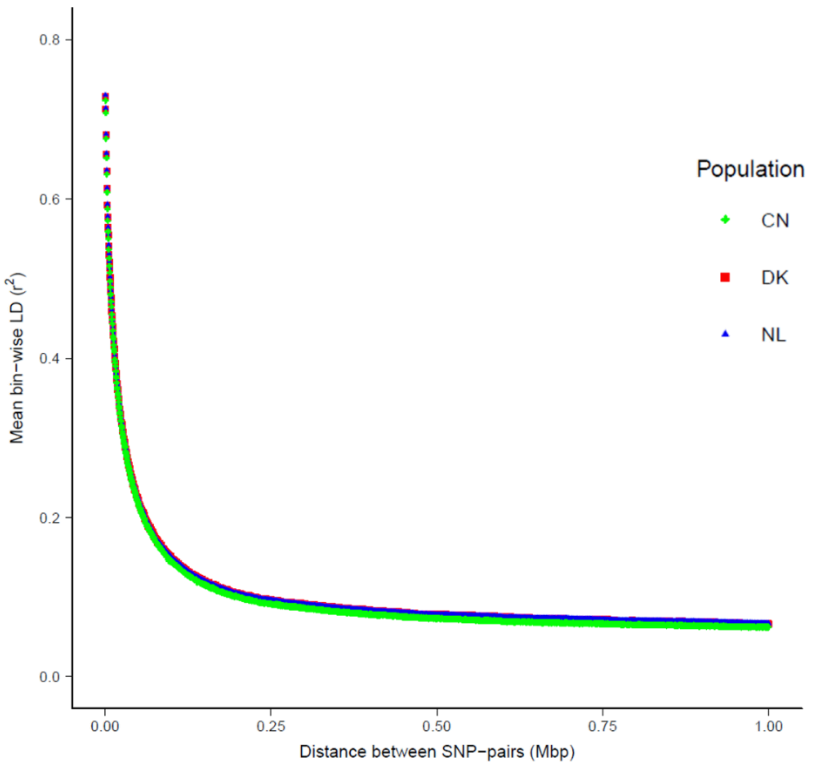

Fig 4.2. Linkage disequilibrium (LD) for the Dutch (blue points), Danish (red points) and Chinese (green points) Holstein Friesian genotypes. Y-axis is the mean bin-wise LD and x-axis the physical distance between pair-wise markers in Mega base pairs (Mbp)

4.3.3 Genomic regions detected across the different analyses

Table 4.3 shows genomic regions significantly associated with at least one FA trait using the different analyses. Using the different scenarios (population-specific, combined population and meta-analyses), a total of 68 genomic regions were found significantly associated with the studied FAs. Regions were identified on all the chromosomes except BTA 18. Only three regions (14a, 14b and 26) were commonly detected with significant association to at least one FA across the different scenarios, i.e., all population-specific analyses, joint GWA, as well as the three different meta-analyses.

The largest number of significantly associated regions was identified using the joint GWA on the combined dataset: 56 regions were detected with significant associations with at least one of the 16 FA traits studied. The detected regions were spread across all the chromosomes except BTA18. Of all regions detected using the joint GWAS, 28 regions were not detected in any of the population-specific analyses, suggesting increased detection power from combining the datasets. Our computation of theoretic detection power, as a function of sample size and proportion of explained variance by a QTL, also shows that a QTL explaining more than 5% of the genetic variance can be detected with a power of 0.97 in the combined dataset compared to a power of 0.57 in the Dutch, 0.08 in the Chinese and 0.05 in the Danish dataset (Figure 4.3).

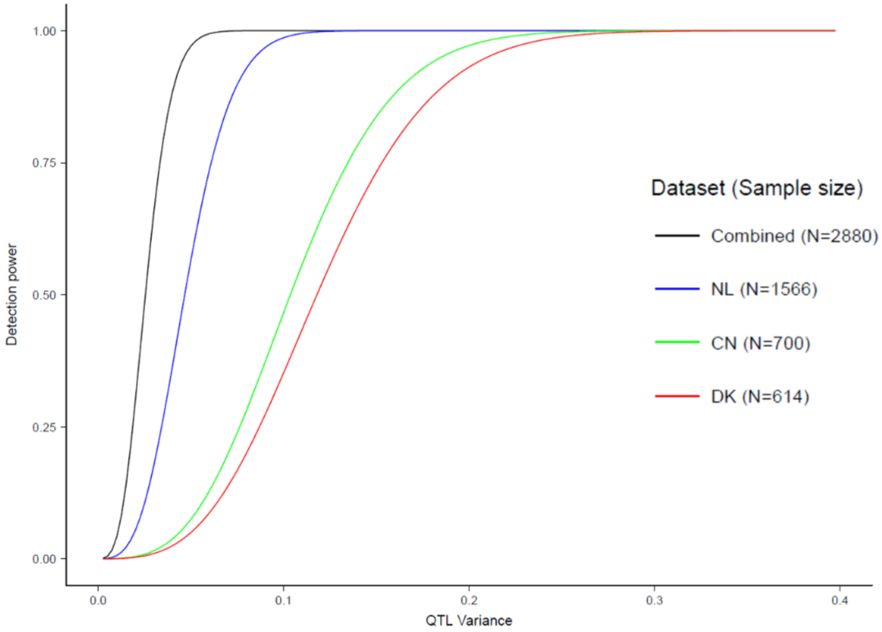

Fig 4.3. Theoretical expectations of GWAS detection power based on sample sizes for the Dutch (blue line), Danish (red line) and Chinese (blue line) datasets and the combined multi-population dataset (black line). On the y-axis is presented the detection power using the different scenarios (sample sizes) and on the x-axis is the simulated QTL variance

Table 4.3. Genomic regions associated with milk FA traits detected using population-specific, combined population and meta-analysis

Region*	Start	End	Number of FAs significantly associated**						
	(Mbp)	(Mbp)	NL	DK	CN	Combined	Meta-Fix	Meta-Rand	Meta-Zsc
1a	19.92	19.93	-	-	-	1	1	1	1
1b	60.0	60.0	-	-	-	-	1	1	1
1c	101.0	101.0	-	-	-	1	-	-	-
1d	132	132	-	2	-	-	-	-	-
1e	141.3	142.5	-	-	-	1	1	1	1
2a	12.5	19.8	-	-	-	2	2	2	2
2b	54.9	59.8	1	-	-	4	3	3	3
2c	64.1	67.8	-	-	-	2	2	2	2
2d	106.5	135.6	1	-	-	4	2	2	3
3a	8.5	8.7	1	-	-	-	-	-	-
3b	104.2	104.2	-	-	-	-	1	-	1
3c	116.2	119.4	-	-	-	2	-	-	1
4	15.59	15.6	-	-	-	1	1	1	1
5a	10.33	10.36	-	-	1	1	1	1	1
5b	65.7	82.8	-	-	1	2	2	2	2
5c	87.4	100.0	7	2	-	10	9	8	9
6a	20.6	20.6	-	-	-	-	1	-	1
6b	41.4	41.4	-	-	-	1	-	-	-
7a	5.05	5.09	1	-	-	-	-	-	-
7b	14.6	15.5	-	-	-	2	2	2	2
7c	78.4	78.4	-	-	1	1	1	1	1
7d	81.6	83.2	-	-	-	2	2	2	2

Table 4.3. *Continued*

Region*	Start (Mbp)	End (Mbp)	NL	DK	CN	Combined	Meta-Fix	Meta-Rand	Meta-Zsc
8a	57.5	59.7	-	-	-	3	2	1	3
8b	79.9	98.4	-	-	-	3	3	2	3
9a	25.5	25.6	-	-	-	1	-	-	1
9b	81.3	81.3	-	-	-	1	1	1	1
9c	97.1	97.2	1	-	-	-	1	1	-
10a	1.1	8.6	2	1	-	2	2	2	2
10b	11.7	12.9	1	-	-	2	1	1	1
10c	73.4	73.5	-	-	-	-	1	1	1
10d	78.1	80.1	1	-	-	1	-	-	1
10e	87.3	93.1	1	-	-	3	2	2	2
11a	24.7	26.7	-	-	-	1	1	1	1
11b	58.81	58.89	-	-	-	1	1	1	1
12a	17.1	17.1	-	-	-	1	1	1	1
12b	24.0	24.8	-	-	-	1	-	-	-
12c	70.0	77.4	-	-	1	2	2	2	2
12d	86.4	86.4	-	-	1	-	-	-	-
13	64.6	65.7	2	-	-	1	2	1	2
14a	1.5	5.0	13	8	4	14	14	13	14
14b	5.2	20	11	5	1	12	12	9	12
14c	44.7	49.9	1	-	-	4	3	3	3
15a	27.2	31.2	-	-	3	3	3	2	2
15b	46.9	65.9	-	-	1	1	5	3	4
16a	23.8	25.22	-	-	-	2	2	2	2
16b	57.53	57.58	2	-	-	2	2	1	2

Table 4.3. *Continued*

Region*	Start (Mbp)	End (Mbp)	NL	DK	CN	Combined	Meta-Fix	Meta-Rand	Meta-Zsc
17a	17.4	22.6	-	-	-	2	2	2	2
17b	27.8	44.1	4	-	-	5	3	2	4
19	37.3	61.3	6	-	2	8	7	6	7
20a	1.8	11.0	2	-	-	-	-	-	-
20b	32.4	34.2	-	-	1	2	2	-	1
20c	36.7	36.9	-	-	1	2	-	-	-
20d	55.3	60.4	-	-	-	2	1	1	1
21	53.8	59.1	1	-	-	4	2	1	2
22	59.12	59.13	-	-	-	1	1	1	1
23a	21.22	21.23	2	-	-	-	-	-	-
23b	26.7	32.7	-	-	-	1	1	1	1
23c	33.5	36.5	2	-	-	1	1	1	1
23d	40.7	43.5	2	-	1	3	2	1	2
24a	6.82	6.85	-	-	1	-	-	-	-
24b	10.2	10.2	-	-	-	1	-	-	1
25a	9.8	9.9	-	-	-	1	-	-	1
25b	24.7	24.7	-	-	-	1	1	1	1
25c	41.4	41.7	-	-	-	2	-	-	1
26	2.9	43.0	6	4	2	11	9	7	9
27	37.0	42.2	-	-	1	1	1	1	1
28	36.6	37.2	-	-	-	2	-	-	-
29	32.9	40.5	2	-	-	2	1	1	1

The separate analysis of the Dutch samples detected 24 regions with SNPs significantly associated with at least one of the 16 FA traits except C18:0. Four of these regions were only detected in the Dutch data and not in any of the other scenarios including the joint GWA and meta-analyses. These regions exclusively detected for the Dutch samples were significantly associated with C18:3n3 (region 3a), C16:0 (region 7a), C16:1 and C16 index (region 20a) and with C18 index and CLA (region 23a). Separate analysis based on the Danish samples resulted in detection of significant associations between the FA traits and SNPs at six regions found on BTA 1, 5, 10, 14 and 26. Significant associations in the Danish sample were limited to nine FA traits, with no significant association detected for C8:0, C10:0, C12:0, C14:0, C18:0, C18 index and CLA. One of the regions detected in the separate analysis for the Danish population (region 1d) was significantly associated with C14:1 and C14 index but was not detected in any of the other scenarios. The separate GWAS for the Chinese population detected 16 regions. Significant associations detected in the Chinese sample were limited to C14:1, C14 Index, C18:0, C18:1 and C18:2n6. Significant associations detected in the separate analysis for the Chinese population with C18:2n6 (region 12d) and C18:0 (region 24a) were not detected in the other population-specific analyses as well as in the combined GWAS and meta-analyses.

Most of the genomic regions detected using the joint GWAS were also detected in the different meta-analyses. Meta-analysis using the fixed-model approach resulted in the detection of 50 regions with significant association to the FA traits. All but five of these regions were among the 56 regions detected using the joint GWAS. The random-model approach resulted in detection of the lowest number of regions, compared to the other two meta-analysis approaches, with detection of 47 regions, of which 3 were not detected in the joint-GWAS. A lower number of significantly associated regions for the random effects model is supported by the higher p-values estimated in general for SNPs with higher heterogeneity statistics in this model compared to the other meta-analysis approaches. Figure 4.4 shows the –log10 p-values of SNP effects estimated using the fixed effects model plotted against the values from the random effects model. The –log10 p-values estimated using the two meta-analysis approaches were similar for SNPs with smaller I^2 values (<10). With increase in heterogeneity statistic estimates ($I^2 > 10$), the random effects model tends to give lower –log10 p-values compared to the fixed-effects model. Meta-analysis using the z-score method resulted in detection of 55 regions significantly associated with the milk FAs, of which 51 were among the regions detected in the joint GWA.

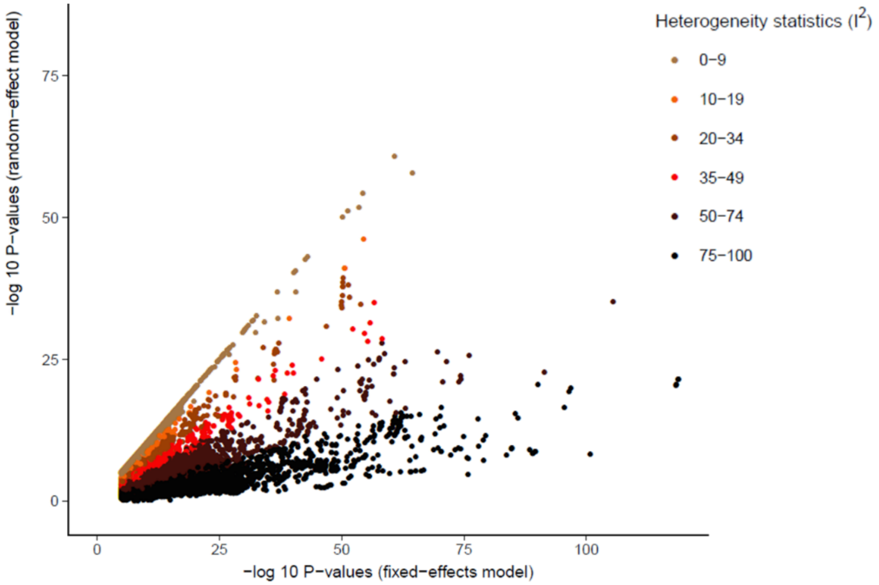

Fig 4.4. Correlation of –log10 p-values from the fixed (x-axis) and random (y-axis) effects meta-analysis models colored by heterogeneity statistics (I^2) computed from the random effects meta-analysis model for each SNP. The values for I^2 range from 0 to 100 proportionally from the least to the highest heterogeneity between population-specific studies at SNP level.

Apart from differences in the number of regions detected for at least one FA across scenarios, there were also differences in the number of FA traits significantly associated with the detected regions (Table 4.4). For region 14a for instance, only four FA traits were found to have significant associations in the Chinese analysis while the analysis in the Dutch sample resulted in detection of significant associations with13 FA traits. The same number of FA traits were found to have significant associations with regions 14a (14 FA) and 14b (12 FA) using the joint analysis, meta-analysis with weighted z-score and fixed effects models, while the random effects model lead to detection of significant association with lower number of FAs. Interestingly, association with C14:1 of region 14b was found only in the separate analysis of the Chinese data. For BTA 26, significant associations were detected with 11 FA traits in the joint GWA compared to 9 FA traits in Meta-analysis with weighted z-score and fixed effects models and 7 FA traits with the random effects model followed by 6 FAs in the Dutch, 4 in the Danish and 2 in the Chinese separate analyses.

Table 4.4 FA traits significantly associated with the genomic regions 14a, 14b and 26 detected using population-specific analyses, joint GWA and meta-analysis.

Scenario	Overlapping regions		
	Region 14a (1.5 -5 Mb)	Region 14b (1.5-5 Mb)	Region 26 (2.9-43.0 Mb)
NL	C8:0, C10:0, C14:0, C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C8:0, C10:0, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C10:0, C14:1, C14 index, C16:1, C16 index, C18 index
DK	C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3	C16:0, C16:1, C16 index, C18:2n6, C18:3n3	C14:1, C14 index, C16:1, C16 index
CN	C14:1, C14 index, C16:1, C18:2n6	C14:1	C14:1, C14 index
Joint GWA	C8:0, C10:0, C14:0, C14:1, C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C8:0, C10:0, C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C8:0, C10:0, C12:0, C14:0, C14:1, C14 index, C16:0, C16:1, C16 index, C18:0, C18 index
Meta-Fix	C8:0, C10:0, C14:0, C14:1, C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C8:0, C10:0, C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C8:0, C10:0, C12:0, C14:0, C14:1, C14 index, C16:1, C16 index, C18 index
Meta-Rand	C8:0, C10:0, C14:1, C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA	C8:0, C10:0, C14:0, C14:1, C14 index, C16:1, C16 index
Meta-Zsc	C8:0, C10:0, C14:0, C14:1, C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C8:0, C10:0, C14 index, C15:0, C16:0, C16:1, C16 index, C18:1c9, C18:2n6, C18:3n3, CLA, C18 Index	C8:0, C10:0, C12:0, C14:0, C14:1, C14 index, C16:1, C16 index, C18 index

4.3.4 SNP effects across the scenarios

Table 4.5 presents the estimated regression coefficients and –log10 p-values for lead SNPs on BTA 14 (SNP within the *DGAT1* gene) and 26 (SNP within the *SCD1* gene) found to have the strongest associations with the studied FA traits. The results also show that the combined-population analysis resulted in substantially increased –log10 p-values for the significant regions in most of the traits compared to the population-specific GWAS. For instance for C14 index, –log10 p-value for the lead SNP on chromosome 26 increased from 70.9 in the Dutch analysis to 126.1 in the combined analysis. These results also show that when the associations were significant, directions of SNP effects were similar for the three populations. Apart from direction of effects, we have compared estimated effects of the *DGAT1* (ARS-BFGL-NGS-4939) and *SCD1* (BovineHD2600005461) loci standardized with the standard deviation of the FA in the combined dataset (by dividing the estimates by the standard deviation of the FA) in the different populations.

Table 4.5. Population-specific and combined-population regression coefficients and –log10 p-values for leadSNPs on BTA 14 and 26.

		Population							
	Trait	NL		DK		CN		Combined	
SNP		*b*$_{(\pm SE)}$	-log10p	*b*$_{(\pm SE)}$	-log10p	*b*$_{(\pm SE)}$	-log10p	*b*$_{(\pm SE)}$	-log10p
ARS-BFGL-NGS-4939 (BTA14)									
	C8:0	0.04$_{(0.006)}$	12.8	0.02$_{(0.01)}$	0.7	0.03$_{(0.01)}$	2.0	0.04$_{(0.006)}$	11.0
	C10:0	0.09$_{(0.02)}$	6.4	0.02$_{(0.03)}$	0.2	0.09$_{(0.03)}$	3.5	0.08$_{(0.01)}$	8.0
	C12:0	0.03$_{(0.02)}$	1.0	0.001$_{(0.04)}$	0.01	0.08$_{(0.03)}$	2.1	0.04$_{(0.02)}$	2.1
	C14:0	-0.26$_{(0.04)}$	12.9	-0.28$_{(0.07)}$	4.8	0.08$_{(0.07)}$	0.7	-0.20$_{(0.03)}$	11.0
	C15:0	0.04$_{(0.006)}$	11.1	0.04$_{(0.009)}$	6.8	0.03$_{(0.008)}$	4.1	0.04$_{(0.004)}$	21.0
	C16:0	1.40$_{(0.10)}$	42.4	1.20$_{(0.16)}$	13.0	0.49$_{(0.12)}$	4.6	1.2$_{(0.07)}$	58.0
	C18:0	0.03$_{(0.06)}$	0.1	-0.14$_{(0.10)}$	0.8	-0.05$_{(0.10)}$	0.2	-0.02$_{(0.05)}$	0.2
	C14:1	0.01$_{(0.01)}$	0.9	0.03$_{(0.02)}$	1.5	0.06$_{(0.01)}$	5.8	0.03$_{(0.007)}$	4.8
	C16:1	0.14$_{(0.01)}$	33.3	0.23$_{(0.03)}$	18.7	0.12$_{(0.02)}$	7.4	0.16$_{(0.01)}$	55.0
	C18:1n9	-1.23$_{(0.09)}$	39.6	-0.87$_{(0.15)}$	8.2	-0.67$_{(0.16)}$	4.3	-1.03$_{(0.07)}$	46.0
	C18:2n6	-0.07$_{(0.006)}$	29.7	-0.05$_{(0.009)}$	9.3	-0.04$_{(0.007)}$	6.7	-0.06$_{(0.004)}$	45.0
	C18:3n3	-0.06$_{(0.007)}$	15.1	-0.06$_{(0.01)}$	7.1	-0.03$_{(0.008)}$	3.6	-0.05$_{(0.005)}$	26.0
	CLA	-0.08$_{(0.01)}$	9.5	-0.07$_{(0.01)}$	5.9	-0.05$_{(0.01)}$	3.2	-0.07$_{(0.007)}$	21.0
	C14 index	0.34$_{(0.07)}$	6.2	0.46$_{(0.12)}$	4.2	0.49$_{(0.11)}$	5.3	0.40$_{(0.05)}$	14.8
	C16 index	0.22$_{(0.03)}$	9.4	0.47$_{(0.07)}$	10.8	0.25$_{(0.06)}$	4.5	0.28$_{(0.03)}$	23.1
	C18 index	-1.37$_{(0.16)}$	17.8	-0.64$_{(0.23)}$	2.2	-0.40$_{(0.21)}$	1.2	-1.01$_{(0.11)}$	19.3

Table 4.5 *Continued*

Trait	NL		DK		CN		Combined	
SNP	$b_{(\pm SE)}$	-log10p	$b_{(\pm SE)}$	-log10p	$b_{(\pm SE)}$	-log10p	$b_{(\pm SE)}$	-log10p
BovineHD2600005461 (BTA26)								
C8:0	$0.03_{(0.007)}$	3.7	$0.03_{(0.01)}$	1.5	$0.002_{(0.01)}$	0.05	$0.02_{(0.006)}$	3.7
C10:0	$0.11_{(0.02)}$	8.5	$0.11_{(0.03)}$	3.7	$0.03_{(0.03)}$	0.70	$0.10_{(0.01)}$	11.9
C12:0	$0.06_{(0.03)}$	1.5	$0.10_{(0.03)}$	2.7	$0.03_{(0.03)}$	0.40	$0.07_{(0.02)}$	3.8
C14:0	$0.15_{(0.04)}$	3.9	$0.24_{(0.06)}$	4.1	$0.20_{(0.07)}$	2.43	$0.20_{(0.03)}$	10.90
C15:0	$0.002_{(0.007)}$	0.1	$-0.002_{(0.009)}$	0.07	$0.002_{(0.008)}$	0.07	$0.0005_{(0.005)}$	0.04
C16:0	$-0.14_{(0.11)}$	0.7	$-0.08_{(0.16)}$	0.2	$-0.10_{(0.12)}$	0.40	$-0.11_{(0.07)}$	1.00
C18:0	$-0.23_{(0.07)}$	3.1	$-0.25_{(0.09)}$	1.6	$-0.02_{(0.11)}$	0.06	$-0.20_{(0.05)}$	4.30
C14:1	$-0.18_{(0.01)}$	56.1	$-0.17_{(0.02)}$	26.7	$-0.10_{(0.01)}$	13.30	$-0.16_{(0.008)}$	98.0
C16:1	$0.15_{(0.01)}$	32.4	$0.13_{(0.02)}$	7.84	$0.06_{(0.02)}$	2.13	$0.12_{(0.01)}$	33.5
C18:1n9	$0.11_{(0.10)}$	0.6	$-0.14_{(0.14)}$	1.7	$-0.01_{(0.17)}$	0.02	$-0.003_{(0.07)}$	0.01
C18:2n6	$0.0003_{(0.006)}$	0.01	$-0.01_{(0.008)}$	0.8	$-0.007_{(0.007)}$	0.40	$-0.005_{(0.004)}$	1.0
C18:3n3	$0.01_{(0.008)}$	1.7	$-0.02_{(0.01)}$	1.7	$-0.01_{(0.009)}$	0.90	$-0.001_{(0.005)}$	0.1
CLA	$0.01_{(0.01)}$	0.2	$0.005_{(0.01)}$	0.1	$-0.003_{(0.01)}$	0.08	$0.008_{(0.007)}$	0.6
C14 index	$-1.42_{(0.08)}$	70.8	$-1.37_{(0.11)}$	34.1	$-1.0_{(0.11)}$	18.10	$-1.3_{(0.06)}$	126.1
C16 index	$0.50_{(0.04)}$	36.0	$0.40_{(0.07)}$	9.1	$0.19_{(0.06)}$	2.70	$0.4_{(0.03)}$	39.8
C18 index	$0.68_{(0.17)}$	4.1	$0.46_{(0.22)}$	1.4	$0.02_{(0.22)}$	0.03	$0.46_{(0.11)}$	4.3

Figure 4.5 shows correlations between standardized effects of DGAT1 marker in the three populations for all FAs except C12:0, C14:1 and C18:0, which are not significantly affected by *DGAT1* in the combined analysis. The plots indicate that not only "directions of SNP effects were similar" but the estimates of *DGAT1* effect in the Dutch and Danish population are very similar (high correlation and regression coefficient of about 1). The Chinese population showed a different pattern: the correlation between effects in the Dutch and Chinese population is high as is the correlation between effects in the Danish and Chinese populations. However, the regression coefficients are approx. 0.5 (0.57 and 0.52) indicating that the standardized effect sizes of *DGAT1* in the Chinese population are about half of that observed in the Dutch and Danish population. Looking at effects at individual FAs, lower SNP effects were consistently estimated for the FAs where phenotypic averages in the Chinese sample significantly differed compared to the other populations i.e., C8:0, C18:1c9 and C18:2n6. Effect of *DGAT1* loci on C8:0 was not significant in the Chinese and Danish datasets, therefore valid comparison cannot be made. For C18:1c9 and C18:2n6 however, the standardize effects of the DGAT1 loci were the lowest in the Chinese dataset compared to the Dutch and Danish samples.

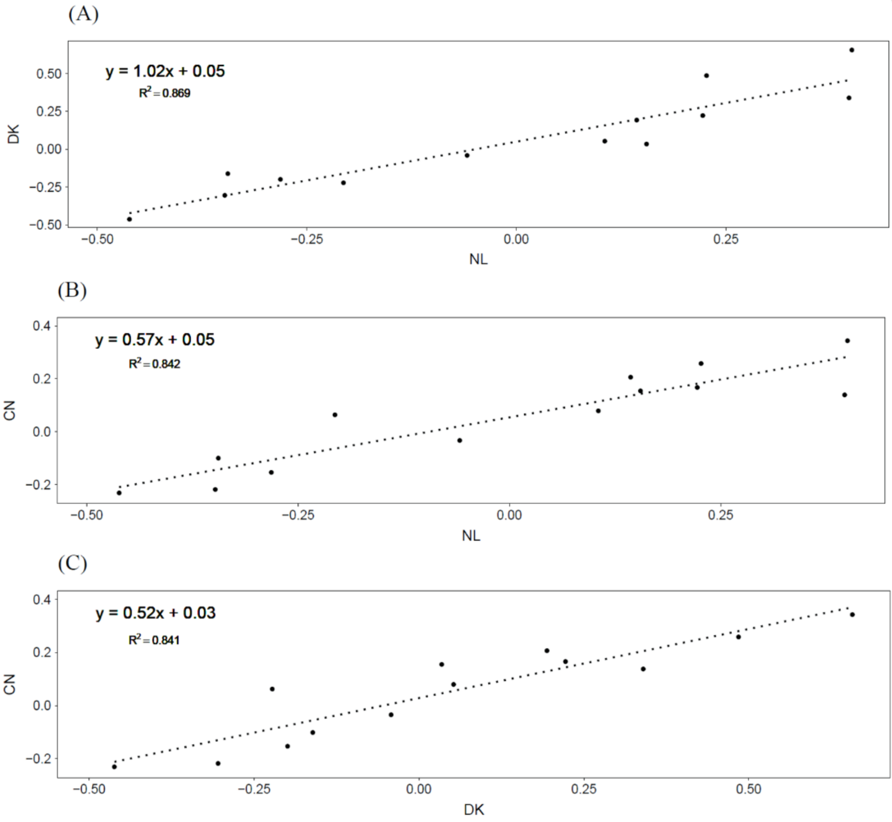

Fig 4.5. Correlation in the effects of DGAT1 locus (ARS-BFGL-NGS-4939) between the three populations standardized by dividing the SNP effects with standard deviation of the FA trait in the combined dataset: (**A**) relationship between the DGAT1 effect in the Danish and Dutch population, (**B**) relationship between the DGAT1 effect in the Chinese and Dutch population, (**C**) relationship between the DGAT1 effect in the Chinese and Danish population

For SCD1, the number of FA detected across analyses and significantly affected in the joint GWA was much lower i.e., five FAs. However, the correlations of loci effects in these FAs showed similar trend with that of DGAT1 effect such that there were high correlations of effects between the populations but lower effect sizes for the Chinese population (Figure 4.6).

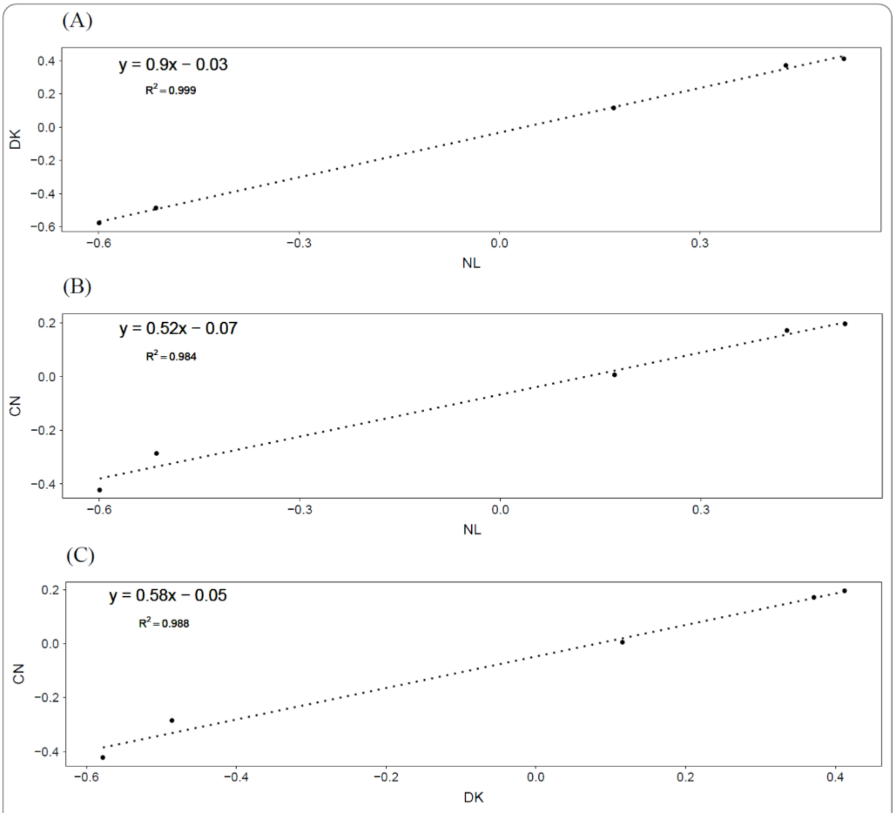

Fig 4.6. Correlation in the effects of SCD1 locus (BovineHD2600005461) between the three populations standardized by dividing the SNP effects with standard deviation of the FA trait in the combined dataset: (A) relationship between the SCD1 effect in the Danish and Dutch population, (B) relationship between the SCD1 effect in the Chinese and Dutch population, (C) relationship between the SCD1 effect in the Chinese and Danish population

4.4 Discussion

4.4.1 Detection of genomic regions under the different scenarios

Our combined-population GWAS resulted in detection of 28 additional regions significantly associated with one or more of the studied FA traits than were detected in the population-specific analyses altogether. We have also shown that –log10 p-values increased up to two folds in the joint GWA compared to the population-specific analyses. Apart from detection of more regions, also more FAs were significantly associated with the identified regions in the joint GWAS compared to the number of FAs associated with similar regions in the population–specific studies. As theoretically expected, under the assumption that traits are genetically the same in the different studies, these results demonstrate that combining datasets can substantially increase detection power.

There were few regions detected within each of the population-specific analyses that were not detected in the remaining populations or in the joint GWAS and meta-analyses. Differences in detections between the three population-specific studies can to a large extent be explained by differences in sample size. This applies to the analysis in the Dutch samples (N=1566) on the one hand (24 regions) and analyses in the Chinese (N=700, 16 regions) and Danish (N=614, 6 regions) samples on the other. As sample sizes in the Chinese and Danish populations are comparable, its relevance to explain the difference between these analyses is minimal. Such differences in detection might also arise from false positives within populations. However, false discovery cannot explain all the eight exclusive detections in our study. Given that there are 16 FAs analyzed in 3 populations, the binomial probability of having 8 false discoveries at FDR value of 5% is only 0.002. This is assuming traits are independent, which is not necessarily true in the case of milk FA traits. Therefore, it is unlikely to have 8 false positives in our analyses.

4.4.2 Heterogeneity between samples

Theoretically, it is expected that combining data, by increasing sample size, will enhance detection power and enable detection of regions with effects that are too small to pass the thresholds in the population-specific studies. This is also demonstrated in our computation of theoretical expectations of detection power as presented in Figure 3. However, these calculations assume that genotypic effects in the different populations are the same. In reality, this assumption of homogeneity might be violated for several reasons.

