Norwegian University of Life Sciences
Faculty of Biosciences

Philosophiae Doctor (PhD)
Thesis 2018:19

# Genomics of bovine milk fat composition

## Genetisk karakterisering av fettsyresammensetning i melk

Tim Martin Knutsen

# Genomics of bovine milk fat composition

Genetisk karakterisering av fettsyresammensetning i melk

Philosophiae Doctor (PhD) Thesis

Tim Martin Knutsen

Norwegian University of Life Sciences
Faculty of Biosciences
Department of Animal and Aquacultural Sciences

Ås (2018)

Thesis number 2018:19
ISSN 1894-6402
ISBN 978-82-575-1499-0

## Acknowledgements

The work presented here has been part of the project "Genome-based improvement of bovine milk fatty acids". The project was coordinated by the Norwegian University of Life Sciences (NMBU) and was a collaboration between NMBU, GENO SA and TINE SA. The project was funded by The Norwegian Research Council.

To my dear supervisors. Sigbjørn, you are constantly supportive, work relentlessly for what you believe in, and it has been an honour to learn from you. Hanne Gro, thanks for the endless hours we have spent discussing this project, your critical sense, and your treasured help in the writing process. Torfinn, thanks for your out-of-this-world-value-for-money bioinformatics coaching and advice. And Matthew, thanks for all your hours polishing my writing, your enthusiasm for cool sequencing technologies and your wonderful sense of humour.

My sincere thanks to all others who have contributed to this work, especially Achim Kohler, Valeria Tafinseva and Morten Svendsen for enabling the fatty acid GWAS with your mathematical wizardry. I also want to express my gratitude to Arne Gvusland for your critical comments to an early draft of this thesis, and for time after time dragging me out to jog at times when I really needed it.

Seven years ago, late for my appointment, I stumbled into Sigbjørn's office to discuss possible master projects. I was a medical laboratory scientist stuck in routine work, with an inherent curiosity about nature and biology. We agreed that fatty acid genetics might be something. Changing course from routine medical lab to the magnificent world of GWAS, reference genomes, SNP-calling and the UNIX shell has profoundly changed my life. I have looked forward to going to work every day since I started.

To my beautiful wife and two crazy little boys. Aksel and Mats, thanks a lot for getting me out of bed (very) early in the mornings. Anne Kjersti, I love you. Without your endless patience and support, I would not have been able to arrive here.

Ås, March 2018.

Tim Knutsen

# Contents

## Summary

Bovine milk is a highly regarded food source. Still, some milk fatty acids may have unfavourable health effects and can influence manufacturing properties of dairy products. Individual milk fatty acids show moderate heritabilities, and better knowledge of the underlying genes may be used to speed up the genetic progress of the traits and provide dairy products of higher quality and nutritional value. In this thesis, mutations underpinning variation in bovine milk fat composition in Norwegian Red cattle was explored, with emphasis on fatty acids produces *de novo* in the mammary gland, and the two dominant acids in bovine milk, palmitic (C16:0) and oleic acid (C18:1*cis*9).

Paper I established the calibration equations to predict the fatty acid profiles from Fourier-transform infrared spectroscopy (FTIR) data used to estimate variance components for individual and groups of fatty acids. Most major fatty acids were predicted rather accurately. Short and medium length saturated acids were, in general, more heritable than longer and unsaturated acids. A genome-wide association analysis performed on both individual acids and groups of acids revealed a region on chromosome 13 with strong influence on levels of the even chain fatty acids C4:0 to C14:0. The association was first thought to be related to the gene *acyl-CoA synthetase 2* (*ACSS2)*, but subsequent fine-mapping highlighted another close-by gene; *nuclear receptor coactivator 6* (*NCOA6).*

Paper II aimed to further explore the genetic basis of the *de novo* synthesised acids, extending the analysis with a larger data set, imputed sequence variants and mammary gene expression data. *Progestagen Associated Endometrial Protein* (*PAEP)* on Bos taurus autosome (BTA)11 was strongly associated with the content of the shortest acid C4:0, *acetoacetyl-CoA synthetase* (*AACS)* on BTA17 was associated with the content of C4:0 and C6:0. *NCOA6* on BTA13 was associated with acids of intermediate chain lengths (especially C8:0), whereas *fatty acid synthase* (*FASN)* was mainly associated with the longest acid, C14:0. All suggested positional candidate genes were expressed in the bovine udder during lactation.

Paper III focused on C16:0 and C18:1*cis*-9, possibly having opposing effects on human cardiovascular health and relevance for dairy manufacturing properties. A set of variants within and close to *PAEP* on BTA11 shown to affect the ratio between the two acids were identified. The variants were further shown associated with *PAEP* gene expression and levels of the translated protein β-lactoglobulin. Breeders may use the Paper III findings to promote milk with a healthier fatty acid profile and positive effect on cheese-making properties.

## Sammendrag

Kumelk er regnet som en god human ernæringskilde. Samtidig kan nivået av enkelte fettsyrer i melk ha uheldige helsekonsekvenser, i tillegg til å kunne påvirke meieriprodukters produksjonsegenskaper. Studier har vist at konsentrasjonen av melkefettsyrer er arvbar, og bedre kunnskap om de underliggende gener og varianter vil kunne utnyttes i avl for å kunne oppnå genetisk fremgang for denne egenskapen. I denne avhandlingen ble mutasjoner med påvirkning på fettsyresammensetningen i melk undersøkt, med fokus på syrer syntetisert *de novo* i jur, og de to vanligste fettsyrene i melk; palmitinsyre (C16:0) og oljesyre (C18:1*cis*9).

Artikkel I etablerte kalibreringslikningene nødvendig for å predikere fettsyreprofiler og beregne fettsyrenes arvbarhet ved bruk av FTIR-spektra. De fleste frekvente melkefettsyrer ble predikert med tilstrekkelig nøyaktighet. Mettede fettsyrer med kort og medium kjedelengde hadde generelt høyere arvbarhet en lengre og umettede syrer. En assosiasjonsstudie, utført med både fettsyregrupper og individuelle fettsyrer, pekte mot en region på kromosom 13 med sterk effekt på nivået av de likekjedede fettsyrene C4:0 til C14:0. Genet *acyl-CoA synthetase 2* (*ACSS2)* ble først pekt ut som det beste kandidatgenet, men videre finkartlegging pekte mot det nærliggende genet *nuclear receptor coactivator 6* (*NCOA6)*.

I artikkel II ble den genetiske bakgrunnen for *de novo*-syntetiserte fettsyrer videre studert. Analysen identifiserte sterke assosiasjoner mellom genene *Progestagen Associated Endometrial Protein* (*PAEP)* og *acetoacetyl-CoA synthetase* (*AACS)* og innhold av C4:0-C6:0, *NCOA6* og syrer med mellomlang kjedelengde (C6:0-C12:0) og *fatty acid synthase* (*FASN)* ble funnet sterkt assosiert til innhold av C14:0. Alle foreslåtte kandidatgener ble funnet uttrykt i jur.

Artikkel III fokuserte på C16:0 og C18:1*cis*9, de to mest frekvente fettsyrene i melk, som har betydning for både human helse og melkeproduksjonsegenskaper. Det ble identifiserte et sett varianter i og nær genet *PAEP* på kromosom 11, med motsatt effekt på palmitin og oljesyre. De samme variantene ble også assosiert til redusert ekspresjon av *PAEP* og redusert nivå av det translaterte proteinet β-lactoglobulin. Funnene fra artikkel III kan utnyttes til å avle frem melkekyr med sunnere melkefettsyreprofil og positive effekter på melkens ysteegenskaper.

## List of papers.

I. 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. and 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**.
*Genetics Selection Evolution*, *49*(1), p.20.

II. Knutsen, T.M., Olsen, H.G., Tafintseva, V., Svendsen, M., Kohler, A., Kent, M.P. and Lien, S., 2018. **Unravelling genetic variation underlying *de novo*-synthesis of bovine milk fatty acids**. *Scientific reports*, *8*(1), p.2179.

III. Knutsen, T.M., Olsen, H.G., Ketto, I.A., Sundsaasen, K.K, Kohler, A., Tafintseva, V., Svendsen, M., Kent, M.P. and Lien, S., 2018. **Genetic variants associated with fatty acid composition offer new opportunities to breed for healthier milk.**
*Manuscript.*

# 1. General introduction.

## 1.1 Why study the genomics of milk fat composition?

Milk is a primary product produced and consumed in almost every country. Its appeal, widespread availability, and versatility as a food product have led milk to become a key nutritional element for billions of people worldwide. Among cows, milk's nutritional component concentration varies, influenced by a complex interplay between genes and environmental factors such as feeding, lactation stage, health status and breed (Jensen 2002; Bionaz & Loor 2011; Maurice-Van Eijndhoven et al. 2011). Although marketed as a healthy, natural product, milk's health effects remain controversial, mainly because 60 to 70 percent of milk fatty acids are saturated (Jensen 2002).

Dietary saturated fatty acids (SFAs) have been linked to cardiovascular and metabolic disease, and it is believed that a reduction in dietary SFAs is beneficial (Hooper et al. 2015). In this context, it will be advantageous to identify individual genes, or preferably causal DNA variation responsible for genetic variation in milk fat composition. Such information is important for understanding bovine fatty acid metabolism and may be used to develop new and innovative dairy products through selective breeding.

## 1.2 Brief overview of milk fat composition.

Bovine milk fat is remarkably complex, containing more than 400 different fatty acids (Jensen 2002). The total fat content in cow's milk is normally between three and five percent, with about 98 percent of the fat present as triglycerides (Jensen 2002). Triglycerides are characterised by three fatty acids attached to a glycerol backbone (Figure 1.1). The fatty acids are composed of carbon chains that differ in length. Short-chain fatty acids (SCFA) refers to acids with five or fewer carbon atoms, medium chain fatty acids (MCFA) six to 14, long chain fatty acids (LCFA) have chains of 15 to 21 carbons, and very long chain fatty acids (VLCFA) have >21 carbons. Most fatty acids are saturated, consisting of an unbranched carbon chain with single bonds between each carbon, but can also have one or several double bonds making an unsaturated fatty acid.

In systematic nomenclature, each unsaturated fatty acid is named according to where the double bond sits in the carbon-bond chain. For example, oleic acid, the mid fatty acid in Figure 1.1 is named cis-Δ9-Octadecenoic acid, or just C18:1*cis*-9.

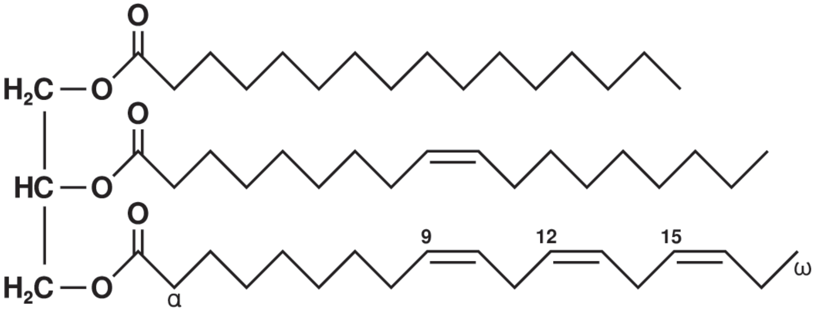


**Figure 1.1 Chemical structure of a triglyceride with a saturated (top), mono-unsaturated (mid) and polyunsaturated fatty acid (bottom) attached to a glycerol backbone.**

The majority of fatty acids in milk are even-chain numbered saturated acids with carbon chains from 4 to 18 (C4:0 - C18:0), along with the unsaturated C18:1*cis*9, which has one double bond in its carbon chain (Table 1.1).

**Table 1.1** Typical composition of bovine milk fatty acids (Adapted from Jensen 2002)

| **fatty acid carbon number** | **fatty acid common name** | **Average range (g/100g fat)** |
|---|---|---|
| 4:0 | Butyric | 2–5 |
| 6:0 | Caproic | 1–5 |
| 8:0 | Caprylic | 1–3 |
| 10:0 | Capric | 2–4 |
| 12:0 | Lauric | 2–5 |
| 14:0 | Myristic | 8–14 |
| 15:0 | Pentadecanoic | 1–2 |
| 16:0 | Palmitic | 22–35 |
| 16:1 | Palmitoleic | 1–3 |
| 17:0 | Margaric | 0.5–1.5 |
| 18:0 | Stearic | 9–14 |
| 18:1 | Oleic | 20–30 |
| 18:2 | Linoleic | 1–3 |
| 18:3 | Linolenic | 0.5–2 |

## 1.3 Milk fat biosynthesis

The complexity of milk fatty acid composition is a consequence of the many pathways and processes by which fatty acids arise. Essentially, in ruminants, they are derived either from direct transport from the diet to the mammary gland via the circulatory system, or by *de novo* synthesised in the mammary gland (reviewed by Bionaz & Loor 2008). The two pathways are represented schematically in Figure 1.2, with central genes shown in green colour.

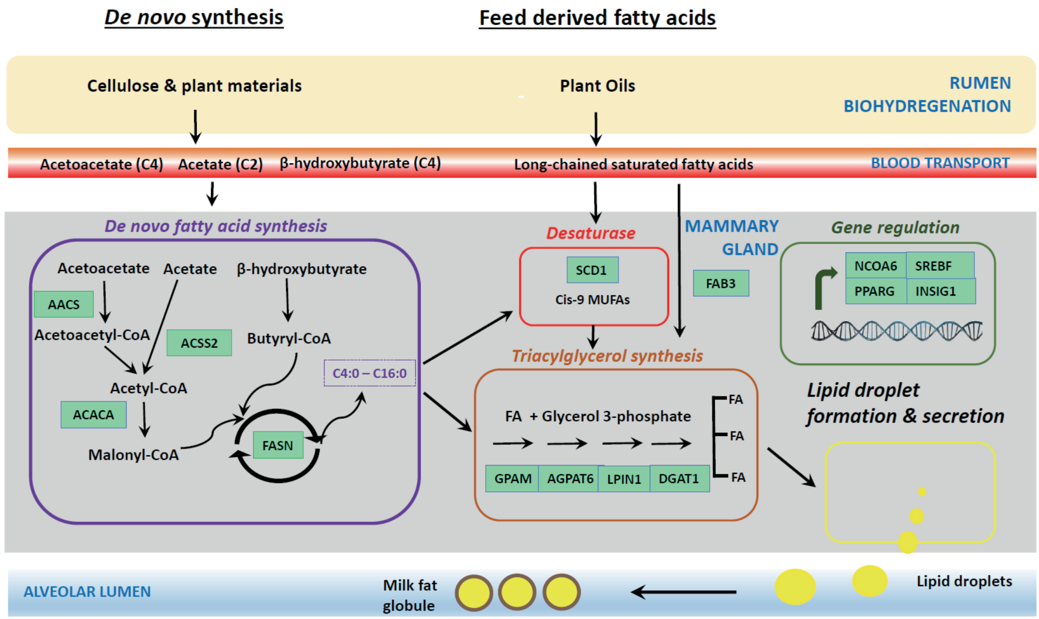


**Figure 1.2. Schematic representation of the metabolism of *de novo* synthesised and feed derived milk fatty acids.**

In the mammary gland, the short- and medium-chained saturated fatty acids C4:0 to C14:0, as well as about half of the palmitic acid (C16:0), are *de novo* synthesised from two and four carbon chain precursors. *De novo* synthesis begins with the uptake of acetate, acetoacetate and a small fraction β-hydroxybutyrate originating from bacterial fermentation of roughage in the rumen. Acetate is activated to acetyl-CoA by the enzyme acyl-CoA synthetase 2 (*ACSS2*). Acetoacetate is first activated by the enzyme acetoacetyl-CoA synthetase (*AACS*) to form acetoacetyl-CoA and then via acetyl-CoA to malonyl-CoA by acetyl-CoA carboxylase alpha (*ACACA*). Malonyl-CoA, along with butyryl-CoA, further serve as precursors for medium-chained acids and C16:0 synthesis. In a cyclic reaction called the malonyl-CoA pathway, the enzyme fatty acid synthase (*FASN*) add two carbon units to the growing fatty acid-chain in each round of the cycle. This cycle's natural endpoint is C16:0. However, in ruminants, SCFAs and MCFAs can leave this cycle at any time by a chain determination mechanism, which gives rise to the relatively high fraction of MCFAs in ruminant milk compared to for example human milk (Barłowska et al. 2011).

Even-numbered LCFAs are transported into the milk from circulating plasma lipids originating either from the diet or lipolysis of adipose tissue triacylglycerol. Odd-numbered SFAs (C15:0 and C17:0) are indirectly derived from feed after first being synthesised by bacteria in the rumen. Once present at the udder, LCFAs enter the mammary cells bound to fatty acid binding protein (*FABP*). Before uptake, most LCFAs have been saturated by the rumen microorganisms. A fraction of these fatty acids is further desaturated by $\Delta^9$-desaturase to their *cis-9* monounsaturated counterparts by the enzyme stearoyl-CoA desaturase (*SCD*). Once inside the cell, the fatty acids are activated (i.e. adding a coenzyme A) by a coordinate activity between fatty acid translocase (*CD36*), fatty acid transporter (*SLC27A*), and acetyl-CoA synthetase (*ACSL*) (genes not shown in Figure 1.1).