In many situations a major cause of heterogeneous genetic effects between populations is genetic distance between the populations. Differences in the pattern of LD structure over chromosomal regions of interest across populations are implicated as a cause of between-study heterogeneity in the genetic effects (Nakaoka and Inoue, 2009). In the presence of marked differences in LD structure, the same QTL might be relevant for the trait of interest in different populations but there could be differences in the markers that are in LD with the QTL. In this study, estimates of pairwise LD within bins of one Mbp size indicate that the genome-wide LD pattern is similar between the Dutch, Danish and Chinese Holstein Friesian. This is in agreement with previous study by Zhou et al. (2013), which reported high consistency in LD between adjacent markers of the Chinese and Danish Holstein populations. These findings are also in line with expectations given the common use of outstanding bulls in the three countries. Therefore, the differences in detections between the populations are less likely to result from differences in LD pattern.

Factors other than genetic distance also cause heterogeneous genetic effects between populations. In our study, phenotypic means and standard deviations were significantly different between the three populations for most of the FAs. Especially the Chinese samples showed larger differences in phenotypic means compared with the Dutch and Danish samples. These differences could result from differences in analytic methods and management of the cows.

4.4.3 Differences in trait measurements

One possible cause of differences in means and standard deviations could be the method of quantifying traits in different populations. For milk FA traits, the GC method is considered the method of choice and was commonly used for quantifying all the samples in this study. However, Contarini et al. (2013) showed the generally low precision of the quantitative evaluation for FAs present in low concentrations. In our study, that would apply especially to C18:3n3 and CLA but to a lesser extent also to C8:0, C14:1, C15:0, C16:1 and C18:2n6. In addition, differences in GC operating conditions (column, temperature profile and integration parameters) applied in the three countries could result in resolution differences in the GC analyses. As an example, comparing different labs with different GC conditions using the same samples, Contarini et al. (2013) indicated that with some of the GC methods quantification of the trans and cis isomers of 18:1 and other unsaturated FA was attained, while with others the quantities of 9c-18:1 were overestimated and quantities of trans-18:1 were underestimated. Such differences in resolution of GC analyses might add noise and reduce the detection power of GWAS, especially in combined analysis. However, medium to high heritability values were estimate for the FA traits from the combined dataset and these estimates are within the ranges reported in previous GC-based studies (e.g., Krag et al., 2013; Bilal et al., 2014). Therefore, we expect the contribution of possible differences in GC operating conditions to detection differences in our study to be minimal.

4.4.4 Consequences of genotype by feed interaction

The larger differences shown in the Chinese dataset compared to the other two samples were mainly for C8:0, C18:1c9 and C18:2n6. These differences might also be partly explained by differences in feeding systems between the sample populations. Feeding systems in North-West Europe are generally characterized by keeping cows indoors and feeding them with concentrates and roughages during the winter while there is some level of grazing/feeding of fresh-cut grass practiced during parts of the summer. Roughages commonly include silage made of rye grass, grass/clover, and/or alfalfa, and maize in some regions. Dairy production in the Beijing area of China, where the farms sampled in our study are located, is based on large-scale intensive dairy farms in densely populated cosmopolitan area with no access to grazing. The main feed sources are concentrates, mainly composed of maize, soybean meal, wheat bran, cottonseed meal and rapeseed meal, and roughages mainly including maize silage, alfalfa, and guinea grass (Beldman et al., 2014). The long chain FA C18:2n-6 mainly originates from maize and concentrates whereas C18:3n-3 originates from fresh grass in the feed (Chilliard et al., 2000). Content of CLA in milk has also been shown to increase with higher levels of fresh grass feeding (e.g. Couvreur et al., 2006). The higher average for C18:2n6 in the Chinese Holstein sample suggests high levels of concentrate feeding with high maize content. This is also supported by the lower phenotypic averages for C18:3n3 and CLA in the Chinese sample, which suggests that the cows are kept under lower levels of grass-based feeding. Higher content of poly-unsaturated FAs like C18:2n6 through the diet of dairy cows is known to decrease the de novo synthesized FA (e.g., Chilliard et al., 2000; Duchemin et al., 2013), explaining the lower averages for the de novo synthesized FAs, including C8:0, in the Chinese sample.

In genetic analysis, known sources of variation can be taken care of by accounting for them in the statistical analysis. In our analyses, phenotypic differences between the populations are accounted for by fitting herd as a fixed effect. Since herds were unique for each country, the effect of herd is expected to also account for differences between countries, including differences in feeding, or GC methods used. However, interaction of genotype with these sources of variation (e.g., genotype by feed interaction) might still have consequences in association analyses. In the presence of genotype interactions, different QTL might be relevant to the investigated trait in one population versus the other. Alternatively, the same QTL might have different effects for the same trait in the different environments. Differences in non-genetic factors such as feed can lead to differences in the expression of genes relevant for the traits of interest. For instances, Tao et al. (2015) has shown that feeding a high concentrate diet in goats down-regulated expression of the *ACACA*, *LPL* and *SCD* genes which play key roles in milk FA composition. Such differences will lead to detection of regions harboring such QTLs in one population but not in the others. Genotype by environment interactions can be quantified by calculating genetic correlations between the populations. This was not possible in our study due to small sample sizes within populations leading to high standard errors of the correlation estimates. The SNP effects in our analyses were similar in direction in the three populations whenever the associations were significant. Comparison of the standardized effects of *DGAT1* and *SCD1* loci on the FA traits also shows strong correlation between estimated effects for both loci in the three population. However, effect sizes seem to be different in the Chinese data as compared to the Dutch/Danish data. Additive genetic variances in the Chinese data also differed from the other two populations in general. For C18:1c9 and C18:2n6, which are derived mainly from the feed and for which phenotypic means in the Chinese dataset showed significant differences to the other populations, standardized effect of the *DGAT1* loci was the lowest in the Chinese dataset. These differences might point to genotype by environment (feed) interaction. However, high correlations and similar direction of SNP effects between the populations suggest that this interaction is mostly due to scaling instead of re-ranking of genotype effects (strong correlation in effects and similar direction of effects). Since there is high correlation between estimated effects, the data from the three populations do not contradict but support each other. There are no indications that the GWA signal might disappear by combining data, due to effects that differ in direction (re-ranking). The estimated SNP effects imply that for at least the *DGAT1* and *SCD1* loci, the value of an observation from the Chinese population contributes less (to the joint GWA signal) than an observation from the other two populations due to the smaller effect sizes.

4.4.5 Meta-analyses

There were slight differences between the different meta-analyses approaches in the number of detected regions. Among the different meta-analyses approaches employed in this study, the weighted z-score method resulted in detection of regions the most comparable to the joint GWAS. The fixed and random effects models resulted in lower numbers of detections compared to the weighted z-score approach, with the random effects model resulting in the smallest number of detections. Similar results were previously reported in various meta-analyses of trans-ethnic GWA studies in human genetics (e.g., Wang et al., 2013) and in livestock multi-breed scenarios (e.g., Van der Berg et al., 2016).

By assuming similar genetic effect size between individual studies, the fixed-effects model is expected to have limited power in the presence of genetic effect heterogeneity (Evangelou and Ioannidis, 2013). However in our study, –log10 p-values tended to be much higher in the fixed effects model when heterogeneity statistics at markers increased. This implies that in case of high heterogeneity between populations, the fixed effects approach estimates SNP effects with smaller confidence intervals, leading to detection of larger numbers of significant associations compared to the random effects model. This is contrary to our expectation and to the fact that the random effects model was developed to specifically account for heterogeneity between studies. It is suggested that the stringent assumptions in the random-effects model, which implicitly assumes heterogeneity under the null hypothesis, cause it to have far more limited power (Han and Eskin, 2011).

4.4.5 Comparison of detections across the scenarios in the DGAT1 and SCD1 regions

While it remains difficult to compare detection power between GWA methods using real data since the “true” effects are unknown, it is possible to compare the different analyses based on example regions with well-established connections to the studied traits. The regions detected on BTA 14 (region 14a) and BTA 26 are known to contain the *DGAT1* and *SCD1* genes, respectively. The *DGAT1* (e.g., Schennink et al., 2008; Bovenhuis et al., 2016) and *SCD1* (e.g., Schennink et al., 2008; Bouwman et al., 2012; Carvajal et al., 2016) genes are frequently reported to have significant associations with most FAs. Detections in the *DGAT1* region were similar for the joint GWA and the meta-analyses using fixed effects and weighted z-score approaches with significant associations established for 14 FA traits. Significant associations were also detected with these traits except C14:0 in the random effects meta-analysis model and C14:1 in the Dutch separate analysis. Previous studies have shown that the K allele of *DGAT1* polymorphisms has a reducing effect in C14:0 (e.g., Schennink et al., 2008; Bovenhuis et al., 2016). Through reduction of C14:0, DGAT1 is also expected to affect the concentrations (%wt/wt) of C14:1. Therefore, significant associations with these FAs were expectable.

The SCD enzyme is involved in the synthesis of monounsaturated FA by introducing a double bond in the delta-9 position of C14:0, C16:0 and C18:0, primarily (Ntambi and Miyazaki, 2003). The significant associations we detected across the different analyses with C14:1, C16:1 and the desaturation indexes of these FAs are thus expected. However, through the desaturation process, *SCD1* also affects the concentrations (%wt/wt) of C14:0, C16:0 and C18:0 and thus significant association with these FAs are also to be expected. Significant associations for C16:0 and C18:0 were, however, only detected using the joint GWAS. Similarly, significant associations with C18 index in this region were only detected using the joint GWAS and the meta-analysis with fixed effects and weighted z-score approaches.

These results indicate that the joint GWA resulted in more power to detect associations in the confirmed *DGAT1* and *SCD1* regions compared to the rest of the scenarios including the meta-analyses. Nonetheless, detection of higher number of regions and significant association of these regions to higher number of FAs in the meta-analyses compared to any of the population-specific analyses emphasizes utility of the methods in the absence of raw multi-population datasets to undertake joint GWA.

While it is similarly difficult to handle heterogeneity between samples in joint GWA as is in meta-analysis, it provides an advantage of flexibility to employ different transformation and standardization strategies on the datasets and allows fitting common model components as opposed to summaries of different studies, often with different model components, in meta-analyses. In our study, such data transformation and standardization strategies have remedied to a certain extent some observed heterogeneity between the samples as discussed below.

4.4.6 Data transformation and standardization for joint GWA

In this study, different data transformation and standardization approaches are implemented to address differences in residuals, standard deviations and differences in stages of lactation between the samples. Residuals of some of the FAs, especially of C18:2n6, C18:3n3 and CLA, tended to increase with the mean, indicating heterogeneity of the residual variances and in this way violating the assumptions underlying significance testing. Logarithmic transformation is thus applied for these traits and the transformed values were used for both the population-specific analyses as well as the joint GWA. Residuals plotted against predicted phenotypes in these FA traits presented in Figure 1 indicate that the problem of heterogeneity is corrected by the transformation applied.

Milk FA traits can also differ between populations due to differences in lactation stages of the cows. In our study, an attempt was made to tackle this source of differences by restricting lactation stages to 60 days in milk and above. There are evidences that effects of genes in early lactation differ from those later in lactation. For instance, Bovenhuis et al. (2015) have reported significant *DGAT1* by lactation stage interaction for milk production traits including fat content and showed that the *DGAT1* effect shows a large increase during early lactation (from the start of lactation to day 50 to 150) and tends to decrease later in lactation.

Some significant differences in standard deviations were also observed between the samples from the three populations. To test sensitivity of such differences, in relation to the joint GWA outcomes, we standardized all the FA measurements within population to have mean of zero and standard deviation of one, and a separate test joint-GWA was undertaken with this standardized dataset. The joint GWAS using the standardized dataset led to detection of significant association with one more FA on BTA 13 but detections in the rest of the regions remained unchanged after the standardization. While detection of significant association with one more FA at one of the 56 regions is still relevant and emphasizes the importance of such standardization, the fact that the rest of detections remained the same with and without standardization shows that our joint GWA is not substantially affected by differences in standard deviations and stresses the stability of our results.

4.5 Conclusion

Joint GWAS using multi-population datasets detected the highest number of regions and the highest number of associated FA traits compared to the population-specific analyses as well as the three different meta-analysis approaches employed. Detection of higher number of regions using the different meta-analyses methods, compared to any of the population-specific analyses, emphasizes utility of these methods in the absence of raw multi-population datasets to undertake joint GWA. Among the meta-analysis methods employed, the weighted z-score method was the closest to the joint GWA in terms of the number of detected regions and the number of FAs associated to each detected region. The random effects model, which is specifically designed to handle presence of heterogeneity among individual studies, resulted in the lowest number of QTL detected.

Authors’ contributions

GG processed the data, implemented the analyses and drafted the manuscript. HB conceived the study and contributed to the discussion of the results. AJB co-planned the study and contributed to the discussion of the results. NAP and HJFV collected the milk samples and contributed to the milk analysis and discussion of the results. MHPWV, DS and QZ contributed to the discussion of the results. All authors read and approved the final manuscript.

Acknowledgements

GG is enrolled in the Erasmus-Mundus joint doctorate European Graduate School in Animal Breeding and Genetics and was also supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research. The study used data from the Dutch Milk Genomics Initiative (www.milkgenomics.nl), High Technology R & D Program of China,2013AA102504 and Beijing Dairy Industry Innovation Team, BAIC06-2018, Danish-Swedish Milk

Genomics Initiative (www.milkgenomics.dk) and the Milk Levy Fund (Denmark) projects: “Phenotypic and genetic markers for specific milk quality parameters” and “The importance of the metagenome for milk composition and quality".

References

Ball R.D. (2004). ldDesign- an R package for design of experiments for detection of linkagedis-equilibrium. http://ftp.auckland.ac.nz/software/CRAN/doc/packages/ldDesign.pdf

Bastin C., Soyeurt H. and Gengler N. (2013). Genetic parameters of milk production traits and fatty acid contents in milk for Holstein cows in parity 1-3. J Anim Breed Genet. 130(2):118-27. doi: 10.1111/jbg.12010

Begum F., Ghosh D., Tseng G.C. and Feingold E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res40(9):3777-84. doi: 10.1093/nar/gkr1255

Beldman Alfons, Bai Junfei, Cao Binbin, Cao Zhijun, Du Wen Fang Xiangming, Guo Huiyuan, et al. (2014). White Paper on China Dairy. Sino-Dutch Dairy Development Center. http://www.sdddc.org/index_en.aspx

Bilal G., Cue R.I., Mustafa A.F. and Hayes J.F. (2014). Short communication: Genetic parameters of individual fatty acids in milk of Canadian Holsteins. J Dairy Sci. 97(2):1150-6. doi: 10.3168/jds.2012-6508

Bouwman A.C., Visker M.H., van Arendonk J.A. and Bovenhuis H. (2012). Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples.BMC Genet. 29;13:93.

Bouwman A.C., Daetwyler H.D., Chamberlain A.J., Ponce C.H., Sargolzaei M., Schenkel F.S. et al. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 50(3):362-367. doi: 10.1038/s41588-018-0056-5

Bovenhuis H., M.H.P.W., van Valenberg H.J., Buitenhuis A.J. and van Arendonk J.A. (2015).Effects of the DGAT1 polymorphism on test-day milk production traits throughout lactation. J Dairy Sci. 98(9):6572-82. doi: 10.3168/jds.2015-9564.

Bovenhuis H., Visker M.H.P.W., Poulsen N.A., Sehested J, van Valenberg H.J.F., van Arendonk J.A.M., Larsen L.B. and Buitenhuis A.J. (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(4):3113-3123. doi: 10.3168/jds.2015-10462

Carvajal A.M., Huircan P., Dezamour J.M., Subiabre I., Kerr B., Morales R. and Ungerfeld E.M. (2016). Milk fatty acid profile is modulated by DGAT1 and SCD1 genotypes in dairy cattle on pasture and strategic supplementation. Genet Mol Res. 9;15(2). doi: 10.4238/gmr.15027057

Chilliard Y., Ferlay A., Mansbridge R.M. and Doreau M. (2000). Ruminant milk fat plasticity: nutritional control of saturated, polyunsaturated, trans and conjugated FA. Ann. Zootech. 49: 181-205.

Consortium-Major Depressive Disorder Working Group of the Psychiatric GWAS, Ripke S., Wray N.R., Lewis C.M., Hamilton S.P., Weissman M.M., et al. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. Molecular psychiatry. 18(4):10.1038/mp.2012.21. doi:10.1038/mp.2012.21

Contarini G., M. Povolo, V. Pelizzola, L. Monti, and G. Lercker. (2013). Interlaboratory evaluation of milk fatty acid composition by using different GC operating conditions. J. Food Compos. Anal. 32:131–140.

Costafreda SG. (2009). Pooling FMRI data: meta-analysis, mega-analysis and multi-center studies. Front Neuroinform. 30;3:33. doi: 10.3389/neuro.11.033.2009

Couvreur S., Hurtaud C., Lopez C., Delaby L. and Peyraud J.L. (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(6):1956-69.

Duchemin S., Bovenhuis H., Stoop W.M., Bouwman A.C., van Arendonk J.A. and Visker M.H. (2013). Genetic correlation between composition of bovine milk fat in winter and summer, and DGAT1 and SCD1 by season interactions. J Dairy Sci. 96(1):592-604. doi: 10.3168/jds.2012-5454

Duchemin S.I., Visker M.H., Van Arendonk J.A., Bovenhuis H. (2014). A quantitative trait locus on Bos taurus autosome 17 explains a large proportion of the genetic variation in de novo synthesized milk fatty acids. J Dairy Sci. 97(11):7276-85.

Evangelou E. and Ioannidis J.P. (2013). Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 14(6):379-89.

Gebreyesus G., Lund M.S., Janss L., Poulsen N.A., Larsen L.B., Bovenhuis H. and Buitenhuis A.J. (2016). Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J Dairy Sci. 99(4):2863-2866. doi: 10.3168/jds.2015-10501

Han B. and Eskin E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. The American Journal of Human Genetics. 88:586–598.

Higgins, J. P. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Stat. Med. 15;21(11):1539-58

Kavvoura F.K. and Ioannidis J.P. (2008). Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. Hum Genet. 23(1):1-14

Krag K., Poulsen N.A., Larsen M.K., Larsen L.B., Janns L. and Buitenhuis B. (2013). Genetic parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a Bayesian approach. BMC Genet 2013, 14:79. doi: 10.1186/1471-2156-14-79.

Li C., Sun D., Zhang S., Wang S., Wu X., Zhang Q., Liu L., Li Y. and Qiao L. (2014). Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. PLoS One. 23;9(5):e96186. doi: 10.1371/journal.pone.0096186

Lin D.Y. and Zeng D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol. 34(1):60-6.

Nakaoka H. and Inoue I. (2009). Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse. J Hum Genet. 54(11):615-23. doi: 10.1038/jhg.2009.95

Ntambi J.M. and Miyazaki M. (2003). Recent insights into stearoyl-CoA desaturase-1. Curr Opin Lipidol. 14:255–61

Poulsen N.A., Gustavsson F., Glantz M., Paulsson M., Larsen L.B. and Larsen M.K. (2012). The influence of feed and herd on fatty acid composition in 3 dairy breeds (Danish Holstein, Danish Jersey, and Swedish Red). J Dairy Sci. 95(11):6362-71.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 81(3):559-75.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rubio Bernal Y.L., Gualdrón Duarte J.L., Bates R.O., Ernst C.W., Nonneman D., Rohrer G.A., King D.A., Shackelford S.D., Wheeler T.L., Cantet R.J. and Steibel J.P. (2015). Implementing meta-analysis from genome-wide association studies for pork quality traits. J Anim Sci. 93(12):5607-17. doi: 10.2527/jas.2015-9502

Sargolzaei M., Chesnais J.P., Schenkel F.S. (2014). A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 17;15:478. doi: 10.1186/1471-2164-15-478.

Schennink A., Heck JM., Bovenhuis H., Visker M.H., van Valenberg H.J. and van Arendonk

J.A. (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(5):2135-43. doi: 10.3168/jds.2007-0825

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

Sung Y.J., Schwander K., Arnett D.K., Kardia S.L., Rankinen T., Bouchard C., Boerwinkle E., Hunt S.C. and Rao D.C. (2014). An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions. Genet Epidemiol. 38(4):369-78.

Tao H., Chang G., Xu T., Zhao H., Zhang K. and Shen X. (2015). Feeding a High Concentrate Diet Down-Regulates Expression of ACACA, LPL and SCD and Modifies Milk Composition in Lactating Goats. PLoS One. 18;10(6):e0130525. doi: 10.1371/journal.pone.0130525

Van den Berg I., Boichard D. and Lund M.S. (2016). Comparing power and precision of within-breed and multi-breed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. J Dairy Sci. 99(11):8932-8945. doi: 10.3168/jds.2016-11073

Veerkamp R.F., Coffey M., Berry D., de Haas Y., Strandberg E., Bovenhuis H., Calus M. and Wall E. (2012). Genome-wide associations for feed utilisation complex in primiparous Holstein-Friesian dairy cows from experimental research herds in four European countries. Animal. 6(11):1738-49.

Wang X., Chua H.X., Chen P., Ong R. T-H., Sim X., Zhang W., et al. (2013). Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. Human molecular genetics. 2013:ddt064

Whitlock M.C. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. J. Evol. Biol. 18(5):1368-73

Yang J., Lee S.H., Goddard M.E. and Visscher P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1): 76-82.

Zhou L., Ding X., Zhang Q., Wang Y., Lund M.S. and Su G. (2013). Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genet Sel Evol. 21;45:7. doi: 10.1186/1297-9686-45-7.

Zimin A.V., Delcher A.L., Florea L., Kelley D.R., Schatz M.C., Puiu D., et al. (2009). A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 10(4):R42. doi: 10.1186/gb-2009-10-4-r42

5

Multi-population GWAS and enrichment analyses reveal novel genomic regions and promising candidate genes underlying bovine milk fatty acid composition

G. Gebreyesus[1,2], A. J. Buitenhuis[1], N. A. Poulsen[3], M. H. P. W. Visker[2], Q. Zhang[4], H. J. F. van Valenberg[5], D. Sun[4], and H. Bovenhuis[2]

[1]Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark; [2]Animal Breeding and Genomics Centre, Wageningen University, PO Box 338, 6700 AH Wageningen, the Netherlands; [3]Department of Food Science, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark; [4]Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture of China, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China; [5]Dairy Science and Technology Group, Wageningen University and Research, P.O. Box 17, 6700 AA Wageningen, the Netherlands

Submitted [BMC Genomics]

Abstract

The power of genome-wide association (GWA) studies is often limited by the sample size available for analysis. Milk fatty acid (FA) traits are scarcely recorded due to expensive and time-consuming analytical techniques. Combining multi-population datasets can enhance the power of GWA studies enabling detection of genomic region explaining medium to low proportions of the genetic variation. GWA studies often detect broader genomic regions containing several positional candidate genes making it difficult to untangle the causative candidates. Post-GWA analyses with data on pathways, ontology and gene expression status on tissues of relevance might allow prioritization among positional candidate genes. Multi-population GWA for 16 FA traits quantified using gas chromatography (GC) in sample populations of the Chinese, Danish and Dutch Holstein with high-density (HD) genotypes detects 56 genomic regions significantly associated to at least one of the studied FAs; some of which have not been previously reported. Pathways and gene ontology (GO) analyses suggest promising candidate genes on the novel regions including *OSBPL6* and *AGPS* on BTA 2, *PRLH* on BTA 3, *SLC51B* on BTA 10, *ABCG5/8* on BTA 11 and *ALG5* on BTA 12. Novel genes in previously known regions, such as *FABP4* on BTA 14, *APOA1/5/7* on BTA 15 and *MGST2* on BTA 17, are also linked to important FA metabolic processes. Detection of such regions and candidate genes will be crucial in understanding the complex genetic control of FA metabolism. The findings can also be used to augment genomic prediction models with regions collectively capturing most of the genetic variation in the milk FA traits.

Key words: Milk fatty acids, multi-population GWA, candidate genes, pathway analysis

5.1 Introduction

Several fatty acids (FAs) of varying carbon chain length (C_4-C_{22}) and degree of saturation are present in milk. FAs in milk can originate either through direct transport from the rumen to the mammary gland via the blood, or from *de novo* synthesis in the mammary gland from acetate, beta-hydroxybutyrate (Bauman and Griinari, 2003) and propionate (Massart-Leën et al., 1983; Vlaeminck et al., 2006). Additionally, FAs in the mammary gland can originate from mobilization of body fat reserves. The short and intermediate chain FAs are mostly synthesized *de novo* in the mammary gland with the exception of C16:0, of which approximately 50% is assumed to be synthesized *de novo*. The long chain FAs, and approximately 50% of C16:0, are suggested to be derived from blood lipids originating from the diet (Chilliard et al., 2000) and mobilization of body fat reserves (Bauman and Griinari, 2003). Considerable genetic variation has been reported for the fat composition of milk (e.g. Stoop et al., 2008; Krag et al., 2013). Genes with major effects such as the *DGAT1* and *SCD1* explain part of this genetic variation (Schennink et al., 2008). In addition, several regions on the bovine genome with suggestive effects on milk fat composition have been reported from GWA studies (e.g. Bouwman et al. 2012; Buitenhuis et al. 2014; Li et al. 2014). Identified genes and genomic regions explain a fraction of 3.6 to 53 % of the total genetic variation in different milk FA traits (e.g. Bouwman et al., 2011, 2012). Detection of additional genomic regions requires availability of larger sample size and high-density markers. GC analysis, the current method of choice to quantify milk FA, requires expensive equipment and is time-consuming, thus limiting measurement of the traits to experimental scale. GWA studies for the milk FA traits so far relied on such smaller datasets within different dairy cattle breeds/populations.

An option to deal with the limitation in sample size could be to combine the available smaller datasets across populations for joint GWA. Such analyses can increase detection power depending on the genetic distance between the populations and the marker density (Lund et al., 2014). In this study, we undertake multi-population GWA for milk FA traits by combining samples from Chinese, Danish and Dutch Holstein Friesians with HD genotypes available. Previous studies show high consistency in the linkage disequilibrium (LD) and minor allele frequencies between the three populations (e.g. Zhou et al. 2013; Li et al. 2015; Gebreyesus et al. submitted). Thus, combining samples from these populations for joint GWA might allow identification of genomic regions explaining even small proportions of the genetic variation in milk FA.

A hurdle is that due to the long range of LD in livestock breeds, GWA studies often result in detection of large genomic regions (de Roos et al., 2008) containing several positional candidate genes. Identifying the actual causative variants, therefore, requires additional evidence on top of the GWA test. Enrichment analysis is commonly undertaken in GWA studies to prioritize positional candidate genes linked to significantly enriched pathways and gene ontology (GO) terms that are believed to be relevant to traits of interest. However, FA synthesis can take place in various mammalian tissues and thus further evidence is needed to determine whether such prioritized genes are relevant particularly to milk FA related mechanisms. Studies have been profiling differential expression of genes in the mammary tissues in various species (e.g. Bionaz et al., 2012; Lemay et al., 2013). Such publicly available data can been used to further prioritize candidate genes. Furthermore, the mammalian phenotype ontology (Smith et al., 2005), which provides annotation of mammalian phenotypes in the context of mutations, is increasingly becoming useful in fine-tuning the link between detected genes and phenotypes associated (e.g. Cai et al., 2018).

To prioritize positional candidates in this study, we use publicly available resources to obtain Gene Ontology (GO) terms and enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, and link results with reported mammary gland gene expression and the mammalian phenotype ontology database (Smith et al., 2005).

5.2 Methods

5.2.1. Animals and phenotypes

The dataset used for the association analysis comprised 700 Chinese, 614 Danish and 1,566 Dutch Holstein cows sampled from 18 herds in China, 22 herds across Denmark and 398 herds in the Netherlands. Stages of lactation of sampled cows ranged from 60 to 700 days in milk in the Chinese population, 60 to 481 days in milk in the Danish population and 60 to 278 days in milk in the Dutch Holstein cows.

FA traits, including C8:0, C10:0, C12:0, C14:0, C14:1, C15:0, C16:0, C16:1, C18:0, C18:1c9, C18:2n6, C18:3n3 and C18:2 cis-9,trans-11 (CLA), were analyzed using the GC method. Details of the quantification methods are as described by Li et al. (2014) for Chinese samples, Poulsen et al. (2012) for Danish samples and Stoop et al. (2008) for Dutch samples. Genomic regions affecting the saturated FAs might show association to the unsaturated forms because the saturated form available for desaturation determines proportion of the unsaturated FAs. Hence, calculation of the desaturation indexes might allow detection of regions particularly associated with the desaturation process. Accordingly, desaturation indexes were calculated based on the FA measurements as: C14 index = C14:1/(C14:1+C14:0) * 100; C16 index = C16:1/(C16:1+C16:0) * 100 and C18 index = C18:1c9/ (C18:1c9+C18:0) * 100.

5.2.2. Genotypes and Imputation

High-density (HD) genotypes, real or imputed, were available for all cows used in the analyses. Sampled cows from the Chinese Holstein were initially genotyped using the BovineSNP50 Beadchip (50K, Illumina). The 50K genotypes were then imputed to HD using reference population of 96 Chinese Holstein bulls, genotyped with the BovineHD Beadchip (777K).

Part of the Danish dataset included cows genotyped with the BovineHD Beadchip, while the remaining Danish cows were genotyped using the BovineSNP50 Beadchip. The HD genotypes available for part of the samples was therfore used as reference to impute the 50K genotypes of the first part of the Danish cows to HD. Details are described in Gebreyesus et al. (2016). Cows in the Dutch dataset were genotyped with a custom 50K SNP Beadchip and subsequently imputed to HD as presented in detail by Duchemin et al. (2014). SNPs with minor allele frequencies (MAF) less than 0.05 or with a count of one of the genotypes less than 10 in each population were excluded from the association analysis. A total of 464,130 SNPs were available for the association analysis. The SNP positions were based on the bovine genome assembly UMD 3.1 (Zimin et al., 2009).

5.2.3. Association analysis

A single-SNP association test was implemented using a mixed linear model in the GCTA program (Yang et al., 2011). Association analysis was carried out using the following statistical model:

$$y_{ijkl} = \mu + parity_i + herd_j + b_1 * DIM_{ijkl} + b_2 * SNP_k + animal_l + e_{ijkl}, \quad (1)$$

Where y_{ijkl} is the phenotype of cow l; μ is the fixed effect of mean; $parity_i$ and $herd_j$ are the fixed effects of parity and herd, respectively; b1 is the regression coefficient for DIM, DIM_{ijkl} is a covariate of days in milk; b2 is the allele substitution effect for SNP, SNP_k is a covariate indicating the number of copies of a specific allele (0, 1 or 2) of the SNP; and animal is the random additive genetic effect. Animal effects were assumed to be distributed as: $\mathrm{N}(0, \mathbf{G}\sigma_a^2)$, where **G** is the genomic relationship matrix constructed using all HD genotypes but excluding the SNPs on the chromosome on which SNP k is located. Residuals were assumed to be distributed as: $\mathrm{N}(0, \mathbf{I}\sigma_e^2)$, where **I** is the identity matrix.

Since only cows with more than 60 days-in-milk were included in the analyses, a linear adjustment for DIM was sufficient. For the FA traits C18:2n6, C18:3n3 and CLA, log transformation was applied prior to the association analysis to account for observed heterogeneity of residual variances.

Significance thresholds were determined using a false discovery rate (FDR). Significance thresholds corresponding to FDR of 5% ranged for different FA from –log10 p-value = 3.4 to –log10 p-value = 5.0. We used a –log10 p-value of 5.0 as the genome-wide significance threshold for all FA composition traits.

5.2.3.1 Determining multiple regions on a chromosome

To determine if a region harbored one or more QTL, iterative approaches fitting the effect of SNPs with the highest –log 10 p-values were employed. In this approach, the SNP with the highest –log 10 p-value for the studied FA trait was considered as the lead SNP. The allelic dosage of such a lead SNP was then fitted as fixed effect for a second round of chromosome-wide analyses. If other SNPs, also significantly associated in the first round GWA, were still found to have -log 10-pvalue > 5 in the second round analysis, the SNP with the highest –log 10 p-value in the second analysis was taken as the second lead SNP and its allelic dosage fitted as fixed effect for a third round of analysis. This procedure was iterated until no further SNP with -log 10-pvalue > 5 was observed. The SNPs that showed significant association in a round of GWA but showed –log 10 p-value < 5 upon fitting the allelic dosage of the lead SNP were then considered as part of a region around that lead SNP. The position of the first and last such SNP before and after the lead SNP were considered as the boundaries of the region.

5.2.3.2 Estimation of genetic variances explained by SNPs

Genetic variance explained by the lead SNP in a region was calculated from the GWA summary as: $2pq\alpha^2$, where p and q are the allele frequencies and α is the allele substitution effect (Park et al., 2010). The proportion of total genetic variance explained by such a lead SNP was then calculated as: $2pq\alpha^2 / \sigma_a^2$,

Where, σ_a^2 is the additive genetic variance estimated using model 1 but without fitting fixed effects of SNP and using **G** constructed using all HD SNPs. Computation of genetic variance explained by SNPs from a GWA summery might lead to overestimation of SNP effects (Beavis 1998) specially for small effect size SNPs that only just reach the significance threshold. Heritability (h^2) estimates were computed as:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}, \qquad (2)$$

Where, σ_e^2 is the residual variance.

5.2.4. Gene assignment and enrichment analyses

Genes found within detected genomic regions were retrieved from the ensemble database using the BioMart web interface based on the UMD 3.1 bovine genome assembly (https://www.ensembl.org/biomart/martview). The DAVID functional annotation tool (https://david.ncifcrf.gov) was then used to analyze overrepresented GO biological terms, which included the terms cellular component (CC), molecular function (MF), biological process (BP) and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. Ontologies in the mammalian phenotype database were accessed and searched for genes connected to abnormalities relevant to FA metabolism using the Mouse Genome Informatics (MGI) web platform (http://www.informatics.jax.org/batch).

5.3 Results

5.3.1. Descriptive statistics and genetic parameters

Table 5.1 presents phenotypic means, additive genetic variances and heritability estimates of the FAs expressed as weight percentage of total fat and the desaturation indexes in the combined multi-population dataset. The 13 FAs studied together amounted to 87.6% of total fat. Of the studied FAs, C18:3n3 and CLA occurred at concentrations less than 1% of total fat in the milk samples. Other FAs including C15:0, C8:0, C14:1 and C16:1 also occurred at low concentrations of total fat (means = 1.09 – 1.49). Coefficients of variation (not shown) of the FA traits ranged between 0.06 % (C18 index) and 0.43 % (CLA). Heritability estimates in the studied FA traits ranged from low (0.18) for C18:2n6 to high (0.53) for C14 index. The dataset used in the current study comprises samples from the Chinese, Danish and Dutch Holstein population and details regarding descriptive statistics and genetic parameters within each population can be found in chapter 4.

Table 5.1. Phenotypic means (with standard deviations, SD) and genetic parameters (with standard errors, SE) in the combined-population dataset

FAs	Mean	SD	σ_a^2	SE	h^2	SE
Saturated FAs[1]						
C8:0	1.18	0.38	0.008	0.04	0.27	0.03
C10:0	2.80	0.58	0.07	0.09	0.39	0.04
C12:0	3.58	0.76	0.09	0.11	0.33	0.04
C14:0	11.00	1.26	0.21	0.17	0.25	0.03
C15:0	1.09	0.18	0.004	0.02	0.23	0.04
C16:0	30.20	3.53	1.80	0.48	0.34	0.04
C18:0	10.30	1.99	0.52	0.29	0.25	0.04
Unsaturated FAs[1]						
C14:1	1.19	0.35	0.03	0.05	0.47	0.04
C16:1	1.49	0.35	0.05	0.07	0.46	0.04
C18:1c9	21.90	4.37	1.38	0.46	0.27	0.04
C18:2n6	1.89	1.19	0.01	0.05	0.18	0.03
C18:3n3	0.48	0.13	0.005	0.01	0.19	0.03
CLA	0.53	0.23	0.004	0.02	0.21	0.04
Desaturation indexes[2]						
C14 index	9.71	2.37	1.57	0.37	0.53	0.03
C16 index	4.70	0.97	0.32	0.19	0.38	0.04
C18 index	67.80	3.98	3.95	0.73	0.31	0.04

[1]Expressed in % wt/wt

[2]Desaturation indexes calculated as unsaturated/(unsaturated + saturated) × 100

5.3.2. Detected genomic regions

Our multi-population GWA resulted in the detection of 56 genomic regions containing SNPs significantly associated with at least one of the studied FA traits (Table 5.2). Significant associations were detected on all chromosomes except bos Taurus autosome (BTA) 18. Most of the FA traits showed significant associations with multiple genomic regions on several chromosomes; particularly for C10:0 (14 regions), C16:0 (12 regions), C16:1 (13 regions), C18:1c9 (11 regions) and C16 index (13 regions). Proportions of genetic variance explained by the lead SNPs in the detected regions ranged between 1.4 % and 45.3 % for the different FA traits studied.

Table 5.2. Genomic regions associated with milk fatty acid traits in the multi-population analysis and suggested candidate genes

Region[*]	Start (Mbp)	End (Mbp)	Traits associated (and % of explained genetic variance)	Candidate genes
1a	19.92	19.93	C16:0$_{(3.1)}$	
1b	101.0	101.0	C18 index$_{(2.8)}$	
1c	141.3	142.5	C15:0$_{(3.9)}$	
2a	12.5	19.8	C8:0$_{(3.7)}$, C10:0$_{(3.0)}$	*OSBPL6, AGPS*
2b	54.9	59.8	C14:1$_{(1.6)}$, C16:0$_{(3.6)}$, C16:1$_{(2.1)}$, C14 index$_{(1.5)}$	
2c	64.1	67.8	C16:1$_{(2.3)}$, C16 index$_{(2.3)}$	
2d	106.5	135.6	C12:0$_{(2.5)}$, C15:0$_{(5.6)}$, C16:0$_{(2.8)}$, C18:1c9$_{(3.8)}$	*MOGAT1, FABP3, MECR*
3	116.2	119.4	C18:3n3$_{(4.3),}$ CLA$_{(3.2)}$	*PRLH*
4	15.59	15.6	C15:0$_{(5.2)}$	
5a	10.33	10.36	C15:0$_{(9.0)}$	
5b	65.7	82.8	C8:0$_{(3.9)}$, C10:0$_{(2.5)}$	*CHPT1*
5c	87.4	99.0	C8:0$_{(4.3)}$, C10:0$_{(3.2)}$, C12:0$_{(2.6)}$, C14:1$_{(1.7)}$, C16:0$_{(2.7)}$, C16:1$_{(2.1)}$, C18:1c9$_{(5.6)}$, CLA$_{(3.2)}$, C14 index$_{(2.4)}$, C16 index$_{(4.9)}$	*MGST1, PLBD1, LRP6*
6	41.4	41.4	C18 index$_{(2.9)}$	
7a	14.6	15.5	C8:0$_{(3.3)}$, C10:0$_{(2.2)}$	
7b	78.4	78.4	C18:2n6$_{(3.3)}$	
7c	81.6	83.2	C12:0$_{(3.0)}$, C15:0$_{(6.0)}$	
8a	57.5	59.7	C15:0$_{(6.1)}$, C16:1$_{(2.0)}$, C16 index$_{(2.5)}$	*PIGO, STOML2*
8b	79.9	98.4	C14:0$_{(3.9)}$, C18:0$_{(4.1)}$, CLA$_{(3.3)}$	
9a	25.5	25.6	C14:1$_{(1.7)}$	
9b	81.3	81.3	C15:0$_{(5.0)}$	
10a	1.1	8.6	C10:0$_{(2.0)}$, C12:0$_{(3.5)}$	

Table 5.2. *Continued*

Region[*]	Start (Mbp)	End (Mbp)	Traits associated (and % of explained genetic variance)	Candidate genes
10b	12.9	12.9	C14:1$_{(1.6)}$, C18:0$_{(3.6)}$	*SLC51B*
10c	78.1	80.1	C18:3n3$_{(4.9)}$	*PIGH*
10d	87.5	93.1	C18:0$_{(4.1)}$, CLA$_{(3.4)}$, C18 index$_{(2.5)}$	
11a	24.7	26.7	C16:0$_{(2.6)}$	*ABCG5, ABCG8*
11b	58.81	58.89	C16:0$_{(2.8)}$	
12a	17.1	17.1	C18:1c9$_{(3.5)}$	
12b	24.0	24.8	C14:1$_{(1.8)}$	*ALG5*
12c	70.0	77.4	CLA$_{(3.5)}$, C16 index$_{(2.5)}$	
13	64.6	65.7	C10:0$_{(2.4)}$	*NCO6, ACSS2*
14a	1.5	5	C8:0$_{(7.8)}$, C10:0$_{(3.6)}$, C14:0$_{(8.8)}$, C14:1$_{(2.1)}$, C15:0$_{(16.3)}$, C16:0$_{(33.8)}$, C16:1$_{(7.8)}$, C18:1c9$_{(34.1)}$, C18:2n6$_{(34.3)}$, C18:3n3$_{(24.2)}$, CLA$_{(14.6)}$, C14 index$_{(4.5)}$, C16 index$_{(11.3)}$, C18 index$_{(11.4)}$	*DGAT1, GPAA1*
14b	5.2	20	C8:0$_{(4.3)}$, C10:0$_{(2.7)}$, C15:0$_{(5.2)}$, C16:0$_{(11.2)}$, C16:1$_{(6.6)}$, C18:1c9$_{(10.5)}$, C18:2n6$_{(15.2)}$, C18:3n3$_{(12.8)}$, CLA$_{(4.7)}$, C14 index$_{(1.8)}$, C16 index$_{(3.4)}$, C18 index$_{(4.4)}$	*ST3GAL1*
14c	44.7	49.9	C14:1$_{(2.0)}$, C16:1$_{(1.9)}$, C14 index$_{(1.6),}$ C18 index$_{(2.7)}$	*PMP2, FABP9, FABP4 FABP12*
15a	27.2	31.2	C10:0$_{(2.3)}$, C14:0$_{(4.6)}$, C18:0$_{(4.6)}$	*APOA1, APOA4, APOA5, DPAGT1*
15b	46.9	65.9	C10:0$_{(2.8)}$	*CAT, ELF5*
16a	23.8	25.22	C18:0$_{(3.8)}$, C16 index$_{(2.3)}$	
16b	57.53	57.58	C16:1$_{(1.7)}$, C16 index$_{(2.1)}$	
17a	17.4	22.6	C16:1$_{(3.0)}$, C16 index$_{(2.1)}$	*MGST2*
17b	27.8	44.1	C8:0$_{(5.9)}$, C10:0$_{(3.0)}$, C16:1$_{(2.6)}$, C18:3n3$_{(4.8)}$, C16 index$_{(2.3)}$	*LARP1B*

Table 5.2 *Continued*

Region*	Start (Mbp)	End (Mbp)	Traits associated (and % of explained genetic variance)	Candidate genes
19	37.3	61.3	C8:0$_{(7.6)}$, C10:0$_{(12.6)}$, C12:0$_{(13.6)}$, C14:0$_{(22.3)}$, C16:0$_{(4.6)}$, C18:1c9$_{(3.9)}$, C14 index$_{(3.1),}$ C18 index$_{(2.5)}$	*ACLY, BRCA1, FASN, STAT5A,*
20a	32.4	34.2	C16:1$_{(1.9)}$, C18:0$_{(4.3)}$	*PLCXD3, PRKAA1*
20b	36.7	36.9	C14:1$_{(1.6)}$, C18:1c9$_{(3.9)}$	
20c	55.3	60.4	C14 index$_{(1.6)}$, C18 index$_{(2.8)}$	
21	53.8	59.1	C10:0$_{(2.3)}$, C12:0$_{(2.9)}$, C14:0$_{(3.3)}$, C18:1c9$_{(4.1)}$	
22	59.12	59.13	C14 index$_{(1.6)}$	
23a	26.7	32.7	CLA$_{(4.3)}$	*AGPAT1, ATAT1*
23b	33.5	36.5	C15:0$_{(5.8)}$	
23c	40.7	43.5	C18:1c9$_{(3.4)}$, C16 index$_{(2.1)}$, C18 index$_{(2.6)}$	
24	10.2	10.2	C18:0$_{(4.2)}$	
25a	9.8	9.9	C12:0$_{(3.1)}$	
25b	24.7	24.7	C18:1c9$_{(3.5)}$	
25c	41.4	41.7	CLA$_{(3.0)}$ C14 index$_{(1.4)}$	
26	2.9	43.0	C8:0$_{(3.7)}$, C10:0$_{(5.5)}$, C12:0$_{(3.3)}$, C14:0$_{(8.0)}$, C14:1$_{(39.0)}$, C16:0$_{(2.4),}$ C16:1$_{(13.6)}$, C18:0$_{(4.5)}$, C14 index$_{(45.3)}$, C16 index$_{(19.7)}$, C18 index$_{(3.3)}$	*SCD, ELOVL3, ACSL5, GPAM*
27	37.0	42.2	C16:0$_{(2.9)}$	
28	36.6	37.2	C16:1$_{(2.3)}$, C16 index$_{(2.5)}$	
29	32.9	40.5	C16:0$_{(2.5)}$, C18:1c9$_{(3.2)}$	*TKFC*

Peak sizes (highest –log 10 p-value) across FA traits ranged from a –log 10 p-value of 6.9 for C18:0 to a –log 10 p-value of 126 for C14 index. Figures 5.1 – 5.4 present Manhattan plots for all FAs according to the different FA groups i.e., *de novo* FAs (Figure 5.1), intermediate to long-chain saturated FAs (Figure 5.2), the unsaturated FAs (Figure 5.3), and desaturation indexes (Figure 5.4). The strongest association for C8:0 (-log10 p-value=11.39), C15:0 (-log10 p-value=21), C16:0 (-log10 p-value=58), C16:1 (-log10 p-value=55), C18:1c9 (-log10 p-value=46), C18:2n6 (-log10 p-value=29), C18:3n3 (-log10 p-value=24.8), CLA (-log10 p-value=18.1) and C18 index (-log10 p-value=19.3) was observed at two variants on BTA 14 (ARS-BFGL-NGS-4939 and BovineHD1400000216). This region (14a) was significantly associated with all studied FA traits except C12:0. The lead SNP in this region explained up to 34 % of the genetic variation in C18:1c9 and C18:2n6. Two other regions on BTA 14 remained significantly associated with multiple FA traits after accounting for the fixed effect of the lead SNP from region 14a (ARS-BFGL-NGS-4939). The second region (14b) was also significantly associated with most FA traits except C12:0. The third region on BTA 14 (14c), was significantly associated with C14:1, C16:1, C14 index and C18 index. The lead SNP in this region explained 2.7 % of the genetic variation in C18 index and 1.6 % in C14 index.

Strongest association for C10:0 (-log10 p-value=24.3), C12:0 (-log10 p-value=22) and C14:0 (24.2) was detected with two variants on BTA 19 (BovineHD1900014372 and BovineHD1900014348). Significant associations were also detected for C8:0, C16:0, C18:1c9, C14 index and C18 index with SNPs located between 37.3 to 61.3 Mbp on chromosome 19. Particularly for C14:0, 22.3 % of the genetic variation was explained by the lead SNP in this region.

The strongest association for C14:1 (-log10 p-value=98.8), C14 index (-log10 p-value=126) and C16 index (-log10 p-value=39.8) was found with SNPs on chromosome 26 (BovineHD2600005461). Significant associations were also detected for C8:0, C10:0, C12:0, C14:0, C16:0, C16:1, C18:0 and C18 index. The lead SNP in this region explained 39.0 % of the genetic variation in C14:1.

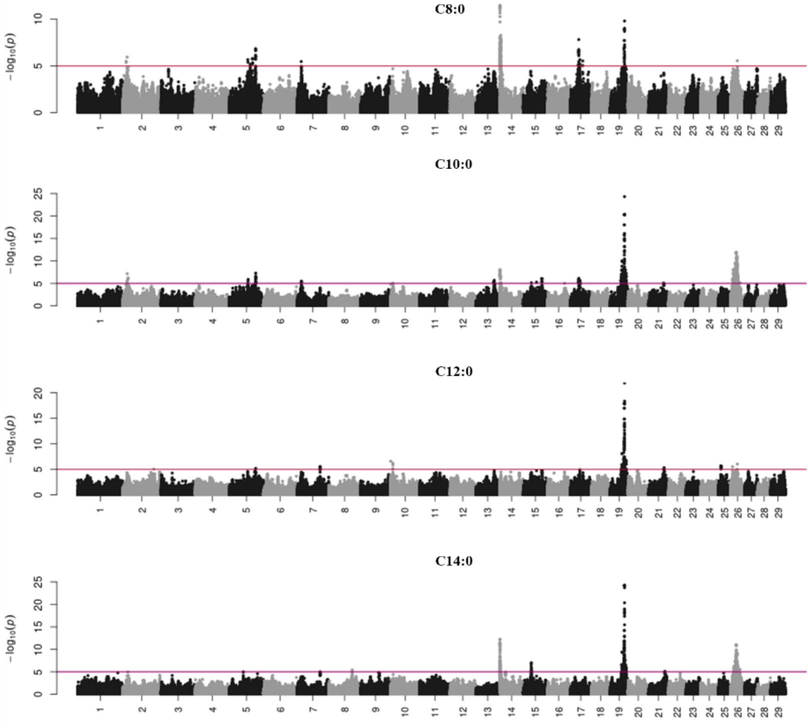

Fig 5.1. Manhattan plots showing BTAs on the x-axis and -log 10-p values on the y-axis for the de novo synthesized FAs. Red line indicates the significance threshold (log 10 p-value =5.0).

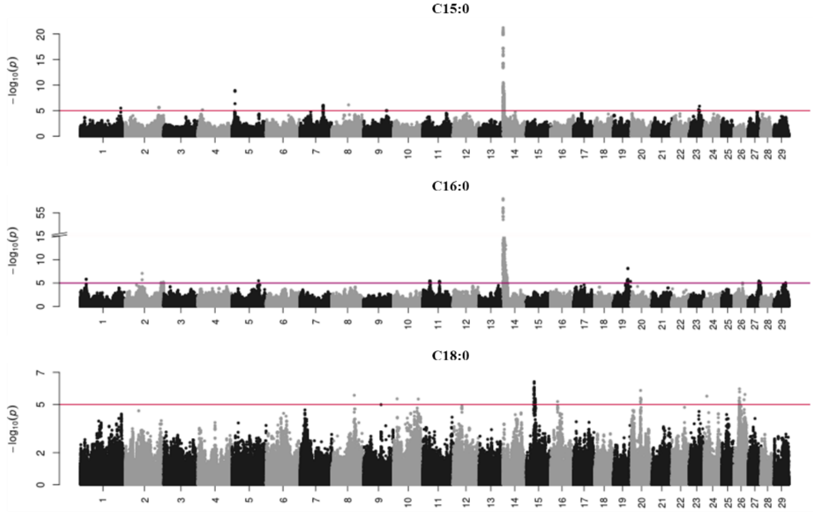

Fig 5.2. Manhattan plots showing BTAs on the x-axis and -log 10-p values on the y-axis for the intermediate and long chain saturated FAs. y-axis for C16:0 has breaks at –log 10-p-value =15

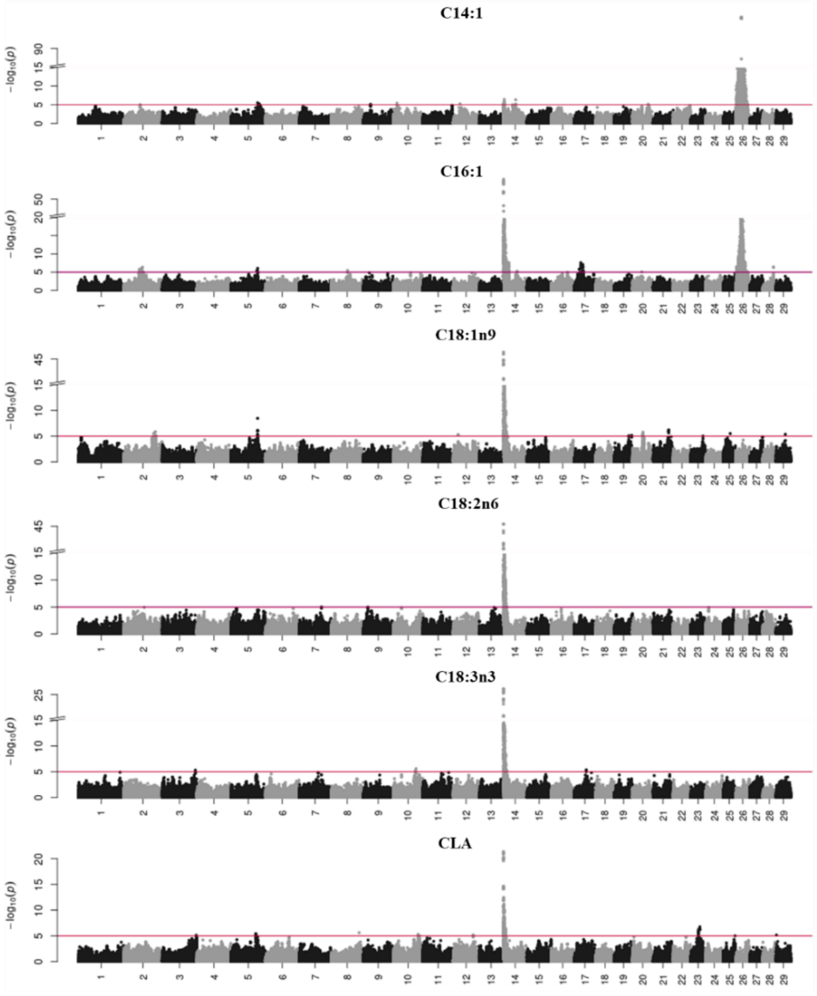

Fig 5.3. Manhattan plots showing BTAs on the x-axis and -log 10-p values on the y-axis for the unsaturated FAs. Y-axis breaks at –log 10-p-value =20 for C16:1 and at –log 10-p-value = 15 for C14:1, C18:1n9 and C18:2n6.

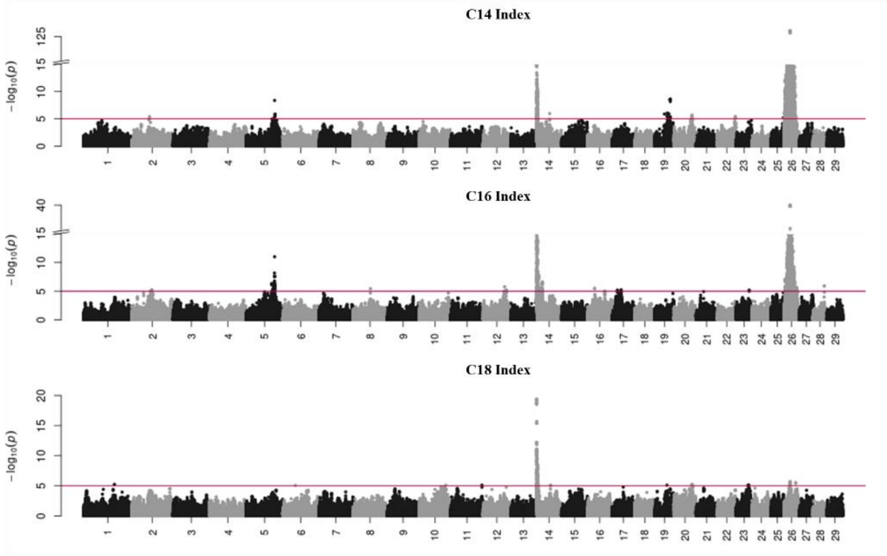

Fig 5.4. Manhattan plots showing BTAs on the x-axis and -log 10-p values on the y-axis for the desaturation indexes with y-axis breaks at –log 10 p-values = 15 for C14 Index and C16 Index

5.3.3. Gene assignment and functional annotation

Several genes positioned within the detected genomic regions were retrieved from the ensemble database. These positional candidate genes were further prioritized using enrichment analyses implemented in the DAVID web platform, which resulted in different significantly enriched GO terms and KEGG pathways relevant to FA related mechanisms (Table 5.3).

Among the enriched GO terms and pathways were biosynthesis related, such as 'GO:0006633~FA biosynthetic process' (7 genes) and 'bta01040:biosynthesis of unsaturated FAs' (3 genes), binding and transport related, such as 'GO:0008289~lipid binding (8 genes) and 'GO:0006869~lipid transport' (3 genes), and metabolism, such as 'GO:0006631~FA metabolic process' (21 genes) and 'bta00591:linoleic acid metabolism' (4 genes).

Some among the set of genes in all significantly enriched pathways and GO terms were also found to be expressed in mammary tissues and epithelial cells across different species. Furthermore, some of the prioritized candidate genes were linked to abnormalities related to FA metabolism in the mammalian phenotype database including 'increased circulating triglyceride levels' (MP:0001552), 'abnormal lipid homeostasis' (MP:0002118) and 'abnormal phospholipid level' (MP:0004777). Apart from genes, also non-coding genomic features such as micro RNAs were located within the detected genomic regions.

Table 5.3 List of significantly enriched pathways and GO terms

Category	Term	Count	Pvalue
GOTERM_BP_DIRECT	GO:0006633~fatty acid biosynthetic process	7	<0.001
GOTERM_MF_DIRECT	GO:0008289~lipid binding	8	<0.001
GOTERM_BP_DIRECT	GO:0070328~triglyceride homeostasis	5	<0.001
GOTERM_BP_DIRECT	GO:0008610~lipid biosynthetic process	4	<0.001
GOTERM_BP_DIRECT	GO:0016042~lipid catabolic process	15	<0.001
GOTERM_BP_DIRECT	GO:0045717~negative regulation of fatty acid biosynthetic process	4	<0.001
GOTERM_BT_ALL	GO:0010876~lipid localization	12	0.001
GOTERM_MF_DIRECT	GO:0005543~phospholipid binding	5	0.001
GOTERM_BP_DIRECT	GO:0006631~fatty acid metabolic process	21	0.005
GOTERM_BP_DIRECT	GO:0006629~lipid metabolic process	4	0.01
GOTERM_BP_DIRECT	GO:0006869~lipid transport	3	0.02
GOTERM_BP_DIRECT	GO:0006750~glutathione biosynthetic process	3	0.02
GOTERM_CC_DIRECT	GO:0043190~ATP-binding cassette (ABC) transporter complex	2	0.01
GOTERM_MF_DIRECT	GO:0036041~long-chain fatty acid binding	2	0.02
GOTERM_BP_DIRECT	GO:0045796~negative regulation of intestinal cholesterol absorption	2	0.02
GOTERM_MF_DIRECT	GO:0004623~phospholipase A2 activity	3	0.03
GOTERM_MF_DIRECT	GO:0070653~high-density lipoprotein particle receptor binding	2	0.03
GOTERM_BP_DIRECT	GO:0046486~glycerolipid metabolic process	3	0.03
GOTERM_BP_DIRECT	GO:0055114~oxidation-reduction process	8	0.04
UP_KEYWORDS	Acyltransferase	10	<0.001
INTERPRO	IPR016181:Acyl-CoA N-acyltransferase	6	<0.001
KEGG_PATHWAY	bta00564:Glycerophospholipid metabolism	9	<0.001
KEGG_PATHWAY	bta04975:Fat digestion and absorption	10	<0.001
KEGG_PATHWAY	bta00565:Ether lipid metabolism	7	<0.001
KEGG_PATHWAY	bta00062:Fatty acid elongation	4	0.004
KEGG_PATHWAY	bta05204:Chemical carcinogenesis	3	0.004
KEGG_PATHWAY	bta00591:Linoleic acid metabolism	4	0.01
KEGG_PATHWAY	bta01040:Biosynthesis of unsaturated fatty acids	3	0.03
KEGG_PATHWAY	bta04977:Vitamin digestion and absorption	3	0.03
KEGG_PATHWAY	bta04919:Thyroid hormone signaling	5	0.05

5.4 Discussion

5.4.1 Agreement between detected regions and previous reports

Our multi-population GWA resulted in detection of large numbers of genomic regions significantly associated with at least one of the 16 milk FA traits studied, indicating the complexity of the milk FA synthesis pathways. Most of the detected genomic regions have been previously reported in connection to milk FA traits, e.g. genomic regions on BTA 14, BTA 19 and BTA 26 (e.g. Schennink et al., 2009a; Bouwman et al., 2012; Li et al., 2014).

On BTA 14, our analysis indicates three distinct regions significantly associated with several FA traits. The first region is known to contain the *DGAT1* gene, of which the effects are well established for multiple FA traits (e.g. Grisart et al., 2002; Bovenhuis et al. 2016). The second region was previously reported to show significant associations with milk fat percentage (Jiang et al., 2010). The boundaries of these two regions (14a and 14b) are in close proximity of each other (1.5-5 Mbp and 5.2-20 Mbp) and the regions appear to be highly correlated in terms of associated FA traits and proportions of genetic variance explained for these traits. While our analysis indicates two distinctive regions, Bouwman et al. (2012), based on part of the dataset used in our study, reported a single, broader region (0.0 - 26.3 Mbp) with significant associations with several FA traits. Our hypothesis is that different QTLs underlie these two regions (14a and 14b) but that estimated effects of the two QTL could be confounded, because the high LD at the start of BTA 14 (Arias et al., 2009) makes it difficult to disentangle the effects of multiple QTL.

The third region on BTA 14 (44.7 – 49.9 Mbp) was exclusively associated with C14:1 and C16:1 as well as C14 index and C18 index. This region was previously reported for significant associations with C16:1 (Bouwman et al., 2012) and milk fat percentage (Cole et al., 2011). The region contains the fatty acid binding proteins *FABP4*, *FABP9* and *FABP12* as well as the peripheral myelin protein (*PMP2*), enriching the GO terms of FA (GO:0006631) and lipid (GO:0006629) metabolic processes as well as lipid binding activities (GO:0008289). A study by Nafikov et al. (2013) reported a *FABP4* haplotype negatively associated with saturated milk FAs and the ratio between saturated and unsaturated FAs while having positive effects on the unsaturated FAs. Marchitelli et al. (2013) also reported that the *FABP4* affected the ratio of monounsaturated/saturated FA in milk. Additionally, variation in *FABP4* is reported to affect other milk production traits such as milk yield (Zhou et al., 2015). Therefore, results of our analysis and previous studies suggest a role of this region in desaturation of C14:0, C16:0 and C18:0 with the FABP4 as the most likely candidate gene.

Broader regions were detected on BTA 19 (37.3 – 61.3 Mbp) and BTA 26 (2.9 – 43.0 Mbp). The genes FASN on BTA 19 (e.g. Schennink et al., 2009b) and *SCD1* on BTA 26 (e.g. Mele et al., 2007) have previously been suggested as the likely candidate genes for FA traits. However, our enrichment analysis indicate additional genes in these regions connected to important FA metabolism processes including the *ACLY*, *STAT5a*, *PLCXD3*, *PRKAA1, GH* on BTA 19 and *ELOVL3*, *ACLS5* on BTA 26. Significant associations were previously reported between variants within some of these genes and some milk FA traits (e.g. Bouwman et al., 2011; Strillacci et al., 2014).

In our study, more FA traits have been found to have significant associations with the *DGAT1* and *SCD1* regions than previous GWA studies using different parts of the multi-population dataset used in the current analysis (e.g. Bouwman et al., 2011, 2012; Buitenhuis et al. 2014; Li et al., 2014, 2015). These previous studies might not be considered as independent of the current analysis; however, more associations in the current analysis can be an indication of improved detection power from combining the populations. This was also demonstrated in our previous study (Chapter 4) in which results of population-specific analyses versus multi-population joint GWA were compared. Effects of the *DGAT1* (ARS-BFGL-NGS-4939) and *SCD1* (BovineHD2600005461) loci were similar in direction and highly correlated between the three populations but estimated effects in the Chinese sample were consistently lower across the FAs compared to the Dutch and Danish Holstein samples.

The three regions detected on BTA 5 overlap with previously reported regions for milk FA traits (Bouwman et al., 2012; Buitenhuis et al., 2014; Littlejohn et al., 2016). For region 5c, *MGST1* was suggested as the most likely candidate gene (Littlejohn et al., 2016). In our analysis, the lead SNP in the region was located within the *MGST1* gene. However, our enrichment analysis did not establish any connection to *MGST1* with significantly enriched FA related GO terms and pathways. Additionally, *PLBD1* and *LRP6* genes were connected to several pathways including lipid localization (GO:0010876) and transport (GO:0006869) suggesting that the significant association observed in the region with 10 FA traits might not be limited to the *MGST1* effect.

The region on BTA 13 was previously detected in the Dutch Holstein population (Bouwman et al., 2011 and 2012) and in Danish Jersey (Buitenhuis et al., 2014) with both studies suggesting the *ACSS2* as the highly likely candidate gene. Meanwhile, using IR predicted phenotypes for the *de novo* FAs, Olsen et al. (2017) suggested that the *NCOA6*, not the *ACSS2*, is responsible for significant associations in the region. Our enrichment analysis however links *ACSS2* with several significantly enriched pathways while no such links were established for the *NCOA6* gene.

Similarly, the first region on BTA 15 (27.2 – 31.2 Mbp) has been reported in previous studies including a joint Chinese-Danish Holstein population (Li et al., 2015). Several genes enriching FA related pathways were detected in the region including *APOA1*, *APOA4*, *APOA5*, and *DPAGT1*. The apolipoproteins *APOA1/4/5* enriched glycerolipid metabolic process (GO:0046486), fat digestion and absorption (bta04975) as well as negative regulation of FA biosynthetic process (GO:0045717) while the *DPGAT1* was involved in lipid biosynthetic process (GO:0046486). The strongest associations observed in the region were between C18.0 and variants within the alipoprotein genes, which showed opposite direction of effects on C10:0 and C14:0. Although effects were not significant, the lead SNP in the region also showed moderate effects on the other *de novo* FAs including C8:0 (-log 10 p-value = 2.96) and C12:0 (-log 10 p-value = 2.96) with direction of effects similar to C10:0 and C14:0. The alipoproteins *APOA1/4/5* are thus collectively suggested as the candidates underlying the strong effect on C18:0 observed in the region. The opposing effects on the *de novo* FAs might be directly through involvement of the alipoproteins in negative regulation of FA biosynthesis or indirectly through the effect on C18:0, which suppresses *de novo* synthesis.

The two regions detected on BTA 17 are also in agreement with previous findings. The regions detected by Bouwman et al. (2012) (15.0 – 23.9 Mbp) and Li et al., (2014) (19.5 – 22.5 Mbp) overlap with the first region (17a) detected in our study. In the region, *MGST2* significantly enriched GO terms that included FA (GO:0006631) and lipid (GO:0006629) metabolic processes and FA biosynthetic process (GO:0006633). The *MGST2* is previously linked to intramuscular FA composition in pigs (Muñoz et al, 2013) and shown to be expressed in all stages of lactation in humans (Lemay et al., 2013). Therefore, the MGST2 is suggested as the likely candidate gene underlying effects on the first region of BTA 17. Using a subset of the dataset used in the current study to fine map BTA 17, Duchemin et al. Duchemin et al. (2017) suggested the *LARP1B* as a primary candidate gene in the second region (17b). However, our enrichment analysis did not result in significant enrichment of any of the FA pathways and ontology terms for genes in the region.