The next step is shared by both feed derived, and *de novo* synthesised acids, where Acyl-CoA synthetase activate the fatty acids before they enter the triacylglyceride synthesis pathway. In this pathway, the fatty acids are attached to a glycerol 3-phosphate backbone in several steps catalysed by the enzymes glycerol-3-phosphate acyltransferase (*GPAM*), 6-acylglycerol-3-phosphate O-acyltransferase (*AGPAT6*), lipin (*LPIN1*) and diacylglycerol O-acyltransferase (*DGAT1*). Once formed, the triacylglycerides are inserted into the intra-leaflet of the endoplasmic reticulum membrane, forming lipid droplets coated with proteins and polar lipids. Upon secretion from the cell to the milk, the lipid droplets are enveloped with the cell plasma membrane. This plasma membrane called the milk fat globule membrane consists mainly of polar lipids and membrane-bound proteins. The size and composition of the milk fat globule membrane have impact on the stability and technological properties of milk (Lindmark Månsson 2008).

All these fatty acid metabolism steps are catalysed and regulated by a network of genes encoding a set of enzymes, transcription regulators and nuclear factors. Among the key regulators are nuclear receptor coactivator 6 (*NCOA6*), peroxisome proliferator activated receptor gamma (*PPARG*), insulin induced gene 1 (*INSIG1*) and sterol regulatory element binding transcription factor 1 (*SREBF1*) (Bionaz & Loor 2008).

## 1.4 Milk fat and human health.

Cow´s milk and milk derived dairy products constitute approximately 20 percent of the total fat consumed in a typical western diet. Health authorities in many countries advise people to reduce dietary saturated fat (Montagnese et al. 2015; Mozaffarian & Ludwig 2015), and since the fraction of SFA in bovine milk can be as high as 70 percent, peoples perception of milk and dairy products has developed unfavourably in recent decades.

While some epidemiological studies have indicated a protective effect of milk against coronary heart disease, stroke, diabetes and certain cancers (Haug et al. 2007), there is also evidence for adverse effects of individual fatty acids. SFAs with 14 or 16 carbons (C14:0 and C16:0) have been associated with increased low-density lipoprotein cholesterol levels (German and Dillard, 2006) which in turn are associated with increased risk of cardiovascular disease (Mensink et al. 2003). In contrast, SFAs shorter than C12 and longer than C18 are considered to have neutral or positive effects on cardiovascular diseases (Mensink et al. 2003). Among LCFAs, particular attention has been given to the conjugated linoleic acids (CLAs) and the omega-6:omega-3 ratio (Haug et al. 2007). CLAs are interesting because of their supposed role in plasma lipid modulation, anti-carcinogenic and anti-inflammatory effects (Haug et al., 2007). Western diets are believed to have an unfavourable high omega-6 to omega-3 ratio (10:1 – 14:1) linked to heart disease and insulin resistance (Bartsch et al. 1999). Bovine milk, on the other hand, can have a ratio close to the optimal 2:1, depending on feeding regime (Thorsdottir et al. 2004).

The conclusion concerning the health effects of milk fat, especially on cardiovascular disease, has yet to be drawn. Nevertheless, increased understanding of milk fat synthesis and its heritable component can be used to optimise the lipid profile of milk products.

## 1.5 The genetic basis of bovine milk fat composition.

Trait heritability measures the fraction of a trait’s phenotypic variation that is due to genetics. Previous heritability estimates for fatty acid concentration (g fat/100g fat)  range from 20 to 70 percent depending on breed, season, and fatty acids investigated (Soyeurt et al. 2007; Bobe et al. 2008; Stoop et al. 2008; Garnsworthy et al. 2010; Krag et al. 2013).

Experimental strategies to identify genetic variants associated with a trait like fatty acid composition include candidate gene studies and genome-wide association studies (GWAS). Candidate gene studies examine genetic variants of pre-selected genes for association to fatty acid concentrations. One example is the detection of a single nucleotide polymorphism (SNP) within the diacylglycerol O- acyltransferase 1 (*DGAT1*) gene shown to explain more than 50 percent of the genetic variance of milk fat percentage (Grisart et al. 2002). Another example is the detection of variants within *DAGT1* affecting the fatty acid indexes mono-unsaturated fatty acids (MUFA) and MCFAs (Roy et al. 2006; Morris et al. 2007; Rincon et al. 2012).

While the candidate gene approach relies on pre-existing biochemical knowledge, GWAS provide a way to identify chromosome regions affecting a trait of interest without any prior understanding of underlying biology or associated genes (Goddard & Hayes 2009). In a GWAS, one searches for associations between SNPs evenly distributed throughout the genome (e.g. 50,000) and trait animal recordings, preferably in the thousands. The success of GWAS relies on the existence of linkage disequilibrium (LD) between the causative genetic variants and those variants empirically tested in the experiment.

For milk fatty acid composition, previous GWAS have reported multiple significant regions, called quantitative trait loci (QTL). Stoop et al. (2009) found significant QTLs affecting short- and medium chained fatty acids on Bos taurus autosome (BTA)6, 14, 19 and 26, and suggestive QTLs on 21 other chromosomes. The same group revealed significant QTLs associated with LCFA on BTA14, 15 and 16 as well as suggestive QTLs on 16 additional chromosomes (Schennink et al. 2009), indicating that fatty acid composition is affected by many genes (i.e. being polygenic as opposed to monogenic or oligogenic). While the importance of *DGAT1*, *SCD1*, and *FASN* was confirmed in these and subsequent studies (Bouwman et al. 2011; Bouwman et al. 2014), genes with previously unknown effects have also been revealed. Duchemin et al. (2014) found a highly significant region on bovine chromosome 17 affecting *de novo* synthesised fatty acids, which included the progesterone receptor membrane component 2 (*PGRMC2*) gene not previously described in the context of milk fatty acid composition.

### 1.6 Identifying putative causative variants.

While the GWAS approach efficiently identifies both novel and known loci affecting milk fatty acid concentration, our ability to identify the underlying causal variants is hampered by the long-range LD found in most modern cattle breeds with low effective population size (Ne). Low Ne is caused by intensive historical selection (Sodeland et al. 2011; Kemper et al. 2015), and leave long, unrecombined segments of DNA to segregate in the population. Long-range LD makes GWAS with relatively low-density marker maps possible, but at the same

time makes it challenging to separate underlying causal variants from other variants co-segregating with the QTL  (Goddard & Hayes 2009; Sodeland et al. 2011). In response to this, researchers have adopted an approach where a selection of key reference animals are genotyped using costly high-density SNP-arrays, while the remainder of the population is genotyped using affordable mid-density (e.g. 50K) arrays. In this instance, the issue of LD becomes an asset, enabling genotypes of the high-density SNP markers to be imputed throughout the mid-density genotyped samples (Scheet & Stephens 2006) thereby creating an opportunity to identify markers in closer LD to the causal variant and reducing the list of potential causative genes.
A natural extension of imputation from SNP-array genotypes is to use resequencing data as a source of SNP loci and genotypes in the reference animals. Even the highest density arrays are limited to containing only a fraction of the factual SNPs in a bovine genome, and many novel, breed specific markers, or low frequency markers are likely to be missing from a consortium developed commercial array. In the 1000 bull genomes project (Daetwyler et al. 2014), partners have volunteered re-sequencing data from, at the time of writing, more than 2000 cattle. The intention for this data is that it may serve as a multi-breed reference allow partners to obtain (impute) full genome sequence for bulls and cows within their study population that have been genotyped with SNP-arrays (Goddard 2017).

Although the principles on which imputation is based are relatively simple, factors such as imputation errors, statistical sampling errors and extensive LD make the tests uncertain, and it is necessary to filter the result based on the likely functional effect of each SNP. Various pipelines exist which can predict whether a SNP can lead to a frameshift mutation, introduction of a stop-codon, change an amino-acid, or reside within a region potentially involved in promoter activity (e.g. the Ensembl variant effect predictor; VEP (McLaren et al. 2016). Indeed, while SNPs that change the protein sequence are obvious targets of associations studies, most significant variants found in GWAS does not alter proteins, but rather the expression levels of the gene. The Functional Annotation of Animal Genomes project (FAANG) aims to produce comprehensive maps of functional elements in domesticated animal species genomes (Andersson et al. 2015) and promises to provide a basis for the regulatory annotation of candidate variants. Beyond the predictive modelling performed by VEP and enabled by FAANG data, the analysis of data from RNA-sequencing and proteomics can endorse causal variants and allow us to discard non-causal, co-segregating variants.

## 1.7 FTIR calibration and measurement of milk fat composition

Any effort to improve our understanding of a trait's genetics would benefit from a fast and inexpensive method of phenotyping. Fatty acid profiling is usually done by gas chromatography (GC). However, while this method is accurate, it is also time-consuming and expensive, and therefore not so well suited for high throughput screening. An alternative approach is to use Fourier-transform infrared spectroscopy (FTIR) for fatty acid profiling of milk samples. This is a fast and inexpensive method already routinely used in the dairy industry to quantify milk components such as fat and protein percentage, casein contents, lactoferrin and antibiotics (Afseth et al. 2010). Soyeurt et al. (2006) demonstrated that the most frequent fatty acids in cow's milk could be predicted with acceptable accuracy using calibration equations developed utilising pairwise GC and FTIR measurements.

The equations are founded on the absorption of infrared light at specific wavelengths is proportional to the concentration of a given fatty acid in the sample. FTIR analysis of a milk sample yields a spectrum of absorption signal which is mathematically converted to interpretable spectral data using Fourier-transformation, which enable the spectra to represent the absorptions at different wavenumbers ($cm^{-1}$) for each distinct fatty acid chemical composition of the sample analysed (Coates 2000). Since 2006, several studies investigating milk fatty acid composition applied this quantification method (Soyeurt et al. 2007; Rutten et al. 2009; Afseth et al. 2010; Wang et al. 2016), which confirms its potential for use in regular milk recording. With fatty acid composition estimates for each cow, it becomes possible to quantify the genetic contribution to fatty acid concentration and facilitate genome-based selection to improve the nutritional quality of cow milk.

## 2. Methodological overview

A schematic representation of the methodological workflow underlying much of this thesis is presented in Figure 1.3. Between 2007 and 2014, more than 8 million FTIR recordings were obtained from routine milk samplings conducted in Norway and stored in a relational database management system. Using pairwise GC and FTIR measurements, calibration models were developed and applied to predict fatty acid profiles for all milk samples. The predicted fatty acid profiles were used further to calculate daughter yield deviations (DYDs) for progeny tested bulls.  A DYD value describes the average performance of a sire's daughters corrected by their environmental and other non-genetic effects (Szyda et al. 2008). After obtaining high-resolution genotypes for the bulls with imputation, the DYDs were used in a GWAS to identify chromosome regions, genes and genetic variants associated with variation in milk fatty acid composition.

DYDs were calculated for 991 bulls in Paper I and 1811 bulls in Papers II and III. In addition to calculating fatty acid DYDs, the heritability of each fatty acid was estimated.

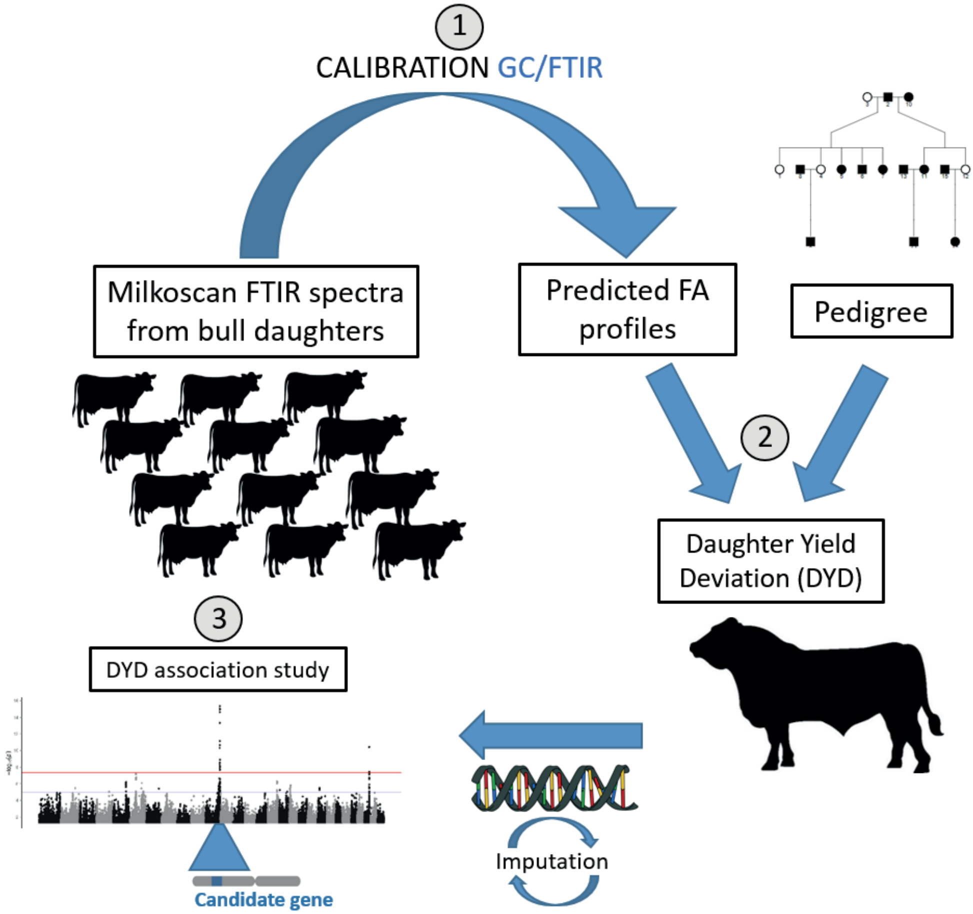


**Figure 1.3 Schematic representation of how milk FTIR data was utilised for prediction of fatty acid composition and GWAS.** 1). Development of fatty acid calibrations using GC measurements and FTIR spectra from milk samples for the prediction of fatty acid profiles. 2). Fatty acid heritabilities and bull DYDs calculated using the fatty acid profiles of the bull's daughters and pedigree. 3) Association studies using imputed genotypes from the bulls and the DYDs for individual milk fatty acids as phenotype.

# 3. Aims of the thesis

The primary objective of this thesis was to identify DNA-variation underlying bovine milk fat composition. The primary objective can be subdivided into the following specific aims:

1. Utilise a national database containing more than 8 million FTIR spectra to predict fatty acid phenotypes for GWAS and estimate fatty acid trait heritabilities. (Paper I)
2. Evaluate the FTIR-based fatty acid calibrations in context of genome-based improvement of milk fat composition by assessing the quality of the calibration equations developed. (Paper II)
3. Fine-map associated variants and identify candidate genes and causative variants underlying the observed variation in milk fatty acid levels, using whole genome sequence imputation, gene expression data and milk protein level measurements. (Papers II and III)

# 4. Brief summary of Papers I-III

### Paper I

In Paper I, milk fatty acid composition was predicted from the nation-wide recording scheme using Fourier transform infrared (FTIR) spectroscopy data and applied to estimate heritabilities for 24 individual and 12 combined fatty acid traits.

Twenty-nine traits had a prediction accuracy in the form of $R^2$CV above 0.5 which we considered sufficient for further analysis. Heritability estimates for the studied traits ranged from 0.09 for C18:1trans-11, to 0.35 for C4:0. Short and medium length fatty acid were somewhat more heritable than longer and unsaturated fatty acid, while heritability for the polyunsaturated index (PUFA) was slightly higher than that of the MUFA and saturated (SAT) indexes, being 0.171, 0.130 and 0.137, respectively.

The recordings were used to generate daughter yield deviations that were first applied in a GWAS with 17,343 markers to identify QTL affecting fatty acid composition. The GWAS revealed 200 significant associations, with the strongest QTLs located on BTA1, 13 and 15. The results on BTA13 were followed up with high-density genotyping and sequence data. The most significant signals were found close to *ACSS2*, which is considered a good functional candidate gene for *de novo* synthesis of short- and medium-chained SFAs. The fine-mapping identified another nearby candidate gene, *NCOA6*. *NCOA6* is a nuclear receptor known to interact with transcription factors such as *PPARγ*, which is a master regulator of bovine milk fat synthesis.

### Paper II

In Paper II, we sought to explore the genetic basis of *de novo* synthesis by doubling the number of predicted fatty acid recordings for the GWAS and utilising whole genome sequence data from 153 Norwegian Red cattle. Most of the sequenced animals were elite sires; key ancestors of the Norwegian Red cattle population. BTA 11, 13, 17 and 19 were imputed to sequence density for 1811 elite artificial insemination (AI) bulls and significant regions from the initial SNP array-based GWAS were fine mapped. RNA-sequence data obtained from somatic cells in milk were used to assess expression of the candidate genes in the mammary gland. The results of the GWAS and subsequent fine mapping using sequence imputed genotypes, revealed the involvement of the genes *PAEP* on BTA11, *AACS* on

BTA17, *NCOA6* or *ACSS2* on BTA13 and *FASN* on BTA19. Among these, polymorphisms in *PAEP* and *AACS* seem to mostly affect *de novo* synthesis of the shortest acids (C4:0-C6:0), in *NCOA6* or *ACSS2* the synthesis of fatty acids of intermediate chain lengths (C6:0-C14:0), and variants in *FASN* to affect the longest acid (C14:0). In all cases, the effect of the underlying polymorphism was expected to regulate gene expression rather than changing the amino acid sequence. Expression analyses performed on mRNA isolated from milk samples revealed that all suggested candidate genes were expressed in the bovine mammary gland during lactation.

### Paper III

C16:0 is the predominant SFA in milk, and it may be possible to counteract its implied adverse health effects by replacing it with higher levels of unsaturated fatty acids, such as C18:1*cis*-9. Paper III utilizes dense genotyping, whole genome sequence data, mRNA transcript profiling and protein analyses to reveal genetic variants underlying levels of C16:0 and C18:1*cis*-9. The initial whole genome scan exposed significant associations on 17 chromosomes. We further dissected a strong QTL located at ≈103 Mb on chromosome 11 showing opposite effects on the content of C16:0 and C18:1. The QTL region covered a tightly linked cluster of significant genetic variants in coding and regulatory regions of *PAEP*. The favourable haplotype, linked to reduced levels of C16:0 and increased C18:1*cis*-9, was also associated with a marked reduction in *PAEP* gene expression and levels of protein. *PAEP* encodes *β*-lactoglobulin, an abundant milk protein, whose level in milk affect important dairy production parameters such as cheese yield. The genetic variants detected in this paper can potentially be exploited in breeding programs to lead to milk with a healthier fatty acid profile and positive effect on cheese making properties.

# 5. Discussion

## 5.1 Predicting individual fatty acid profiles using FTIR data

A keystone methodology of this thesis was the use of large-scale FTIR-data to predict milk fatty acid composition. More than 4 million fatty acid profiles from ≈640,000 cows were generated after applying calibration models to infrared spectra collected as part of the Norwegian Dairy Herd recording system between 2007 and 2014. The calibration equations were produced from ≈900 milk samples measured with both FTIR and GC. The equations were developed using the  partial least squares regression method by Indahl et al., (2005), which utilises all the spectral data for the calibration and takes the covariance between the predictor (spectral data) and response variables (GC-measured fatty acid compositions) into account when the models are established (Frank et al. 1984; Martens & Næs 1989).

The optimal number of informative components used in the equation was determined using 20-fold cross-validation. As shown in Paper I, applying the calibration equations on all FTIR/GC measured sample pairs, 18 of the 21 individual fatty acid achieved prediction accuracies ($R^2$CV) above 0.5. Paper III focus on *de novo* synthesised C16:0 and C18:1*cis*9 fatty acid, which all have had $R^2$CV well above 0.7. Together, fatty acids with $R^2$CV exceeding 0.7 represented more than 70 percent of the total fat content in the reference samples. Paper I therefore conclude that the majority of milk fat components could be satisfactorily predicted from FTIR data.

Although not significantly investigated in papers I-III, the $R^2$CV for poly- and mono-desaturation indexes might be of particular interest, as both these indexes could serve as markers for milk with properties beneficial to heart health (Haug et al. 2007; Hooper et al. 2015). The MUFA index had an $R^2$CV of 0.96, while the PUFA index was 0.72. The MUFA index seems heavily influenced by C18:1cis-9 ($R^2$CV = 0.94), which constitute about 80 percent of it, while the PUFA index may benefit from the effect of grouped measurements since the $R^2$CV of the index exceeds that achieved for any of the individual fatty acids it contains. While Paper I encompassed a range of fatty acids, Papers II and III explore specific classes of fatty acids in more detail. A consistent finding was that short and medium *de novo* synthesised fatty acid (C6:0-C14:0) and the SAT and MUFA group indexes were all well predicted with a $R^2$CV above 0.86.

Levels of the *de novo* synthesised acids are known to be highly correlated, which seems logical since they are all products of the same reaction governed by the multifunctional enzyme *FASN*.

The high internal correlations yield stronger signals in the spectral data for each fatty acid and give better predictions than if they were independent of each other. Afseth et al. (2010) noted that if these internal correlations were stable also for future samples, they could be utilised to improve prediction equations.

The concentration of milk fatty acids is affected by the total milk fat percentage in the sample (Eskildsen et al. 2014). This relationship can lead to fatty acid predictions being influenced by a sample's total fat percentage, rather than reflecting the true concentration of each acid (Soyeurt et al. 2006). To account for this, we assessed fatty acid concentrations as percentages of total fat instead of gram-acid-per-unit-of-milk. As a result, predicted fatty acid levels were more effectively disconnected from total fat percentage with no individual correlations exceeding 0.3. In none of the cases were the $R^2$CV of a single fatty acid or index higher than the squared correlations between total fat and the trait, which suggest that the predicted concentrations were due to real absorbance values specific to the fatty acid (Soyeurt et al. 2006; Paper I: Table 1). Furthermore, we observed a general trend for long unsaturated fatty acids to be negatively correlated to total fat and short- and medium-chain fatty acids to be positively correlated to total fat. This is supported by literature claiming that a diet rich in polyunsaturated fatty acids affect the cow's ability to synthesise fatty acids *de novo* (MacLeod et al. 2016).

The trait heritabilities we obtained were in general somewhat lower than those reported by other studies using infrared spectroscopy (Soyeurt et al. 2007; Stoop et al. 2008; Bastin et al. 2013; Lopez-Villalobos et al. 2014). Still, the reported heritabilities of these studies vary considerably and factors such as sample size, breed, and chosen mathematical model, which may explain some of this discrepancy. The estimates of predictability ($R^2$CV) and heritabilities presented in Paper I largely agree with what has been reported elsewhere and most major fatty acids were considered predictable and showed substantial heritability. Our results underline that, with the widespread use of FTIR instruments and their speed and efficiency considered, FTIR data coupled with modern genomics tools can provide ways to genetically improve milk fat composition as well as to identify milk fatty acid QTLs using GWAS.

Even though most major fatty acids were predicted with high accuracy, the methodology did not provide satisfactory prediction equations for fatty acids present in low concentrations. Thus, improved calibration methods are needed to quantify the whole range of fatty acid composition in bovine milk. Afseth et al. (2010) showed that milk samples on dry-film could be used to produce feasible calibrations ($R^2$CV from 0.78 to 0.93) for the low concentration fatty acids such as CLA (18:2cis-9, trans-11), PUFA, and the summed 18:1transisomers. They conclude that it is possible to perform dry-film measurements in mass scale, but the method is not implemented in the Norwegian Dairy Herd Recording system. While Afseth et al. (2010) showed potential for enhanced FTIR measurements, others have demonstrated that calibration models can be improved by preselecting informative wavelengths and thus avoiding errors linked the spectra (Ferrand-Calmels et al. 2014).

### 5.2 Candidate genes in light of fatty acid metabolism

Milk fatty acid metabolism is a complex process involving multiple pathways, transcription factors and enzymes. Paper I focused on a wide array of short and long-chained, branched and unbranched acids. We found that the relatively frequent short and medium chained SFA were predicted most accurately. Paper I and II focus on identifying the genes involved in the synthesis of fatty acids C4:0 to C14:0. Paper III, focus on C16:0 and C18:1*cis*9. These are fatty acids derived mainly from circulating blood lipids, which may suggest the involvement of genes related to transportation and cellular trafficking.

The most prominent candidate genes for fatty acid composition detected in this thesis were *PAEP* on BTA11 (discussed in papers II and III), *NCOA6* and/or *ACCS2* on BTA13 (papers I and II), *AACS* on BTA17 (Paper II) and *FASN* on BTA19 (Paper II). Variants within *AACS* showed the strongest association to the short fatty acids C4:0 and C6:0. Polymorphisms within *PAEP* were also associated with levels of C4:0 but, in addition, associated with the inverse effect relationship seen for C16:0 and C18:1*cis*-9. Variants in *NCOA6* or *ACSS2* were related to synthesis of acids with intermediate chain lengths (especially C8:0), while the *FASN* variants were associated with levels of the longest DNS fatty acids (C14:0). All these genes have largely defined roles in bovine milk fat synthesis, and operate across the core pathways responsible for DNS and triacylglycerol (TAG) metabolism (Figure 1.2). Early in DNS, *ACSS2* facilitates the conversion of acetate to acetyl-CoA (Bionaz & Loor 2008). Alternatively, acetyl-CoA may be derived from acetoacetyl-CoA in the process beginning with the production of acetoacetate-CoA from acetoacetate by *AACS* (Buckley & Williamson 1975). Later, *FASN* oversees a process whereby palmitate (C16:0) is synthesised from acetyl-CoA and malonyl-CoA in a repeated, cyclic reaction. Importantly, intermediate length acids (C4:0 to C14:0) can leave the elongation cycle before the chain reaches full length (Knudsen & Grunnet 1982). The entire lipid synthesis machine is regulated by a network of genes encoding transcription factors and nuclear receptors. One of these, *peroxisome proliferator-activated receptor gamma* (*PPARG*), is a well-described transcriptional regulator affecting lipid storage (Bionaz & Loor 2008; Liu et al. 2016), while *NCOA6*, being a ligand for *PPARG* and *PPARA* (Caira et al. 2000; Lemay et al. 2007) is a transcriptional coactivator enhancing the activity of, among other things, *PPARG*. *PAEP* encodes the milk protein β-lactoglobulin which is abundant in bovine milk. Although the effect of *PAEP* alleles on several milk production traits including fat yield and fat percentage is well documented (Tsiaras et al. 2005; Berry et al. 2010), *PAEP*s role in milk fat synthesis is poorly understood. β-lactoglobulin bind both saturated and unsaturated fatty acids in vitro, especially C16:0, which may suggest a role in fatty acid transport.

For all detected candidate genes, most of the top-ranking variants were found in putatively regulatory regions such as the promoter, in untranslated regions, or in regions of uncertain function such as introns and intergenic regions. The only exceptions are the two nonsynonymous SNPs within *PEAP* encoding the well characterised A and B protein variants of the β-lactoglobulin protein. As shown in Paper III, for Norwegian red cattle (and well documented in other breeds) these SNPs are in strong LD with several variants in the *PEAP* promoter. Considering this fact together with the large transcription and protein level differences seen between haplotypes, and presented in Paper III, we propose that the effects of *PEAP* are caused by variants within regulatory regions rather than by variants within the protein coding region.

The amount of data material, especially the marker density, increased markedly from Paper I to Papers II and III. While the GWAS of Paper I involved only 17,000 SNPs, more than 600,000 markers were included in the GWAS of Papers II and III. Despite this, the findings in the three papers are quite similar. QTLs on BTA13 and BTA17 were detected in the same region in both Papers I and II. In Paper I we first proposed the BTA13 QTL to be caused by variants within ACSS2 but later fine-mapped it to NCOA6. In Paper II, the QTL was mapped to a region that spanned both these two genes, but we were not able to identify the underlying causal variant or variants. The reason for this is somewhat unclear since the LD among the significant markers were not particularly high. Further, the QTL affecting C4:0 at AACS was also mapped to approximately the same position in Paper I as in Paper II.

In contrast, the QTLs located near *PAEP* and *FASN* were not detected in Paper I. We believe this is most likely because of the lower marker density and fewer animals with phenotypes (≈900 vs ≈1800) used in the first paper. Paper I on the other hand reports associations for several *de novo* synthesized acids close to very interesting functional and positional candidate genes on BTA1 and 15, but these were not confirmed in Paper II and Paper III. A possible explanation is that as the number of tests increases, so does the significance threshold, leading to these variants being filtered as non-significant markers in Paper II.

Papers II and III included DYD estimates using spectra from a much larger number of cows compared to Paper I. The number of genotyped bulls with DYDs was doubled, and the marker density of Paper I was a fraction (<3%) of that used in the GWA studies of papers II and III. With these differences in mind, we conclude that agreement among the three papers was good.

The results presented in this thesis are generally well supported by literature. Previous studies have found a QTL near *ACSS2* and *NCOA6* with effect on *de novo* synthesis of C6:0, C8:0 and C10:0 in Dutch Holstein Friesian (Bouwman et al. 2011) and in Danish Jersey cattle (Buitenhuis et al. 2014). The same region has also been associated with several C16 and C18 fatty acids in Chinese Holstein (Li et al., 2014). Several authors have reported significant associations within or near *FASN* on BTA19 (Bouwman et al. 2014; Li et al. 2014). *FASN* is an obvious candidate gene because of its well documented role in fat synthesis and has been extensively studied in candidate gene studies for fat content in milk and adipose tissue (Roy et al. 2006; Zhang et al. 2008; Abe et al. 2009; Schennink et al. 2009; Li et al. 2012; Oh et al. 2012). PAEP is a novel candidate gene in the context of milk fatty acid composition in cattle, but variants of β-lactoglobulin was found to affect the concentration of C16:0 and other fatty acids in dairy ewes (Mele et al. 2007), as well as associated to a number of milk traits in cattle, including total fat yield and fat percentage (Tsiaras et al. 2005; Berry et al. 2010).

The fact that AACS and *PAEP* have yet not been detected in GWA studies focusing on bovine milk fatty acids might have several explanations. Breed differences between Norwegian Red cattle and breeds studied in other studies will affect the allele frequencies of the underlying causal polymorphisms. Hence, regulatory SNPs in LD with the *PAEP* protein variants in Norwegian Red cattle might be fixed in Holsteins for example. Another aspect is the wide array of methods used. For instance, may a small difference in significance levels cause an association to be detected in one study and not in others.

Most genome scans performed in other cattle breeds have reported strong associations between milk fatty acids and the genes *diacylglycerol acyltransferase 1* (*DGAT1*) on BTA14 and *stearoyl-coenzyme A desaturase 1* (*SCD*) on BTA26. *DGAT1* encodes an enzyme that catalyses the final stage of triacylglyceride synthesis (Cases et al. 1998), while *SCD* on BTA26 is involved in the synthesis of monounsaturated fatty acids by introducing a double bond in the delta-9 position of C14:0 and C16:0, primarily, thus producing the *cis*-9 variant of these acids (Ntambi & Miyazaki 2003). No genome-wide significant associations were detected near these genes in our studies of Norwegian Red cattle. We have not found any animals that carry the K variant of the *DGAT1* K232A polymorphism (unpublished results), suggesting that the K2342A polymorphism is missing in the Norwegian Red population. The *SCD* A293V polymorphism that is the suspected causal variant (Schennink et al. 2008) does segregate in our breed, but this SNP was not significantly associated with any fatty acid in our studies. However, C14:1 and C16:1 were poorly predicted by our FTIR approach, which most likely hindered the possibility to detect significant associations for these fatty acids.

## 6. Concluding remarks and future perspectives

A critical goal of the current thesis was to develop an efficient workflow to facilitate genome-based selection for fatty acid composition in Norwegian Red cattle. FTIR data is, as of today, routinely gathered as a part of the national milk recording system in Norway. Even though there is room for improvement to the presented prediction qualities, we believe the work presented in this thesis has shown that millions of FTIR-predicted fatty acid profiles gathered over several years can serve as a fast and inexpensive method that, coupled with high-density genotype data, can be implemented to breed for improved milk fatty acid composition in Norway. Moreover, we have shown that the methodology can contribute to the biological understanding of milk fat metabolism in cattle, and with proper management of the spectral database, will continue to do so in the future.

If we assume that the increasing availability of high-quality sequence data will enable the identification of large proportions of the quantitative traits causal variants, it may also be possible to specifically improve breeding schemes by gene editing techniques like the CRISPR-Cas9 system.  Furthermore, today's sequencing technologies are versatile and can be used for both quantitative and qualitative analysis of the transcriptome, and for DNA-methylation analysis, while other technological refinements have significantly improved accuracy and precision of high-resolution proteome quantification. In combination with genome information, supplementary functional genomics data will contribute to a more complete understanding of the biology underlying milk fatty acid composition.

# REFERENCES

Abe, T. et al., 2009. Novel mutations of the FASN gene and their effect on fatty acid composition in japanese black beef. *Biochemical Genetics*, 47(5–6), pp.397–411.

Afseth, N.K. et al., 2010. Predicting the Fatty Acid Composition of Milk: A Comparison of Two Fourier Transform Infrared Sampling Techniques. *Applied Spectroscopy*, 64(7), pp.700–707. Available at: http://asp.sagepub.com/lookup/doi/10.1366/000370210791666200 [Accessed May 26, 2017].

Andersson, L. et al., 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology*, 16(1), p.57. Available at: http://genomebiology.com/2015/16/1/57 [Accessed May 26, 2017].

Barłowska, J. et al., 2011. Nutritional value and technological suitability of milk from various animal species used for dairy production. *Comprehensive Reviews in Food Science and Food Safety*, 10(6), pp.291–302.