Some of the regions detected in our analysis overlap with results from some of the recently published GWA studies that are based on IR predicted FA phenotypes (e.g. Olsen et al., 2017; Knutsen et al., 2018). Interestingly, some of the well-established genomic regions in connection to GC-based FA traits, which were also detected in our analysis, have not been found to have significant associations with any of the milk FA phenotypes in these studies. For instance, GWA studies of Olsen et al. (2017) and Knutsen et al. (2018) using the FTIR predicted FA phenotypes in Nordic Red cattle did not detect any significant association in the *DGAT1* and *SCD1* regions. Lack of segregation of the A variant of the DGAT1 K232A polymorphism has been suggested as the potential reason for the lack of association in the *DGAT1* region. Additionally, Wang et al. (2016a) showed that the *SCD1* polymorphism did not significantly affect any of the milk IR wavenumbers in samples from the Dutch Holstein population. These findings suggest that IR predicted FA phenotypes are not preferred for GWA studies. While some FAs can be accurately predicted based on IR (Soyeurt et al., 2011), low prediction accuracies (e.g. De Marchi et al., 2011) and low genetic correlations with GC measured FA (e.g. Poulsen et al., 2014) have been reported for other FAs. Especially FAs found in low concentrations in milk were shown to have low IR prediction accuracies (Rutten et al. 2009). Apparently, the power to detect QTL can be severely restricted by the IR prediction accuracy.

5.4.2 Novel genomic regions and candidate genes

Of the 56 genomic regions significantly associated with at least one FA trait in this study, regions located on BTA 2, 3 10, 11, 12 and 21 appear to be novel regions that have not been previously connected to milk FA traits. The lead SNPs in these regions explained between 1.4 % and 5 % of the genetic variation in at least one of the FA traits studied.

5.4.2.1 BTA 2

Two genes retrieved for region 2a enriched GO terms related to fatty acids. The *OSBPL6* gene belonging to the oxysterol-binding protein (*OSBP*) family, a group of intracellular lipid receptors, is shown to be involved in lipid binding (GO:0008289) and transport (GO:0006869) processes. The *OSBPL6* gene is shown to be expressed in the human mammary gland during several stages of lactation (Lemay et al., 2013). The human *OSBPL6* gene is also shown to have a binding site for miR-33a/b (Ouimet et al. 2016), which is a microRNA shown to have targeting effects on genes regulating β-oxidation of FAs (Gerin et al., 2010), leading to significantly lower levels of β-hydroxybutyrate (Goedeke et al., 2013). Another gene located in the region (*AGPS*) also enriched GO terms related to FA synthesis including lipid biosynthesis process (GO:0008610) and lipid metabolic process (GO:0006629). In the mammalian phenotype database, mutation in the *AGPS* gene in mice has been linked to abnormal lipid levels (MP:0001547), which is a rather broad term in the database referring to any anomaly in the concentrations of fat-soluble substances in the body, including circulating triglyceride and free FAs. Thus, our enrichment analysis indicate that both the *OSBPL6* and *AGP*S might have roles on *de novo* synthesis of FAs. Pattern of SNP effects in the region is also in agreement with enrichment analysis such that strongest association was estimated with C8:0 and C10:0 while moderate, but not significant effect was measured for C12:0 and C14:0 (-log 10 p-value = 4.2). Opposing direction of the lead SNP effect were also observed for the *de novo* synthesized FAs, except C15:0, on the one hand and most of the long chain FAs on the other (Figure 5 A). Therefore, both the *OSBPL6* and *AGPS* are considered as likely candidates in the region. .

5.4.2.2 BTA 3

On the detected novel region of BTA 3, the prolactin releasing hormone (*PRLH*) was shown to be involved in lipid metabolic process (GO:0006629). Mutations on the *PRLH* gene in mice have been associated with increased circulating triglyceride levels (MP:0001552) and increased total body fat amount (MP:0010024) in the mammalian phenotype database. In mammals, the *PRLH* gene is known to stimulate prolactin release and regulate its expression. Prolactin, which is a polypeptide hormone, has been shown to stimulate the expression of genes involved in milk protein synthesis and lipid metabolism (Houdebine et al., 1985; Matusik and Rosen, 1980; Rudolph et al., 2011) and induce lipogenesis in many tissues (Barber et al., 1991). Moreover, prolactin has been shown to have a wide-range of effects on lactation including growth and development of the mammary gland, promotion of milk synthesis and maintenance of milk secretion (Shiu and Friesen, 1980; Akers et al., 1981; Lamberts and Macleod, 1990). Therefore, the *PRLH* gene, through regulation of prolactin release might have effects on milk yield. The pattern of SNP effects in the region suggest a connection with the poly-unsaturated fatty acids (PUFAs) with strongest associations observed for C18:3n3 and CLA. The direction of effects of the lead SNP was similar for all unsaturated FAs as well as all the desaturation indexes, while opposing effects were estimated for the *de novo* synthesized FAs and C16:0 (Figure 5 B). C16:0 is shown to have strong negative genetic correlation with milk yield, while moderate positive correlations were reported for the PUFAs (Stoop et al., 2008). Therefore, the PRLH is suggested as the candidate gene in the region; the effect of which might be indirect through its effect on milk yield, affecting the concentration of the PUFAs and C16:0.

5.4.2.3 BTA 10

The second region on BTA 10 contains the solute carrier family 51-beta subunit (*SLC51B*) gene, implicated in lipid transport (GO:0006869) and localization (GO:0010876) processes. Pattern of effects in the region show strong effect on

C14:1 and moderate effects on with C14 index (-log 10 p-value = 4.5) and C18 index (-log 10 p-value = 3.6) in direction opposite to the strong effect on C18:0 (Figure 5C). This pattern suggests a reduction in desaturation when C18:0 increases. C18:0 in milk is largely derived via direct transport through the blood from the rumen where is it formed from bio-hydrogenation of dietary C18:2n6 and C18:3n3. Therefore, the effect of *SLC51B* is highly likely through its involvement in the FA transport processes. Dietary poly-unsaturated FAs, such as C18:2n6, are known to suppress *SCD1* activity, thereby reducing its desaturation activity (Jeffcoat and James, 1978). Thus, we hypothesize that *SLC51B* underlies the effect on C18:0, while observed opposite effects on the unsaturated FAs and desaturation indexes are rather due to the correlation in C18:0 in milk and dietary PUFA, which suppress desaturation.

5.4.2.4 BTA 11

Among the genes located in the first region of BTA 11, the ATP binding cassette subfamily G5 (*ABCG5*) and *ABCG8* enriched several pathways and processes including fat digestion and absorption pathway (KEGG~bta04975) and the GO terms of lipid localization (GO:0010876) and transport (GO:0006869). In the mammalian phenotype database, the *ABCG5* and *ABCG8* genes are linked to increased circulating triglyceride level (MP:0001552), abnormal lipid homeostasis (MP:0002118) and abnormal phospholipid level (MP:0004777). In humans, mutations in *ABCG5/8* have been linked to conditions characterized by abnormal accumulation of sterols in blood and tissues (e.g. Berge et al., 2000; Lee et al., 2001) implicating them in lipid absorption and transport. The KEGG pathway for fat digestion and absorption involves absorption of lipid from the rumen to the blood stream and from the blood stream to the mammary gland. Viturro et al. (2006) previously reported high expression levels of both *ABCG5* and *ABCG8* genes in bovine liver, mammary gland, digestive tract and blood samples. Expression of *ABCG5/8* in bovine mammary gland might indicate that apart from absorption and transport of lipids from the digestive tract, *ABCG5/8* might also be involved in the secretion of lipids from the mammary gland into the milk. Significant association in the region was limited to C16:0. Although not significant, this region was also associated with C16:1 (-log 10-pvalue = 3.8) and CLA (-log 10-pvalue = 2.2), with directions of effects on CLA opposite to the effects on C16:0, C16:1, and C16 index (Figure 5 D). The GO term of lipid localization and association almost exclusively with C16:0, which is one of the FAs that is highly mobilized from body reserves during negative energy balance, might also indicate a role in the mechanism of body fat reserve mobilization.

5.4.2.5 BTA 12

The dolichyl-phosphate beta glucosyltransferase (*ALG5*) gene located on the second region of BTA 12 was shown to enrich the lipid biosynthesis process (GO:0008610) and glycerolipid metabolic process (GO:0046486). The *ALG5* gene has previously been shown to be differentially expressed during the different stages of lactation in bovine (Bionaz et al., 2012) and human (Lemay et al, 2013). Significant effects in the region were limited to C14:1. The lead SNP also showed moderate effect on C14 index (-log 10 p-value=3.07) and C18:0 (-log 10 p-value=3.43) where opposite direction of effects were observed for C18:0 (Figure 5 E). Therefore, the *ALG5* is suggested as promising candidate for further characterization for potential role in desaturation process.

5.4.2.6 BTA 21

Significant associations were detected on BTA 21 with C10:0, C12:0, C14:0 and C18:1c9. Effects estimated for the lead SNP were generally positive in the *de novo* synthesized FAs and C16:0 while they were negative for the long-chain FAs and desaturation indexes (Figure 5 F). Significant associations have previously been reported with bovine milk yield and milk protein yield (Kolbehdari et al., 2009) as well as cow fertility traits (Nayeri et al., 2016). However, our enrichment analysis show no gene implicated in the significantly enriched pathways and GO terms. Despite lack genes implicated on FA related pathways, moderate effects observed for multiple traits in the region are of particular interest. QTL detected through GWA might be located in non-coding regions. Such QTLs might be involved in regulation of expression of other genes affecting the traits of interest. Therefore, the region might be of interest for eQTL based GWA studies in milk FA traits.

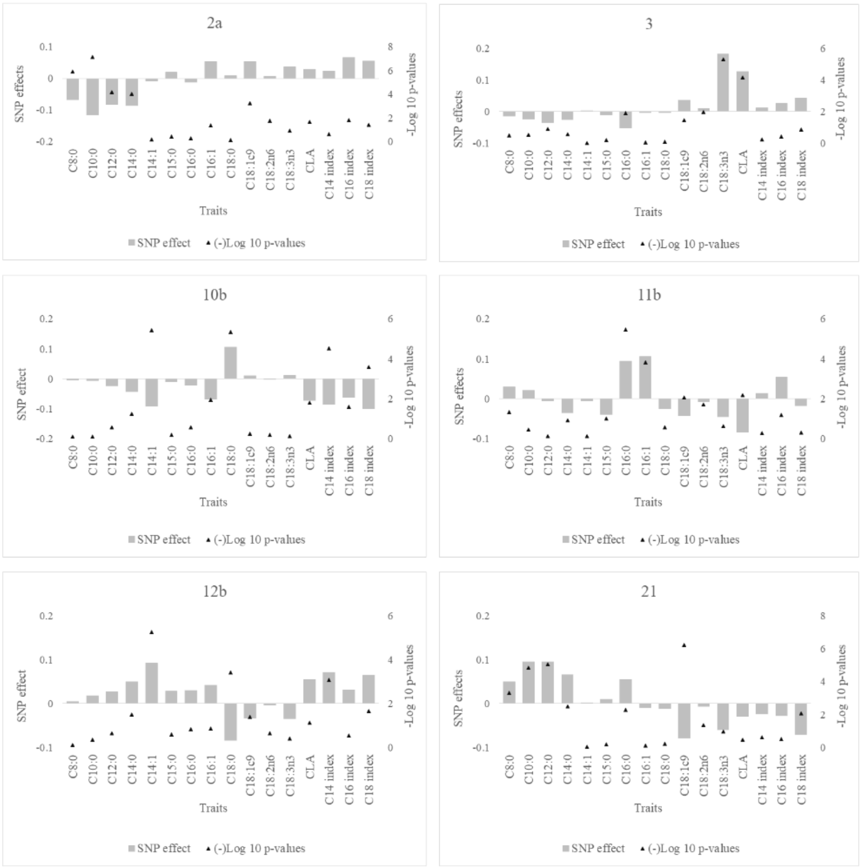

Fig 5.5. Effects of lead SNPs on regions 2a, 3, 10b, 11b, 12b and 21 standardized by dividing the SNP effects with standard deviation of the respective FA trait

5.4.3 Regulatory elements within detected genomic regions

Apart from coding genes, retrieved genes from the detected regions included regulatory elements, most commonly microRNAs (miRNAs). MiRNAs are small RNAs that regulate the expression of complementary messenger RNAs (Ambros, 2004). Several studies have reported possible roles of miRNAs in lipid and fatty acid metabolisms and in mammary gland development and lactation in several species (e.g. Dávalos et al., 2011; Li et al., 2016; Wang et al., 2016b). Some of the miRNAs in the detected genomic regions in our study were previously linked to regulatory roles on genes related to FA metabolism and synthesis. Of these, bta-mir-27b, on BTA 8 (region 8b) was shown to target known FA synthesis genes such as *FASN* and *SCD1* (Zhang et al., 2017) as well as mRNAs involved in lipid metabolism (Vickers et al., 2013) and shown to be highly expressed during different stages of bovine lactation (Do et al., 2017). Among the genes located on BTA 2 (region 2d), the bta-mir-26b was shown to be expressed in bovine milk cells and mammary gland Li et al., 2016. Wang et al., (2016b) showed that downregulation of miR-26a/b and their host genes decreased the expression of genes related to fatty acid synthesis, including *DGAT1* and *SCD1*.

5.5 Conclusion

Multi-population GWA for GC-quantified FA traits resulted in the detection of 56 genomic regions significantly associated to at least one of the studied FAs, including novel regions explaining relatively smaller fractions of the genetic variation. Enrichment analysis of genes harbored in detected regions reveals promising candidate genes some of which have not been previously linked to milk FA traits, including *OSBPL6* and *AGPS* on BTA 2, *PRLH* on BTA 3, *SLC51B* on BTA 10, *ABCG5/8* on BTA 11 and *ALG5* on BTA 12. Post-GWA analyses using multiple data sources on pathways, ontology terms and tissue-specific gene expression status enabled prioritization of highly likely causative candidate genes among several positional candidates on detected regions. Use of such data in combination to patterns of effects across the milk FA spectrum allowed linking some of the candidates to specific FA synthesis mechanisms. Detection of several novel regions and candidate genes will be contribute to the knowledge base on genetics underlying the bovine milk FA composition.

Authors' contributions

GG processed the data, implemented the analyses and drafted the manuscript. HB conceived the study and contributed to the discussion of the results. AJB co-planned the study and contributed to the discussion of the results. NAP and HJFV collected the milk samples and contributed to the milk analysis and discussion of the results. MHPWV, DS and QZ contributed to the discussion of the results. All authors read and approved the final manuscript.

References

Akers RM., Bauman DE., Capuco AV., Goodman GT., Tucker HA. (1981). Prolactin regulation of milk secretion and biochemical differentiation of mammary epithelial cells in periparturient cows. Endocrinology. 109(1):23-30.

Ambros V. (2004). The functions of animal microRNAs. Nature. 431:350-355.

Arias JA., Keehan M., Fisher P., Coppieters W., Spelman R. (2009). A high density linkage map of the bovine genome. BMC Genet. 24;10:18.

Barber M.C., Finley E., Vernon RG. (1991). Mechanisms whereby prolactin modulates lipogenesis in sheep mammary gland. Horm Metab Res. 23(3):143-5.

Beavis W.D. (1998). QTL analyses: power precision and accuracy. In Molecular dissection of complex traits. Edited by AH P. New York: CRC Press: 145–162.

Berge K.E., Tian H., Graf G..A., Yu L., Grishin NV., Schultz J., Kwiterovich P., Shan B., Barnes R., Hobbs HH. (2000). Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. Science. 290(5497):1771-5.

Bionaz M., Periasamy K., Rodriguez-Zas S.L., Hurley W.L., Loor J.J. (2012). A novel dynamic impact approach (DIA) for functional analysis of time-course omics studies: validation using the bovine mammary transcriptome. PLoS One. 7(3):e32455.

Bovenhuis H., Visker M.H.P.W., Poulsen N.A., Sehested J., van Valenberg H.J.F, van Arendonk J.A.M., Larsen L.B., Buitenhuis A.J. (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(4):3113-3123.

Bouwman A.C., Bovenhuis H., Visker M.H., van Arendonk J.A. (2011). Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet. 11;12:43.

Bouwman A.C., Visker M.H., van Arendonk J.A., Bovenhuis H. (2012). Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. BMC Genet. 29;13:93. doi: 10.1186/1471-2156-13-93.

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

Cai Z., Guldbrandtsen B., Lund M.S., Sahana G. (2018). Dissecting closely linked association signals in combination with the mammalian phenotype database can identify candidate genes in dairy cattle. BMC Genet. 11;19(1):30.

Chilliard Y., Ferlay A., Mansbridge R.M., Doreau M. (2000). Ruminant milk fat plasticity: nutritional control of saturated, polyunsaturated, trans and conjugated FA. Ann. Zootech. 49: 181-205.

Cole J.B., Wiggans G.R., Ma L., Sonstegard T.S., Lawlor TJ. Jr., Crooker B.A., Van Tassell C.P., Yang J., Wang S., Matukumalli L.K., Da Y. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics. (11)12:408.

Dávalos A., Goedeke L., Smibert P., Ramírez C.M., Warrier N.P., Andreo U., Cirera-Salinas D., Rayner K., Suresh U., Pastor-Pareja J.C., Esplugues E., Fisher E.A., Penalva L.O., Moore K.J., Suárez Y., Lai E.C., Fernández-Hernando C. (2011). miR-33a/b contribute to the regulation of fatty acid metabolism and insulin signaling. Proc Natl Acad Sci U S A. 31;108(22):9232-7.

De Marchi M., Penasa M., Cecchinato A., Mele M., Secchiari P., Bittante G. (2011). Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal. 5(10):1653-8.

de Roos A.P., Hayes B.J., Spelman R.J., Goddard M.E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 179(3):1503-12.

Do D.N., Li R., Dudemaine P.L., Ibeagha-Awemu E.M. (2017). MicroRNA roles in signaling during lactation: an insight from differential expression, time course and pathway analyses of deep sequence data. Sci Rep. 20;7:44605.

Duchemin S.I., Visker M.H., Van Arendonk J.A., Bovenhuis H. (2014). A quantitative trait locus on Bos taurus autosome 17 explains a large proportion of the genetic variation in de novo synthesized milk fatty acids. J Dairy Sci. 97(11):7276-85.

Duchemin S.I., Bovenhuis H., Megens H.J., Van Arendonk J.A.M., Visker M.H.P.W. (2017). Fine-mapping of BTA17 using imputed sequences for associations with de novo synthesized fatty acids in bovine milk. J Dairy Sci. 100(11):9125-9135.

Gebreyesus G., Lund M.S., Janss L., Poulsen N.A., Larsen L.B., Bovenhuis H., Buitenhuis A.J. (2016). Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J Dairy Sci 99(4):2863-2866.

Gerin I., Clerbaux L.A., Haumont O., Lanthier N., Das A.K., Burant C.F., Leclercq I.A., MacDougald O.A., Bommer G.T. (2010). Expression of miR-33 from an SREBP2 intron inhibits cholesterol export and fatty acid oxidation. J Biol Chem. 29;285(44):33652-61.

Goedeke L., Vales-Lara F.M., Fenstermaker M., Cirera-Salinas D., Chamorro-Jorganes A., Ramírez C.M., Mattison J.A., de Cabo R., Suárez Y., Fernández-Hernando C. (2013). A regulatory role for microRNA 33* in controlling lipid metabolism gene expression. Mol Cell Biol.;33(11):2339-52.

Grisart B., Coppieters W., Farnir F., Karim L., Ford C., Berzi P., Cambisano N., Mni M., Reid S., Simon P., Spelman R., Georges M., Snell R. (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. 2002 Feb;12(2):222-31.

Houdebine L.M., Djiane J., Dusanter-Fourt I., Martel P., Kelly P.A., Devinoy E., Servely J.L. (1985). Hormonal action controlling mammary activity. J Dairy Sci. 1985 Feb;68(2):489-500. Review.

Jeffcoat R., James AT. (1978). The control of stearoyl-CoA desaturase by dietary linoleic acid. FEBS Lett. 1978 Jan 1;85(1):114-8.

Jiang L., Liu J., Sun D., Ma P., Ding X., Yu Y., Zhang Q. (2010). Genome wide association studies for milk production traits in Chinese Holstein population. PLoS One. 27;5(10):e13661.

Knutsen T.M., Olsen H.G., Tafintseva V., Svendsen M., Kohler A., Kent M.P., Lien S. (2018). Unravelling genetic variation underlying de novo-synthesis of bovine milk fatty acids. Sci Rep. 1;8(1):2179. doi: 10.1038/s41598-018-20476-0.

Kolbehdari D., Wang Z., Grant J.R., Murdoch B., Prasad A., Xiu Z., Marques E., Stothard P., Moore S.S. (2009). A whole genome scan to map QTL for milk production traits and somatic cell score in Canadian Holstein bulls. J Anim Breed Genet. 2009 Jun;126(3):216-27.

Lamberts S.W., Macleod R.M. (1990). Regulation of prolactin secretion at the level of the lactotroph. Physiol Rev. 1990 Apr;70(2):279-318. Review.

Lee M.H., Lu K., Hazard S., Yu H., Shulenin S., Hidaka H., Kojima H., Allikmets R., Sakuma N., Pegoraro R., Srivastava A.K., Salen G., Dean M., Patel S.B. (2001). Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. Nat Genet. 27(1):79-83.

Lemay D.G., Ballard O.A., Hughes M.A., Morrow A.L., Horseman N.D., Nommsen-Rivers L.A. (2013). RNA sequencing of the human milk fat layer transcriptome reveals distinct gene expression profiles at three stages of lactation. PLoS One. 5;8(7):e67531.

Li C., Sun D., Zhang S., Wang S., Wu X., Zhang Q., Liu L., Li Y., Qiao L. (2014). Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. PLoS One. 23;9(5):e96186

Li X., Buitenhuis A.J., Lund M.S., Li C., Sun D., Zhang Q., Poulsen N.A., Su G. (2015). Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. J Dairy Sci. 98(11):8152-63.

Li R., Dudemaine P.L., Zhao X., Lei C., Ibeagha-Awemu E.M. (2016). Comparative Analysis of the miRNome of Bovine Milk Fat, Whey and Cells. PLoS One. 21;11(4):e0154129.

Littlejohn M.D., Tiplady K., Fink T.A., Lehnert K., Lopdell T., Johnson T., Couldrey C., Keehan M., Sherlock R.G., Harland C., Scott A., Snell R.G., Davis S.R., Spelman R.J. (2016). Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. Sci Rep. 5;6:25376.

Lund M.S., Su S., Janss L., Guldbrandtsen B., Brøndum R.F. (2014). Genomic evaluation of cattle in a multi-breed context. Livestock Science: (166)101–110.

Marchitelli C., Contarini G., De Matteis G., Crisà A., Pariset L., Scatà M.C., Catillo G., Napolitano F., Moioli B. (2013). Milk fatty acid variability: effect of some candidate genes involved in lipid synthesis. J Dairy Res. 80(2):165-73.

Massart-Leën A.M., Roets E., Peeters G., Verbeke R. (1983). Propionate for fatty acid synthesis by the mammary gland of the lactating goat. J Dairy Sci. 66(7):1445-54.

Matusik R.J., Rosen J.M. (1980). Prolactin regulation of casein gene expression: possible mediators. Endocrinology. 106(1):252-9.

Mele M., Conte G., Castiglioni B., Chessa S., Macciotta N.P., Serra A., Buccioni A., Pagnacco G., Secchiari P. (2007). Stearoyl-coenzyme A desaturase gene polymorphism and milk fatty acid composition in Italian Holsteins. J Dairy Sci. 90(9):4458-65.

Muñoz M., Rodríguez M.C., Alves E., Folch J.M., Ibañez-Escriche N., Silió L., Fernández A.I. (2013). Genome-wide analysis of porcine backfat and intramuscular fat fatty acid composition using high-density genotyping and expression data. BMC Genomics. 2;14:845.

Nafikov R.A., Schoonmaker J.P., Korn K.T., Noack K., Garrick D.J., Koehler K.J., Minick-Bormann J., Reecy J.M., Spurlock D.E., Beitz D.C. (2013). Association of polymorphisms in solute carrier family 27, isoform A6 (SLC27A6) and fatty acid-binding protein-3 and fatty acid-binding protein-4 (FABP3 and FABP4) with fatty acid composition of bovine milk. J Dairy Sci. 96(9):6007-21.

Nayeri S., Sargolzaei M., Abo-Ismail M.K., May N., Miller S.P., Schenkel F., Moore S.S., Stothard P. (2016). Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. BMC Genet. 10;17(1):75.

Olsen H.G., Knutsen T.M., Kohler A., Svendsen M., Gidskehaug L., Grove H., Nome T., Sodeland M., Sundsaasen K.K., Kent M.P., Martens H., Lien S. (2017). Genome-wide association mapping for milk fat composition and fine mapping of a QTL for de novo synthesis of milk fatty acids on bovine chromosome 13. Genet Sel Evol. 13;49(1):20. doi: 10.1186/s12711-017-0294-5.

Ouimet M., Hennessy E.J., van Solingen C., Koelwyn G.J., Hussein M.A., Ramkhelawon B., Rayner K.J., Temel R.E., Perisic L., Hedin U., Maegdefessel L., Garabedian M.J., Holdt L.M., Teupser D., Moore K.J. (2016). miRNA Targeting of Oxysterol-Binding Protein-Like 6 Regulates Cholesterol Trafficking and Efflux. Arterioscler Thromb Vasc Biol. 36(5):942-951.

Park J.H., Wacholder S., Gail M.H., Peters U., Jacobs K.B., Chanock S.J., Chatterjee N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010 Jul;42(7):570-5.

Poulsen N.A., Gustavsson F., Glantz M., Paulsson M., Larsen L.B., Larsen M.K. (2012). The influence of feed and herd on fatty acid composition in 3 dairy breeds (Danish Holstein, Danish Jersey, and Swedish Red). J Dairy Sci. 95(11):6362-71.

Poulsen N.A., Eskildsen C.E.A., Skov T., Larsen L.B., Buitenhuis A.J. (2014). Comparison of genetic parameters estimation of fatty acids from gas chromatography and FT-IR in Holsteins. In proceedings: 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC Canada.

Rudolph M.C., Russell T.D., Webb P., Neville M.C., Anderson S.M. (2011). Prolactin-mediated regulation of lipid biosynthesis genes in vivo in the lactating mammary epithelial cell. Am J Physiol Endocrinol Metab. 2011 Jun;300(6):E1059-68.

Rutten M.J., Bovenhuis H., Hettinga K.A., van Valenberg H.J., van Arendonk J.A. (2009). Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J Dairy Sci. 92(12):6202-9.

Schennink A., Heck J.M., Bovenhuis H., Visker M.H., van Valenberg H.J., van Arendonk J.A. (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(5):2135-43.

Schennink A., Stoop W.M., Visker M.H., van der Poel J.J., Bovenhuis H., van Arendonk J.A. (2009a). Short communication: Genome-wide scan for bovine milk-fat composition. II. Quantitative trait loci for long-chain fatty acids. J Dairy Sci. 92(9):4676-82. doi: 10.3168/jds.2008-1965.

Schennink A., Bovenhuis H., Léon-Kloosterziel K.M., van Arendonk J.A.M., Visker M.H.P.W. (2009b). Effect of polymorphisms in the FASN, OLR1, PPARGC1A, PRL and STAT5A genes on bovine milk-fat composition. Anim Gen. 40:909-916.

Shiu R.P., Friesen H.G. (1980). Mechanism of action of prolactin in the control of mammary gland function. Annu Rev Physiol. 42:83-96. Review.

Smith C.L., Goldsmith C.A., Eppig J.T. (2005). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol. 6(1):R7.

Soyeurt H., Dehareng F., Gengler N., McParland S., Wall E., Berry D.P., Coffey M., Dardenne P. (2011). Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J Dairy Sci. 94(4):1657-67.

Stoop W.M., van Arendonk J.A., Heck J.M., van Valenberg H.J., Bovenhuis H. (2008). Genetic parameters for major milk fatty acids and milk production traits of Dutch Holstein-Friesians. J Dairy Sci. 91(1):385-94.

Stoop W.M., Bovenhuis H., Heck J.M., van Arendonk J.A. (2009). Effect of lactation stage and energy status on milk fat composition of Holstein-Friesian cows. J Dairy Sci. 92(4):1469-78. doi: 10.3168/jds.2008-1468.

Strillacci M.G., Frigo E., Canavesi F., Ungar Y., Schiavini F., Zaniboni L., Reghenzani L., Cozzi M.C., Samoré A.B., Kashi Y., Shimoni E., Tal-Stein R., Soller M., Lipkin E., Bagnato A. (2014). Quantitative trait loci mapping for conjugated linoleic acid, vaccenic acid and Δ(9) -desaturase in Italian Brown Swiss dairy cattle using selective DNA pooling. Anim Genet. 45(4):485-99.

Vickers K.C., Shoucri B.M., Levin M.G., Wu H., Pearson D.S., Osei-Hwedieh D., Collins F.S., Remaley A.T., Sethupathy P. (2013). MicroRNA-27b is a regulatory hub in lipid metabolism and is altered in dyslipidemia. Hepatology. 57(2):533-42.

Viturro E., Farke C., Meyer H.H., Albrecht C. (2006). Identification, sequence analysis and mRNA tissue distribution of the bovine sterol transporters ABCG5 and ABCG8. J Dairy Sci. 89(2):553-61.

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

Wang Q., Hulzebosch A., Bovenhuis H. (2016a). Genetic and environmental variation in bovine milk infrared spectra. J Dairy Sci. 99(8):6793-6803.

Wang H., Luo J., Zhang T., Tian H., Ma Y., Xu H., Yao D., Loor J.J. (2016b). MicroRNA-26a/b and their host genes synergistically regulate triacylglycerol synthesis by targeting the INSIG1 gene. RNA Biol. 3;13(5):500-10.

Yang J., Lee S.H., Goddard M.E., Visscher P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76-82.

Zhang M., Sun W., Zhou M., Tang Y. (2017). MicroRNA-27a regulates hepatic lipid metabolism and alleviates NAFLD via repressing FAS and SCD1. Sci Rep. 3;7(1):14493.

Zhou L., Ding X., Zhang Q., Wang Y., Lund M.S., Su G. (2013). Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genet Sel Evol. 21;45:7.

Zhou H., Cheng L., Azimu W., Hodge S., Edwards G.R., Hickford J.G. (2015). Variation in the bovine FABP4 gene affects milk yield and milk protein content in dairy cows. Sci Rep. 12;5:10023. doi: 10.1038/srep10023.

Zimin A.V., Delcher A.L., Florea L., Kelley D.R., Schatz M.C., Puiu D., Hanrahan F., Pertea G., Van Tassell C.P., Sonstegard T.S., Marçais G., Roberts M., Subramanian P., Yorke J.A., Salzberg S.L. (2009). A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 10(4):R42

6

Accuracy of genomic prediction for milk fatty acid composition using multi-population reference and incorporating GWAS findings

G. Gebreyesus[1,2], H. Bovenhuis[2], M. S. Lund[1], N. A. Poulsen[3], D. Sun[4] and A. J. Buitenhuis[1]

[1]Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark; [2]Animal Breeding and Genomics Centre, Wageningen University, PO Box 338, 6700 AH Wageningen, the Netherlands; [3]Department of Food Science, Aarhus University, Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark; [4]Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture of China, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China; [5]Dairy Science and Technology Group, Wageningen University and Research, P.O. Box 17, 6700 AA Wageningen, the Netherlands

Submitted [Genetics Selection Evolution]

Abstract

Large-scale phenotyping for milk fatty acid (FA) composition is difficult due to expensive and time-consuming analytical techniques. Reliability of genomic prediction is often low for traits that are expensive/difficult to measure and for breeds with small reference population sizes. An effective method to increase reference population size could be to combine datasets from different populations. Prediction models might also benefit from incorporation of information on biological underpinnings of quantitative traits. Genome-wide association studies (GWAS) show that genomic regions on BTA 14, 19 and 26 underlie substantial proportions of the genetic variation in milk FA traits. Genomic prediction models incorporating such findings could enable improved prediction accuracy despite limited reference population sizes. In this study, we combine gas chromatography (GC) quantified FA samples from the Chinese, Danish and Dutch, Holstein populations and implement a genomic-features best linear unbiased prediction (GFBLUP) model incorporating variants on BTA 14, 19 and 26 as genomic features for which random genetic effects are estimated separately. Prediction accuracies were compared to traditional GBLUP models. Our prediction using multi-population reference with traditional GBLUP model resulted on average gains in prediction reliability of 11 percentage points in the Dutch, 9 in the Danish and 1 percentage point in the Chinese prediction compared to predictions with population-specific references. Implementation of GFBLUP model with multi-population reference led to further increases in prediction reliability of up to 38 percentage points in the Dutch, 26 percentage points in the Danish and 4 percentage points in the Chinese population compared to the traditional GBLUP. Prediction reliabilities from the GFBLUP model were moderate to high across the FA traits. With our results, we show that it is possible to predict genetic merits for milk FA traits with reasonable accuracy by combining related populations of a breed and with models incorporating GWAS findings.

Key words: Genomic prediction, Milk fatty acid, Multi-population, GWAS

6.1 Introduction

Milk contains several FAs, which can be grouped into different categories depending on the length of carbon chains, degree of unsaturation and isomerization. Some groups of FAs in milk have been linked to various health risks, while others have been suggested as beneficial for human health. Such links have long triggered interests to alter the FA profile of bovine milk. Several studies have reported substantial genetic variation in bovine milk FA traits (e.g. Stoop et al., 2008; Krag et al., 2013) presenting an opportunity to alter the milk FA composition through selective breeding. Genomic selection has become the main strategy in livestock selective breeding allowing selection of candidate bulls at younger ages (Hayes et al. 2009). However, prediction accuracy for traits that are difficult and expensive to measure is still limited due to small reference population sizes. So far, genomic prediction accuracy has not been reported for milk FA composition traits despite the growing interest to include these in the dairy cattle breeding goals (Boichard and Brochard, 2012). This is mainly due to the difficulty to record milk FA traits at large-scale. Gas chromatography (GC), the current method of choice in quantifying milk FA traits at high accuracy, requires expensive equipment and time-consuming techniques challenging large-scale phenotyping.