Bartsch, H., Nair, J. & Owen, R.W., 1999. Dietary polyunsaturated fatty acids and cancers of the breast and colorectum: emerging evidence for their role as risk modifiers. *Carcinogenesis*, 20(12), pp.2209–2218. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10590211.

Bastin, C., Soyeurt, H. & Gengler, N., 2013. Genetic parameters of milk production traits and fatty acid contents in milk for Holstein cows in parity 1 - 3. *Journal of Animal Breeding and Genetics*, 130(2), pp.118–127.

Berry, S.D. et al., 2010. Mapping a quantitative trait locus for the concentration of beta-lactoglobulin in milk, and the effect of beta-lactoglobulin genetic variants on the composition of milk from Holstein-Friesian x Jersey crossbred cows. *New Zealand veterinary journal*, 58(1), pp.1–5.

Bionaz, M. & Loor, J.J., 2011. Gene networks driving bovine mammary protein synthesis during the lactation cycle. *Bioinformatics and Biology Insights*, 5, pp.83–98.

Bionaz, M. & Loor, J.J., 2008. Gene networks driving bovine milk fat synthesis during the lactation cycle. *BMC genomics*, 9, p.366.

Bobe, G. et al., 2008. Estimates of genetic variation of milk fatty acids in US Holstein cows. *J Dairy Sci*, 91. Available at: https://doi.org/10.3168/jds.2007-0252.

Bouwman, A.C. et al., 2014. Fine mapping of a quantitative trait locus for bovine milk fat composition on Bos taurus autosome 19. *Journal of dairy science*, 97(2), pp.1139–49. Available at: http://www.sciencedirect.com/science/article/pii/S0022030213008187.

Bouwman, A.C. et al., 2011. Genome-wide association of milk fatty acids in Dutch dairy cattle. *BMC Genetics*, 12(1), p.43. Available at: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-12-43.

Buckley, B.M. & Williamson, D.H., 1975. Acetoacetyl-CoA synthetase; a lipogenic enzyme in rat tissues. *FEBS Letters*, 60(1), pp.7–10.

Buitenhuis, B. et al., 2014. Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. *BMC Genomics*, 15(1), p.1112. Available at: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-1112.

Caira, F. et al., 2000. Cloning and characterization of RAP250, a novel nuclear receptor coactivator. *The Journal of biological chemistry*, 275(8), pp.5308–17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10681503.

Cases, S. et al., 1998. Identification of a gene encoding an acyl CoA:diacylglycerol acyltransferase, a key enzyme in triacylglycerol synthesis. *Proceedings of the National Academy of Sciences*, 95(22), pp.13018–13023. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.95.22.13018.

Coates, J., 2000. Interpretation of infrared spectra, a practical approach. *Encyclopedia of analytical chemistry*.

Daetwyler, H.D. et al., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, 46(8), pp.858–865. Available at: http://www.nature.com/doifinder/10.1038/ng.3034 [Accessed January 30, 2017].

Duchemin, S.I. et al., 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. *Journal of dairy science*, 97(11), pp.7276–85. Available at: http://www.sciencedirect.com/science/article/pii/S0022030214006377.

Eskildsen, C.E. et al., 2014. Quantification of individual fatty acids in bovine milk by infrared spectroscopy  and chemometrics: understanding predictions of highly collinear reference variables. *Journal of dairy science*, 97(12), pp.7940–7951.

Ferrand-Calmels, M. et al., 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. *Journal of Dairy Science*, 97(1), pp.17–35. Available at: http://linkinghub.elsevier.com/retrieve/pii/S002203021300790X.

Frank, I.E. et al., 1984. Prediction of Product Quality from Spectral Data Using the Partial Least-Squares Method. *Journal of Chemical Information and Computer Sciences*, 24(1), pp.20–24.

Garnsworthy, P.C. et al., 2010. Short communication: Heritability of milk fatty acid composition and stearoyl-CoA desaturaes indices in dairy cows. *J Dairy Sci*, 93. Available at: https://doi.org/10.3168/jds.2009-2695.

Goddard, M.E., 2017. Can we make genomic selection 100% accurate? *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie*, 134(4), pp.287–288.

Goddard, M.E. & Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature reviews. Genetics*, 10(6), pp.381–391.

Grisart, B. et al., 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 research*, 12(2), pp.222–231.

Haug, A., Høstmark, A.T. & Harstad, O.M., 2007. Bovine milk in human nutrition – a review. *Lipids in Health and Disease*, 6(1), p.25. Available at: http://lipidworld.biomedcentral.com/articles/10.1186/1476-511X-6-25.

Hooper, L. et al., 2015. Reduction in saturated fat intake for cardiovascular disease. *The Cochrane database of systematic reviews*, (6), p.CD011737.

Indahl, U., 2005. A twist to partial least squares regression. *Journal of Chemometrics*, 19(1), pp.32–44.

Jensen, R.G., 2002. The composition of bovine milk lipids: January 1995 to December 2000. *Journal of dairy science*, 85(2), pp.295–350. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022030202740794.

Kemper, K.E. et al., 2015. How old are quantitative trait loci and how widely do they segregate? *Journal of Animal Breeding and Genetics*, 132(2), pp.121–134. Available at: http://doi.wiley.com/10.1111/jbg.12152 [Accessed May 26, 2017].

Knudsen, J. & Grunnet, I., 1982. Transacylation as a chain-termination mechanism in fatty acid synthesis by mammalian fatty acid synthetase. Synthesis of medium-chain-length (C8-C12) acyl-CoA esters by goat mammary-gland fatty acid synthetase. *The Biochemical journal*, 202(1), pp.139–43. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1158083&tool=pmcentrez&rendertype=abstract.

Krag, K. et al., 2013. Genetic parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a Bayesian approach. *BMC genetics*, 14, p.79.

Lemay, D.G. et al., 2007. Gene regulatory networks in lactation: identification of global principles using bioinformatics. *BMC Systems Biology*, 1(1), p.56. Available at: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-1-56.

Li, C. et al., 2012. Association analyses of single nucleotide polymorphisms in bovine stearoyl-CoA desaturase and fatty acid synthase genes with fatty acid composition in commercial cross-bred beef steers. *Animal Genetics*, 43(1), pp.93–97.

Li, C. et al., 2014. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. *PLoS ONE*, 9(5), p.e96186.

Lindmark Månsson, H., 2008. Fatty acids in bovine milk fat. *Food & Nutrition Research*, 52(1), p.1821. Available at: https://www.tandfonline.com/doi/full/10.3402/fnr.v52i0.1821.

Liu, L. et al., 2016. Regulation of peroxisome proliferator-activated receptor gamma on milk fat synthesis in dairy cow mammary epithelial cells. *In Vitro Cellular & Developmental Biology - Animal*, 52(10), pp.1044–1059. Available at: http://link.springer.com/10.1007/s11626-016-0059-4 [Accessed March 8, 2017].

Lopez-Villalobos, N. et al., 2014. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in New Zealand dairy cattle. *The Journal of dairy research*, 81(3), pp.340–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25052435.

MacLeod, I.M. et al., 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*, 17(1), p.144. Available at: http://www.biomedcentral.com/1471-2164/17/144.

Martens, H. & Næs, T., 1989. *Multivariate calibration*, John Wiley & Sons.

Maurice-Van Eijndhoven, M.H.T., Hiemstra, S.J. & Calus, M.P.L., 2011. Short communication: milk fat composition of 4 cattle breeds in the Netherlands. *Journal of dairy science*, 94(2), pp.1021–1025.

McLaren, W. et al., 2016. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), p.122. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4.

Mele, M. et al., 2007. Relationship between beta-lactoglobulin polymorphism and milk fatty acid composition in milk of Massese dairy ewes. *Small Ruminant Research*, 73(1–3), pp.37–44. Available at: http://www.sciencedirect.com/science/article/pii/S0921448806003117 [Accessed April 10, 2017].

Mensink, R.P. et al., 2003. Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: A meta-analysis of 60 controlled trials. *American Journal of Clinical Nutrition*, 77(5), pp.1146–1155. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12716665 [Accessed April 11, 2017].

Montagnese, C. et al., 2015. European food-based dietary guidelines: a comparison and update. *Nutrition (Burbank, Los Angeles County, Calif.)*, 31(7–8), pp.908–915.

Morris, C.A. et al., 2007. Fatty acid synthase effects on bovine adipose fat and milk fat. *Mammalian genome : official journal of the International Mammalian Genome Society*, 18(1), pp.64–74.

Mozaffarian, D. & Ludwig, D.S., 2015. The 2015 US Dietary Guidelines: Lifting the Ban on Total Dietary Fat. *JAMA*, 313(24), pp.2421–2422.

Ntambi, J.M. & Miyazaki, M., 2003. Recent insights into stearoyl-CoA desaturase-1. *Current opinion in lipidology*, 14(3), pp.255–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12840656.

Oh, D. et al., 2012. Fatty acid composition of beef is associated with exonic nucleotide variants of the gene encoding FASN. *Molecular Biology Reports*, 39(4), pp.4083–4090.

Rincon, G. et al., 2012. Polymorphisms in genes in the SREBP1 signalling pathway and SCD are associated with milk fatty acid composition in Holstein cattle. *The Journal of dairy research*, 79(1), pp.66–75.

Roy, R. et al., 2006. Association of polymorphisms in the bovine FASN gene with milk-fat content. *Animal Genetics*, 37(3), pp.215–218.

Rutten, M.J.M. et al., 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. *Journal of Dairy Science*, 92(12), pp.6202–6209. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022030209713396.

Scheet, P. & Stephens, M., 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics*, 78(4), pp.629–644.

Schennink, A. et al., 2009. Effect of polymorphisms in the FASN, OLR1, PPARGC1A, PRL and STAT5A genes on bovine milk-fat composition. *Animal Genetics*, 40(6), pp.909–916.

Schennink, A. et al., 2008. Milk fatty acid unsaturation: genetic parameters and effects of stearoyl-CoA desaturase (SCD1) and acyl CoA: diacylglycerol acyltransferase 1 (DGAT1). *Journal of dairy science*, 91(5), pp.2135–2143.

Sodeland, M. et al., 2011. Recent and historical recombination in the admixed Norwegian Red cattle breed. *BMC genomics*, 12, p.33.

Soyeurt, H. et al., 2006. Estimating Fatty Acid Content in Cow Milk Using Mid-Infrared Spectrometry. *Journal of Dairy Science*, 89(9), pp.3690–3695. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022030206724092.

Soyeurt, H. et al., 2007. Estimation of heritability and genetic correlations for the major fatty acids in bovine milk. *Journal of dairy science*, 90(9), pp.4435–4442.

Stoop, W.M. et al., 2008. Genetic parameters for major milk fatty acids and milk production traits of Dutch Holstein-Friesians. *Journal of Dairy Science*, 91(1), pp.385–394. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18096963.

Thorsdottir, I., Hill, J. & Ramel, A., 2004. Omega-3 fatty acid supply from milk associates with lower type 2 diabetes in men and coronary heart disease in women. *Preventive medicine*, 39(3), pp.630–634.

Tsiaras, A.M. et al., 2005. Effect of kappa-casein and beta-lactoglobulin loci on milk production traits and reproductive performance of Holstein cows. *Journal of dairy science*, 88(1), pp.327–334.

Wang, Q., Hulzebosch, A. & Bovenhuis, H., 2016. Genetic and environmental variation in bovine milk infrared spectra. *Journal of Dairy Science*, 99(8), pp.6793–6803. Available at: http://linkinghub.elsevier.com/retrieve/pii/S002203021630248X.

Zhang, S. et al., 2008. DNA polymorphisms in bovine fatty acid synthase are associated with beef fatty acid composition. *Animal Genetics*, 39(1), pp.62–70.

# Paper I

Olsen *et al. Genet Sel Evol (2017) 49:20*
DOI 10.1186/s12711-017-0294-5

Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

# 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

Hanne Gro Olsen[1*], Tim Martin Knutsen[1], Achim Kohler[2,3], Morten Svendsen[4], Lars Gidskehaug[5], Harald Grove[1], Torfinn Nome[1], Marte Sodeland[6,7], Kristil Kindem Sundsaasen[1], Matthew Peter Kent[1], Harald Martens[8] and Sigbjørn Lien[1]

## Abstract

**Background:** Bovine milk is widely regarded as a nutritious food source for humans, although the effects of individual fatty acids on human health is a subject of debate. Based on the assumption that genomic selection offers potential to improve milk fat composition, there is strong interest to understand more about the genetic factors that influence the biosynthesis of bovine milk and the molecular mechanisms that regulate milk fat synthesis and secretion. For this reason, the work reported here aimed at identifying genetic variants that affect milk fatty acid composition in Norwegian Red cattle. Milk fatty acid composition was predicted from the nation-wide recording scheme using Fourier transform infrared spectroscopy data and applied to estimate heritabilities for 36 individual and combined fatty acid traits. The recordings were used to generate daughter yield deviations that were first applied in a genome-wide association (GWAS) study with 17,343 markers to identify quantitative trait loci (QTL) affecting fatty acid composition, and next on high-density and sequence-level datasets to fine-map the most significant QTL on BTA13 (BTA for *Bos taurus* chromosome).

**Results:** The initial GWAS revealed 200 significant associations, with the strongest signals on BTA1, 13 and 15. The BTA13 QTL highlighted a strong functional candidate gene for de novo synthesis of short- and medium-chained saturated fatty acids; *acyl-CoA synthetase short-chain family member 2*. However, subsequent fine-mapping using single nucleotide polymorphisms (SNPs) from a high-density chip and variants detected by resequencing showed that the effect was more likely caused by a second nearby gene; *nuclear receptor coactivator 6* (*NCOA6*). These findings were confirmed with results from haplotype studies. NCOA6 is a nuclear receptor that interacts with transcription factors such as PPARγ, which is a major regulator of bovine milk fat synthesis.

**Conclusions:** An initial GWAS revealed a highly significant QTL for de novo-synthesized fatty acids on BTA13 and was followed by fine-mapping of the QTL within *NCOA6*. The most significant SNPs were either synonymous or situated in introns; more research is needed to uncover the underlying causal DNA variation(s).

*Correspondence: hanne-gro.olsen@nmbu.no
[1] Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, PO Box 5003, 1432 Ås, Norway
Full list of author information is available at the end of the article

## Background

While bovine milk is generally regarded as being highly nutritious for humans and serving as an important source of proteins, fat, minerals, vitamins and bio-active lipid components, the net effect of dairy fat on human health is strongly debated. This is because saturated fatty acids (FA), which constitute roughly 60 to 70% of the FA in milk, have been associated with cardiovascular disease and obesity, while mono-and polyunsaturated FA have been associated with positive effects on both cardiovascular health and diabetes (see e.g., [1] for a review).

Biosynthesis of bovine milk fat is a complex process, which is regulated by a network of genes that encode a set of transcription regulators and nuclear factors [2]. In essence, milk FA are derived via one of two major pathways: either by de novo synthesis in the mammary gland, or by direct transport from rumen to mammary gland through blood. More specifically, short- and medium-chained saturated FA C4:0–C14:0, as well as approximately 50% of C16:0, are synthesized de novo in the mammary gland from C2 and C4 precursors. The remaining C16:0 and long-chained saturated FA are derived from circulating plasma lipids which originate from the diet or from lipolysis of adipose tissue triacylglycerols. Long-chained FA are mainly saturated in the rumen. Both the long- and the medium-chained acids can be desaturated by $\Delta^9$-desaturase to their *cis-9* monounsaturated counterparts.

Milk FA composition varies among individuals, as well as within individuals depending on their lactation stage [3, 4]. It is highly affected by environmental factors such as feeding, udder health and season, but is also genetically influenced. Substantial genetic variation associated with bovine milk fat composition has been reported [5–10], with estimated heritabilities for individual FA being low to moderate (usually in the range from 0.05 to 0.40). This raises the possibility to improve nutritional properties of milk fat by selective breeding.

Traditionally, detailed milk fat composition is determined by gas chromatography (GC) analysis. This is an accurate but expensive method and is not suitable for routine milk recording. Recent studies showed that Fourier transform infrared spectroscopy (FTIR) data, calibrated against gas chromatography with flame ionization detector (GC–FID) reference data from the same samples, has the potential to provide detailed prediction of milk fat composition [11–19]. An advantage of this approach is that the millions of records obtained by routine recording of cows can be used to estimate genetic parameters and improve traits by breeding. In this study, we used such data to perform a genome-wide association analysis (GWAS) in Norwegian Red cattle to search for genes that affect milk fat composition. A candidate region on BTA13 (BTA for *Bos taurus* chromosome) that influences de novo synthesis of short- and medium-chained FA was fine-mapped and re-analyzed for novel single nucleotide polymorphisms (SNPs) that were detected by re-sequencing in order to attempt to identify the underlying causal DNA variation.

## Methods

### Estimation of bovine milk fat composition from FTIR spectroscopy data

To obtain a calibration model for FTIR spectra, 262 milk samples obtained from a feeding experiment [14] and 616 samples from field sampling were analyzed in parallel by FTIR spectroscopy and GC–FID reference analysis. All samples were from Norwegian Red (NR) cows. FTIR analyses were performed using an FT-IR MilkoScan Combifoss 6500 instrument (Foss, Hillerød, Denmark). Samples were homogenized and temperature-regulated before entering a cuvette (37 µm) for transmission measurements in the spectral range from 925 to 5011 $\text{cm}^{-1}$. The instrument was equipped with a DTGS detector. All spectra were transformed from transmittance to absorbance units. Absorbance spectra were preprocessed by taking the second derivative using Savitzky–Golay algorithm with a polynomial of degree 2 and a window size of 9 channels followed by extended multiplicative signal correction [20] in order to correct for baseline variations and multiplicative effect [21]. FTIR spectra (regressors) were subsequently calibrated against GC–FID reference values (regressands) by using powered partial least squares regression (PPLSR, [22]). Regressands were presented as percentages of GC–FID fatty acid values to total fat in order to reduce to a minimum value the correlation between the FA and total fat in milk samples. Calibration was assessed by 20-fold cross-validation, i.e. the calibration data was divided randomly into 20 segments and each of them was used as an independent test set at a time. The number of components was selected automatically by evaluating if the improvement of the cross-validated prediction of the regressands was significant when the number of PLS components (linear channel combinations) increased in the reduced-rank PPLSR model. If improvement of the calibration model was not significant when moving from component number $A$ to component number $A + 1$, $A$ was chosen as the optimal number of components. However, in order to avoid overfitting, the maximum number of components was set to 25.

The traits that were calibrated in this study included 24 individual FA and 12 combined traits. Individual FA included seven short- and medium-chained, even-numbered saturated FA (C4:0, C6:0, C8:0, C10:0, C12:0, C14:0, C16:0), two long-chained saturated FA (C18:0, C20:0), two odd-numbered saturated FA (C15:0, C17:0), seven monounsaturated FA (C14:1*cis*-9, C16:1*cis*-9, C18:1*cis*-9, C18:1*cis*-11, C18:1*trans*-9, C18:1*trans*-10, C18:1*trans*-11) and six polyunsaturated FA [C18:2*cis*-9,*cis*-12, C18:3*cis*-9,*cis*-12,*cis*-15, arachinonic acid (ARA), conjugated linoleic acid (CLA), docosahexaenoic acid (DHA) and eicosapentaenoic acid (EPA)]. The combined traits were CIS (% of FA with *cis* bonds), TRANS (% of FA with *trans* bonds), TRANS:CIS (*trans*:*cis* ratio), N3 (total amount of omega-3 FA), N6 (total amount of omega-6 FA), N3:N6 (omega-3:omega-6 ratio), DNS (de novo FA synthesis, i.e., sum of the short-chained FA C6:0–C12:0), SAT (% of saturated FA), MUFA (% monounsaturated FA), PUFA (% polyunsaturated FA), TOTAL (total fat yield), and iodine value. NEFA (free FA) and UREA were also included in the GWAS, but these traits have built-in prediction equations in the FT-IR instrument and are stored as a routine procedure in the Norwegian Dairy Herd recording system as parameters of milk quality and feeding, and were therefore not calibrated in this study.

### Estimation of variance components and daughter yield deviations

The obtained calibration models were applied to about 1,650,000 infrared spectra from the Regional Laboratories of the Norwegian Herd recording system for the periods February to November 2007 and July 2008 to March 2009 (spectra from November 2007 to July 2008 were missing due to technical problems with the storage of data during that period). Predicted values of bimonthly test day samples were used for further statistical analyses. The ~1,650,000 FTIR-based FA profile predictions for individual cows (Y) were related to the pedigree structure of the NR population. To condense the information for genetic analyses, only a subset of the data was used. The cows had to be in 1st to 4th lactation and the test-days between 10 and 320 days after calving. The milk yield at the test-day had to be between 5 and 50 kg, and the fat percentage between 1.75 and 7.0. These criteria were designed to remove obvious outliers. Finally, the sire had to be an artificial insemination (AI) NR bull. Milk samples were recorded on a bimonthly basis. This left 950,170 profiles from 300,126 cows that were daughters from 1095 sires, with a total number of animals in the pedigree of 871,455 animals.

The data were analyzed with the following mixed linear animal repeatability model:

$$\mathrm{Y} = \mathrm{RYM_i} + \mathrm{RPL_j} + \mathrm{htd_k} + \mathrm{pe_l} + \mathrm{a_m} + \mathrm{e_{ijklm}},$$

where RYM is the fixed effect of region (9 regions) by year and month of the test-day, with i ranging from 1 to 170; RPL is the fixed effect of region by lactation number by 10-day period in lactation of the test-day, with j ranging from 1 to 1116; htd is the random effect of herd by test-day, with k ranging from 1 to 83,850; pe is the random permanent environmental effect of the cow on her repeated records, with l ranging from 1 to 300,126; a is a random additive genetic effect of the animal, with m ranging from 1 to 871,455; and e is a random residual effect.

The distributional assumptions for the random effects were the following: htd ~ $N(\mathbf{0}, \mathbf{I}\sigma^2_{\mathrm{htd}})$, pe ~ $N(\mathbf{0}, \mathbf{I}\sigma^2_{\mathrm{pe}})$, a ~ $N(\mathbf{0}, \mathbf{A}\sigma^2_{\mathrm{a}})$, and e ~ $N(\mathbf{0}, \mathbf{I}\sigma^2_{\mathrm{e}})$, where **0** is a null vector, **I** an identity matrix and **A** is the additive genetic relationship matrix.

The variance components were estimated by using the DMU software [23] and an average information algorithm. Given the variance components, breeding values and fixed effects were estimated by the DMU software using an iteration on data algorithm.

Daughter yield deviations (DYD) for the GWAS were then derived from these results as the sire averages of daughters' predicted FA compositions, which were each corrected for her fixed effects, non-genetic random effects and half of her dam's genetic effect. The number of bulls with DYD and genotype information varied from step to step as described below, mainly because genotyping on the SNP chips (see below) was performed on animals with trait data for many of the traits in the breeding goal, and was not specific to animals with DYD for the milk FA. The average number of daughters per bull was ~300 in all steps.

### Genotypes for genome-wide association analyses

Initial genotyping for the GWAS was performed on 2552 NR AI bulls using the Affymetrix 25K bovine SNP chip (Affymetrix, Santa Clara, CA, USA) as described in [24]. SNP filtering reduced the number of useful SNPs to 17,343 (see [24] for details). SNPs were positioned on the genome by using the UMD 3.1 assembly [25]. DYD were available for 991 of the 2552 bulls.

### Construction of a high-density SNP dataset with 16,567 SNPs on BTA13

A dense SNP map for fine-mapping on BTA13 was constructed by combining genotypes from the Affymetrix 25K SNP chip with genotypes from Illumina's BovineSNP50 (54K) and BovineHD (777K) BeadChips (Illumina, San Diego, CA, USA). A total of 1575 NR bulls were genotyped with the 54K chip, 536 of these bulls were also among the 2552 animals genotyped with the 25K chip. Next, 384 of the 1575 bulls were genotyped with the 777K chip. The three datasets were filtered to remove SNPs with a minor allele frequency lower than 0.05 and all remaining SNPs were positioned according to the UMD 3.1 assembly. The 25K dataset was imputed to 54K before the combined 54K dataset was imputed to

777K. All imputations and phasing were performed using BEAGLE v3.3.1 [26] with default options. Phase information of the imputed haplotypes was used to identify double recombinants and if possible correct or remove these. The resulting dataset consisted of 3289 NR bulls and 16,567 SNPs on BTA13. DYD were available for 1024 of the bulls, with an average of 278 daughters per son. The 991 bulls used in the previous GWAS step were among these 1024 bulls.

### Genome re-sequencing and construction of a sequence-level SNP dataset for the candidate gene region

Whole-genome re-sequencing data were obtained for five NR elite bulls on an Illumina Genome Analyzer GAIIx instrument (Illumina, San Diego, CA, USA) with 2× 108 paired end reads. The five bulls were selected based on their large numbers of offspring and minimum relationships and therefore represented the genetic diversity of the population. Library preparation was performed using a TruSeq SBS V2-GA kit (Illumina, San Diego, CA, USA). Adaptor- and quality-trimming of raw reads in FASTQ-format was performed using the FASTX-toolkit v0.0.13 [27]. The reads were aligned against BTA13 (bovine reference genome assembly UMD 3.1) using Bowtie v0.12.7 [28] with default parameters. Sorting, marking of PCR duplicates and indexing of the resulting SAM files were performed using Samtools v0.1.17 [29]. Between 98.7 and 99.7% of the reads were mapped to the bovine reference genome assembly UMD 3.1, including all chromosomes and unplaced scaffolds. The average whole-genome sequence coverage for each animal was estimated by dividing the total number of sequenced fragments times read length by the length of the bovine genome (3 gigabases). Two bulls in the dataset had an average whole-genome sequence coverage of about 10×, while three bulls had an average coverage of 4×. Variant calling was performed with Freebayes v0.1.0 [30] with a minimum read coverage of 2 and a minimum alternate allele count of 1. The settings were chosen to maximize calling sensitivity given the relatively low sequence coverage for three of the samples.

Since the parameters for variant calling were set to detect as much variation as possible, rather strict criteria for selecting a novel SNP for further genotyping were set. A total of 1260 SNPs were found within the two genes *nuclear receptor coactivator 6* (*NCOA6*) and *acyl-CoA synthetase short-chain family member 2* (*ACSS2*) or within 2000 bp on either side of these genes. Among these 1260 SNPs, all SNPs in exons and UTR were selected for genotyping together with intronic SNPs that were present in the dbSNP database [31] and co-segregated with the most significant SNPs from the analyses of the high-density data on BTA13. This approach resulted in 71 SNPs that were used to genotype 570 animals. However, as expected given the relatively relaxed SNP detection criteria applied initially, several of these SNPs were found to be monomorphic and hence to be false positives after genotyping. Only 17 SNPs passed all the steps. Of these, two exonic and 11 intronic SNPs were positioned within *NCOA6*, one exonic and two intronic SNPs were located within *ACSS2*, and one SNP was found in the neighboring gene *GSS*. In order to include missing genotypes, to include bulls with trait data that were not genotyped, and to also cover the regions outside the two genes, the 17 novel SNPs together with SNPs from the BovineHD array positioned in the QTL region were imputed by using BEAGLE v3.3.1 [26]. Hence, the final map consisted of 204 SNPs that were located between 63,488,876 and 65,786,868 bp. Of these, 15 and 9 SNPs were located within *NCOA6* and *ACSS2*, respectively. The total number of bulls with genotypes for the 204 SNPs and trait data in the dataset was equal to 782, and the average number of daughters per bull was equal to 362. This dataset was used to fine-map the candidate gene region and for haplotype analyses. Names, positions and primer sequences for the 17 novel SNPs detected by re-sequencing are in Additional file 1: Table S1.

### Single-marker association studies

A single-marker association model was used for the GWAS, the re-sequenced BTA13 map and the candidate gene map. The model that was fitted to the performance data for each trait and each SNP was as follows:

$$\mathrm{DYD_i} = \mu + \mathrm{m} + \mathrm{a_i} + \mathrm{e_i},$$

where $\mathrm{DYD_i}$ is performance of bull i, μ is the overall mean, m is a random SNP effect, $\mathrm{a_i}$ is a random polygenic effect of bull i, and $\mathrm{e_i}$ is a residual effect. We used a random SNP effect because since we performed a REML likelihood ratio test using REML, it was necessary to have the same fixed effects in H1 and H0 (i.e., the model with and without the SNP effect) for the two models to be comparable. Alleles were coded as numbers from 1 to 4 (i.e., A = 1, C = 2, G = 3 and T = 4). A random polygenic effect was included to account for putative genetic differences among bulls other than the SNP effect. The DYD were weighed by the number of daughters. The variances were estimated from the data. The SNP effect m was assumed to follow a normal distribution $\sim N(\mathbf{0}, \sigma^2_{\mathrm{m}})$, where $\sigma^2_{\mathrm{m}}$ is the SNP variance. The polygenic effect a was assumed to follow a normal distribution $\sim N(\mathbf{0}, \mathbf{A}\sigma^2_{\mathrm{a}})$, where **A** is the relationship matrix among the analyzed bulls derived from the pedigree, and $\sigma^2_{\mathrm{a}}$ is the additive genetic variance. The residual effect e was assumed to follow a normal distribution $\sim N(\mathbf{0}, \mathbf{W}\sigma^2_{\mathrm{e}})$, where $\sigma^2_{\mathrm{e}}$ is the environmental variance and **W** is the matrix of weights computed by ASReml based on the number of daughters in the DYD mean.

Significance levels for the random SNP effects were obtained from the log-likelihoods (logL) of a model that includes the SNP effect [LogL(H1)] as well as those of a model without this SNP effect [LogL(H0)], which were both calculated for each SNP using the ASREML package version 2.0 [32]. A likelihood ratio test-statistic (LRT) was calculated as LRT = 2 * [LogL(H1) − LogL(H0)]. Following Baret et al. [33], the distribution of the LRT under the null hypothesis can be seen as a mixture of two Chi square distributions with 0 and 1 degree of freedom, respectively. The significance levels are then obtained from a Chi square distribution with 1 degree of freedom but doubling the probability levels. Due to the amount of multiple-testing performed, we required a rather stringent significance threshold of $p = 0.00025$. Thus, the corresponding LRT were obtained from a Chi square distribution with 1 degree of freedom and $p = 0.0005$, and must be equal to 12.12 or more.

### Correction for the most significant QTL

In order to determine if more than one QTL was segregating in the candidate region, the effect of the most significant SNP from the single-marker analyses of the candidate gene region was corrected for by including it as a fixed effect in the single-marker model and repeating the analysis for all other SNPs in the candidate gene region.

### Haplotype analyses

Pair-wise LD measure ($r^2$) was estimated for all SNP pairs in the candidate gene region on BTA13 using Haploview 4.2 [34]. Haploptype blocks were defined manually. Block 1 was a narrow *NCOA6* block that contained the most significant SNPs (SNPs 98–102), block 2 was a wider *NCOA6* block (SNPs 98–108), block 3 spanned *ACSS2* (SNPs 114–122), while block 4 included SNPs that were present in both *NCOA6* and *ACSS2* (SNPs 98–125). For each of the defined blocks, haplotypes for each sire were determined from the phased genotypes. Since very few sires were homozygous for the least frequent haplotypes, sires with one or two copies of the haplotype were grouped and a two-sample *t* test was performed in R [35] to test for differences in mean phenotypic value between this group and the remaining sires.

## Results and discussion

### FTIR spectroscopy and variance component estimation

A key requirement of this study was to be able to estimate FA composition in milk samples based on FTIR spectroscopy data using a GC–FID reference analysis method [14]. The results showed that 29 of the FA, together representing more than 90% of the total fat content, achieved cross-validated squared Pearson product-moment correlation coefficients ($R^2$CV) above 0.5; these FA were therefore considered predictable and included in the further analyses. As shown in Table 1 and Additional file 2: Table S2, the highest concentrations of individual FA were found for C16:0, C18:1*cis*-9, C18:0 and C14:0 (mean concentrations equal to 25.25, 21.4, 11.29 and 11.21% of total fat, respectively). The best combined trait predictions were obtained for SAT, CIS and MUFA ($R^2$CV = 0.96), while the best predictions for individual FA were found for C18:1*cis*-9 ($R^2$CV = 0.94) and for C8:0 to C12:0 ($R^2$CV = 0.91). The results showed that most major FA were predicted rather accurately, however with lower $R^2$CV for C16:0, C14:0 and C18:0 ($R^2$CV = 0.77, 0.73 and 0.54, respectively). The ability to predict a FA with high confidence depended strongly on its concentration, and FA with concentrations less than 1% generally showed low $R^2$CV and were considered unpredictable (Table 1). There were exceptions to this with a few low-frequency FA that achieved high $R^2$CV, which is most likely due to cross-correlation with more frequent, predictable FA. Correlations between predicted FA and total fat percentage were low to moderate (Table 1) and showed a general trend for negative correlations for longer unsaturated FA, and positive correlations for shorter saturated FA. Mean concentrations of each trait from the GC–FID reference analyses, $R^2$CV, correlation coefficients between each predicted FA and total fat percentage as well as heritabilities are in Table 1, while all the results for the PPLSR calibration and the GC–FID reference values and variance components are in Additional file 2: Table S2.