A strategy that is increasingly getting attention in genomic prediction for numerically small breeds or traits difficult to measure is to combine datasets from different breeds/populations (Lund et al., 2014, van den Berg et al., 2017). Benefits of combining data for genetic analysis are highly dependent on the genetic distance between the populations used in different studies and the marker density (Lund et al., 2014). In this study, we combine samples for 16 FA traits quantified by GC method in the Chinese, Danish and Dutch Holstein populations genotyped using high density (HD) single nucleotide polymorphism (SNP) arrays for genomic prediction. Given the common use of outstanding North American bulls in the Chinese, Danish and Dutch Holstein breeding population, high genetic similarities are to be expected between these populations. Previously studies also show high consistency in linkage disequilibrium (LD) patterns between the Danish and Chinese Holstein (Zhou et al., 2013; Li et al., 2015) and between the Dutch, Danish and Chinese Holstein (Gebreyesus et al., submitted).

While genomic prediction allows using all markers genome-wide without the need for mapping quantitative trait loci (QTLs), incorporation of biological information might further improve accuracy for scarcely recorded traits. Methods have been suggested to weigh variants according to prior knowledge of their effect on the traits (e.g. Brøndum et al., 2015; MacLeod et al., 2016), with reports of some gain in prediction accuracies (e.g., Edwards et al., 2016; Sarup et al., 2016). GWAS have for long been used as a powerful tool for investigating the genetic background of quantitative traits and diseases. Incorporation of GWAS detections in genomic prediction models might improve genomic prediction accuracy (e.g. Spindel, 2016), especially when predication accuracy is limited by reference size. GWAS on milk FA traits have frequently reported significant associations on broader regions of BTA 14, 19 and 26 (Bouwman et al., 2011,2012; Buitenhuis et al., 2014; Li et al., 2014). Further characterization studies also suggest large effects of these regions on most milk FA traits (Ntambi et al., 2003; Mele et al., 2007; Schennink et al., 2007,2008; Bovenhuis et al., 2016; Pegolo et al., 2016;). In addition, several other regions explaining relatively smaller proportions on the genetic variations in multiple FA traits are also reported across the bovine genome (e.g. Bouwman et al., 2012; Buitenhuis et al., 2014; Gebreyesus et al., Submitted). Information on such major regions underlying the genetic variation might improve genomic prediction accuracy for the scarcely recorded milk FA traits.

Traditionally, GBLUP model (VanRaden, 2008) is based on the assumption that many QTL explaining small fractions of the genetic variance underlie quantitative traits. In implementation, genetic effects are estimated based on realized relationship matrix computed from genome-wide markers (VanRaden, 2008). Often, contribution of genetic markers to the genomic relationship is not weighted according to explained proportion of the genetic variance. Such approach where all markers contribute equally to the relationship matrix, despite differences in association with the traits, might cause “dilution” of effects of major regions. In this context, Sørensen et al. (2014) suggested extension of the GBLUP model to allow incorporation of available information regarding the biological mechanism underlying quantitative traits. To implement such extensions, Sørensen et al. (2014) suggested a genomic features BLUP approach (GFBLUP), where variants can be categorized according to biological information, such as chromosome, genes, genes grouped in pathways, to allow differentiation between groups of SNPs in their explained variance and size of effects for genomic prediction.

In this study, we implement a GFBLUP approach in which GWAS reported regions; i.e., BTA 14, BTA 19, BTA 26, are fitted as genomic features of interest to predict genomic breeding values (GBVs) for the FA traits. Prediction reliabilities are then compared with the traditional GBLUP model, which assumes common variance for markers throughout the genome. The objectives of this study were to: 1) Study genomic prediction reliabilities for 16 milk FA traits in three Holstein populations; 2) Investigate gains in genomic prediction reliability from combining multi-population reference sets and incorporating biological information based on GWAS findings.

6.2 Methods

This study compares prediction reliabilities from scenarios where the Dutch, Danish and Chinese Holstein sample cows were used as reference populations separately or combining these populations for a common reference. Moreover, the study compares prediction reliabilities from a traditional GBLUP model versus GFBLUP model where GWAS identified genomic regions are fitted as features explaining different.

6.2.1. Animals and phenotypes

Milk samples were obtained from 700 Chinese, 614 Danish and 1566 Dutch Holstein cows. The sampling of cows involved 18 herds in China, 22 herds across Denmark and 398 herds in the Netherlands. Sampled cows were found in different stages of lactation ranging between 60 to 700 days in milk in the Chinese population, 60 to 481 days in milk in the Danish population and 60 to 278 days in milk in the Dutch Holstein cows.

The GC method was used to quantify 13 FA traits (presented in Table 6.1) with details on methods as described by Li et al. (2014) for the Chinese samples, Poulsen et al. (2012) for the Danish samples and Stoop et al. (2008) for the Dutch samples. Furthermore, desaturation indexes were calculated based on the FA measurements as: C14 index = C14:1/(C14:1+C14:0) * 100; C16 index = C16:1/(C16:1+C16:0) * 100 and C18 index = C18:1c9/ (C18:1c9+C18:0) * 100.

6.2.2. Genotypes and Imputation

Real and/or imputed high-density (HD) genotypes were available for all the sample cows. All cows in the Chinese dataset were genotyped using the BovineSNP50 Beadchip (50K, Illumina). A population of 96 Chinese Holstein bulls, genotyped using the BovineHD Beadchip (777K), was used as reference to impute the 50K genotypes of the cows to HD. Some of the cows (N=278) in the Danish dataset were genotyped using the BovineSNP50 Beadchip. The rest of Danish cows were genotyped using the BovineHD Beadchip and used as reference to impute the 50K genotypes of the first part of the Danish cows to HD as described in Gebreyesus et al. (2016). A custom 50K SNP Beadchip was used to genotype cows in the Dutch dataset. A reference population consisting of 1333 Dutch Holstein cows and 55 bulls genotyped using BovineHD Beadchip (777K) was used to subsequently impute the 50K genotypes of the Dutch samples to HD as presented in detail in Duchemin et al. (2014).

Quality controls were undertaken on SNPs within each population. Accordingly, SNPs with minor allele frequencies (MAF) less than 0.05 or with a count of one of the genotypes less than 10 in each population were excluded from both the population-specific as well as combined-population predictions. A total of 464,130 SNPs were available in common for all the populations and scenarios.

6.2.3. Models

Traditional and "genomic features" GBLUP models were implemented to estimate genomic breeding values (GBVs).

6.2.3.1 Traditional GBLUP

GBLUP models were implemented using DMU (Madsen and Jensen, 2010) considering two scenarios: 1) a population-specific reference sets within the Chinese, Danish and Dutch samples and; 2) a combined reference set of the three populations. The general model used for the traditional GBLUP, both population-specific as well as combined population reference sets, was:

$$y_{ijkl} = \mu + parity_i + herd_j + b_1\ DIM_k$$

$$+ b_2 * exp^{-0.05*DIM_k} + g_l + e_{ijkl}\ , \qquad (1)$$

where y_{ijkl} is phenotype of cow *l* in parity *i*, and herd *j,* µ is the fixed mean effect; b_1 is the regression coefficient for DMI k, which is a covariate describing the effect of days in milk, b_2 is the regression coefficient for the Wilmink adjustment ($exp^{-0.05*DIM_l}$) of DMI, e_{ijkl} is a random residual effect assumed normally distributed with $\boldsymbol{e} \sim N(0,\ \mathbf{I}\ \sigma_e^2)$, where $\mathbf{I}$ is an identity matrix. The effect of g_l is a random additive genetic effect of cow *l* with distribution $N(0, \mathbf{G}\sigma_a^2)$, where **G** is the genomic relationship matrix between individuals and σ_a^2 is the genetic variation. The genomic relationship matrix used in the GBLUP models was calculated as described in the first method presented by VanRaden (2008).

6.2.3.2 GFBLUP

A GFBLUP model was implemented using a combined population reference sets to estimate GBV for population specific training-sets. In the traditional GBLUP model, single random genetic effect based on genomic relationship matrix constructed using all markers was considered. In contrast, four random genetic effects were considered in the GFBLUP approach according to the following model:

$$y_{ijkl} = \mu + parity_i + herd_j + b_1\ DIM_k$$

$$+b_2 * exp^{-0.05*DIM_k} + \boldsymbol{g}_{14} + \boldsymbol{g}_{19} + \boldsymbol{g}_{26} + \boldsymbol{g}_{\mathrm{R}} + e_{ijkl},\ (2)$$

Where $\boldsymbol{g}_{14}$ is vector of random additive genetic effects based on relationships matrix ($\mathbf{G_{14}}$) constructed using markers on BTA 14, with distribution $N(0, \mathbf{G_{14}}\sigma_{14}^2)$; where σ_{14}^2 is the genetic variation explained by markers on BTA 14. Similarly, $\boldsymbol{g_{19}}$ and $\boldsymbol{g_{26}}$ are vectors of random additive genetic effects based on relationships matrices computed using variants on BTA 19 and 26, respectively with similar distributional assumptions as in $\boldsymbol{g}_{14}$; while $\boldsymbol{g_R}$ is the vector of additive genetic effects based genomic relationship matrix constructed using all the rest of variants excluding those on BTA 14, 19 and 26, with distribution $N(0, \mathbf{G_R}\sigma_R^2)$. Variants used to calculate the relationship matrices include 13,033 SNPs for BTA 14, 12,603 SNPs for BTA19 and 9,703 SNPs for BTA 26. The different genomic relationship matrixes for the GFBLUP model were computed following the first method of VanRaden (2008). Other model (2) components were as in model (1).

The total genomic value was calculated as: $\boldsymbol{g} = \boldsymbol{g_{14}} + \boldsymbol{g_{19}} + \boldsymbol{g_{26}} + \boldsymbol{g_R}$.

Proportion of the genomic variance explained by each genetic effect component of the GFBLUP model was computed as:

$$\%var_{feature} = \sigma_{feature_i}^2 \Big/ \sigma_{total}^2,$$

Where $\sigma_{feature_i}^2$ is σ_{14}^2, or $\sigma_{19,}^2$ or σ_{26}^2, and σ_{total}^2 is the total additive genetic variance computed as: $\sigma_{total}^2 = \sigma_{14}^2 + \sigma_{19}^2 + \sigma_{26}^2 + \sigma_R^2$.

To study similarity of the LD structures in the three populations on the BTAs taken as features, pair-wise linkage disequilibrium (LD) was calculated between the SNP markers within a 1 Mbp window on BTAs 14, 19 and 26 with the r^2 as a measure in the Plink program (Purcell et al., 2007).

6.2.4. Training and validation populations

For all the scenarios, a resampling strategy was used to create five validation sets of 100 cows for each of populations. The general principle was to avoid sibling relationships among validation sets and with the reference population for each validation set. Table 6.1 shows the reference population sizes used in the within- as well as the combined-population genomic prediction for each trait and population. For the Danish population, subset of cows which had no siblings within the dataset, were first selected (n = 197). In each of the resampled analyses, 100 of these cows were randomly sampled for the validation set, while the remaining of these cows were included back to the reference population. In the Dutch and Chinese population, all the sampled cows had at least one half-sib in the dataset. In the Dutch sample, all cows belonged to one of three sire-groups whereas in the Chinese dataset, the majority of the cows was from five different sires. Hence, a sire-group (group of cows with common sire) was randomly selected for each validation set. As each sire-group contained more than 100 cows in both the Chinese and Dutch dataset, further random sampling of 100 cows with undertaken within selected sire-groups and the remaining cows in the group were excluded for the reference population. Main reason for limiting the validation to 100 cows is to have comparable reference population sizes in the Chinese and Dutch datasets as in the Danish population.

6.2.5. Prediction reliability

For all models, prediction reliability for cows was computed as the squared correlation between estimated GBV and the phenotype corrected for fixed effects and scaled by dividing with heritability estimates. Corrected phenotypes were computed based on single-population traditional GBLUP as in model model (1) and used commonly for all scenarios. Heritability estimates used to scale the reliabilities were from the traditional GBLUP approach computed as:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}. \qquad (3)$$

Accordingly, for population-specific genomic prediction, heritability estimates from traditional GBLUP model within each population were considered. Similarly, for genomic prediction using combined-population reference, heritability estimates from traditional GBLUP computed with the combined dataset were used to scale reliabilities in all the validations.

Table 6.1 Number of cows in the reference sets for each FA trait in the Chinese (CN), Danish (DK), Dutch (NL) and the combined population genomic prediction.

	CN		DK		NL	
Trait	Single	Combined	Single	Combined	Single	Combined
C8:0	584	2764	518	2771	892	2188
C10:0	585	2767	520	2775	892	2192
C12:0	585	2765	519	2774	892	2190
C14:0	586	2766	519	2774	892	2191
C15:0	583	2751	516	2760	887	2181
C16:0	583	2762	518	2769	892	2186
C18:0	587	2762	518	2771	889	2178
C14:1	584	2761	516	2769	890	2187
C16:1	583	2755	519	2763	887	2185
C18:1c9	585	2765	518	2773	892	2190
C18:2n6	585	2760	518	2768	889	2188
C18:3n3	583	2750	518	2759	885	2180
CLA	580	2750	518	2758	886	2178
C14index	583	2758	515	2767	890	2184
C16index	580	2750	517	2757	887	2177
C18index	585	2758	516	2767	889	2185

6.3 Results

6.3.1 Descriptive statistics and genetic parameters

Table 6.2 presents phenotypic means, coefficient of variation (%) and the heritability estimates for the FA traits in the different populations and the combined dataset. Generally, phenotypic means were comparable between the Danish and Dutch samples while the Chinese dataset showed larger differences in some of the FA traits. Such differences between the Chinese data on one hand and the Danish and Dutch on the other were specially observed for C8:0, C18:2n6 and C18:1c9. Larger differences were also observed in coefficient of variation estimates between the populations for some of the studied traits. In the combined dataset, coefficient of variation ranged between 5.6% (C18 index) and 63.0% (C18:2n6). Similarly, some differences were also observed in heritability estimates, as heritabilities were generally higher within the Dutch sample followed by the Danish estimates.

Table 6.2. Phenotypic means and coefficient of variation (%) for FA traits across populations and combined dataset

FAs	CN			DK			NL			Combined		
	Mean	CV	h^2	Mean	CV	h^2	Mean	CV	h^2	Mean	CV	h^2
Saturated FAs[1]												
C8:0	0.58	37.9	0.06	1.47	15.0	0.33	1.31	13.0	0.48	1.18	32.2	0.27
C10:0	2.22	18.0	0.16	3.22	17.4	0.36	2.87	15.7	0.51	2.80	20.7	0.39
C12:0	2.94	16.7	0.21	3.69	18.4	0.30	3.79	19.0	0.40	3.58	21.2	0.33
C14:0	10.10	11.3	0.22	11.60	11.7	0.14	11.10	9.5	0.39	11.00	11.5	0.25
C15:0	0.99	13.1	0.10	1.11	17.1	0.27	1.11	17.1	0.29	1.09	16.5	0.23
C16:0	32.90	5.6	0.27	30.10	11.6	0.12	29.10	12.0	0.48	30.20	11.7	0.34
C18:0	12.00	14.1	0.25	9.84	19.4	0.23	9.84	17.7	0.37	10.30	19.3	0.25
Unsaturated FAs[1]												
C14:1	0.86	24.4	0.35	1.01	27.7	0.49	1.38	19.6	0.55	1.19	29.4	0.47
C16:1	1.64	20.1	0.26	1.58	26.6	0.42	1.39	20.9	0.65	1.49	23.5	0.46
C18:1c9	28.30	8.6	0.24	19.60	14.5	0.07	20.20	13.8	0.41	21.90	20.0	0.27
C18:2n6	3.99	11.5	0.26	1.74	15.5	0.17	1.11	22.5	0.27	1.89	63.0	0.18
C18:3n3	0.42	14.3	0.05	0.50	18.0	0.05	0.50	32.0	0.27	0.48	27.1	0.19
CLA	0.41	22.0	0.15	0.57	26.3	0.11	0.56	46.4	0.32	0.53	43.4	0.21
Desaturation indexes [2]												
C14 index	7.84	20.8	0.36	7.98	23.7	0.59	11.0	16.6	0.62	9.71	24.4	0.53
C16 index	4.74	19.6	0.24	4.97	22.3	0.37	4.6	19.8	0.55	4.70	20.6	0.38
C18 index	70.20	4.7	0.21	66.60	5.9	0.26	67.3	5.8	0.49	67.80	5.87	0.31

[1] Expressed in % wt/wt

[2] Desaturation indexes calculated as unsaturated/(unsaturated + saturated) × 100

All parameter estimates for C18:2n6, C18:3n3 and CLA are computed on raw data before the log-transformation

6.3.2 Consistency in LD between the populations on BTAs 14, 19 and 26

Estimation of pair-wise LD (r^2) on BTAs 14, 19 and 26 indicates consistent LD structures between the populations on the regions considered for the genomic features prediction model (Figure 6.1). Furthermore, the minimum and maximum average pairwise LD values for SNPs within bins of 1Mbp sizes were similar between the populations in the three BTAs.

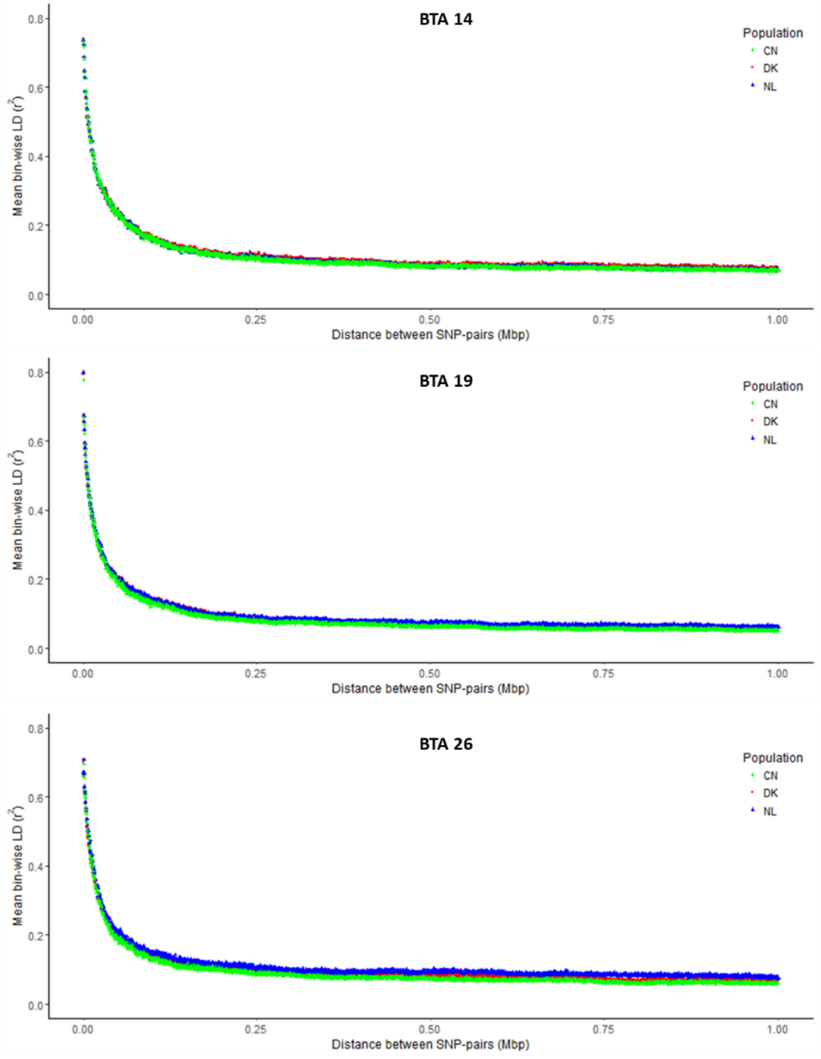

Fig 6.1. Mean bin-wise linkage disequilibrium (LD) for the Dutch (blue points), Danish (red points) and Chinese (green points) Holstein Friesian genotypes on BTA 14, 19 and 26. Y-axis is the mean bin-wise LD and x-axis the physical distance between pair-wise markers in Mega base pairs (Mbp).

Prediction accuracy with traditional GBLUP models

Table 6.3 presents reliabilities of predictions for the studied FA traits in the three populations using GBLUP model based on population-specific reference sets. Prediction reliabilities using reference populations separately were in general low across the FA traits and populations. Prediction reliabilities were especially low for the Chinese validation followed by the prediction in the Danish sets. In the Chinese single-population prediction, the highest reliability was observed for C18 index (0.15) followed by CLA (0.12). Prediction reliabilities were very low for the *de novo* synthesized FAs in general and for C10:0, C12:0 and C14:0 in particular in the Chinese validation set. Similarly, genomic prediction using the Danish reference population separately resulted in low reliabilities across the traits. The highest reliability was observed for CLA (0.14) followed by C14:0 (0.11), whereas the lowest values were observed for C16:0 and C18:1c9. Low to moderate prediction reliabilities were calculated for the Dutch reference population. The highest genomic prediction reliability using the Dutch separate reference population was computed for C14 index (0.43) followed by C14:1 (0.39).

Results of genomic prediction using combined-population reference show some increase in reliability across the traits and populations compared to the predictions using separate reference populations (Table 6.3). Genomic prediction using multi-population reference resulted on an average increase in prediction reliability of 12 percentage points for the Dutch validation population compared to single-population genomic prediction. The higher increases in prediction reliability were observed for C16 index (Δ = 0.26), C18:2n6 (Δ = 0.21) and C18 index (Δ = 0.20), whereas no improvement in prediction reliability was observed for C18:0. In the Danish validation, an average increase in prediction reliability of 8 percentage points was observed as a result of adding the Dutch and Chinese reference populations. The improvements in reliability were especially higher for C10:0 (Δ = 0.22), C8:0 (Δ = 0.20) and C12:0 (Δ = 0.14). Sizable improvements in reliability were also observed for C14 and C18 indexes. No improvements were shown for C14:0 and reliability declined for C18:3n3 in the multi-population prediction for the Danish validation. With an average increase in reliability of 1 percentage point, little or no increase in reliability was observed for most FA traits in the Chinese validation. In the Chinese validation, highest benefits of adding the Danish and Dutch reference populations were observed for C18:2n6 (Δ = 0.09), followed by C10:0 (Δ = 0.07) and C18:1c9 (Δ = 0.07).

Table 6.3 Genomic prediction reliabilities with traditional GBLUP based on within- and combined-population references.

Trait	CN		DK		NL	
	Single	Combined	Single	Combined	Single	Combined
C8:0	0.06	0.06	0.06	0.26	0.11	0.18
C10:0	0.003	0.07	0.06	0.28	0.17	0.27
C12:0	0.001	0.03	0.04	0.18	0.25	0.30
C14:0	0.0004	0.001	0.11	0.11	0.25	0.36
C15:0	0.03	0.03	0.04	0.19	0.03	0.06
C16:0	0.19	0.18	0.001	0.06	0.13	0.19
C18:0	0.19	0.11	0.06	0.07	0.05	0.05
C14:1	0.11	0.15	0.06	0.16	0.39	0.49
C16:1	0.07	0.11	0.09	0.13	0.22	0.38
C18:1c9	0.07	0.14	0.005	0.09	0.13	0.26
C18:2n6	0.06	0.15	0.01	0.05	0.10	0.31
C18:3n3	0.08	0.05	0.10	0.01	0.06	0.23
CLA	0.12	0.14	0.14	0.21	0.16	0.33
C14 index	0.10	0.16	0.05	0.18	0.43	0.56
C16 index	0.10	0.13	0.09	0.14	0.12	0.30
C18 index	0.15	0.05	0.02	0.15	0.12	0.19

6.3.3 Prediction reliability with GFBLUP model

Substantial increases in prediction reliability were obtained using GFBLUP model for multi-population reference sets compared to the combined GBLUP model (Figure 6.2). Accordingly, further average gain in prediction reliability of 13.9 percentage points was observed in the Dutch validation using the GFBLUP model compared to the traditional GBLUP. The increase in prediction reliability from the GFBLUP model reached as high as 40 percentage points (C14:0). Large increases in prediction reliabilities were also observed for C18:1c9 (Δ= 0.38) and C16:0 (Δ=0.33) in the validation for the Dutch Holstein. Similarly, gains in reliability of 10 percentage points in the Danish validation and 2.6 percentage points in the Chinese validation populations were achieved using the GFBLUP model. High improvements in prediction reliability were obtained for C15:0 (Δ = 0.26), C14:0 (Δ = 0.22), for C16 index (Δ = 0.22) and C18 index (Δ = 0.22). In the Chinese validation, the largest increase in reliability from the GFBLUP model was observed for C18 index (0.14) followed by C14 index (0.10).

Whereas the increases in prediction reliability from the GFBLUP model varied between populations, sizable improvements for specific FAs were consistent across the populations. For C14:1, for instance, prediction reliability was improved by 15 percentage points in the Dutch, 18 percentage points in the Danish and 10 percentage points in the Chinese validation populations.

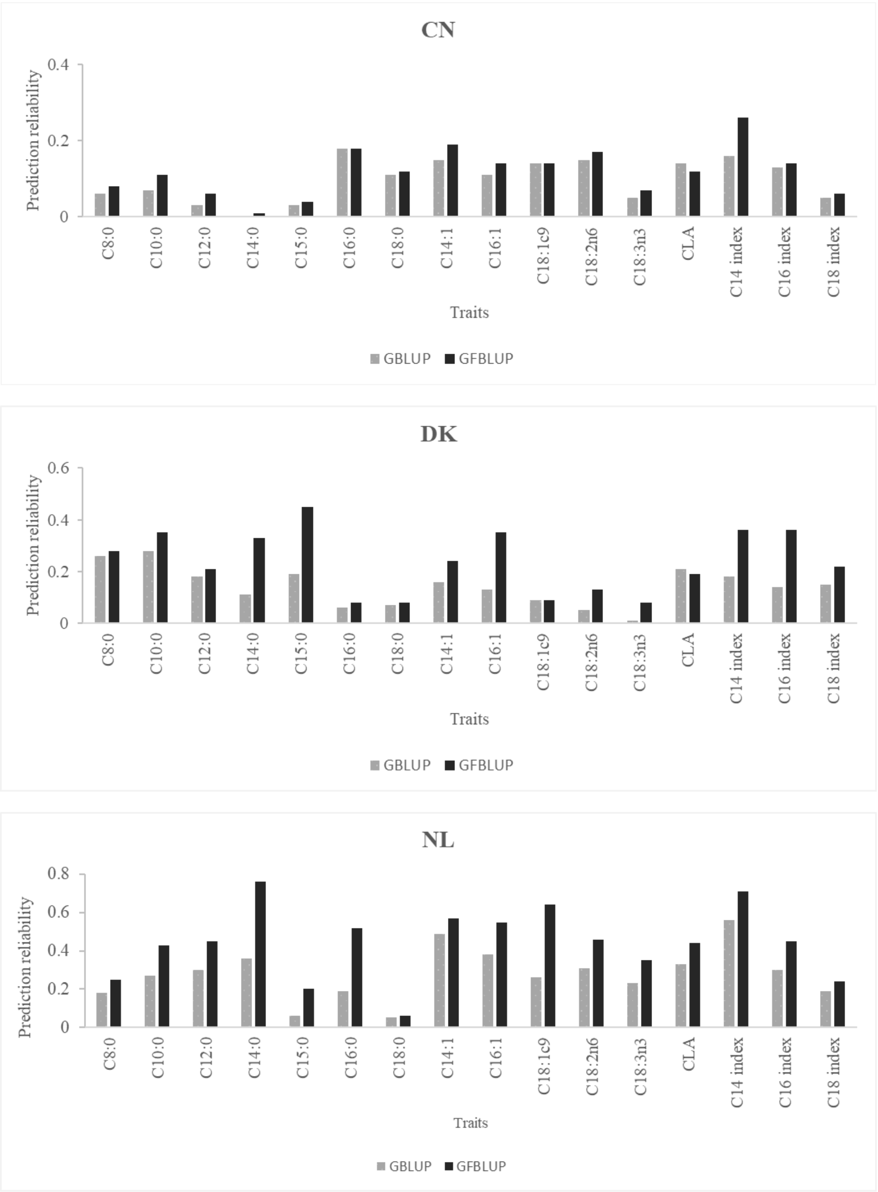

Fig 6.2. Genomic prediction reliabilities using combined references populations with traditional GBLUP and GFBLUP models in the Chinese, Danish and Dutch validation population

Genomic features fitted in the GFBLUP model (BTA 14, BTA 19 and BTA 26) collectively explained substantial proportions of the total genetic variation across the FA traits (Figure 6.3). For instance in C14 index, 57% of the genetic variance was explained by the genomic features with BTA 26 alone explaining 38.3 % of the genetic variation. Similarly, sizable proportion of the total genetic variation for C14:0 (33.4%) and C10:0 (20%) of the genetic variation was explained by BTA 19 alone. Variants on BTA 14 collectively explained 38.3% of the genetic variation for C16:0 and 36.7% of the genetic variation in C18:1n9 and C18:2n6. On the contrary for C18:0, BTA 14 and 26 explained 5% of the genetic variation each.

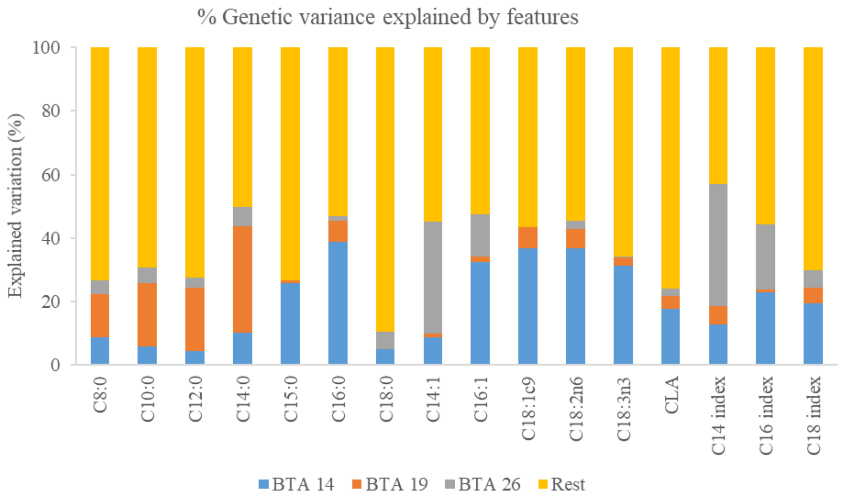

Fig 6.3. Proportion (%) of the genomic variance explained by genomic features fitted in the GFBLUP and the Rest of genome-wide variants

In general, the gain in prediction reliability from the GFBLUP model across the FAs showed patterns of correlation with the proportional of genomic variance explained by the fitted genomic features. Such trends were consistent across the validation populations for some traits, for instance C18:0, CLA and C14 index. For others, like C18:1c9, lack of improvement in reliability observed in the Chinese and Danish dataset were in contrast to moderate proportion of genetic variance explained by the features for C18:1c9. The relationship between the proportions of genetic variance explained by the features and gains in prediction reliability from the GFBLUP model (vis-à-vis traditional GBLUP) are presented in Figure 6.4.