Several studies investigated the effectiveness of mid-infrared spectroscopy to predict bovine FA composition [11–19], and reported that accuracies vary due to differences in the number of samples, breeds, spectra pretreatments, reference methods and units of measure. In agreement with our findings, prediction accuracies are generally best for FA with high concentrations and for the short and medium-chained FA, C18:1*cis*-9, and for SAT and MUFA. Prediction accuracies were in general better when FA concentrations were expressed as a quantity per unit of milk rather than a quantity of total milk fat, which is most likely because FA concentrations are correlated to total fat, and predicting FA in milk on the basis of FTIR is the combined effect of predicting fat content and fat composition [11, 13, 16]. However, these correlations should be lower when FA concentrations are expressed as quantity of total milk fat when models are developed on the basis of fat as in our study. Soyeurt et al. [11] suggested that the predicted concentrations were not due to real absorbance values specific to FA if

**Table 1 Mean concentrations, cross-validated squared correlation coefficients, correlations to total fat, and heritabilities for all calibrated traits**

| Trait | Cons | $R^2$CV | Corr (SE) | $h^2$ (SE) |
|---|---|---|---|---|
| C4:0 | 4.16 | 0.73 | 0.111 (0.039) | 0.353 (0.004) |
| C6:0 | 2.48 | 0.89 | 0.104 (0.039) | 0.231 (0.003) |
| C8:0 | 1.48 | 0.91 | 0.040 (0.039) | 0.187 (0.003) |
| C10:0 | 3.2 | 0.91 | 0.034 (0.039) | 0.171 (0.003) |
| C12:0 | 3.55 | 0.91 | 0.045 (0.039) | 0.179 (0.003) |
| C14:0 | 11.21 | 0.86 | 0.077 (0.039) | 0.109 (0.003) |
| C14:1*cis*-9 | 0.98 | 0.52 | 0.089 (0.039) | 0.222 (0.003) |
| C15:0 | 1.0 | 0.59 | 0.071 (0.039) | 0.146 (0.003) |
| C16:0 | 25.25 | 0.77 | 0.433 (0.035) | 0.145 (0.003) |
| C16:1*cis*-9 | 1.17 | 0.51 | 0.392 (0.036) | 0.146 (0.003) |
| C17:0 | 0.49 | 0.43 | 0.146 (0.039) | 0.142 (0.003) |
| C18:0 | 11.29 | 0.54 | −0.279 (0.038) | 0.175 (0.003) |
| C18:1*trans*-9 | 0.24 | 0.74 | −0.521 (0.033) | 0.141 (0.002) |
| C18:1*trans*-10 | 0.36 | 0.56 | −0.543 (0.033) | 0.171 (0.003) |
| C18:1*trans*-11 | 1.33 | 0.67 | −0.318 (0.037) | 0.092 (0.002) |
| C18:1*cis*-9 | 21.4 | 0.94 | −0.186 (0.038) | 0.127 (0.003) |
| C18:1*cis*-11 | 0.79 | 0.73 | −0.357 (0.037) | 0.146 (0.003) |
| C18:2*cis*-9,*cis*-12 | 1.39 | 0.61 | −0.409 (0.036) | 0.172 (0.003) |
| C18:2*cis*-9,*trans*-11 | 0.62 | 0.65 | −0.325 (0.037) | 0.120 (0.002) |
| C18:3*cis*-9,*cis*-12,*cis*-15 | 0.54 | 0.42 | −0.231 (0.038) | 0.190 (0.003) |
| C20:0 | 0.2 | 0.39 | −0.336 (0.037) | 0.161 (0.003) |
| ARA | 0.07 | 0.46 | −0.052 (0.039) | 0.236 (0.004) |
| EPA | 0.06 | 0.16 | 0.088 (0.039) | 0.173 (0.003) |
| DHA | 0.02 | 0.62 | −0.014 (0.039) | 0.159 (0.003) |
| SAT | 64.31 | 0.96 | 0.308 (0.037) | 0.137 (0.003) |
| MUFA | 26.28 | 0.96 | −0.229 (0.038) | 0.130 (0.003) |
| PUFA | 2.7 | 0.72 | −0.491 (0.034) | 0.171 (0.003) |
| Iodine value | 25.51 | 0.95 | −0.241 (0.038) | 0.144 (0.003) |
| CIS | 26.43 | 0.96 | −0.198 (0.038) | 0.138 (0.003) |
| TRANS | 2.56 | 0.73 | −0.419 (0.036) | 0.103 (0.002) |
| TRANS:CIS | 0.1 | 0.64 | −0.377 (0.036) | 0.096 (0.002) |
| DNS | 10.72 | 0.92 | 0.048 (0.039) | 0.165 (0.003) |
| N3 | 0.62 | 0.37 | −0.211 (0.038) | 0.191 (0.003) |
| N6 | 1.47 | 0.62 | −0.386 (0.036) | 0.170 (0.003) |
| N3:N6 | 0.44 | 0.42 | 0.143 (0.039) | 0.193 (0.003) |
| Total | 93.29 | 0.59 | 0.377 (0.036) | 0.106 (0.002) |

Mean concentration from the GC–FID reference analyses (Cons), cross-validated squared Pearson product-moment correlation coefficients ($R^2$CV), Pearson correlation coefficients between the predicted fatty acids and total fat percentage (corr) and standard errors of the correlation, heritabilities ($h^2$) and standard errors of the heritability for all calibrated traits. The concentration is expressed as percentage by weight of total fatty acid content (on a fatty acid methyl ester basis), except iodine value, which is expressed as g $I_2$/100 g of total fatty acid content

the calibration correlations were not higher than the correlations between total fat and FA. As shown in Additional file 2: Table S2, the squared correlations between a FA and total fat percentage were markedly lower than the $R^2$CV for all FA and combined traits groups in our study, which indicated that the predicted concentrations are due to real absorbance values of the FA rather than to correlations to total fat only. Moreover, prediction accuracies for C8:0, C10:0, C12:0, C18:1*cis*-9, SAT and MUFA were as high as those reported with milk-based models [13, 15, 17–19]. C4:0 and C14:0 were predicted with somewhat poorer accuracies than those usually obtained with milk-based models, but with better accuracies than those obtained with fat-based models [11, 13, 19]. Predictions of C16:0 were comparable to those obtained with fat-based models [11, 13, 19].

In general, the selected number of components was large, but since the PPLSR model is very selective for each component, a larger number of selected components is expected than with a conventional PLSR model. In addition, the complexity of the calibration reference data used in this study was considerably higher and the level of variation of the data was much higher compared to those for the data reported in Afseth et al. [14], and thus the model is expected to be more complex. Compared to the reference data used in Afseth et al. [14], the current calibration set contains many samples with a considerable higher proportion of unsaturated acids.

Relatively high heritabilities were estimated from the FTIR predictions (Table 1). Estimates for the predictable FA ranged from 0.09 for C18:1*trans*-11 to 0.35 for C4:0. Short and medium length FA were slightly more heritable than longer and unsaturated FA. This is as expected since the shorter saturated FA are mainly synthesized by the animal, while longer unsaturated FA originate predominately from the diet. The heritability for the sum of polyunsaturated FA (PUFA) was somewhat higher than that for the sum of monounsaturated (MUFA) and saturated (SAT) FA ($h^2 = 0.171$, 0.130 and 0.137, respectively). These results can be explained by the fact that all three indices (SAT, MUFA and PUFA) reflect a combination of genetic and environmental factors, and that the prediction accuracy and concentration of individual FA are expected to affect the estimates for the indices. Estimated heritabilities for the sum of *trans* FA (TRANS) were lower than for the sum of *cis* FA (CIS), and this was also reflected in the individual FA.

In the literature, estimated heritabilities for bovine milk FA composition vary largely among studies depending on sample size, breed, and method. Our estimates were generally lower than those from other studies in which FA concentrations were predicted with mid-infrared spectroscopy [5, 7, 8, 10], but they were in the same range as in the study of Krag et al. [9] in which GC was used. Our observation that individual saturated FA have higher heritabilities than unsaturated FA has been previously reported by several authors [5, 7, 9], whereas estimated heritabilities for groups of FA varied among studies. Whereas many studies support the general pattern of higher heritabilities for saturated FA than for unsaturated FA [5, 6, 8, 10], the highest estimates were found for MUFA in the study of Krag et al. [9], and for PUFA in the current study. The disparity in these results most likely reflects differences in concentrations and prediction accuracies of the FA included in the different FA groups.

### Genome-wide association studies

Phenotypic records for the 29 traits considered to be predictable, together with pre-existing records for urea and NEFA, were tested for their association with ~17,000 genome-wide distributed SNPs using a single-marker association model. We detected 200 significant marker-trait associations and the most significant associations were clustered on BTA1, 13 and 15. These QTL are further discussed below and compared with findings from other studies. All significant results are in Additional file 3: Table S3.

#### *BTA13*

In our study, the most relevant QTL were detected between 55.4 and 66.1 Mb on BTA13. These QTL affected the content in all short- and medium-chained, saturated de novo synthesized milk FA (i.e.; C4:0–C14:0 and DNS). Among these, the highest LRT was detected between SNP rs29018443 and C8:0 (LRT = 26.98), and this same SNP was also highly associated with C6:0, C10:0, C12:0, C14:0 and DNS. A strong candidate gene, *acyl-CoA synthetase short-chain family member 2* (*ACSS2*), lies nearby this SNP and encodes an enzyme that catalyzes the activation of acetate for de novo synthesis of short-chained FA [36]. *ACSS2* was also suggested as a candidate gene that affects de novo synthesized FA (C6:0, C8:0 and C10:0) in Dutch Holstein–Friesian [37] and Danish Jersey cattle [38], and several C16 and C18 FA in Chinese Holstein [39].

#### *BTA1*

In our study, the most significant association (LRT = 33.94) was between SNP rs29019625 located at 144.4 Mb on BTA1 and C18:2*cis*-9,*cis*-12. This SNP was also significantly associated to N6, C18:1*trans*-11, C15:0 and PUFA. The QTL region spanned the ~126.3–144.4 Mb region and included also significant associations to C6:0–C12:0, DNS and DHA. SNP rs29019625 lies approximately 20 kb away from the *SLC37A1* gene, which encodes a membrane bound protein that is involved in the translocation of glycerol-3-phosphate into the endoplasmic reticulum [40]. Other positional candidate genes are *ABCG1* and *AGPAT3.* The former is located at 144 Mb and is involved in macrophage cholesterol and phospholipid transport and may regulate cellular lipid homeostasis in other cell types (e.g., [41]), while *AGPAT3* is located at 146.7 Mb and encodes an acyltransferase that has a role in the de novo phospholipid biosynthetic pathway [42].

A connection between BTA1 and predominantly long-chained FA was reported in several studies. Schennink et al. [43] observed significant associations between markers on BTA1 and C18:0, C18-index and CLA-index at ~125 cM (which corresponds roughly to ~140 Mb according to their map published in Schopen et al. [44]). Bouwman et al. [37] reported a QTL region for C14:0 that is located between ~121 and 130 Mb and for C16:1 between ~146 and 161 Mb in the Dutch Holstein–Friesian population. Li et al. [39] detected significant associations with markers on BTA1 for C10:0 and C12:0 at 132 Mb and for C18:0 and C18 index at 146 Mb in Chinese Holstein. Furthermore, Li et al. [45] reported associations between BTA1 and C18 index at 142.2 Mb in Chinese Holstein and C18:0 at 146.3 Mb in a joint analysis of Chinese and Danish Holstein.

#### *BTA15*

The QTL region that was detected on BTA15 (between 22.6 and 29.0 Mb) affects C8:0–C14:0, DNS, C18:0, C18:1*cis*-9, CIS, *trans:cis* ratio, iodine value and total fat yield, with the highest LRT being for DNS (LRT = 25.8). This QTL is situated close to the genes encoding the following apolipoproteins APOA1, APOA3, APOA4 and APOA5 at 27.9 Mb. This QTL region is frequently cited in the literature. Bouwman et al. [37] detected associations between QTL in the region that lies from 20.5 to 27 Mb on BTA15 and two de novo synthesized FA (C10:0 and C14:0) in Dutch Holstein–Friesian. Within the same region, associations to C18:0 and C18 index in Chinese Holstein [39] and to C12:0, C14:0, and C18:1*cis*-9 in Danish Jersey [38] were reported. Furthermore, Li et al. [45] reported associations to C18:0 and C18 index at position 28.6 Mb in Chinese Holstein and at 27.3–32.8 Mb in a joint analysis of Chinese and Danish Holstein.

GWAS studies frequently report strong associations between milk FA and the genes *diacylglycerol acyltransferase 1* (*DGAT1*) on BTA14 and *stearoyl-coenzyme A desaturase 1* (*SCD*) on BTA26. *DGAT1* encodes an enzyme that catalyzes the final stage of triacylglycerol synthesis (e.g. [46]), while *SCD* 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 [47]. No significant associations in the vicinity of *DGAT1* were detected in our study. Subsequent re-sequencing of 147 NR animals showed that they were all homozygous for the *A* variant of the *DGAT1* K232A polymorphism (not shown). In contrast to the *A* variant, the *K* variant is associated with increased fat yield, fat percentage and protein percentage and decreased milk yield and protein yield. Selection may have favored the *A* variant in the NR population, because most selection pressure was put on milk and protein yield in the breeding goal. In contrast, both allele variants of an important *SCD1* polymorphism (A293V) were found to be relatively common in the sequenced NR individuals with a MAF of 0.25 (data not shown); however a follow-up study that examined the *SCD1* region by including additional SNPs did not detect any significant associations near *SCD1* (unpublished results).

### Fine-mapping using a high-density SNP dataset on BTA13

Subsequent analyses were performed to fine-map the BTA13 QTL that affects de novo synthesized FA and to identify potential causal variations. We began by reanalyzing the associations between all the high-density SNPs on BTA13 (n = 16,567) and the traits that were significant in the initial GWAS (i.e. C4:0–C14:0 and DNS). Somewhat surprisingly, this analysis did not point towards the prime candidate gene *ACSS2* as the most likely position of the QTL, but to a nearby gene i.e. *nuclear receptor coactivator 6* (*NCOA6*) that encodes a transcriptional co-activator, which interacts with nuclear hormone receptors. The most significant result was found for SNP rs41700740 at 64,650,276 bp which is a synonymous variant located within this gene. The LRT for this SNP ranged from 62.6 for C8:0 to 24.5 for C14:0. Significant LRT were found for ~500 SNP/trait combinations in the QTL region. As an example, results for DNS are in Fig. 1, while LRT for all SNP/trait combinations are in Additional file 4: Table S4.

### Fine-mapping using SNPs in the *NCOA6* and *ACSS2* genes at the sequence level

Since our analyses pointed towards *NCOA6* and not *ACSS2* as the most likely positional candidate gene underlying the QTL, both genes were investigated in more detail. A dataset consisting of 15 SNPs within *NCOA6* and nine SNPs within *ACSS2* as well as 180 SNPs in the regions surrounding these genes was constructed by combining sequence-level polymorphisms with SNPs from the Bovine HD BeadChip. Both C6–C14 as well as DNS were reanalyzed for these SNPs using the single-SNP model. The results showed that, for C6:0–C12:0 and DNS, the highest LRT was found for SNP 99, i.e. rs41700742 at 64,648,620 bp, which is a synonymous SNP located within *NCOA6*. High LRT were also detected for SNP 100 (rs41700740 at 64,650,276 bp), SNP 102 (rs41700737 at 64,655,588 bp) and SNP 98 (rs41700745 at 64,639,392 bp). All these SNPs are localized within *NCOA6*; the former and the latter are synonymous exonic SNPs whereas rs41700737 at 64,655,588 bp is an intronic SNP. For C14:0, SNP 161 (rs43711970) at 65,246,092 bp was slightly more significant (24.2 vs. 23.8) than SNP 99. SNP 161 is located within the gene *UQCC*, which is almost 400 kb away from *NCOA6* on the telomeric side. Complete results for all traits and SNPs are in Additional file 5: Table S5. As an example, results for DNS are in Fig. 2.

In order to determine if more than one QTL segregated in the detected region, the DNS traits were re-analyzed by including the effect of SNP rs41700742 as a fixed term (not shown). The results showed that this SNP explained all the variation, which indicates that only one QTL is segregating for the DNS traits, and the signals detected

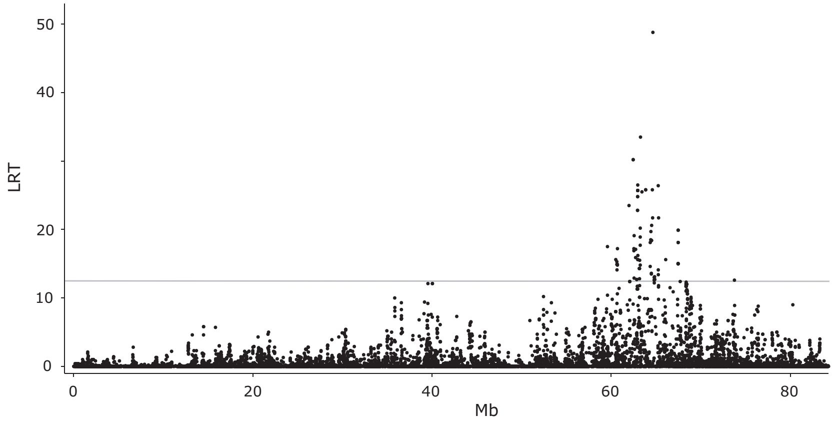


**Fig. 1** Association analysis of FA synthesized de novo (DNS) for SNPs on BTA13 from the BovineHD BeadChip. The ordinate denotes the LRT, while the abscissa denotes SNP positions in Mb. The *grey line* indicates the significance threshold (LRT = 12.12)

for the remaining SNPs were merely due to LD between SNPs.

### Haplotype analyses

Finally, to better characterize the BTA13 QTL, all the SNPs within the QTL region were grouped into haplotype blocks in order to identify the haplotypes that displayed the strongest associations to C8:0, which is a proxy for DNS. Pair-wise LD measure ($r^2$) for all SNP pairs in the candidate gene region are in Fig. 3 along with four manually-constructed haplotype blocks. Within each block, each haplotype with a frequency higher than 0.01 was

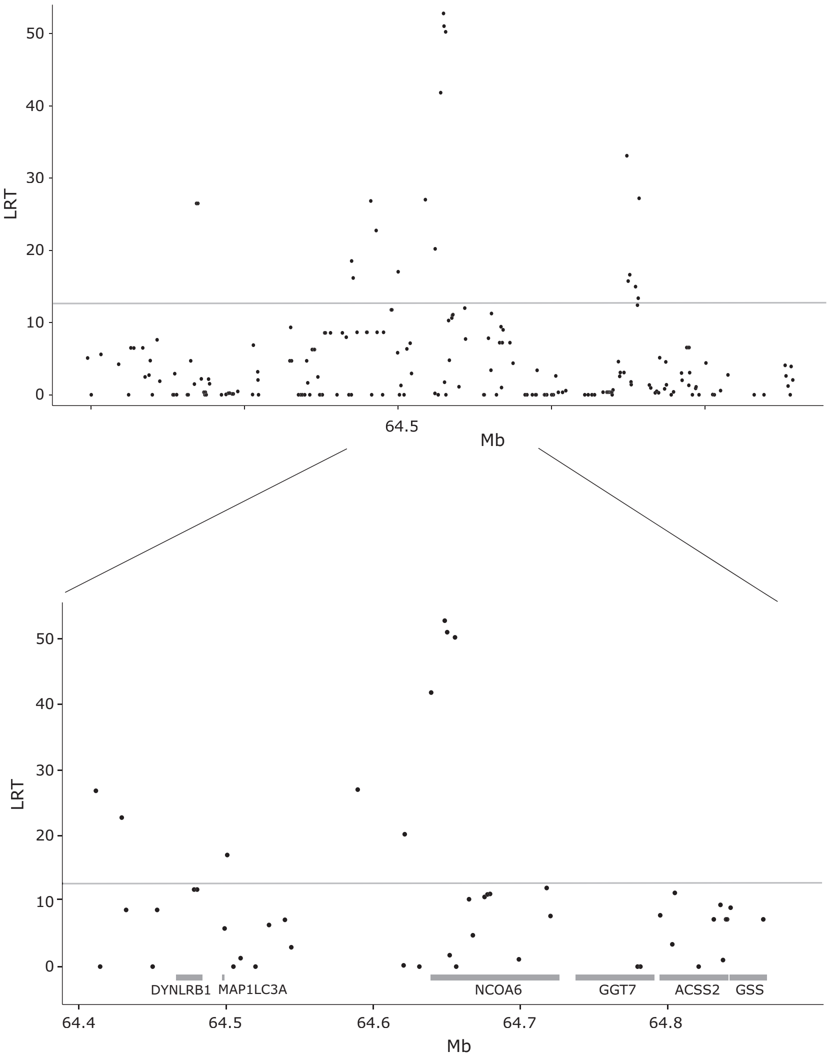


**Fig. 2** Association analysis of FA synthesized de novo (DNS) in the candidate gene region. *Top* results for the entire candidate gene region. The ordinate denotes the LRT, while the abscissa denotes SNP positions in bp. *Bottom* zoom on the region between 64.4 and 64.9 Mb. The positions of the genes in the region are indicated with *grey boxes*. The *grey line* indicates the significance threshold (LRT = 12.12)

tested against the mean of the remaining haplotypes. Results for haplotypes with a frequency of 0.05 or more are in Table 2. The most significant effects were detected in the narrow *NCOA6* block (block 1 that included SNPs 98 to 102), which displays eight haplotypes. A frequent haplotype (denoted 1.1) was associated with higher content of short-chained FA ($p = 0.00037$), while haplotypes 1.2 and 1.4 were associated with lower FA content ($p = 0.0000048$ and 0.027, respectively). When the haplotype block was extended to include SNPs 98 to 108 in the broader *NCOA6* block (block 2, which also consisted of eight haplotypes), the differences between haplotypes were less marked. Haplotype 2.1 within this block had an identical frequency and p value as in the narrow block. The two negative haplotypes from block 1 were split into several less frequent haplotypes, with the most frequent being haplotypes 2.4 ($p = 0.038$) and 2.6 ($p = 0.09$). Block 3 covered *ACSS2* (SNPs 114 to 122) and produced even less significant results. A larger block that contained the SNPs located within both *NCOA6* and *ACSS2* (block 4, including SNPs 98 to 125 with eight haplotypes), the differences between haplotypes became more marked again. The most frequent haplotype (4.1) showed a stronger effect than the remaining haplotypes with a p value of 0.00046.

In summary, the strongest associations were found for haplotypes within a rather narrow region that contained *NCOA6*. Neither the haplotypes within a larger block that included both *NCOA6* and *ACSS2* nor the block that contained only *ACSS2* were significant. Thus, the results of the haplotype analyses also suggest that *NCOA6* is a stronger positional candidate for the observed variation in de novo FA synthesis than *ACSS2*.

### *NCOA6*

*NCOA6*, or *nuclear receptor coactivator 6*, encodes an essential, non-redundant multifunctional coactivator for nuclear hormone receptors and certain other transcription factors [48]. The gene is expressed in a variety of tissues, such as testis, brain, ovary, liver, fat and heart [48] and also in the mammary gland [49]. *NCOA6* is essential for embryonic development [50], it is involved in cell survival, growth, wound healing and energy metabolism [51], and is important for normal mammary gland development [52]. Different *NCOA6* isoforms are expressed in the mouse mammary gland at different developmental stages including adult virgin, pregnancy, lactation and involution [48].

To the best of our knowledge, no studies have specifically investigated the role of *NCOA6* in milk fat synthesis.

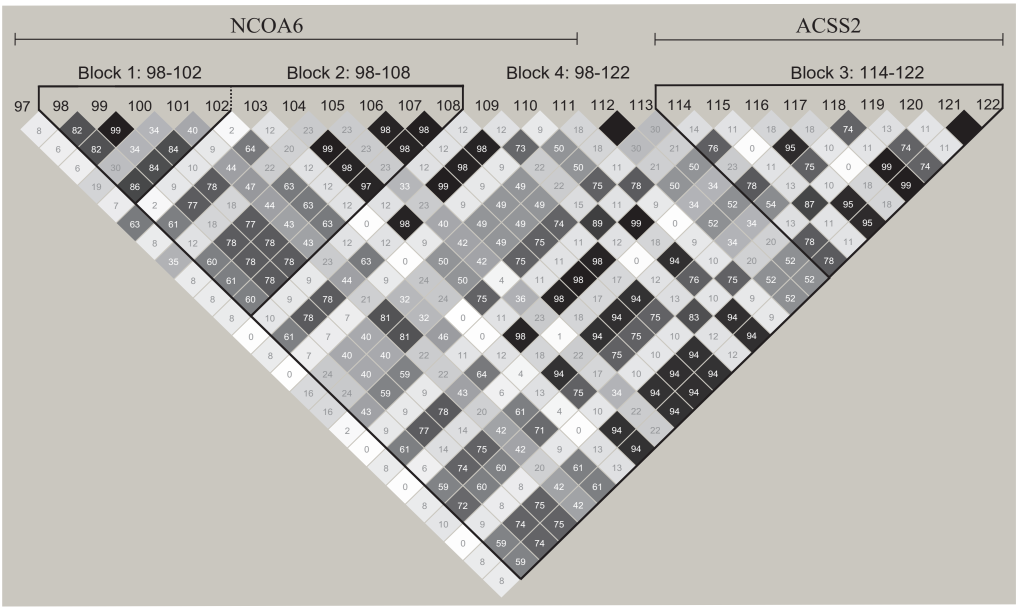


**Fig. 3** Haploview plot illustrating LD between pairwise combinations of SNPs within and between *NCOA6* and *ACSS2*. Genes are shown together with the blocks used in the haplotype analyses. *Numbers above the triangle* denote marker number in the candidate gene region. *Numbers within the triangle* are pair-wise LD between markers in the form of $r^2 * 100$

**Table 2 Haplotype analyses**

| Block | Hap number | Haplotype | Effect | Freq | p value |
|---|---|---|---|---|---|
| 1 | 1.1 | GAACA | + | 0.56 | 0.00037 |
| | 1.2 | AGGCG | − | 0.18 | 0.0000048 |
| | 1.4 | AGGAG | − | 0.19 | 0.027 |
| 2 | 2.1 | GAACAAAGAAG | + | 0.56 | 0.00037 |
| | 2.4 | AGGCGAGGCGA | − | 0.12 | 0.038 |
| | 2.6 | AGGAGAGGCGA | − | 0.09 | 0.09 |
| | 2.2 | AGGCGAAGAAG | − | 0.05 | 0.000000059 |
| | 2.5 | AGGAGAGACGA | − | 0.05 | 0.017 |
| 3 | 3.1 | GACGGGGAC | + | 0.62 | 0.125 |
| | 3.3 | AAAGACGGA | − | 0.12 | 0.043 |
| | 3.5 | AAAAACGGA | − | 0.08 | 0.032 |
| | 3.4 | AGAGACAGA | − | 0.06 | 0.08 |
| 4 | 4.1 | GAACAAAGAAGGACAAGACGGGGACGGC | + | 0.56 | 0.00046 |
| | 4.4 | AGGCGAGGCGAGTCAAAAAGACGGAAAA | − | 0.12 | 0.043 |
| | 4.6 | AGGAGAGGCGAGTCGGAAAAACGGAAAA | − | 0.08 | 0.034 |
| | 4.2 | AGGCGAAGAAGGACAAGACGGGGACGGC | − | 0.05 | 0.000000059 |

Block number, haplotype number (Hap number), haplotype, effect of haplotype (+ is higher content of C8:0 as compared to mean of remaining haplotypes in the block, − is lower C8:0 content), frequency and p-value of each haplotype. Block 1: *NCOA6*, SNPs 98 to 102. Block 2: *NCOA6*, SNPs 98 to 108. Block 3: *ACSS2*, SNPs 114 to 122. Block 4: *NCOA6* and *ACSS2*, SNPs 98 to 125

However *NCOA6* is known to be a ligand for transcription factors such as PPARα and PPARγ [53], and thus, its effect could be through these. PPARγ affects expression of genes that are involved in fatty acid transport such as *LPL*, *CD36* and *ACSL1* [54], and is proposed as a major regulator of bovine milk fat synthesis [2]. In a study on the gene regulatory networks in lactation, *NCOA6* (in that study denoted *PRIP*) was identified as one of the factors involved in PPARα/RXRα signaling [55]. Therefore, *NCOA6* could be a functional as well as a positional candidate for the QTL on BTA13.

Our study did not identify any candidate causal polymorphisms underlying the QTL. The three SNPs with the highest LRT are either synonymous or intronic and therefore do not directly alter the protein sequence. However, introns can harbor important regulating elements such as binding sites for transcription factors and sites that affect alternative splicing. Synonymous SNPs are also suggested to have important biological roles, as they may have an impact on critical cis-regulating sequences, alter mRNA structure and influence translational speed [56]. Further analyses will be undertaken in order to investigate the nature of the QTL on BTA13 and other QTL that have an effect on bovine milk FA composition.

## Conclusions

Using a combined dataset of high-resolution genotypes and FTIR phenotypes, our GWAS detected significant QTL for milk fatty acids on BTA1, 13 and 15. On BTA13, the QTL for de novo fatty acid synthesis mapped close to a known candidate gene (*ACSS2*), but subsequent refined analyses highlighted that *ACSS2* had little effect and that SNPs within the nearby *NCOA6* gene were responsible for the observed QTL. To date, the functional role of *NCOA6* in milk fatty acid synthesis is unclear, but one possible effect could be that it is a ligand for the transcription factor PPARγ, which is suggested to be a major regulator of milk fat synthesis.

## Additional files

**Additional file 1: Table S1.** SNPs genotyped for the candidate gene map, with rs numbers, positions in base pairs and primer sequences.

**Additional file 2: Table S2.** Results for calibration of FTIR-spectra against GC-FID reference data, correlations between each FA and total fat percentage, and estimates of variance components with standard errors.

**Additional file 3: Table S3.** GWAS results.

**Additional file 4: Table S4.** Results from single-marker association analyses on BTA13 data from the BovineHD BeadChip.

**Additional file 5: Table S5.** Results from single-marker association analyses on SNPs in the candidate gene region.

**Authors' contributions**

HGO carried out the association analyses and drafted the manuscript. TMK contributed to the genome re-sequencing, SNP selection and writing the manuscript. AK, LG and HM performed the prediction of milk fatty acid concentrations. MS estimated genetic variance components and daughter yield deviations. HG performed the imputation and phasing. TN contributed to the association analyses. MS wrote the script for association analyses and haplotype analyses. KKS performed the genotyping. MPK assisted in planning

the experiment and contributed to writing the manuscript. SL conceived and planned the experiment and contributed to writing the manuscript. All authors read and approved the final manuscript.

**Author details**

[1] Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, PO Box 5003, 1432 Ås, Norway. [2] Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, PO Box 5003, 1432 Ås, Norway. [3] Centre for Biospectroscopy and Data Modeling, Nofima AS, Osloveien 1, 1430 Ås, Norway. [4] Geno Breeding and AI Association, 1432 Ås, Norway. [5] CAMO Software AS, Nedre Vollgate 8, 0158 Oslo, Norway. [6] Institute of Marine Research, Flødevigen, 4817 His, Norway. [7] Department of Natural Sciences, Faculty of Engineering and Science, University of Agder, PO Box 422, 4604 Kristiansand, Norway. [8] Department of Engineering Cybernetics, Norwegian University of Science and Technology, 7034 Trondheim, Norway.

**Acknowledgements**

The authors thank Valeria Tafintseva for useful advice.

**Competing interests**

The authors declare that they have no competing interests.

**Availability of data and material**

The sequence data of the animals will be submitted to the European Nucleotide Archive, http://www.ebi.ac.uk/ena.

**Funding**

The project was funded by the Norwegian Research Council (Project Number 225173) with additional funding from GENO Breeding and AI Organization and Tine Norwegian Dairy Company.

Received: 29 June 2016   Accepted: 3 February 2017
Published online: 13 February 2017

**References**

1. Haug A, Hostmark AT, Harstad OM. Bovine milk in human nutrition—a review. Lipids Health Dis. 2007;6:25.
2. Bionaz M, Loor JJ. Gene networks driving bovine milk fat synthesis during the lactation cycle. BMC Genomics. 2008;9:366.
3. Palmquist DL, Beaulieu AD, Barbano DM. Feed and animal factors influencing milk-fat composition. J Dairy Sci. 1993;76:1753–71.
4. Jensen RG. The composition of bovine milk lipids: January 1995 to December 2000. J Dairy Sci. 2002;85:295–350.
5. Soyeurt H, Gillon A, Vanderick S, Mayeres P, Bertozzi C, Gengler N. Estimation of heritability and genetic correlations for the major fatty acids in bovine milk. J Dairy Sci. 2007;90:4435–42.
6. Bobe G, Bormann JAM, Lindberg GL, Freeman AE, Beitz DC. Estimates of genetic variation of milk fatty acids in US Holstein cows. J Dairy Sci. 2008;91:1209–13.
7. Stoop WM, van Arendonk JAM, Heck JML, van Valenberg HJF, Bovenhuis H. Genetic parameters for major milk fatty acids and milk production traits of Dutch Holstein–Friesians. J Dairy Sci. 2008;91:385–94.
8. Bastin C, Soyeurt H, Gengler N. Genetic parameters of milk production traits and fatty acid contents in milk for Holstein cows in parity 1-3. J Anim Breed Genet. 2013;130:118–27.
9. Krag K, Poulsen NA, Larsen MK, Larsen LB, Janss LL, Buitenhuis B. Genetic parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a Bayesian approach. BMC Genet. 2013;14:79.
10. Lopez-Villalobos N, Spelman RJ, Melis J, Davis SR, Berry SD, Lehnert K, et al. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in New Zealand dairy cattle. J Dairy Res. 2014;81:340–9.
11. Soyeurt H, Dardenne P, Dehareng F, Lognay G, Veselko D, Marlier M, et al. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J Dairy Sci. 2006;89:3690–5.
12. Soyeurt H, Dardenne P, Dehareng F, Bastin C, Gengler N. Genetic parameters of saturated and monounsaturated fatty acid content and the ratio of saturated to unsaturated fatty acids in bovine milk. J Dairy Sci. 2008;91:3611–26.
13. Rutten MJM, Bovenhuis H, Hettinga KA, van Valenberg HJF, Van Arendonk JAM. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J Dairy Sci. 2009;92:6202–9.
14. Afseth NK, Martens H, Randby A, Gidskehaug L, Narum B, Jorgensen K, et al. Predicting the fatty acid composition of milk: a comparison of two Fourier transform infrared sampling techniques. Appl Spectrosc. 2010;64:700–7.
15. Soyeurt H, Dehareng F, Gengler N, McParland S, Wall E, Berry DP, et al. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J Dairy Sci. 2011;94:1657–67.
16. De Marchi M, Penasa M, Cecchinato A, Mele M, Secchiari P, Bittante G. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal. 2011;5:1653–8.
17. Ferrand M, Huquet B, Barbey S, Barillet F, Faucon F, Larroque H, et al. 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. 2011;106:183–9.
18. Maurice-Van Eijndhoven MHT, Soyeurt H, Dehareng F, Calus MPL. Validation of fatty acid predictions in milk using mid-infrared spectrometry across cattle breeds. Animal. 2013;7:348–54.
19. Bonfatti V, Degano L, Menegoz A, Carnier P. Short communication: mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. J Dairy Sci. 2016;99:8216–21.
20. Martens H, Stark E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. J Pharm Biomed Anal. 1991;9:625–35.
21. Zimmermann B, Kohler A. Optimizing Savitzky–Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. Appl Spectrosc. 2013;67:892–902.
22. Indahl U. A twist to partial least squares regression. J Chemom. 2005;19:32–44.
23. Madsen P, Jensen J. DMU: a user's guide. A package for analysing multivariate mixed models. Version 6, release 4.7. Foulum: Danish Institute of Agricultural Sciences; 2008.
24. Olsen HG, Hayes BJ, Kent MP, Nome T, Svendsen M, Larsgard AG, et al. Genome-wide association mapping in Norwegian Red cattle identifies quantitative trait loci for fertility and milk production on BTA12. Anim Genet. 2011;42:466–74.
25. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 2009;10:R42.
26. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84:210–23.
27. Hannon Lab. FASTX-Toolkit. 0.0.13 2010. http://hannonlab.cshl.edu/fastx_toolkit/. Accessed 31 Oct 2010.
28. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
30. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN]; 2012.
31. The database of Short Genetic Variation (dbSNP). National Center for Biotechnology Information, National Library of Medicine. http://www.ncbi.nlm.nih.gov/SNP/. Accessed 5 Jan 2015.
32. Gilmour A, Gogel B, Cullis M, Thompson R. ASReml user guide release 2.0. Hemel Hempstead: VSN International Ltd; 2006.
33. Baret PV, Knott SA, Visscher PM. On the use of linear regression and maximum likelihood for QTL mapping in half-sib designs. Genet Res. 1998;72:149–58.
34. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.
35. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014.