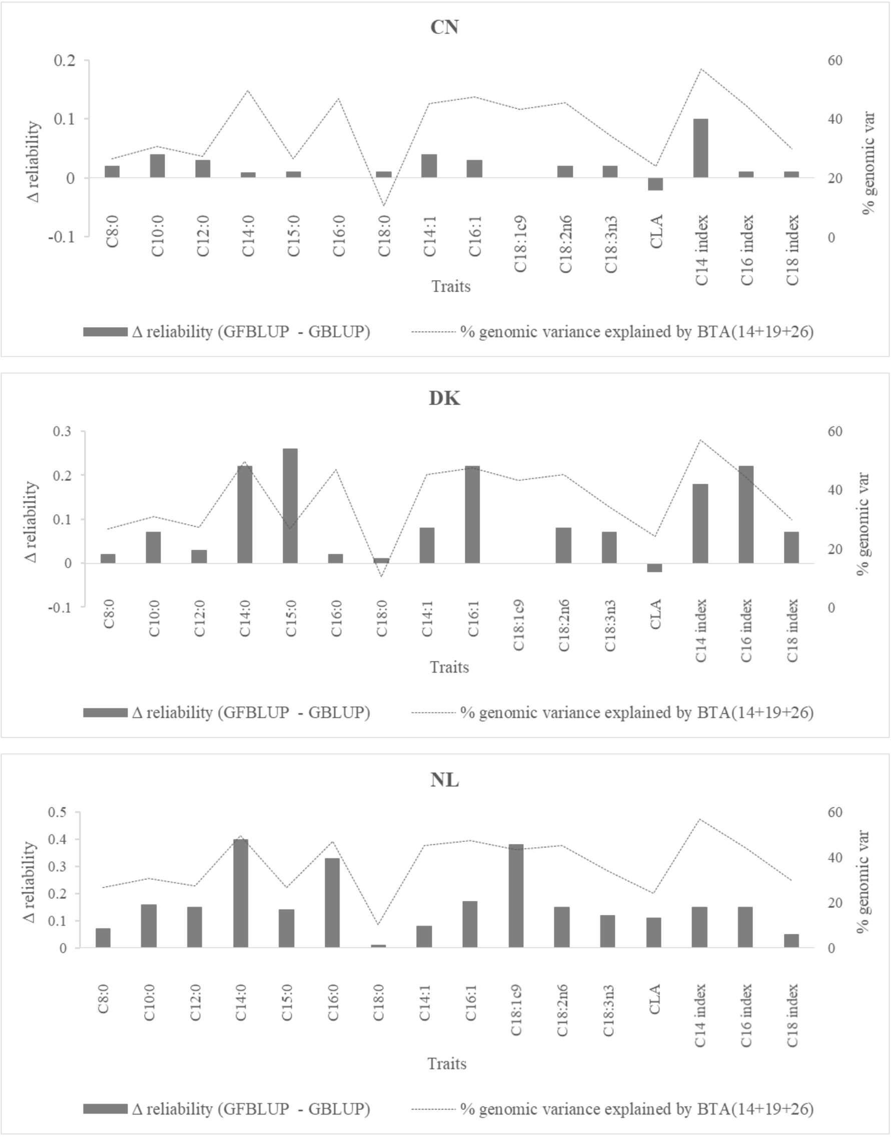

Fig 6.4. Relationship between the proportion of genetic variance explained by BTA 14, 19 and 26 summed together (dotted lines, with values on right x-axis) and the change in prediction reliability using GFBLUP model vis-à-vis traditional GBLUP (bars, with values on left x-axis) in the Chinese, Danish and Dutch validation population

6.4. Discussion

6.4.1. Combining reference populations

The population-specific genomic prediction for the FA traits generally resulted in low prediction reliabilities in the Chinese and Danish population. In the Dutch validation, moderate prediction reliabilities were achieved for some of the FA traits. In general, the low prediction reliabilities with population-specific reference reflect the well-established impact of a small reference population in genomic prediction reliability (Daetwyler et al., 2008; Goddard, 2009). However, prediction reliabilities were in general lower in the Chinese single-population prediction compared to the Danish reference despite similar reference sizes used. This might be partially explained by the lower heritability estimates in the Chinese dataset for most of the FAs compared to the other populations. The effect of small reference population size has a larger impact on traits of low heritability where a relatively large number of records will be required in the reference population to achieve high accuracies of GBV in unphenotyped animals (Hayes et al., 2010).

Combining the reference populations resulted in relatively sizable improvements in prediction reliability in the Danish and Dutch validation. Previous studies using simulation (e.g., de Roos et al., 2009) as well as real data (e.g. Lund et al., 2011) suggest advantages of pooling data from different populations for genomic prediction accuracy. However, such advantages might be affected by the genetic distance in the populations (Lund et al., 2014), marker density used, genetic architecture of the traits (e.g., Zhou et al., 2014) and inconsistencies in allele substitution effects (Wientjes et al., 2015). A gain in prediction reliability of up to 9 percentage points was obtained for some traits in the Chinese validation, whereas little or no increase in prediction reliability was observed for some of the traits from adding the Dutch and Danish Holstein reference populations. This is in contrast to our expectations given the genetic similarity and high consistency in genome-wide LD shown to exist between the populations combined (Zhou et al., 2013; Li et al., 2015; Gebreyesus et al., submitted). A possible cause for observed lack of benefits from combining the reference populations for the Chinese validation could be differences in SNP effects between the Chinese population on one hand and the Dutch and Danish populations on the other. In a joint GWAS using the same dataset, we previously (Gebreyesus et al., submitted) reported that SNP effects in the *DGAT1* and *SCD1* polymorphisms on BTA 14 and 26, respectively, are smaller in the Chinese dataset compared to the Dutch and Chinese population. *DGAT1* and *SCD1* polymorphisms underlie substantial proportion of the genetic variation in most milk

FA traits. In addition, we show that LD structures on BTA 14, 19 and 26 are consistent in the three populations. Therefore, such differences in SNP effects are less likely to be caused by incomplete LD. In our previous GWA study, we have suggested differences in feeding systems as the most likely source of significant differences in phenotypic means between the Chinese dataset on one hand and the Dutch and Danish datasets on the other. Fitting herd as fixed effect accounts for differences due to management systems. However, such differences can introduce feed by genotype interactions resulting in differences in SNP effect sizes. Inconsistencies in SNP effects have been shown to reduce the advantages from multi-population genomic prediction (Wientjes et al., 2015). SNP effects estimated in a multi-population reference, dominated by the Dutch Holstein (n=1566) compared to the Danish (n=614) and Chinese (n=586) populations, are used to predict breeding values for the 100 Chinese validation animals. The corrected phenotypes for these validation animals reflect the SNP effects in the Chinese population. With such differences in SNP effect estimates for the Chinese population compared to others, the correlations between breeding values estimated using the multi-population reference population and the corrected phenotype, i.e., prediction accuracy, can thus be expected to be low. This also leads us to expect lower contributions from the Chinese reference population for the gains in prediction reliability observed for the Dutch and Danish validations using the combined reference predictions.

6.4.2 Incorporating biological information in prediction models

In this study, we also implemented a GFBLUP model considering BTA 14, 19 and 26 as genomic features and allowing separate random genetic effect components for these regions. Implementation of the GFBLUP model generally resulted in further improvements in prediction ability in most of the traits across the validation populations. However, amount of improvement varied across the populations in which case the lowest improvements were generally observed for the Chinese validation.

Gain in prediction reliability under the GFBLUP model correlated with the proportion the genetic variance collectively explained by the features considered. Lowest increase in prediction reliability across the populations was computed for C18:0 which also had the lowest proportion of variance estimated by BTA14, 19 and 26. Similarly, the gain in prediction reliability from the GFBLUP model was highest for C14 index for which more than half of genomic variance was explained by BTA14, 19 and 26.

Different methods have been suggested to incorporate biological information in genomic prediction models. For instance, MacLeod et al. (2016) introduced Bayes RC model, an extension of the Bayes R (Erbe et al., 2016), to allow incorporating biological information by defining classes of SNP likely to be enriched for causal variants. Similarly, Brøndum et al., (2012) presented Bayesian prediction models based on genome position specific priors, whereas Gebreyesus et al. (2017) introduced hierarchical Bayesian models based on grouping of adjacent SNPs to exploit heterogeneous (co)variance patterns. However, most of these proposed models are implemented in Bayesian frameworks that are computationally demanding. Hence, applicability in routine evaluations is limited. GBLUP is straightforward to implement and estimated GBVs accuracies are similar to those estimated in BLUP approach (Hayes et al., 2009) since the method is equivalent to the BLUP approach used in traditional breeding programs (VanRaden et al, 2009). Thus, these simplicity and less computational burdens have made GBLUP a method of choice in routine genetic evaluations. Therefore, implementing biological information-augmented approaches in GBLUP models are closer to practical implementation in the breeding industry.

For some FA traits, improvement in prediction reliability from multi-population prediction using the GFBLUP model was substantial and consistent across the populations. These include C14 index and to some extent C14:1 and C16:1 FAs. Saturated FAs in milk, in particular C12:0, C14:0 and C16:0 are frequently connected to increases in serum cholesterol in human, which has been the background for development of the atherogenicity index: (C12+4·C14+C16)/(MUFA + n3 PUFA + n6 PUFA) (Ulbricht and Southgate, 1991). In this study, we show that with data pooling and incorporating biological information, it is possible to predict genetic merits for the composition of such FAs with reliabilities as high as 0.76 despite limited reference populations used. These findings highlight the possibility of implementing selective breeding to alter the bovine milk FA composition, despite challenges in large-scale phenotyping. Our findings also indicate that genomic prediction for scarcely recorded traits might benefit with international collaborations for data access across populations.

6.5 Conclusion

Accuracies of genomic prediction for the detailed milk fat composition traits using a multi-population reference and a model incorporating GWAS findings are compared with traditional GBLUP models in a single population scenario. Our results indicate that pooling multi-population data and implementation of prediction models augmented with biological information can enable prediction of genetic merits for the scarcely recorded bovine milk FA composition with reasonable accuracies. High prediction reliabilities were estimated for some of the FA traits using the multi-population reference and a GFBLUP model, including 0.76 for C14:0, 0.71 for C14 index and 0.64 for C18:1c9, indicating the possibility of altering milk fatty acid composition through selective breeding despite the current limitation in large-scale phenotyping.

Acknowledgements

GG was supported by grants from the European Commission, within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG", and the Center for Genomic Selection in Animals and Plants (GenSAP), funded by The Danish Council for Strategic Research. The study used data from the Dutch Milk Genomics Initiative (www.milkgenomics.nl), High Technology R & D Program of China, 2013AA102504 and Beijing Dairy Industry Innovation Team, BAIC06-2018, Danish-Swedish Milk Genomics Initiative (www.milkgenomics.dk) and the Milk Levy Fund (Denmark) projects: "Phenotypic and genetic markers for specific milk quality parameters" and "The importance of the metagenome for milk composition and quality".

References

Boichard D, Brochard M. (2012). New phenotypes for new breeding goals in dairy cattle. Animal. 6(4):544-50. doi: 10.1017/S1751731112000018.

Bouwman A.C., Bovenhuis H., Visker M.H., van Arendonk J.A. (2011). Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet. 11;12:43.

Bouwman A.C., Visker M.H., van Arendonk J.A. and Bovenhuis H. (2012). Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. BMC Genet. 29;13:93.

Bovenhuis H., Visker M.H.P.W., Poulsen N.A., Sehested J, van Valenberg H.J.F., van Arendonk J.A.M., Larsen L.B. and Buitenhuis A.J. (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(4):3113-3123. doi: 10.3168/jds.2015-10462

Brøndum R.F., Su G., Lund M.S., Bowman P.J., Goddard M.E., Hayes B.J. (2012). Genome position specific priors for genomic prediction. BMC Genomics. 10;13:543. doi: 10.1186/1471-2164-13-543.

Brøndum R.F., Su G., Janss L., Sahana G., Guldbrandtsen B., Boichard D., Lund M.S. (2015) Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 98(6):4107-16. doi: 10.3168/jds.2014-9005.

Buitenhuis B., Janss L.L., Poulsen N.A., Larsen L.B., Larsen M.K., Sørensen P. (2014). Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. BMC Genomics. 15;15:1112. doi: 10.1186/1471-2164-15-1112.

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

de Roos A.P., Hayes B.J., Goddard ME. (2009). Reliability of genomic predictions across multiple populations. Genetics. 183(4):1545-53. doi: 10.1534/genetics.109.104935.

Duchemin S.I., Visker M.H., Van Arendonk J.A., Bovenhuis H. 2014. A quantitative trait locus on Bos taurus autosome 17 explains a large proportion of the genetic variation in de novo synthesized milk fatty acids. J Dairy Sci. 97(11):7276-85. doi: 10.3168/jds.2014-8178.

Edwards S.M., Sørensen I.F., Sarup P., Mackay T.F., Sørensen P. (2016). Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. Genetics. 203(4):1871-83. doi: 10.1534/genetics.116.187161.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A., Goddard ME. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95, 4114 – 4129. (doi:10.3168/jds.2011-5019)

Gebreyesus G., Lund M.S., Janss L., Poulsen N.A., Larsen L.B., Bovenhuis H. and Buitenhuis A.J. (2016). Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J Dairy Sci. 99(4):2863-2866. doi: 10.3168/jds.2015-10501

Gebreyesus G., Lund M.S., Buitenhuis B., Bovenhuis H., Poulsen N.A., Janss L.G. (2017). Modeling heterogeneous (co)variances from adjacent-SNP groups improves genomic prediction for milk protein composition traits. Genet Sel Evol. 5;49(1):89. doi: 10.1186/s12711-017-0364-8

Goddard M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica136,245–257, http://dx.doi.org/10.1007/s10709-008-9308-0.

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J., Goddard M.E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6:e1001139

Hayes B.J., Bowman P.J., Chamberlain A.J., Goddard M.E. (2009a). Invited review: Genomic selection in dairy cattle: progress and challenges. (2009). J Dairy Sci. 92(2):433-43. doi: 10.3168/jds.2008-1646. Review. Erratum in: J Dairy Sci. 92(3):1313.

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J., Goddard M.E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 23;6(9):e1001139. doi: 10.1371/journal.pgen.1001139.

Krag K., Poulsen N.A., Larsen M.K., Larsen L.B., Janss L.L., Buitenhuis B. (2013). Genetic parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a Bayesian approach. BMC Genet. 11;14:79. doi:10.1186/1471-2156-14-79.

Li C., Sun D., Zhang S., Wang S., Wu X., Zhang Q., Liu L., Li Y., and Qiao L. (2014). Genome Wide Association Study Identifies 20 Novel Promising Genes Associated with Milk Fatty Acid Traits in Chinese Holstein. PLoS ONE. 9:e96186.

Li X., Buitenhuis A.J., Lund M.S., Li C., Sun D., Zhang Q., Poulsen N.A., Su G. (2015). Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. J. Dairy Sci. 98(11):8152-8163.

Lund M.S., Roos A.P., Vries A.G., Druet T., Ducrocq V., Fritz S., Guillaume F., Guldbrandtsen B., Liu Z., Reents R., Schrooten C., Seefried F., Su G. (2011). A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol. 12;43:43. doi:10.1186/1297-9686-43-43.

Lund M.S., Su G., Janss L., Guldbrandtsen B., and Brøndum R.F. (2014). Genomic evaluation of cattle in a multi-breed context. Liv. Sci. 166: 101-110.

MacLeod I.M., Bowman P.J., Vander Jagt C.J., Haile-Mariam M., Kemper K.E., Chamberlain A.J., Schrooten C., Hayes B.J., Goddard M.E. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 27;17:144. doi: 10.1186/s12864-016-2443-6.

Madsen P. and Jensen J. (2010). A User's Guide to DMU. Version 6, Release 5.0. University of Aarhus, Faculty Agricultural Sciences (DJF), Department of Genetics and Biotechnology, Research Centre Foulum, Tjele, Denmark.

Mele M., Conte G., Castiglioni B., Chessa S., Macciotta N.P., Serra A., Buccioni A., Pagnacco G., Secchiari P. (2007). Stearoyl-coenzyme A desaturase gene polymorphism and milk fatty acid composition in Italian Holsteins. J Dairy Sci. 90(9):4458-65.

Ntambi J.M. and Miyazaki M. (2003). Recent insights into stearoyl-CoA desaturase-1. Curr Opin Lipidol. 14:255–61

Pegolo S, Cecchinato A, Mele M, Conte G, Schiavon S, Bittante G. Effects of candidate gene polymorphisms on the detailed fatty acids profile determined by gas chromatography in bovine milk. J Dairy Sci. 2016 Jun;99(6):4558-4573. doi: 10.3168/jds.2015-10420.

Poulsen N.A., Gustavsson F., Glantz M., Paulsson M., Larsen L.B. and Larsen M.K. (2012). The influence of feed and herd on fatty acid composition in 3 dairy breeds (Danish Holstein, Danish Jersey, and Swedish Red). J Dairy Sci. 95(11):6362-71. doi: 10.3168/jds.2012-5820

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 81(3):559-75.

Sarup P., Jensen J., Ostersen T., Henryon M., Sørensen P. (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. BMC Genet. 2016;17:11

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

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

Sørensen P., Edwards S.M., Jensen P. (2014). Genomic feature models. In: 10th World Congress of Genetics Applied to Livestock Production, Vancouver

Spindel J.E., Begum H., Akdemir D., Collard B., Redoña E., Jannink J.L., McCouch S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity (Edinb). 116(4):395-408. doi: 10.1038/hdy.2015.113.

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

Ulbricht T.L.V, Southgate D.A.T. (1991). Coronary Heart-Disease - 7 Dietary Factors. Lancet 338(8773): 985-992.

van den Berg I., Bowman P.J., MacLeod I.M., Hayes B.J., Wang T., Bolormaa S., Goddard M.E. (2017). Multi-breed genomic prediction using Bayes R with sequence data and dropping variants with a small effect. Genet Sel Evol. 21;49(1):70. doi:10.1186/s12711-017-0347-9.

VanRaden P.M. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91(11):4414-23. doi: 10.3168/jds.2007-0980.

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

Wientjes Y.C.J., Veerkamp R.F., Bijma P., Bovenhuis H., Schrooten C., Calus M.P.L. (2015). Empirical and deterministic accuracies of across population genomic prediction. Genet Sel Evol. 47:5.

Zhou L., Ding X., Zhang Q., Wang Y., Lund M.S., Su G. (2013). Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. GSE. 45:7.

Zhou L., Lund M.S., Wang Y., Su G. (2014). Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. J Anim Breed Genet. 131(4):249-57. doi:10.1111/jbg.12089.

7

General discussion

7.1. Introduction

The protein and fatty acid (FA) composition of milk affect its suitability for further processing into high-value products, thus determining profitability of the dairy processing industry. With the ascendency of the fast food industry, that requires processed dairy products as inputs, the demands for products such as cheese and butter is increasing. In addition, there are growing consumer preferences towards certain protein and FA component of milk, and concerns to others, mainly on health grounds. Such economic and consumer pressures are pushing interests to alter the detailed milk protein and FA composition. Despite such interests and increasing availability of the tools for setting up and running selective breeding for new phenotypes, the detailed milk fat and protein composition are not yet included in the dairy cattle breeding goals anywhere. Implementation of selective breeding for new phenotypes requires, among others, definition of the breeding objectives, understanding the extent of genetic variation and the genetic correlations with other breeding goal traits as well as accurate prediction of breeding merits. Genomic prediction accuracy is generally limited by reference population size, which in turn requires availability of cheap and rapid phenotyping methods. The current standard analytical methods for detailed milk protein and FA composition require expensive equipment and time-consuming techniques; thus, large-scale phenotyping is a challenge. Rapid advances in quantitative methods are increasingly allowing high accuracies in parameter estimation and prediction of breeding merits. Therefore, development of efficient methods to allow accurate genetic analysis with limited information is one option to deal with traits that are expensive to measure. Alongside such efforts, exploring for alternative high-throughput and cheaper, yet reasonably accurate, phenotyping strategies should be a priority.

The main objective of this PhD thesis is to improve accuracy of genomic prediction and genetic parameter estimation as well as enhance our understandings on the genetic backgrounds of the detailed milk protein and FA composition traits. To this end, the study focused on investigating efficient quantitative approaches that allow such accurate genetic analyses in scarcely recorded traits mainly through exploiting additional information from other correlated traits, multiple cattle populations as well as the genetic architecture underlying the traits. In chapter 2, we investigate the advantages of multi-trait analyses for estimation of genetic parameters in the milk protein composition traits. In chapter 3, we present novel Bayesian single and multi-trait genomic prediction models assuming heterogeneous correlation structures between the detailed milk protein traits, measured at small-scale, and protein yield from a large population of proven bulls. Chapter 4 investigates advantages of pooling multi-population datasets for genome-wide association (GWA) in the milk FA traits. In chapter 5, genomic regions underlying milk FA traits are detected using multi-population GWA. Implementing post-GWA analyses, using multiple data sources for pathways and ontology information, detected novel regions are further characterized and promising candidate genes suggested. In chapter 6, we implement genomic prediction for the milk FA traits using multi-population reference and linear models that allow incorporation of GWA findings.

Here in this chapter, contributions of the different component studies to the knowledge base and to the broader context of implementing selective breeding for milk protein and FA composition traits are discussed. Moreover, other methodologies that could have been implemented are pointed out with detailed discussions on potential advantages and challenges. The chapter also highlights currently available information regarding genetic correlations of the detailed protein and FA traits with other breeding goal traits as well as status of the evolving strategy of IR prediction of detailed protein and FA composition of milk.

7.2. Multi-trait approaches

7.1.1. Multi-trait parameter estimation and genomic prediction

Traditionally, genetic parameter estimation and genomic prediction relied on models that fit only a single trait at a time. In contrast, simultaneous analysis of multiple traits might provide additional power to estimation of variance components and marker effects by utilizing information from between-trait genetic correlations. This is particularly of interest in the case of traits that are expensive to measure such as the detailed protein and FA composition of milk. Multi-trait analysis also allows utilization of information from relatives, which might not have phenotypes for the trait of interest but for other correlated trait(s). In both cases, the potential advantages of multi-trait analysis have been shown to largely depend on the genetic correlation between the traits (Calus et al., 2011; Jia and Jannick, 2012). In fact, information from correlated traits can be as useful as information from repeated records of the same trait to the extent of the magnitude of genetic correlation between traits. Substantial genetic correlation exists among some group of proteins and FA traits that are thought to arise into the milk via highly interrelated mechanisms. In chapter 2 (Gebreyesus et al., 2016), we show high genetic correlations among some of the milk proteins; for instance correlation of -0.70 between α_{S2}-CN and β-CN, and 0.75 between α-LA and β-LG. Similarly, strong genetic correlations are reported between milk FAs of similar origin. Stoop et al. (2008) reported high genetic correlations (0.76 – 0.96) among the *do novo* FAs (C6:0 – C14:0) and moderate to high genetic correlations (0.35 – 0.95) among the PUFA in the Dutch Holstein Friesian. Similarly, Krag et al. (2013) reported high genetic correlations between C6:0 and C8:0 (0.86), C8:0 and C10:0 (0.86) as well as C10:0 and C12:0 (0.91) in the Danish Holstein.

Furthermore, almost all published studies regarding the genetics of milk protein and FA composition looked into multiple protein or FA traits. Thus, while limited sample size is used across studies, in terms of number of individuals sampled, records are often available on multiple protein and FA traits. Therefore, given substantial genetic correlations between some groups within the milk protein and FA trait spectra, and that records on multiple protein and FA traits are usually available, multi-trait approach is set to be the most useful statistical tool in the genetic analysis of the scarcely recorded milk protein and fat composition traits. This is also reflected in our findings in chapter 2, where we show that, despite limited dataset (N=650), genetic parameters can be estimated with lower standard errors (0.08 - 0.10) with multi-trait analysis.

Multi-trait analysis also allows using information from relatives that have records for correlated traits that are recorded at large-scale. In chapter 3, we implement multi-trait genomic prediction for milk protein composition traits measured in small number of cows together with protein yield derived for large population of proven bulls in novel Bayesian models. Previous studies show that the gain in accuracy from multi-trait genomic prediction is limited when the genetic correlation between traits considered is weak (Calus et al., 2011; Jia and Jannick, 2012). However, the implicit assumptions behind traditional multi-trait models, on which these studies were based, is that the correlation structure between traits is homogenous throughout the genome and the genome-wide average estimates are used to declare correlations as weak or strong. In chapter 3, we show that while genome-wide average correlation between traits might be weak, there exist genomic regions where higher correlations than the genome-wide average are explained. Therefore, the limitation of current multi-trait models can be overcome by accounting for such heterogeneous correlation structures when genome-wide correlation is weak.

7.2.1. Multi-trait genome-wide association analyses

While utility of multi-trait analyses for GWA was not investigated in our studies, such approach might be very useful in the effort to unravel the genetic backgrounds of milk protein and FA traits. Genome-wide association testing has largely been limited to single-trait scenario to detect variants associated with a particular phenotype or disease risk. Simultaneous evaluation of correlated traits might improve statistical power of association tests. Genetic correlation between two traits might be attributed to QTLs affecting both traits i.e., in what is known as pleotropic effect. Genetic correlations might also be caused by different QTL, affecting each trait, but that are in LD with each other. Studies based on simulated and real data have shown that multi-traits GWA of correlated traits resulted in increased power, compared to single-trait analyses, to detect pleiotropic QTL affecting more than one trait (e.g. Korte et al., 2012) or even when the QTL effected only one of the traits (e.g. Galesloot et al. 2014).

Different models have been proposed for multi-variate GWA including the extensions of the linear regression (e.g. Wu and Pankow, 2018), linear mixed models (e.g. Korte et al., 2012; Zhou and Stephens, 2014), Bayesian multi-marker methods (e.g. Kemper et al., 2018) and meta-analysis of single-trait GWA summaries (e.g. Gai and Eskin, 2018; Turley et al., 2018). In livestock GWA studies, linear mixed models are commonly applied to estimate the fixed effects of variants one at a time with a relatedness matrix, constructed from pedigree information or marker data, fitted as random effect. Extension of this approach for multi-trait scenario requires modeling between-trait covariances as additional random effects. Such additional between-trait covariance structure in some cases allows estimation of fixed effects of the variants with improved reliability. Therefore, even if a QTL is affecting only one of the traits, multi-trait analysis allows more reliable estimation of its effects, compared to single-trait analysis due to additional information from between-trait covariances. Such additional information might highly benefit GWA for the milk protein and FA traits, where high genetic correlations exist among related group of traits. Additionally, multi-trait analysis also allows utilization of data on related traits measured at large-scale on other related individuals. In chapter 3, we present novel Bayesian models to utilize information from large-scale recorded traits for prediction of genetic merits in related small-scale recorded traits. Such models can be further extended for GWA test to benefit scarcely recorded traits using information from routinely recorded traits even while the genome-wide correlation is weak.

Apart from increased power in detection of trait-specific QTLs, multi-trait GWA also allow detection of pleotropic QTLs and untangling the pleotropic structure. The effect of pleotropic QTLs might be different in magnitude as well as direction between the correlated traits. Detecting such QTL and structure of the pleiotropic effect provides additional insight into the physiological pathways of interrelated spectrum of traits such as milk FA composition. While it is still possible to detect polymorphisms significantly associated with two different traits using single-trait analyses, resolving whether such associations with more than one trait are due to a QTL with pleiotropic effects or two linked QTLs, or a variant in LD with two different QTLs has proven challenging (Solovieff et al., 2013; Bolormaa et al., 2014).

Additionally, significant association established using single-trait GWA for a variant with more than one trait might be due to the “phenotypic dependency” of one of the traits on the other. The case of milk FA traits can be a very good example to elaborate such phenomenon. The milk FA composition traits are interrelated. Expressed as percentage of the total fat, the proportion of short-chain FAs, part of which are elongated into the intermediate chain FAs, affects the proportion the intermediate chain FAs. Likewise, the proportion of saturated FAs available for desaturation affects the proportion of corresponding unsaturated forms. The concentration of long chain FAs is also known to inhibit *de novo* synthesis. Hence, phenotypic dependency exists between the different group of FAs. A QTL affecting some of the short-chain FAs might not be a causal QTL for the other intermediate FAs but indirectly affect the intermediate chain FAs through its effect on the short-chain FAs. Similarly, QTLs affecting the long chain FAs might be indirectly affecting the short-chain FAs through the suppression effects of the long chain FAs on the de novo synthesis. Solovieff et al. (2013) describes such phenomenon as mediated-pleiotropy. If the aim of a GWA is to study genetic background of a specific mechanism, for instance desaturation or elongation process, such mediated-pleiotropic QTLs might be a noise rather than signal. Single trait GWA is expected to result in significant association of the variant with both the first trait it truly affects as well as the second trait that is affected by the first trait. However, such indirect associations might not be detected if the analysis is conditioned on the first trait (Hackinger and Zeggini, 2017).

In general, our studies as well as other previous studies indicate that multi-trait analyses can benefit genetic parameter estimation, genomic prediction and GWA studies of scarcely recorded traits. A limitation to the application of multi-trait analyses is the requirement to estimate the additional covariances, which is prone to inaccuracies unless sufficient data is available. Potential inaccuracies in estimation of the parameters are directly proportional to the number of traits included and inversely related to available dataset for each trait. Therefore, benefit from multi-trait analyses requires delicate balance between number of traits analyzed simultaneously and data available for each trait.

Additional challenge in implementing multi-trait analyses, especially for routine evaluations, is the longer run time and convergence problems when many traits are included. Convergence issues under the REML setting in multi-trait analysis is especially problematic when highly correlated traits are considered. In chapter 3, our comparison of the BayesAS and GBLUP models was limited to bivariate scenarios due to convergence issue with the GBLUP model when more than two traits, including the DRP, are simultaneously fitted. In our multi-trait BayesAS model presented in chapter 3, latent variables are used to model the correlation of SNP-group effects between traits. The BayesAS model taking whole genome as a segment might in fact be considered as an approximation of the GBLUP approach. Although results are not presented in chapter 3, for the sake of comparison with GBLUP models, it was possible to simultaneously run the ten milk protein traits without convergence issues in the BayesAS approach. Similarly, Bouwman et al. (2014) showed that it was possible to run 14 different milk FA traits, some of which are highly correlated, using a latent variable approach to fit (co)variance structures in a Bayesian mixed model setting.

7.3. Multi-population approaches

Genetic analysis for scarcely recorded traits might benefit from combining datasets available in different populations/breeds of livestock. Such approaches are shown to be beneficial when measurements from closely related breeds or populations are combined (e.g. VanRaden et al., 2012; Zhou et al., 2014). Apart from genetic distance between combined populations, marker density used also affects the potential benefit from multi-population analyses. Using samples from the Chinese, Danish and Dutch Holstein with HD genotypes, we show in chapter 4 the advantages of combining multi-population dataset for GWA in the milk FA traits. Similar findings are presented in chapter 6 where genomic prediction based on multi-population reference sets resulted in improved prediction reliabilities compared to within-population analyses. Amount of gains in prediction reliability was not however uniform across the validation populations considered. Specifically, gain in prediction reliability was slight or none for the Chinese dataset compared to the Danish and Dutch validation population. This is in agreement with our report in chapter 4 that estimated effect sizes for some of the major QTLs, including *DGAT1* and *SCD1*, were lower in the Chinese sample compared to the two European populations. Such differences in SNP effects are in contrary to our observation of similar LD patters across the genome and high correlations in allele frequencies, as has also been shown in previous reports (e.g. Zhou et al., 2013; Li et al., 2015). Our hypothesis is that such differences in allele substitution effects arise due to feed by genotype interaction as our Chinese dataset comes from a highly intensive production system characterized with concentrate feeding. Therefore, our study indicate that apart from genetic distance and marker density, differences in management, with consequences of genotype interactions, should be put to consideration when combining multi-population datasets. However, further sensitivity analysis we undertook excluding the Chinese dataset (not included in chapter 6) show slightly lower gains in the Danish and Dutch validation sets than when the Chinese dataset is included. This indicates that while differences in allele substitution effects might have limited the gain in reliability for the Chinese validation, information from the Chinese sample still contributed to the observed gain in prediction accuracy in the other populations. Such gains might increase with increased data availability within each population. We also show that data standardization and transformation might be helpful for combining datasets from different populations in the presence of heterogeneous residual variances or differences in standard deviations.

In general, our findings on multi-population approaches indicate large benefits for GWA and genomic prediction in the detailed milk composition traits. Several research groups have been working on limited samples of the detailed milk protein and FA traits quantified with the “golden standard” methods. Such data can be pooled for robust analyses compared to within-population studies that use numerically small data. International collaborations, allowing access to multi-population data for mega-analyses, should thus be the way forward in genetic analyses for the detailed milk protein and FA composition traits.

7.4. Including biological information in genomic prediction

The main principle in genomic prediction is that detection of underlying QTL(s) is no longer necessary to predict breeding merits of selection candidates due to use of all markers available genome-wide. This approach has proven to be the most successful of attempts to use genetic markers with quantitative models in the prediction of genetic merits and future phenotypes. Therefore, an important question arises regarding the importance of QTL detection in the context of genomic prediction. While genomic prediction has allowed high accuracy of selection of breeding candidates at early ages, prediction accuracy for scarcely recorded traits is still limited by the small reference population size. Requirement for large number of reference animals is also a major challenge for genomic prediction in breeds with small population size. Moreover, in principle, genomic prediction accuracies can go as high as 100% (1). This is currently not yet achieved even for larger breeds or routinely recorded traits. Therefore, the "black box" approach of genomic prediction has its limitations and incorporation of information on the underlying biology might further improve its accuracy.

7.4.1. Accounting for genetic architecture

Genetic architecture of the traits, especially the number of QTL affecting the trait and the distribution of their effects influence accuracy of genomic prediction (Daetwyler et al., 2010). Hence, prediction models taking into account the genetic architecture might result in improved prediction accuracies. While different genomic prediction models showed slight or no difference in prediction accuracies for most traits (e.g. Daetwyler et al., 2010; Clark et al., 2011), Bayesian models showed superiority to the BLUP-based approaches with larger differences in prediction accuracies for traits controlled by few QTL (e.g. Cole et al., 2009; Legarra et al., 2011). This mainly attributes to the differences between the Bayesian and GBLUP-based approaches in the assumption of distribution of QTL effects. The GBLUP-based models are built on the implicit assumption that quantitative traits are controlled my many QTLs each with small effects. While such assumptions might hold for majority of the traits, it is often violated when it comes to traits controlled by few QTL with large effects. On the other hand, the Bayesian approaches assume non-normal distribution of QTL effects and allow the variance of SNP effects to differ among loci (VanRaden, 2008). In doing so, the Bayesian methods reduce dilution of the effects of major QTL by other variants. However, a challenge related to such locus-specific approaches in genomic prediction is the requirement to estimate too many parameters, often with limited data, given common use of tens of thousands of markers. This is especially problematic for scarcely recorded traits, for which often numerically small dataset is available to estimate too many parameters.

In chapter 3, we show that clustering of SNPs, according to positions in the genome, can be implemented to reduce estimable parameters when data is limited. In our BayesAS models (Gebreyesus et al., 2017), different scenarios in clustering adjacent-SNPs were considered i.e., single SNP, 50, 100 or 200 SNPs, each chromosome or the whole genome. The assumption behind our SNP-grouping approach is that adjacent SNPs are very likely to be in linkage disequilibrium (LD) with the same QTL and thus explain similar (co)variances (Gebreyesus et al., 2017). Therefore grouping such SNPs can reduce dimensionality without causing dilution of effects. Such assumptions can only be met given optimum clustering such that group sizes, in terms of number of SNPs to cluster, are not too large to violate assumption of similar effect within cluster, yet large enough for meaningful reduction in parameters. Substantial improvements in prediction reliability were observed depending on the segment size considered and the genetic architecture of the traits. Highest improvements in prediction reliability were obtained when considering segment size of 100SNPs and for the traits where few regions explained large proportion of the genetic variance. Large gains in accuracies we show indicate that these models can be useful for genomic prediction in other scarcely recorded traits or species. A limitation in our study is that no objective SNP clustering criteria was suggested. LD structure is not the same across cattle breeds or livestock species. Therefore, optimum grouping could vary depending on the studied population. Similarly, number of SNPs with certain level of LD will also vary depending on the density of SNP array used and the optimum SNP-group size reported in our study using 50K SNP arrays might not work well for other SNP arrays with higher density. Further studies are required to develop well-defined criteria and strategy for optimum SNP clustering. A possible approach could be to establish threshold pair-wise LD levels between “reference” SNPs and nearby “candidate” SNPs to consider for clustering. Therefore, size of SNP-groups might differ across the genome but pair-wise LD among SNPs within each group can be comparable across groups. Such “reference” SNPs can be established on the basis of position, for instance the first SNP in the first chromosome being reference SNP1 and all subsequent SNPs with pair-wise LD value included in a cluster with SNP1. Reference SNPs can also be established based on other evidences, such as GWA findings.

7.4.2. Detection of QTLs for genomic prediction

Approaches to using priors that are more informative in genomic prediction models can be further extended to include information on the underlying biology such as causative QTL(s), genome annotation, gene, protein or metabolite expression as well as pathway information. Such approaches will be the major research topic in the area of genomic prediction for several reasons. Primarily, information is increasingly available from GWA, expression and pathway studies. Second, full sequence genotype data is becoming increasingly affordable which in principle should contain the causative QTLs; meaning prediction should no longer rely on LD. However, use of full sequence data in genomic prediction has not so far resulted in satisfactory gains in accuracy (e.g. Van Binsbergen et al. 2015; Calus et al., 2016). While full sequence data includes causative variants, millions of other variants are also included adding noise and making genomic prediction computationally cumbersome, especially with Bayesian models. Thus, biological information can be used to reduce dimensionality of genotype data for more accurate and computationally efficient prediction. Incorporation of biological information can be implemented in different ways. Most studies have used either filtering or clustering approaches based on biological information. Filtering approaches use biological information to select fewer variants for genomic prediction. An example of filtering approach could be to undertake GWA and select most associated variants for genomic prediction. However, variant selection through GWA comes with its own challenges in terms of precision, due to LD, and is prone to bias. Genomic prediction using only variants selected from GWA on full sequence data was previously shown to result in decreased prediction accuracy and increased bias (e.g. Veerkamp et al., 2016). Therefore, bias and precision of detection are major challenges in filtering variants for genomic prediction based on GWA results. In contrast, clustering approaches use biological information to group all available variants into different classes, for instance group of significant and non-significant SNPs from GWA or SNPs assigned to different pathways (e.g. MacLeod et al., 2016; Sarup et al., 2016). In chapter 6, we implement similar approach where linear prediction models (GFBLUP) were fitted with additional random genetic effect components for GWA detected genomic regions. Following the findings in chapter 5 as well as previous studies, BTA 14, 19 and 26 were considered as genomic features with established biological links to the FA traits. The GFBLUP model accordingly included separate genetic effects for the genomic features considered based on additional relationship matrices built using only variants in each feature.

However, information provided by GWA findings might not be strictly considered “biological” in the sense that most GWA results point at genomic positions for which the biological relation with the trait of interest remain unclear. Especially in cattle breeds, long-range LD often results in detection of broader regions containing multiple genes. Therefore, additional evidences are needed to establish biological links between detected regions and traits. Post-GWA analyses with multiple data sources might help refine GWA findings. Functional annotations making use of publicly accessible pathway and ontology databases is increasingly common in post-GWA analysis. In chapter 5, we use pathway information, including the KEGG database, as well as ontology resources, including the gene ontology (GO) terms and the mammalian phenotype ontology database (Smith et al., 2005) to disentangle highly likely genes from many positional candidates. However, available ontology or pathway databases are not complete list of genes with established links to the physiological pathways. For instance for milk protein and FA traits, some major genes such as the *MGST1* are not assigned to any of the pathways linked to protein and FA synthesis.

Tissue-specific information, such as expression of genes or RNAs are increasingly available (e.g. Moore et al., 2016) and might provide valuable additional evidences to prioritize among positional candidate genes. Studies are increasingly using information on differential expression of genes across lactation in the mammary gland to determine causal status of detected genes for milk production traits in general (e.g. Ron et al., 2007) and milk FAs in particular (e.g., Duchemine et al., 2014; Knutsen et al., 2018; Pegolo et al., 2017). In GWA studies, it is also common to detect significant associations with variants located in genomic regions where there are no known genes, i.e., non-coding regions of the genome. It is highly likely that such regions might be involved in regulatory roles affecting expression of other regions linked to the trait of interest, also known as expression QTL (eQTL). Early attempts integrating GWA with eQTL analysis for milk FA composition traits have shown promising results (e.g. Littlejohn et al., 2016). Further eQTL studies might reveal valuable insight into the genetic background of the milk composition traits. However, expression studies in relation to milk production traits in general and the FA traits in particular exclusively focused on the mammary gland. Control of traits such as the milk FAs involve complex and interconnected pathways involving different tissues and systems including digestion, absorption, lipid transport and storage. Hence, genes involved in FA synthesis via such pathways might not be expressed in the mammary gland. In such cases, differential expression of genes in the blood across lactation might provide additional valuable insight into the genetic background of milk FA traits. Only few studies used gene expression data from multiple tissues to assess expression status of genes implicated in lipid and FA mechanisms. These studies indicate that some of the genes implicated in lipid and FA metabolism that are expressed in the mammary gland are also expressed in the blood, among few other tissues (e.g. Viturro et al., 2006). Blood is also the easiest to sample tissue and mostly non-evasive techniques are available. Thus, it provides an additional opportunity to generate large dataset for more powerful analysis in gene expression and eQTL studies to unravel genetic background of complex traits.

7.5. Alternative large-scale phenotyping strategies

Studies in the PhD thesis focused solely on the milk protein and FAs traits quantified using the “golden standard” methods. While investigating efficient quantitative models for such expensive-to-measure, yet highly accurate, phenotypes is imperative, it is equally important to simultaneously explore alternative cheaper and accurate methods for large-scale phenotyping. In this regard, infrared (IR) spectroscopy prediction of detailed phenotypes is becoming one of the major topics in dairy science (De Marchi et al., 2014). Several studies have been investigating the possibility of IR prediction of the detailed milk protein (e.g. De Marchi et al., 2009; Bonfatti et al., 2011; Rutten et al., 2011) and FA traits (e.g. Soyeurt et al., 2006, 2011; Ferrand et al., 2011; De Marchi et al., 2011; Rodriguez et al., 2014; Fleming et al., 2017). Such IR predicted detailed phenotypes are also used to study genetic parameters for the milk protein (e.g. Sanchez et al., 2017a) and FA traits (e.g. Soyeurt et al., 2007; Bastin et al., 2011; Hein et al., 2018). Studies also used IR predicted protein and FA phenotypes to study associations with other milk production and heath traits (e.g. Bobbo et al., 2017; Fleming et al., 2018), to predict other difficult-to-measure traits such as methane (e.g. Shetty et al., 2017) and as indicators of reproductive stages such as pregnancy status (e.g. Toledo-Alvarado et al., 2018). Recent studies have also used IR predicted phenotypes for GWA tests in milk protein (e.g. Sanchez et al., 2017b) and FA traits (e.g. Olsen et al., 2017; Knutsen et al., 2018). However, the IR prediction of detailed milk protein and FA traits is still an evolving area. Thus, suitability of the IR predicted phenotypes for such applications as GWA studies and genomic prediction requires critical re-examination before normalization.

IR prediction of phenotypes rely on infrared spectrum caused by the absorptions of electromagnetic radiation at frequencies correlated to the vibrations of specific chemical bonds of molecule in irradiate sample (Coates, 2000) therefore the spectrum representing these absorptions at different wavenumbers for a specific chemical composition (Smith, 1996; Soyeurt et al., 2006). IR prediction (calibration) models are then developed to process the spectra data. The Partial least squares (PLS) method have mostly been used for such calibration (Soyeurt et al., 2006; De Marchi et al., 2011) while Bayesian models have also been suggested (e.g. Ferragina et al., 2015).

Compared to the FA traits, limited studies developed IR prediction equations for the detailed protein composition traits. In what can be seen as an initial attempt, De Marchi et al. (2009) reported validation R^2 values ranging between 0.29 and 0.58 in IR prediction of α-LA, β-LG, α_{S1}-CN, whey protein and total casein. Rutten et al. (2011) reported even lower IR prediction accuracies with validation R^2 values ranging between 0.18 (α_{S1}-CN) and 0.56 (β-LG) for the detailed milk protein composition. Similarly, Bonfatti et al. (2011) reported low to moderate IR prediction accuracies for the milk proteins with validation R^2 values ranging from 0.09 to 0.66. Sanchez et al., (2017a) reported the highest IR prediction accuracies for the detailed milk protein traits so far with validation R^2 values ranging between 0.59 (α-LA) and 0.92 (β-CN).

Relatively, higher prediction accuracies have been reported for the milk FA traits compared to the detailed protein traits. In general, better IR prediction accuracies are reported for FAs found in high concentrations compared to the minor FAs (Soyeurt et al., 2006; Rutten et al., 2009; De Marchi et al., 2011). Milk FA measured with GC and phenotypes predicted with IR are not identical (Rodriguez et al., 2014), but are expected to be interchangeable in the genetic sense, i.e., highly genetically correlated. Not many studies are available comparing genetic parameters between the IR predicted and GC quantified FA traits using the same individual samples, data size or genetic models. Poulsen et al. (2014) showed that heritability estimates were comparable for some FA traits between IR predicted and GC measured phenotypes, while larger differences were shown for others. Interestingly, Poulsen et al. (2014) reported some of the lowest genetic correlations and larger differences in genetic parameter estimates for some FAs found in high concentrations in milk including C16:0. This is contrary to some studies reporting high IR prediction accuracies for such FAs (e.g. Rutten et al., 2009).

Rutten et al. (2010) used relatively large number of calibration samples (n= 1917) and reported estimates of genetic correlations between GC-based and IR predicted FA phenotypes ranging from 0.77 to 0.99. In addition, using different subsets of the GC-based samples to calibrate IR prediction, Rutten et al. (2010) showed that estimates of genetic correlation between GC and IR predicted phenotypes is affected by the number of calibration samples used. Number of calibration samples used in IR prediction of FAs is known to strongly affect accuracy of the prediction (Rutten et al., 2009). Thus with larger calibration samples from GC, it might be possible to predict FA phenotypes more accurately and attain high genetic correlation with GC-based phenotypes. This elevates the prospect of using IR predicted FA phenotypes from prediction models based on large number of calibration samples for routine genetic evaluation. However, Eskildsen et al. (2014) showed that IR prediction of FAs is based on covariation of the FAs with the total fat content rather than directly with absorption bands associated to the individual FA; meaning that the predictions will be inaccurate if the covariance structures in the initial calibration are not conserved in future samples. IR prediction of FAs is the combined effect of predicting fat content and fat composition and it is performed on milk samples, while GC measurement is performed on fat extracted from milk (De Marchi et al., 2011), thus the variation in fat percentages affects the relationship between GC measurement and IR prediction of FAs (Soyeurt et al., 2006). The consequence of these for routine measurement is that routine re-calibration, with large number of calibration samples, is required to obtain high prediction accuracies and genetic correlations with GC-based FA phenotypes.

For the milk protein traits, Rutten et al. (2011) reported moderate genetic correlation between capillary zone electrophoresis determined phenotypes with the IR predicted protein fractions including β-CN (0.62), for α_{S1}-CN (0.66) and α-LA (0.69).

Recently published GWA studies used IR predicted milk protein (e.g. Sanchez et al., 2017b) and FA phenotypes (e.g. Olsen et al., 2017; Knutsen et al., 2018). The study of Sanchez et al. (2017b) was based on more than 800,000 milk samples from 156,660 cows and led to the re-detection of the known major regions for the milk protein composition traits. In contrast, GWA studies for IR predicted FA traits showed lack of detections on some of the well-established genomic regions associated with most FA traits. For instance, the studies by Olsen et al. (2017) and Knutsen et al. (2018) reported that no significant association was detected between the *DGAT1* region and any of the FTIR predicted FA phenotypes using milk samples from 878 samples in the Nordic Red cattle. Knutsen et al. (2018) suggested that lack of detection could be due the A variant of the DGAT1 K232A polymorphism not segregating in the Nordic Red cattle. However, significant association was also not established between the *SCD1* region and any of the predicted FA traits in the studies of Olsen et al. (2017) and Knutsen et al. (2018) despite that the *SCD1* allele is known to segregate in the Nordic Red cattle (Knutsen et al., 2018). Similarly, GWA study on the milk IR wavenumbers by Wang et al. (2018) using samples from the Dutch Holstein reported that no significant association was detected between the *SCD1* region and any of the wave numbers. The *DGAT1* and *SCD1* are major genes for milk FA traits explaining up to 50% of the genetic variation for some FAs (e.g. Bouwman et al., 2011, 2012). Hence, association tests can detect such regions with relatively high power. Studies of Buitenhuis et al. (2014) and Li et al. (2015) have reported significant associations on the *DGAT1* and *SCD1* with most FAs using sample size smaller than the studies of Olsen et al. (2017) and Knutsen et al. (2018). Therefore, lack of detection of major regions with GWA based on IR phenotypes cast doubt over the reliability of such phenotypes to study the genetic backgrounds of the "standard" traits. However, re-detection of the major regions for milk proteins reported by Sanchez et al. (2017b), despite generally lower IR prediction accuracies for the milk protein traits compared to the FAs, might indicate that it is possible to approximate reference phenotypes with the use of hundreds of thousands of samples.

Additional challenge to prospect of using IR predicted phenotypes for genetic evaluation is the apparent difference in prediction accuracy depending on the expression of the traits. Studies have shown that it is even less accurate to predict the detailed milk proteins/FAs expressed on a protein/fat basis compared to per milk basis (De Marchi et al., 2014). Most of the studies on IR predictions of the milk proteins and FAs used traits expressed per unit of milk with few exceptions using proteins traits expressed per total protein basis (e.g. Bonfatti et al., 2011) and the FAs on a fat basis (e.g. Soyeurt et al., 2006; Rutten et al., 2009; Hein et al., 2018). Soyeurt et al. (2006) and Rutten et al. (2009) reported that accuracy of the IR prediction models were lower for FAs expressed per fat compared to per unit of milk. This will be problematic if the breeding objective is to change the protein and fat composition instead of increasing/decreasing yield of specific milk proteins or FAs. Expression of the traits have also been shown to impact genetic parameter estimates including correlation with other traits. Such differences in genetic parameters in connection to expression of the traits are also reported for the reference method quantified traits (e.g. Fang et al., 2018). Therefore, regardless of the phenotyping method, the choice of trait definition requires further comprehensive investigations.

In general, IR predicted phenotypes are not identical to the phenotype based on the "golden standard" and therefore, GWA results should be interpreted with great care. Compared to milk protein traits, relatively higher IR prediction accuracies are possible for the milk FA phenotypes. However, prediction accuracy is still lower for FAs expressed on a fat basis. Thus, lower IR prediction accuracies will remain a challenge in prospects of using IR phenotypes for genetic evaluations if the breeding objective is to change the FA composition.

7.6. Genetic correlations with other breeding goal traits

Considerations to include new phenotypes requires understanding the magnitude and direction of genetic correlations with other breeding goal traits. Not many studies are available reporting genetic correlations of the detailed protein and FA composition of milk with other milk production and fitness traits. In the case of milk protein composition, some studies have reported genetic correlations with protein percentage and protein and milk yield. In chapter 2 (Gebreyesus et al., 2016), we report generally low estimates of genetic correlations (-0.03 – 0.38) between the detailed protein composition and the total milk protein percentage in agreement with other studies (e.g. Schopen et al., 2009). Likewise, we report in chapter 3 (Gebreyesus et al., 2017), overall weak genetic correlations between milk protein yield on one hand and the detailed milk protein composition traits and protein percentage on the other. Previous studies have also reported low genetic correlations between milk protein yield and the detailed milk protein composition traits (e.g. Schopen et al., 2009) as well as protein percentage (e.g. Chauhan and Hayes, 1991; Roman and Wilcox, 2000; Shahbazkia et al., 2010). Similarly, Bobe et al. (2007) showed that differences in genetic merit of cows for milk production was not correlated with differences in milk protein composition indicating low genetic relationships between milk yield and the detailed protein composition. Fang et al. (2018) also reported weak to moderate genetic correlations between the concentration of α_{S1}-CN and α_{S2}-CN phosphorylation isoforms and milk yield. These findings suggest that selection for the detailed milk protein traits will not substantial impact over the genetic progress in milk yield and protein percentage. This also indicates that selective breeding focusing mainly on milk yield in the past decades might not have any substantial effect on the detailed milk protein composition traits. However, studies show that estimates of genetic correlations between the milk protein traits and other milk production traits depend on how the traits are expressed. For instance, Fang et al. (2018) reported moderate to high genetic correlations between yields of individual α_{S1}-CN and α_{S2}-CN phosphorylation isoforms and the total protein and milk yield, while the corresponding estimates were low to moderate when the phosphorylation isoforms were expressed per protein basis (wt/wt). Therefore, decisions on expression of the traits should take into account the genetic correlations with other breeding goal traits.

For the milk FA traits, reported genetic correlations with milk production traits varied according to the studied breeds/populations, methods used to quantify the FA traits i.e. GC or IR predicted, and the expression of traits (per milk or fat basis). Table 7.1 summarizes reports of different studies regarding genetic correlations of both GC quantified as well as IR predicted FA phenotypes with milk production traits. Studies using both the GC as well as IR phenotypes show strong positive correlations between the saturated group of FAs and fat percentage. This suggest that selection based on fat percentage can increase the saturated FAs in milk.

Regarding the genetic correlations between the FA traits and milk yield, reports from the IR based studies indicate negative genetic correlations, for all FAs/groups, ranging from low (-0.20) to moderate (-0.62). These reports are not however supported by the studies based on GC measured phenotypes which reported positive genetic correlations between most of the FA traits and milk yield ranging between low (0.01 for C6:0) to high (0.77 for C18:2n6) estimates, except C12:0, C16:0, C18:0 and desaturation indexes, for which low to moderate negative correlations were reported. All of the IR based studies presented in Table 7.1 used FA traits expressed per unit of milk yield except the study of Hein et al. (2018), while traits were expressed as percentage of the total fat in all the GC-based studies.

Table 7.1. Genetic correlations of the milk FA traits with the milk production traits based on GC as well as IR predicted phenotypes

FAs/ groups	Genetic correlation with GC or IR quantified FAs [and the reporting studies]					
	GC			IR		
	Milk (kg)	Fat (kg)	Fat (%)	Milk (kg)	Fat (kg)	Fat (%)
SFA	−0.16[1]	0.56[1]	0.94[1]	−0.26[5]	0.29[6]	0.97[5]
				−0.40[6]	0.53[7]	0.97[6]
				−0.38[7]	0.50[8]	0.98[7]
				−0.62[8]	0.34[9]	0.99[8]
				−0.36[10]		0.99[10]
MUFA	0.15[1]	−0.54[1]	−0.89[1]	−0.21[5]	0.15[6]	0.74[5]
			0.01[2]	−0.39[6]	0.31[7]	0.76[6]
				−0.45[7]	−0.33[9]	0.84[7]
				−0.32[10]		0.79[10]
PUFA	0.19[1]	−0.41[1]	−0.80[1]	−0.39[6]	0.13[6]	0.72[6]
				−0.40[7]	0.24[7]	0.67[7]
				−0.38[10]	−0.26[9]	0.52[10]
C4:0	0.05[1]	0.24[1]	0.03[1]	−0.27[6]	0.34[6]	0.81[6]
	0.09[3]	0.30[3]	0.16[3]			
C6:0	0.01[3]	0.58[3]	0.46[3]	−0.30[6]	0.34[6]	0.87[6]
C8:0	0.03[3]	0.45[3]	0.34[3]	−0.31[6]	0.31[6]	0.85[6]
C10:0	0.10[3]	0.24[3]	0.09[3]	−0.32[6]	0.27[6]	0.81[6]
C12:0	−0.26[1]	0.04[1]	0.43[1]	−0.36[5]	0.24[6]	0.91[5]
	0.09[3]	0.15[3]	0.00[3]	−0.34[6]		0.81[6]
C14:0	−0.09[1]	−0.09[1]	0.04[1]	−0.29[5]	0.28[6]	0.80[5]
	0.30[3]	−0.11[3]	−0.40[2]	−0.37[6]	0.06[9]	0.89[6]
			−0.43[3]			

Table 7.1 *Continued*

	GC			IR		
FA	Milk (kg)	Fat (kg)	Fat (%)	Milk (kg)	Fat (kg)	Fat (%)
C16:0	0.10[1]	0.57[1]	0.72[1]	−0.25[5]	0.28[6]	0.95[5]
	−0.50[3]	0.18[3]	0.74[2]	−0.37[6]	0.17[9]	0.92[6]
			0.65[3]	−0.35[10]		0.98[10]
C18:0	−0.20[1]	−0.28[1]	−0.10[1]	−0.28[5]	0.17[6]	0.97[5]
	0.15[3]	0.18[3]	0.28[2]	−0.37[6]	−0.14[9]	0.75[6]
			0.01[3]	−0.28[10]		0.86[10]
C14:1	0.05[1]	−0.01[1]	−0.05[1]	-	-	-
			0.10[2]			
C16:1	0.09[1]	0.31[1]	0.24[1]	-	-	-
			0.34[2]			
C18:1c9	0.13[1]	−0.42[1]	−0.85[1]	−0.39[5]	0.13[6]	0.75[5]
	0.32[3]	−0.36[3]	0.02[2]	−0.35[6]	−0.26[9]	0.68[6]
			−0.63[3]	−0.29[10]		0.80[10]
C18:2n6	0.24[1]	−0.25[1]	−0.69[1]	−0.28[5]	-	0.75[5]
	0.77[3]	0.04[3]	−0.70[3]			
C18:3n3	0.15[1]	−0.25[1]	−0.55[1]	-	-	-
	0.53[3]	−0.28[3]	−0.75[3]			
CLA	0.35[1]	−0.06[1]	−0.68[1]	-	-	-
	0.33[3]	−0.30[3]	−0.55[2]			
			−0.58[3]			
C14index	-0.39[4]	−0.13[4]	0.34[2]	-	-	-
			0.31[4]			
C16index	-0.37[4]	−0.21[4]	0.40[2]	-	-	-
			0.17[4]			
C18index	0.01[4]	−0.36[4]	−0.35[2]	-	-	-
			−0.35[4]			

[1] Bilal et al. (2014). Canadian Holstein;
[2] Mele et al. (2009). Italian Holstein;
[3] Stoop et al. (2008). Dutch Holstein;
[4] Schennink et al. (2008). Dutch Holstein;
[5] Soyeurt et al. (2007). Multi-Breed (Belgian);
[6] Bastin et al. (2011). Belgian (Walloon) Holstein;
[7] Tullo et al. (2014). Italian Holstein;
[8] Fleming et al. (2018). Canadian Holstein;
[9] Hein et al. (2018). Danish Holstein;
[10] Petrini et al. (2016). US Holstein

In general, available literature regarding the genetic correlations of the protein and FA composition of milk with other dairy production, fertility, fitness and conformation traits is very limited. In this regard, there is critical need for future studies estimating such correlations in different dairy breeds.

7.7. Conclusions

In this study, different quantitative approaches were explored to improve accuracy of parameter estimation and genomic prediction for the detailed milk protein and fatty acid composition. We show that information from correlated traits, related populations and the underlying biology can largely benefit genomic prediction for scarcely recorded traits, but efficient models are required to fully exploit the advantages. Limitation of traditional multi-trait models for traits with weak genome-wide correlations can be overcome by disentangling heterogeneous correlation structure and using information from regions where there is higher genetic correlation.

Our studies also show that combining multi-population dataset is advantageous for GWA and genomic prediction in the milk FA traits. International collaborations allowing access to multi-population data can thus benefit the study on genetics of milk protein and FA traits. Information is limited regarding the genetic correlation of the detailed milk protein and FA traits with other traits in the dairy cattle breeding goals. Future studies are thus required regarding genetic correlations of the detailed protein and FA composition of milk with other breeding goal traits or indexes (such as the Nordic total merit).

References

Bastin C., Gengler N., Soyeurt H. (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(8):4152-63.

Bilal G., Cue R.I., Mustafa A.F. and Hayes J.F. (2014). Short communication: Genetic parameters of individual fatty acids in milk of Canadian Holsteins. J Dairy Sci. 97(2):1150-6. doi: 10.3168/jds.2012-6508

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

Bobbo T., Ruegg P.L., Stocco G., Fiore E., Gianesella M., Morgante M., Pasotto D., Bittante G., Cecchinato A. (2017). Associations between pathogen-specific cases of subclinical mastitis and milk yield, quality, protein composition, and cheese-making traits in dairy cows. J Dairy Sci. 2017 Jun;100(6):4868-4883.

Bolormaa S., Pryce J.E., Reverter A., Zhang Y., Barendse W., Kemper K., Tier B., Savin K., Hayes B.J., Goddard M.E. (2014). A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLoS Genet. 27;10(3):e1004198.

Bonfatti V., Di Martino G., Carnier P. (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., Bovenhuis H., Visker M.H., van Arendonk J.A. (2011). Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet. 11;12:43.

Bouwman A.C., Visker M.H., van Arendonk J.A., Bovenhuis H.(2012). Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. BMC Genet. 29;13:93. doi: 10.1186/1471-2156-13-93.

Bouwman A.C., Valente B.D., Janss L.L., Bovenhuis H., Rosa G.J. (2014). Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. Genet Sel Evol. 17;46:2. doi: 10.1186/1297-9686-46-2.

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

Calus M.P., Veerkamp R.F. (2011). Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol. 43:26.

Calus M.P.L., Bouwman A.C., Schrooten C., Veerkamp R.F. (2016). Efficient genomic prediction based on whole genome sequence data using split-and merge Bayesian variable selection. Genet Sel Evol. 2016;48:49.

Chauhan V.P., Hayes J.F. (1991). Genetic parameters for first milk production and composition traits for Holsteins using multivariate restricted maximum likelihood. J Dairy Sci. 74:603-10.

Clark S.A., Hickey J.M., van der Werf J.H. (2011). Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol. 17;43:18.

Coates J., 2000. Interpretation of infrared spectra, a practical approach. Pages 10815–10837 in Encyclopedia of Analytical Chemistry. R. A. Meyers, ed. John Wiley & Sons, New York, NY.

Cole J.B., VanRaden P.M., O'Connell J.R., Van Tassell C.P., Sonstegard T.S., Schnabel R.D., Taylor J.F., Wiggans G.R. (2009). Distribution and location of genetic effects for dairy traits. J Dairy Sci. 92(6):2931-46.

Daetwyler H.D., Pong-Wong R., Villanueva B., Woolliams J.A. (2010). The impact of genetic architecture on genome-wide evaluation methods. Genetics. 185(3):1021-31. doi: 10.1534/genetics.110.116855.

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

De Marchi M., Penasa M., Cecchinato A., Mele M., Secchiari P., Bittante G. (2011). Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal. 2011 Aug;5(10):1653-8.

De Marchi M., Toffanin V., Cassandro M., Penasa M. (2014). Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J Dairy Sci. 97(3):1171-86. doi: 10.3168/jds.2013-6799.

Duchemin S.I., Visker M.H., Van Arendonk J.A., Bovenhuis H. (2014). A quantitative trait locus on Bos taurus autosome 17 explains a large proportion of the genetic variation in de novo synthesized milk fatty acids. J Dairy Sci. 97(11):7276-85.

Eskildsen C.E., Rasmussen M.A., Engelsen S.B., Larsen L.B., Poulsen N.A., Skov T. (2014). Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: understanding predictions of highly collinear reference variables. 2014. JDS 97(12):7940-51.

Fang Z.H., Bovenhuis H., van Valenberg H.J.F., Martin P., Huppertz T., Visker M.H.P.W. (2018). Genetic parameters for α(S1)-casein and α(S2)-casein phosphorylation isoforms in Dutch Holstein Friesian. J Dairy Sci. 101(2):1281-1291. doi:10.3168/jds.2017-13623.

Ferragina A., de los Campos G., Vazquez A.I., Cecchinato A., Bittante G. (2015). Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. J Dairy Sci. 2015 Nov;98(11):8133-51.

Ferrand M., Huquet B., Barbey S., Barillet F., Faucon F., Larroque H., Leray O., Trommenschlager J.M., Brochard M. (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.

Fleming A., Schenkel F.S., Chen J., Malchiodi F., Bonfatti V., Ali R.A., Mallard B., Corredig M., Miglior F. (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(6):5073-5081.

Fleming A., Schenkel F.S., Malchiodi F., Ali R.A., Mallard B., Sargolzaei M., Jamrozik J., Johnston J., Miglior F. (2018). Genetic correlations of mid-infrared-predicted milk fatty acid groups with milk production traits. J Dairy Sci. 101(5):4295-4306. doi: 10.3168/jds.2017-14089.

Gai L., Eskin E. (2018). Finding associated variants in genome-wide association studies on multiple traits. Bioinformatics. 1;34(13):i467-i474

Galesloot T.E., van Steen K., Kiemeney L.A., Janss L.L., Vermeulen S.H. (2014). A comparison of multivariate genome-wide association methods. PLoS One. 24;9(4):e95923. doi: 10.1371/journal.pone.0095923.

Gebreyesus G., Lund M.S., Janss L., Poulsen N.A., Larsen L.B., Bovenhuis H., Buitenhuis A.J. (2016). Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J Dairy Sci. 99(4):2863-2866. doi: 10.3168/jds.2015-10501.

Gebreyesus G., Lund M.S., Buitenhuis B., Bovenhuis H., Poulsen N.A., Janss L.G. (2017). Modeling heterogeneous (co)variances from adjacent-SNP groups improves genomic prediction for milk protein composition traits. Genet Sel Evol. 5;49(1):89. doi: 10.1186/s12711-017-0364-8

Hackinger S., Zeggini E. (2017). Statistical methods to detect pleiotropy in human complex traits. Open Biol. 7(11). pii: 170125. doi: 10.1098/rsob.170125.

Hein L., Sørensen L.P., Kargo M., Buitenhuis A.J. (2018). Genetic analysis of predicted fatty acid profiles of milk from Danish Holstein and Danish Jersey cattle populations. J Dairy Sci. 101(3):2148-2157.

Jia Y., Jannink J.L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 192(4):1513-22.

Kemper K.E., Bowman P.J., Hayes B.J., Visscher P.M., Goddard M.E. (2018). A multi-trait Bayesian method for mapping QTL and genomic prediction. Genet Sel Evol. 24;50(1):10. doi: 10.1186/s12711-018-0377-y.

Knutsen T.M., Olsen H.G., Tafintseva V., Svendsen M., Kohler A., Kent M.P., Lien S. (2018). Unravelling genetic variation underlying de novo-synthesis of bovine milk fatty acids. Sci Rep. 1;8(1):2179. doi: 10.1038/s41598-018-20476-0.

Korte A., Vilhjálmsson B.J., Segura V., Platt A., Long Q., Nordborg M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 44(9):1066-71.

Krag K., Poulsen N.A., Larsen M.K., Larsen L.B., Janns L. and Buitenhuis B. (2013). Genetic parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a Bayesian approach. BMC Genet 14:79.

Legarra A., Robert-Granié C., Croiseau P., Guillaume F., Fritz S. (2011). Improved Lasso for genomic selection. Genet Res (Camb). 2011;93:77–87.

Li X., Buitenhuis A.J., Lund M.S., Li C., Sun D., Zhang Q., Poulsen N.A., Su G. (2015). Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. J Dairy Sci. 98(11):8152-63.

Littlejohn M.D., Tiplady K., Fink T.A., Lehnert K., Lopdell T., Johnson T., Couldrey C., Keehan M., Sherlock R.G., Harland C., Scott A., Snell R.G., Davis S.R., Spelman R.J. (2016). Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. Sci Rep. 5;6:25376.

MacLeod I.M., Bowman P.J., Vander Jagt C.J., Haile-Mariam M., Kemper K.E., Chamberlain A.J., Schrooten C., Hayes B.J., Goddard M.E. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 27;17:144.

Mele M., Dal Zotto R., Cassandro M., Conte G., Serra A., Buccioni A., Bittante G., Secchiari P. (2009). Genetic parameters for conjugated linoleic acid, selected milk fatty acids, and milk fatty acid unsaturation of Italian Holstein-Friesian cows. J Dairy Sci. 92(1):392-400. doi: 10.3168/jds.2008-1445.

Moore S.G., Pryce J.E., Hayes B.J., Chamberlain A.J., Kemper K.E., Berry D.P., McCabe M., Cormican P., Lonergan P., Fair T., Butler S.T. (2016). Differentially expressed genes in endometrium and corpus luteum of Holstein cows selected for high and low fertility are enriched for sequence variants associated with fertility. Biol Reprod. 94(1):19. doi: 10.1095/biolreprod.115.132951.

Olsen H.G., Knutsen T.M., Kohler A., Svendsen M., Gidskehaug L., Grove H., Nome T., Sodeland M., Sundsaasen K.K., Kent M.P., Martens H., Lien S. (2017). Genome-wide association mapping for milk fat composition and fine mapping of a QTL for de novo synthesis of milk fatty acids on bovine chromosome 13. Genet Sel Evol. 13;49(1):20. doi: 10.1186/s12711-017-0294-5.

Pegolo S., Dadousis C., Mach N., Ramayo-Caldas Y., Mele M., Conte G., Schiavon S., Bittante G., Cecchinato A. (2017). SNP co-association and network analyses identify E2F3, KDM5A and BACH2 as key regulators of the bovine milk fatty acid profile. Sci Rep. 11;7(1):17317. doi: 10.1038/s41598-017-17434-7.

Petrini J., Iung L.H., Rodriguez M.A., Salvian M., Pértille F., Rovadoscki G.A., Cassoli L.D., Coutinho L.L., Machado P.F., Wiggans G.R., Mourão G.B. (2016). Genetic parameters for milk fatty acids, milk yield and quality traits of a Holstein cattle population reared under tropical conditions. J Anim Breed Genet. 133(5):384-95. doi: 10.1111/jbg.12205.

Poulsen N.A., Eskildsen C.E.A., Skov T., Larsen L.B., Buitenhuis A.J. (2014). Comparison of genetic parameters estimation of fatty acids from gas chromatography and FT-IR in Holsteins. In proceedings: 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC Canada.

Rodriguez M.A., Petrini J., Ferreira E.M., Mourão L.R., Salvian M., Cassoli L.D., Pires A.V., Machado P.F., Mourão G.B. (2014). Concordance analysis between estimation methods of milk fatty acid content. Food Chem. 1;156:170-5.

Roman R.M., Wilcox C.J. (2000). Bivariate animal model estimates of genetic, phenotypic, and environmental correlations for production, reproduction, and somatic cells in Jerseys. J Dairy Sci. 83:829-35.

Ron M., Israeli G., Seroussi E., Weller J.I., Gregg J.P., Shani M., Medrano J.F. (2007). Combining mouse mammary gland gene expression and comparative mapping for the identification of candidate genes for QTL of milk production traits in cattle. BMC Genomics. 20;8:183.

Rutten M.J., Bovenhuis H., Hettinga K.A., van Valenberg H.J., van Arendonk J.A. (2009). Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J Dairy Sci. 92(12):6202-9.

Rutten M.J., Bovenhuis H., van Arendonk J.A. (2010). The effect of the number of observations used for Fourier transform infrared model calibration for bovine milk fat composition on the estimated genetic parameters of the predicted data. J Dairy Sci. 93(10):4872-82. doi: 10.3168/jds.2010-3157.

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

Sanchez M.P., Ferrand M., Gelé M., Pourchet D., Miranda G., Martin P., Brochard M., Boichard D. (2017a). Short communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds. J Dairy Sci. 100(8):6371-6375. doi:10.3168/jds.2017-12663.

Sanchez M.P., Govignon-Gion A., Croiseau P., Fritz S., Hozé C., Miranda G., Martin P., Barbat-Leterrier A., Letaïef R., Rocha D., Brochard M., Boussaha M., Boichard D. (2017b). Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet Sel Evol. 18;49(1):68.

Sarup P., Jensen J., Ostersen T., Henryon M., Sørensen P. (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. BMC Genet. 17:11.

Shahbazkia H.R., Aminlari M., Tavasoli A., Mohamadnia A.R., Cravador A. (2010). Associations among milk production traits and glycosylated haemoglobin in dairy cattle;importance of lactose synthesis potential. Vet Res Commun. 34(1):1-9.

Shetty N., Difford G., Lassen J., Løvendahl P., Buitenhuis A.J. (2017). Predicting methane emissions of lactating Danish Holstein cows using Fourier transform mid-infrared spectroscopy of milk. J Dairy Sci. 100(11):9052-9060.

Schennink A., Heck J.M., Bovenhuis H., Visker M.H., van Valenberg H.J., van Arendonk J.A. (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(5):2135-43.

Schopen G.C., Heck J.M., Bovenhuis H., Visker M.H., van Valenberg H.J., van Arendonk J.A. (2009). Genetic parameters for major milk proteins in Dutch Holstein-Friesians. Journal of dairy science 92(3):1182-1191.

Smith, B. C. 1996. Fundamentals of Fourier Transform Infrared Spectroscopy. CRC Press, Boca Raton, FL.

Smith C.L., Goldsmith C.A., Eppig J.T. (2005). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol. 6(1):R7.

Solovieff N., Cotsapas C., Lee P.H., Purcell S.M., Smoller J.W. (2013). Pleiotropy in complex traits: challenges and strategies Nat Rev Genet 14: 483–495

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

Soyeurt H., Gillon A., Vanderick S., Mayeres P., Bertozzi C., Gengler N. (2007). Estimation of heritability and genetic correlations for the major fatty acids in bovine milk. J Dairy Sci. 90(9):4435-42.

Soyeurt H., Dehareng F., Gengler N., McParland S., Wall E., Berry D. P., Coffey M., Dardenne P. (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., van Arendonk J.A.M., Heck J.M.L., van Valenberg H.J.F., Bovenhuis H. (2008). Genetic parameters for major milk fatty acids and milk production traits of Dutch Holstein-Friesians. J Dairy Sci, 91:385–394.

Tullo E., Frigo E, Rossoni A., Finocchiaro R, Serra M., Rizzi N., Samorè A.B., Canavesi F, Strillacci M.G., Prinsen R.T.M.M., Bagnato A. (2014). Genetic parameters of fatty acids in Italian Brown Swiss and Holstein cows, Italian journal of animal science; 13:3, 3208. doi:10.4081/ijas.2014.3208

Turley P., Walters R.K., Maghzian O., Okbay A., Lee J.J., Fontana M.A., Nguyen-Viet T.A., Wedow R., Zacher M., Furlotte N.A., Magnusson P., Oskarsson S., Johannesson M., Visscher P.M., Laibson D., Cesarini D., Neale B.M., Benjamin D.J; 23andMe Research Team; Social Science Genetic Association Consortium.

(2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 50(2):229-237. doi: 10.1038/s41588-017-0009-4.

Toledo-Alvarado H., Vazquez A.I., de Los Campos G., Tempelman R.J., Bittante G., Cecchinato A. (2018). Diagnosing pregnancy status using infrared spectra and milk composition in dairy cows. J Dairy Sci. 101(3):2496-2505.

van Binsbergen R., Calus M.P.L., Bink M.C.A.M., van Eeuwijk F.A., Schrooten C., Veerkamp R.F. (2015). Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 47:71.

VanRaden P.M. (2008). Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414–23

VanRaden P.M., Olson K.M., Null D.J., Sargolzaei M., Winters M., van Kaam J.B.C.H.M. (2012). Reliability increases from combining 50,000- and 777,000-marker genotypes from four countries. Interbull. Bull. 46, 75–79.

Veerkamp R.F., Bouwman A.C., Schrooten C., Calus M.P. (2016). Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. Genet Sel Evol. 2016 Dec 1;48(1):95.

Viturro E., Farke C., Meyer H.H., Albrecht C. (2006). Identification, sequence analysis and mRNA tissue distribution of the bovine sterol transporters ABCG5 and ABCG8. J Dairy Sci. 89(2):553-61.

Wang Q., Bovenhuis H. (2018). Genome-wide association study for milk infrared wavenumbers. J Dairy Sci. 101(3):2260-2272.

Wu B., Pankow J.S. (2018). Fast and accurate genome-wide association test of multiple quantitative traits. Comput Math Methods Med. 18;2018:2564531.

Zhou X., Stephens M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 11(4):407-9.

Zhou L., Ding X., Zhang Q., Wang Y., Lund M.S. and Su G. (2013). Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genet Sel Evol. 21;45:7. doi: 10.1186/1297-9686-45-7.

Zhou L., Heringstad B., Su G., Guldbrandtsen B., Meuwissen T.H., Svendsen M., Grove H., Nielsen U.S., Lund M.S. (2014). Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J Dairy Sci. 97(7):4485-96.

Summary

There is an increasing interest in the detailed protein and fatty acid (FA) composition of milk following increasing demands for processed dairy products, coupled with decreasing prices of whole milk, and consumer preferences to certain specific components of milk. Therefore, there is increasing need to alter protein and FA composition of milk to increase yield of dairy products and address consumer preferences. Among possible strategies to alter the protein and FA profile of cow's milk, genetic intervention through selective breeding provides cumulative effects for a one-time investment carried over generations. Emergence of new tools, such as genomic prediction over the past decades has enabled rapid response to selection, through reducing generation interval. Accurate genomic prediction requires availability of a numerically large reference population. The current standard analytical methods for the detailed milk protein and FA phenotypes require costly and time-consuming procedures, limiting measurement to experimental scale. Efficient quantitative methods are therefore required for accurate genetic analysis in scarcely recorded traits. This thesis presents exploration of different quantitative approaches to improve accuracies of genetic parameters estimation, genomic prediction as well as detection power of GWA for scarcely recorded traits focusing mainly on utilization of information from correlated traits, related populations as well as the underlying biology.

Chapter 2 presents multi-trait approaches for estimation of genetic parameters in the detailed milk protein composition traits. Bivariate and multi-trait analyses were implemented in a REML setting with relationship matrix calculated using imputed full sequence data. High heritability estimates, computed with reasonable standard errors using multi-trait analyses, for some of the milk proteins indicate possibility of genetic improvement through selective breeding. Genetic correlations between the milk proteins and milk protein content were generally low, suggesting little or no potential impact on protein content by selecting for detailed protein composition.

In **Chapter 3,** we develop and implement novel univariate and bivariate Bayesian prediction models (BayesAS), to improve accuracy of predicting breeding merits for milk protein composition by dis-entangling heterogeneous (co)variance structures across the genome. Large gains in prediction reliability are shown with implementation of the novel models, in comparison to the traditional GBLUP approach, indicating that efficient statistical models allow accurate genomic prediction for milk protein traits with a numerically small dataset.

In **chapter 4**, we investigate advantages of combining multi-population datasets for genome-wide association (GWA) in the milk FA traits using samples from the Dutch, Danish and Chinese Holstein. Comparing joint GWA, meta-analyses and within-population GWA in terms of number of detected regions and associations in confirmed regions of *DGAT1* and *SCD1*, we show that pooling raw data from different populations for joint GWA allows enhanced detection power for scarcely recorded traits.

Chapter 5 further characterizes major regions detected using the multi-population GWA for the milk FA traits. Proportions of genetic variance explained by detected regions ranged between 1.4% and 45.3% indicating the statistical power of implemented analyses. Post-GWA analysis with multiple data sources on pathway, gene ontology and tissue-specific gene expression data suggest novel promising candidate genes potentially affecting different FA synthesis mechanisms.

In **chapter 6**, we implement findings of chapter 4 and 5 to develop genomic prediction models for milk FA composition using multi-population reference and linear models allowing incorporation of GWA findings. Genomic features-based model (GFBLUP) where separate random genetic effects were considered for BTAs 14, 19 and 26 using relationship matrices constructed separately for variants in the BTAs was implemented. Combined reference population resulted in small to moderate gains in prediction reliability compared to with-in population prediction. GFBLUP model resulted in further gains in prediction reliability for most traits but the amount of gain varied for the different validation populations.

Chapter 7 (general discussion), highlights contributions of the PhD study to the current knowledge base and the broader context of implementing selective breeding to alter detailed milk protein and FA composition. We show that limitation of traditional multi-trait genomic prediction models for traits with weak genome-wide correlation can be overcome by disentangling heterogeneous correlation structure and using information from regions where there is higher genetic correlation. Our studies indicate advantages of pooling data from different population for GWA studies and genomic prediction. Therefore, it is suggested that international collaborations facilitating access to multi-population data are critical to successes in unraveling the genetic backgrounds and implementing selective breeding for the milk protein and FA composition traits. Our studies suggest benefits of incorporating biological information in improving genomic prediction accuracy. It is therefore highlighted that GWA studies will remain to be important in the context of selective breeding. The need for future studies have been stressed especially with regards to genetic correlations of the milk protein and FA composition with other traits in the dairy cattle breeding goal, genetic correlations with infrared predicted proxy phenotypes and regarding appropriate definition of the traits for the breeding objective.

Sammendrag

Der er en stigende interesse for komælks detaljerede sammensætning af protein og fedtsyrer (fatty acid, FA), som følge af en stigende efterspørgsel på forarbejdede mejeriprodukter, koblet med faldende priser på sødmælk samt forbrugernes præferencer for visse specifikke mælkekomponenter. Der er derfor et stigende behov for at ændre mælks protein- og FA-sammensætning med henblik på at øge mejeriprodukters ydelse og håndtere forbrugernes præferencer. Blandt mulige strategier til at ændre protein- og FA-profil af komælk er genetisk intervention gennem selektiv avl, der ved en engangsinvestering giver en akkumuleret effekt over generationer. Fremkomsten af nye værktøjer de seneste årtier, såsom genomisk prædiktion, har muliggjort en hurtigere respons til selektion ved at reducere generationsintervallet. Præcis genomisk prædiktion kræver en talmæssigt stor referencepopulation. De nuværende gængse analysemetoder for mælkeprotein- og FA fænotyper er bekostelige og tidskrævende procedurer, der begrænser målinger til et eksperimentelt omfang. Effektive kvantitative metoder er derfor nødvendige for en præcis genetisk analyse af egenskaber, der kun er sparsomt registrerede. Denne ph.d.-afhandling præsenterer udnyttelsen af forskellige kvantitative tilgange til forbedring af nøjagtigheden i estimeringen af genetiske parametre, genomisk prædiktion såvel som præcisionen af genetiske associationsstudier (GWA) for egenskaber med få registreringer, ved primært at fokusere på anvendelse af information fra korrelerede egenskaber, relaterede populationer samt den underliggende biologi.

Kapitel 2 præsenterer en multi-egenskabs tilgang til vurdering af genetiske parametre for mælkeproteinsammensætningen. Bivariate og multi-egenskabs-analysemetoder blev implementeret i REML med en slægtsskabsmatrice beregnet fra imputerede fuldsekvensdata. Høje estimater for arvelighed, med rimelige estimerede standardfejl, ved at bruge multi-egenskabsanalyser, indikerer at der for nogle af mælkeproteinerne er mulighed for genetisk forbedring gennem selektiv avl. De genetiske korrelationer mellem mælkeproteinerne og mælkeproteinindholdet var generelt lave, hvilket antyder lidt eller ingen mulig indvirkning på proteinindholdet ved at selektere for proteinsammensætning.

I **kapitel 3** udvikler og implementerer vi nye univariate og bivariate Bayesiske prædiktionsmodeller (BayesAS) for at forbedre nøjagtigheden af at forudsige avlsfordele for mælkeproteinsammensætning ved at udrede heterogene (co)variansstrukturer på tværs af genomet. Implementeringen af de nye modeller gav store forbedringer af nøjagtigheden i prædiktionerne, sammenlignet med den traditionelle GBLUP tilgang. Dette indikerer at effektive statistiske modeller tillader

nøjagtig genomisk prædiktion for mælkeprotein-egenskaber med et talmæssigt lille datasæt.

I **kapitel 4** undersøger vi fordele ved at kombinere datasæt fra hollandske-, dansk- og kinesisk- Holstein kvægpopulationer til genetiske associationsstudier for FA. Ved at anvende en kombineret GWA, meta-analyser og intra-populations-GWA med hensyn til antal identificerede regioner, samt associationer i kendte regioner af *DGAT1* og *SCD1* viser vi, at ved at kombinere data fra forskellige populationer i en kombineret GWA giver større statistisk kraft for egenskaber med få registreringer.

Kapitel 5 karakteriserer yderligere væsentlige regioner, der er påvist ved brug af flerpopulations-GWA for mælkefedtsyretrækkene (FA). Andelen af forklaret genetiske varians i de identificerede regionerlå mellem 1.4% og 45.3%, hvilket indikerer øget statistisk styrke. Post-GWA analyser på stofskifteveje, genontologi og vævsspecifik genekspressions indikerer nye lovende kandidatgener, der potentielt kan påvirke forskellige FA-syntesemekanismer.

I **kapitel 6** implementerer vi resultater fra kapitel 4 og 5 til at udvikle genomiske prædiktionsmodeller for mælks fedtsyresammensætning, ved at bruge multi-populationsreference og lineære modeller, der tillader inkorporering af GWA resultater. En GFBLUP (Genomic feature-based) model blev implementeret, hvor separate genetiske effekter for kromosom 14, 19 og 26 blev estimeret ved at konstruere slægtskabsmatricer basseret på genetiske varianter på disse kromosomer. En kombineret referencepopulation resulterede i små til moderate forbedringer i prædiktionsnøjagtigheden, sammenlignet med prædiktion indenfor populationer. GFBLUP-modellen resulterede i yderligere forbedringer i prædiktionsnøjagtigheden for de fleste egenskaber, men størrelsen varierede for de forskellige valideringspopulationer.

Kapitel 7 (generel diskussion), fremhæves bidraget fra denne Ph.d-afhandling til den eksisterende viden, samt den bredere kontekst ved implementering af selektiv avl for at ændre mælkeprotein- og fedtsyresammensætning i komælk. Vi viser, at begrænsningen i traditionelle genomiske prædiktionsmodeller for multi-egenskaber, for egenskaber med svag genetiske korrelationer, kan overvindes ved at udrede den heterogene korrelationsstruktur og bruge information fra regioner, hvor der er højere genetisk korrelation. Vores undersøgelser indikerer fordele ved at samle data fra forskellige populationer til genetiske associationsstudier og genomisk prædiktion. Det foreslås derfor at internationale samarbejder, der letter adgangen til multi-populationsdata, er nødvendig for succes til at udrede den genetiske baggrund for og implementering af selektiv avl for mælkeprotein- og fedtsyresammensætningsegenskaberne. Vores undersøgelser indikerer at ved at inkorporere biologisk information til genomisk prædiktion øges nøjagtigheden. Det

fremhæves derfor at genetiske associationsstudier fortsat vil være vigtige i en selektiv avlskontekst. Behovet for fremtidige undersøgelser er blevet understreget, især med henblik på genetiske korrelationer af mælkeprotein- og fedtsyresammensætning med andre egenskaber i avlsmålene for malkekvæg, genetiske korrelationer med infrarøde, forventede proxy-fænotyper og med henblik på passende definitioner af egenskaberne for avlsmålsætningerne.

Samenvatting

Er is een toenemende belangstelling voor de gedetailleerde eiwit- en vetzuursamenstelling van melk als gevolg van de toenemende vraag naar gedifferentieerde zuivelproducten. Hierdoor is er een toenemende behoefte om de eiwit- en VZ-samenstelling van melk te veranderen om zo aan de vraag en voorkeur van de consument te voldoen. Een van de mogelijke strategieën om het eiwit- en vetzuurprofiel van koemelk te veranderen, is genetische interventie door selectieve fokkerij.

Voorspelling van kenmerken door middel van genetische merkers heeft het mogelijk gemaakt om sneller te selekteren. Een nauwkeurige genomische voorspelling vereist de beschikbaarheid van een grote referentiepopulatie waarin de kenmerken gemeten worden. De huidige standaard analysemethoden voor de gedetailleerde melkeiwitten en vetzuur kenmerken vereisen kostbare en tijdrovende procedures, waardoor de metingen beperkt blijft tot kleinscahalige experimentele populaties. Efficiënte kwantitatieve methoden zijn daarom vereist voor nauwkeurige genetische analyse van deze nieuwe kenmerken. Dit proefschrift presenteert een verkenning van verschillende kwantitatieve benaderingen om de nauwkeurigheid van genetische parameters te verbeteren. Daarbij wordt gekeken naar de voorspelling van de eiwit- en vetzuurkenmerken op basis van genetische DNA merkers op het genoom en het detectievermogen van genetische DNA merkers voor eiwit- en vetzuurkenmerken, waarbij gebruik gemaakt wordt van informatie uit gecorreleerde kenmerken.

Hoofdstuk 2 presenteert multi-trait benaderingen voor het schatten van genetische parameters voor de gedetailleerde melkeiwitsamenstelling. Bivariate- en multi-trait-analyses werden geïmplementeerd in REML. De genomische relatie tussen de dieren in de populatie wordt berekend met een genomische relatiematrix aan de hand van geïmputeerde DNA merkers. De geschatte erfelijkheidsgraad voor de specifieke melkeiwitten wijzen erop dat genetische selektie mogelijk is. De genetische correlatie tussen de specifieke melkeiwitten en het totale eiwitgehalte zijn laag, hetgeen suggereert dat er geen of weinig effect is op het totale eiwitgehalte in de melk wanneer er selektie plaatsvindt op de gedetaillieerde melkeiwitsammenstelling.

In **Hoofdstuk 3** ontwikkelen en implementeren we nieuwe univariate en bivariate Bayesiaanse voorspellingsmodellen (BayesAS) om de nauwkeurigheid van het voorspellen van de specifieke melkeiwitsamenstelling te verbeteren. Hierbij wordt gebruik gemaakt van de heterogene (co) variantie structuren op het genoom. Deze nieuwe Bayesiaanse modellen tonen een grote vooruitgang in de voorspellingsbetrouwbaarheid in vergelijking met de traditionele GBLUP-modellen.

Dit geeft aan dat efficiënte statistische modellen een nauwkeurige genomische voorspelling mogelijk maken voor melkeiwitkenmerken die gemeten zijn in een kleine dataset.

In **hoofdstuk 4** onderzoeken we de voordelen van het combineren van verschillende experimente data-sets voor het vinden van DNA merkers die geassocieerd zijn met specifieke vetzuurkenmerken in de melk. Hierbij wordt gebruik gemaakt van datasets die verzameld zijn van Nederlandse, Deense en Chinese Holstein koeien. Verschillende analyse methoden worden vergeleken (meta-analyse, analyse binnen de individuele populatie en een analyse waarbij alle data gecombineerd wordt) laten we zien dat wanneer alle datasets gezamelijk opnieuw geanalyseerd wordt, de detectie van nieuwe DNA merkers verbeterd.

Hoofdstuk 5 worden de meest belangrijke regio's die zijn gedetecteerd met behulp van de analyse waarbij alle datasets opnieuw gezamelijk geanalyseerd worden, gekarakteriseerd. De genetische variantie die verklaard wordt door de gedetecteerde regio's varieerden tussen 1,4% en 45,3%. Een post-DNA merker associatie-analyse aan de hand van meerdere gegevensbronnen op het gebied van pathway gegevens, genontologie en weefselspecifieke genexpressiegegevens suggereren nieuwe veelbelovende kandidaatgenen die mogelijk van invloed zijn op verschillende vetzuursynthesemechanismen.

In **hoofdstuk 6** implementeren we bevindingen uit hoofdstuk 4 en 5 om genomische voorspellingsmodellen voor melk vetzuursamenstelling te ontwikkelen. Op basis van genomische kenmerken (GFBLUP-model) worden afzonderlijke genetische effecten die gevonden zijn op chromosoom 14, 19 en 26 getest. Het combineren van de Nederlandse, Deense en Chinese Holstein data-sets resulteerde in een kleine tot matige winst in voorspellingsbetrouwbaarheid in vergelijking met voorspelling van de drie individuele populaties. In het algemeen resulteerde het GFBLUP-model in een betere voorspellingsbetrouwbaarheid voor de meeste kenmerken, maar de hoeveelheid winst die behaald wordt varieerde voor de verschillende validatiepopulaties.

De algemene discussie (**Hoofdstuk 7**) belicht de bijdragen van het doctoraatsonderzoek aan de huidige kennis en de plaats de resultaten in een bredere context ten aanzien van de mogelijkheden voor implementatie van genetische selektie om de gedetailleerde eiwit en vetzuursamenstelling in de melk te wijzigen. We laten zien dat beperkingen van traditionele multi-trait genomische voorspellingsmodellen voor kenmerken met een zwakke genoom-brede correlatie kan worden overwonnen door gebruik te maken van de heterogene correlatiestructuur en informatie te gebruiken uit regio's waar er een hogere genetische correlatie is. De studies in dit proefschrift wijzen op de voordelen van

het samenvoegen van gegevens van verschillende (experimentele) data-sets voor DNA merker associatie analyses en genomische voorspelling. Daarom wordt gesuggereerd dat internationale samenwerkingsverbanden die de toegang tot gegevens over meerdere populaties omvatten, zijn van cruciaal belang voor successen bij het ontrafelen van de genetische achtergronden en het implementeren van selectieve fokmethoden voor het melkeiwit- en vetzuursamenstelling. De studies laten voordelen zien van het gebruik van biologische informatie in het verbeteren van de nauwkeurigheid van de genomische voorspelling.

Acknowledgement

Acknowledgement

I would like to take the opportunity to thank, as much as I can, the many people without whose contribution, in one way or another, the completion of this undertaking could not have been possible.

I am exceptionally thankful to my supervisors Bart Buitenhuis, Henk Bovenhuis and Mogens Lund for the opportunity and trust bestowed on me to handle the project and the unreserved supports thereafter. Timely completion of major activities in this endeavor come down to Bart's skills in planning and management together with the persistent support that kept me on track. I have been fortunate to have Bart as my main supervisor with his supervision approach that allowed me the freedom to explore as independently as possible and be in charge of my project, yet consistently providing me with the guidance and support as I stumble across challenging situations. Having Henk as one of my main supervisors and promoter have been very inspirational and a great learning experience. I am always fascinated with new perspectives Henk adds into the different studies and always looked forward excitedly to his comments on my manuscripts. Beyond just inputs to my manuscripts, I see attention to details, dedication to the field, and a level of critical thinking that has been very inspirational in the typically red-colored and exclamation-mark-riddled long lines of comments. My co-supervisor, Mogens, have been very helpful in broadening my view of the PhD project making me think of the practical importance of what I do. Mogens always came up with innovative ideas and helped me see opportunities out of phenomena I otherwise considered hurdles. I have been very lucky to have you all in my supervision team and I cannot thank you enough for the inspirations and lessons that I will carry with me throughout my career.

I would also like to extend my appreciations to all my co-authors in different countries and research groups for the smooth and successful collaborations. Special thanks in this regard goes to Luc Janss, whom I really consider as an additional co-supervisor in light of the consistent support and mentoring that he has always extended, and Nina Poulsen for dedicated contributions in all my manuscripts.

All my colleagues at QGG, I am very much thankful for all the support and friendly working atmosphere. I am very happy that i will continue working with you at QGG after my PhD. I am grateful to late Karin, Louise, Hanne Amtrup and Birgitte Larsen for all the supports at different parts of my PhD here in QGG. Jette Odgaard and

Palle Duun, thank you for the translation works on the summary of the thesis. Thanks go to current/former PhD/postdoc pals at QGG: Ahmed, Lise, Biructawit, Mesbah, Wossenie, Berihu, Margot, Coralia, Helene, Mahlet, Mette, Daniel, Hongding, Mahmoud, Roos and many others I didn’t mention here. Thanks for all the nice discussions, the lunch-table gatherings and social occasions.

Special thanks goes to my friends Gareth Difford and Larissa Zetouni. I enjoyed every moment of my PhD endeavor thanks to your friendship. Starting the PhD at the same time as Gareth and going throughout the journey together with such a friend has been a blessing. Thank you for all the help and the friendship that I will always cherish.

Apart from QGG, thanks go to friends in Foulum, particularly Saman and Mihai for the helps in different occasions.

My warmest appreciation also goes to all my colleagues at ABG of Wageningen University. Maria, Zih-Hua, Sonia, Pascal, Johanna, Floor and Juan thank you for all the good acquaintance and the fantastic parties. Harmen, Zhou, Robert and Langqing, you have been so awesome office mates, thank you. Outside of ABG, special thanks to Meklit Berihun and Shimelis Altaye for all the support and the unforgettable moments during my stay at Wageningen.

Most importantly, my deepest appreciation goes to my family: my beloved daughters Blen and Yabsira and my partner Tehetena Ayele for the love, patience and encouragement. Finally, I would like to thank my mother, Almaz Kassaye and grandmother, Shewaye Cherinet, for all the things that I can’t even count. If anything good exists in me and if any of my undertakings are deemed a success, it is all because of the endless love you gave me and the values you raised me with, thank you!

Curriculum Vitae

About the author

Grum Gebreyesus was born on October 1982 in Dire Dawa, Ethiopia. He did his bachelor's degree in animal sciences at Haramaya University, Ethiopia. Upon completing his bachelor's study, Grum worked at Endasselassie agricultural college as junior instructor and later at Jigjiga University, as an assistant lecturer. In his early teaching experience, Grum mainly focused on courses related to genetics and breeding, including principles of genetics, applied animal breeding, biometrics and statistics. With special interest to genetics, Grum returned to Haramaya University to pursue master's study in animal genetics and breeding. In his master's study, Grum focused on phenotypic characterization of indigenous animal genetic resources with his thesis work exploring the role of indigenous knowledge and communal breeding practices in shaping the local animal genetic resource gene pool. Upon completion of his master's degree, Grum was re-employed by Jigjiga University as a lecturer and researcher where he taught courses including animal genetics and breeding, and participated on various research projects. In 2012, Grum joined the international livestock research institute (ILRI) where he worked as a research assistant in geneticist position. During his stay at ILRI, main activities focused on establishing and running community based breeding programs for indigenous goats in various regions of Ethiopia as part of a broader project entitled: "Harnessing genetic diversity for improving goat productivity in Africa", led by the Biosciences Hub eastern and central Africa (BecA-Hub) and ILRI. In 2014, Grum received the Erasmus Mundus PhD scholarship under the joint program: European Graduate School in Animal Genetics and Breeding (EGS-ABG) in Aarhus University, Denmark and Wageningen University, Netherlands. After completion of the PhD program, Grum will continue his scientific career as postdoc researcher in quantitative genetics and genomics at QGG, Foulum.

Peer reviewed publications

1. **Gebreyesus G.**, Lund MS., Janss L., Poulsen N.A., Larsen L.B., Bovenhuis H., Buitenhuis A.J. Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J Dairy Sci. 2016 Apr;99(4):2863-6. doi: 10.3168/jds.2015-10501
2. **Gebreyesus G.**, Lund M.S., Buitenhuis A.J., Bovenhuis H., Poulsen N.A. and Janss L.G. Modeling heterogeneous (co)variances from adjacent-SNP groups improves genomic prediction for scarcely recorded milk protein composition traits.2017. Genet Sel Evol (2017) 49:89.
3. Buitenhuis B., Poulsen N.A., **Gebreyesus G.**, Larsen L.B. 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 Genet. 2016 Aug 2;17:114.

Conference proceedings, abstracts and presentations

1. **Gebreyesus G**., Buitenhuis A.J., Poulsen N.A., Visker M.W., Zhang Q., van Valenberg H., Sun D., and Bovenhuis H. 2018. Genome-wide association study of the de novo synthesized milk fatty acids based on Dutch, Danish and Chinese Holstein Friesians. In proceedings of the 11th World Congress on Genetics Applied to Livestock Production (WCGALP), Auckland, New Zealand. Paper: 141.
2. **Gebreyesus G.**, Lund M.S., Janss L.G., Bovenhuis H. and Buitenhuis A.J. Fine mapping and genomic prediction for detailed milk protein composition: in proceedings of the 5th International Conference on Quantitative Genetics (ICQG5). June 12- 17, 2016. Madison, Wisconsin, USA.
3. **Gebreyesus G.**, Lund M.S., Janss L.G., Bovenhuis H. and Buitenhuis A.J. Modeling heterogeneous co-variancas for genomic regions in prediction for milk protein compositions: in proceedings of the 67th Annual Meeting of the European Federation of Animal Science. 29 Aug- 2 Sep. Belfast, United Kingdom.
4. Difford G.F., **Gebreyesus G**., Løvendahl P., Buitenhuis A.J., Lassen J., Guldbrandtsen B., and G. Sahana. Can rumen microbes improve prediction of metabolic traits in Dairy cows. in proceedings of the 67th Annual Meeting of the European Federation of Animal Science

Individual Training Plan (ITP)

Training (33 ECTS)		
Mandatory courses (9)	Place	Year
EGS-ABG welcome course	WUR	2014
EGS-ABG Fall school	WUR	2017
EGS-ABG Fall school	SLU	2016
EGS-ABG Fall school	AgroParisTech	2018
Research ethics and integrity in animal sciences	WUR	2017
Advanced scientific courses (18 ECTS)		
Introduction to genomic selection	WUR	2014
Linear models in animal breeding	AU	2015
Gene mapping	AU	2015
Genomic prediction in livestock	Iowa	2015
Design of breeding programs with genomic selection	Iowa	2015
Genomic selection in the era of genomic sequencing	Madison	2016
Statistical genetics of quantitative traits and complex diseases	Madison	2016
Professional skill support courses (6)		
QGG research skill course	AU	2015
Programing in animal science	AU	2015

Knowledge dissemination		
Teaching		
Teaching assistance: Genetics course	AU	2016
International conferences		
The 11th World Congress on Genetics Applied to Livestock Production (WCGALP)	Newzealand	2018
The 5th International Conference on Quantitative Genetics (ICQG5)	Madison	2016
The 67th annual meeting of the European Federation of Animal Science (EAAP)	Belfast	2016
Seminars and workshop		
Annual Gensap meeting	Denmark	2015, 2017
Nordic cattle genomic selection workshop	Denmark	2017, 2018

Colophon

The PhD study was supported by grants from the European commission, within the framework of the Erasmus-Mundus joint doctorate program (EGS-ABG), and the Center for Genomic Selection in Animals and Plants (GenSAP), funded by The Danish Council for Strategic Research.

The studies used data from the Dutch Milk Genomics Initiative (www.milkgenomics.nl); High Technology R & D Program of China,2013AA102504 and Beijing Dairy Industry Innovation Team, BAIC06-2018; Danish-Swedish Milk Genomics Initiative (www.milkgenomics.dk) and the Milk Levy Fund (Denmark) projects: “Phenotypic and genetic markers for specific milk quality parameters” and “The importance of the metagenome for milk composition and quality".

This thesis was printed by Digisource, Viborg, Denmark