36. Luong A, Hannah VC, Brown MS, Goldstein JL. Molecular characterization of human acetyl-CoA synthetase, an enzyme regulated by sterol regulatory element-binding proteins. J Biol Chem. 2000;275:26458–66.
37. Bouwman AC, Bovenhuis H, Visker MHPW, van Arendonk JAM. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet. 2011;12:43.
38. Buitenhuis B, Janss LLG, Poulsen NA, Larsen LB, Larsen MK, Sorensen P. Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. BMC Genomics. 2014;15:1112.
39. Li C, Sun DX, Zhang SL, Wang S, Wu XP, Zhang Q, et al. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. PLoS One. 2014;9:e96186.
40. Bartoloni L, Wattenhofer M, Kudoh J, Berry A, Shibuya K, Kawasaki K, et al. Cloning and characterization of a putative human *glycerol 3-phosphate permease* gene (*SLC37A1* or *G3PP*) on 21q22.3: mutation analysis in two candidate phenotypes, DFNB10 and a glycerol kinase deficiency. Genomics. 2000;70:190–200.
41. Klucken J, Büchler C, Orso E, Kaminski WE, Porsch-Ozcurumez M, Liebisch C, et al. *ABCG1* (*ABC8*), the human homolog of the Drosophila white gene, is a regulator of macrophage cholesterol and phospholipid transport. Proc Natl Acad Sci USA. 2000;97:817–22.
42. Lu B, Jiang YJ, Zhou YL, Xu FY, Hatch GM, Choy PC. Cloning and characterization of murine 1-acyl-sn-glycerol 3-phosphate acyltransferases and their regulation by PPAR alpha in murine heart. Biochem J. 2005;385:469–77.
43. Schennink A, Stoop WM, Visker MHPW, van der Poel JJ, Bovenhuis H, van Arendonk JAM. Genome-wide scan for bovine milk-fat composition. II. Quantitative trait loci for long-chain fatty acids. J Dairy Sci. 2009;92:4676–82.
44. Schopen GCB, Koks PD, van Arendonk JAM, Bovenhuis H, Visker MHPW. Whole genome scan to detect quantitative trait loci for bovine milk protein composition. Anim Genet. 2009;40:524–37.
45. Li X, Buitenhuis AJ, Lund MS, Li C, Sun D, Zhang Q, et al. Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. J Dairy Sci. 2015;98:8152–63.
46. Cases S, Smith SJ, Zheng YW, Myers HM, Lear SR, Sande E, et al. Identification of a gene encoding an acyl CoA: diacylglycerol acyltransferase, a key enzyme in triacylglycerol synthesis. Proc Natl Acad Sci USA. 1998;95:13018–23.
47. Ntambi JM, Miyazaki M. Recent insights into stearoyl-CoA desaturase-1. Curr Opin Lipidol. 2003;14:255–61.
48. Li QT, Xu JM. Identification and characterization of the alternatively spliced nuclear receptor coactivator-6 isoforms. Int J Biol Sci. 2011;7:505–16.
49. Lemay DG, Lynn DJ, Martin WF, Neville MC, Casey TM, Rincon G, et al. The bovine lactation genome: insights into the evolution of mammalian milk. Genome Biol. 2009;10:R43.
50. Zhu YJ, Crawford SE, Stellmach V, Dwivedi RS, Rao MS, Gonzalez FJ, et al. Coactivator PRIP, the peroxisome proliferator-activated receptor-interacting protein, is a modulator of placental, cardiac, hepatic, and embryonic development. J Biol Chem. 2003;278:1986–90.
51. Mahajan MA, Samuels HH. Nuclear receptor coactivator/coregulator NCoA6(NRC) is a pleiotropic coregulator involved in transcription, cell survival, growth and development. Nucl Recept Signal. 2008;6:e002.
52. Qi C, Kashireddy P, Zhu YWT, Rao SM, Zhu YJ. Null mutation of peroxisome proliferator-activated receptor-interacting protein in mammary glands causes defective mammopoiesis. J Biol Chem. 2004;279:33696–701.
53. Caira F, Antonson P, Pelto-Huikko M, Treuter E, Gustafsson JA. Cloning and characterization of *RAP250*, a novel nuclear receptor coactivator. J Biol Chem. 2000;275:5308–17.
54. Desvergne B, Michalik L, Wahli W. Transcriptional regulation of metabolism. Physiol Rev. 2006;86:465–514.
55. Lemay DG, Neville MC, Rudolph MC, Pollard KS, German JB. Gene regulatory networks in lactation: identification of global principles using bioinformatics. BMC Syst Biol. 2007;1:56.
56. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. Exposing synonymous mutations. Trends Genet. 2014;30:308–21.

# Paper II

SCIENTIFIC REPORTS

OPEN

# Unravelling genetic variation underlying *de novo*-synthesis of bovine milk fatty acids

Received: 18 July 2017
Accepted: 18 January 2018
Published online: 01 February 2018

Tim Martin Knutsen[1], Hanne Gro Olsen[1], Valeria Tafintseva[2], Morten Svendsen[3], Achim Kohler[2], Matthew Peter Kent[1] & Sigbjørn Lien[1]

**The relative abundance of specific fatty acids in milk can be important for consumer health and manufacturing properties of dairy products. Understanding of genes controlling milk fat synthesis may contribute to the development of dairy products with high quality and nutritional value. This study aims to identify key genes and genetic variants affecting *de novo* synthesis of the short- and medium-chained fatty acids C4:0 to C14:0. A genome-wide association study using 609,361 SNP markers and 1,811 animals was performed to detect genomic regions affecting fatty acid levels. These regions were further refined using sequencing data to impute millions of additional genetic variants. Results suggest associations of *PAEP* with the content of C4:0, *AACS* with the content of fatty acids C4:0-C6:0, *NCOA6* or *ACSS2* with the longer chain fatty acids C6:0-C14:0, and *FASN* mainly associated with content of C14:0. None of the top-ranking markers caused amino acid shifts but were mostly situated in putatively regulating regions and suggested a regulatory role of the QTLs. Sequencing mRNA from bovine milk confirmed the expression of all candidate genes which, combined with knowledge of their roles in fat biosynthesis, supports their potential role in *de novo* synthesis of bovine milk fatty acids.**

Bovine milk is an important source of many nutrients including proteins, fat, minerals, vitamins and bioactive lipid components. The relative abundance and concentration of individual fatty acids (FAs) in milk affect both human health and the manufacturing properties of dairy products. Myristic (C14:0) and palmitic acid (C16:0) are associated with cardiovascular disease through increased level of blood cholesterol[1], while shorter chain saturated FAs (C4:0 to C12:0) have been associated with positive health effects such as antiviral, antibacterial and anticancer activities[2–4]. The difference in melting point between saturated and unsaturated acids also affects the softness, flavour and colour of dairy products such as butter and cheese[5,6].

By improving our understanding of the pathways in bovine milk FA synthesis and identifying the genes and genetic polymorphisms associated with variation in milk FA content, it may be achievable through genome-based selection methods[7] to optimally balance individual FAs allowing industry to satisfy consumer demands for healthy food of high quality. The short- and medium-chain length acids C4:0 to C14:0 are potential targets for this purpose. In contrast to the bulk of long-chained milk FAs and around half of C16:0 which are largely derived from the cow's diet, C6:0 to C14:0 and a fraction of C4:0 are synthesized *de novo* in the bovine mammary gland[8]. These acids occur in milk in relatively high concentrations and show moderately high heritabilities (usually in the range of 0.10 to 0.50)[9–12] and are therefore well suited for genetic analyses such as a genome-wide association study (GWAS).

The synthesis of short- and medium-chained FAs is founded upon C2 and C4 precursors absorbed from the diet. After being transported to the mammary gland, acetate and acetoacetate are converted to acetyl-CoA and then to malonyl-CoA which, along with butyryl-CoA (from plasma β-hydroxybutyrate and C2), are used as precursors for cytosolic *de novo*-synthesis. The process of carbon chain elongation from C2:0 or C4:0 to C16:0 involves a cyclic reaction[13] which also generates intermediate products, C4:0 to C14:0, via a chain termination mechanism[14]. Newly synthesised FAs are transported from the cytosol to the endoplasmic reticulum where they

[1]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Faculty of Life Sciences (BIOVIT), Norwegian University of Life Sciences (NMBU), PO Box 5003, Ås, Norway. [2]Faculty of Science and Technology (RealTek), Norwegian University of Life Sciences (NMBU), PO Box 5003, Ås, Norway. [3]Geno Breeding and AI Association, N-1432, Ås, Norway. Correspondence and requests for materials should be addressed to H.G.O. (email: hanne-gro.olsen@nmbu.no)

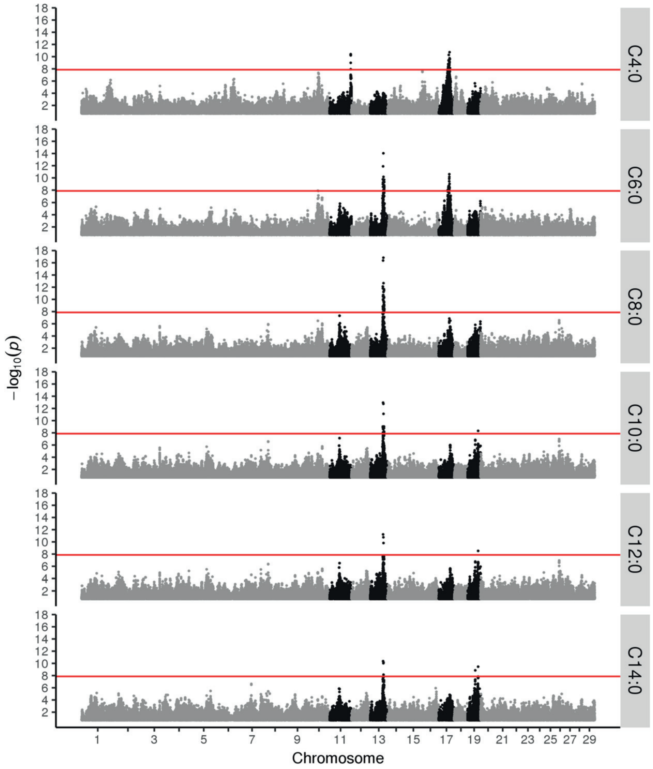


**Figure 1.** Manhattan plots showing results from genome-wide association analyses of C4:0 to C14:0 on high-density marker data. Chromosomes are shown along the abscissa while the ordinate denotes the $-\log_{10}$(p-value) for each marker – trait association. Chromosomes showing genome-wise significant associations for one or more of the tested acids are highlighted with black points. The red line denotes the genome-wide significance level.

are linked to a glycerol 3-phosphate backbone to form triacylglycerols, a final series of steps sees them secreted into the milk in the form of milk fat globules.

The current study explores genetic variation associated with the *de novo*-synthesis of short- and medium-chained FAs (C4:0 to C14:0). Milk fatty acid composition was predicted from Fourier transform infrared spectroscopy (FTIR) using prediction equations derived from GC/FTIR calibration sets. This method has been shown to provide fast and cheap large-scale phenotyping of the breeding population, especially for acids with relatively high concentration and heritability such as the *de novo*-synthesized FAs[12,15–17]. These phenotypes were combined with array-based single nucleotide polymorphism (SNP) genotypes in a genome-wide association study to identify chromosomal regions (quantitative trait loci - QTLs) with substantial effects on the traits under investigation. QTL regions identified on bovine chromosomes (BTA) 11, 13, 17 and 19 were re-analysed using a higher density of sequence variants (SNPs and indels) imputed from re-sequencing data in an attempt to identify putative functional polymorphisms. Moreover, mRNA sequence analysis of mammary epithelial cells from 36 milk samples was used to verify that the candidate genes indicated by GWAS were expressed in the mammary gland during milk production.

## Results

### Genome-wide association analyses for FA concentration.

Our analysis began with combining daughter yield deviations (DYDs) for C4:0 to C14:0 from 1,811 bulls with genotypes from 609,361 autosomal SNPs to perform a GWAS and identify chromosomal regions with a major impact on *de novo* synthesis of these acids.

As shown in Fig. 1, we found the most significant associations on BTA11, BTA13, BTA17 and BTA19. Results for all significant marker and trait combinations are provided in Supplementary Table S1. The QTL on BTA11 was most significant for the shortest of the tested acids; C4:0, while the one on BTA17 was significant for both C4:0 and C6:0. As FA chain length increases, these regions become less important, while the significance of the QTL on BTA13 increases. This QTL was most significant for acids with intermediate chain lengths (especially C8:0) with

| FA | BTA | rs number | Top variant (bp) | Ref allele | Alt allele | MAF | p-value |
|---|---|---|---|---|---|---|---|
| C4:0 | 11 | rs109837926 | 103,300,697 | C | A | 0.34 | 3.47e-9 |
| C4:0 | 13 | — | 62,280,697 | A | G | 0.13 | 7.37e-7 |
| C4:0 | 17 | rs477658921 | 53,078,216 | GAAAGTGA | G | 0.08 | 8.09e-11 |
| C4:0 | 19 | rs797503644 | 52,884,766 | G | A | 0.28 | 2.17e-8 |
| C6:0 | 13 | rs41700742 | 64,648,620 | A | G | 0.45 | 6.82e-16 |
| C6:0 | 17 | rs379029510 | 51,669,903 | T | C | 0.46 | 2.05e-10 |
| C6:0 | 19 | rs476079746 | 37,421,626 | G | GAAAAAA | 0.40 | 5.86e-9 |
| C8:0 | 13 | rs381037433 | 64,523,817 | G | GA | 0.21 | 1.08e-18 |
| C8:0 | 17 | rs456738710 | 51,161,184 | A | T | 0.22 | 1.31e-8 |
| C8:0 | 19 | rs457952543 | 51,334,328 | C | CT | 0.03 | 1.26e-6 |
| C10:0 | 13 | rs381037433 | 64,523,817 | G | GA | 0.21 | 7.6e-16 |
| C10:0 | 17 | rs456738710 | 51,161,184 | A | T | 0.22 | 2.47e-8 |
| C10:0 | 19 | rs109016955 | 51,381,233 | G | C | 0.04 | 5.21e-9 |
| C12:0 | 13 | rs381037433 | 64,523,817 | G | GA | 0.21 | 4.79e-14 |
| C12:0 | 17 | rs456738710 | 51,161,184 | A | T | 0.22 | 2.17e-7 |
| C12:0 | 19 | rs109016955 | 51,381,233 | G | C | 0.04 | 7.5e-9 |
| C14:0 | 13 | rs381037433 | 64,523,817 | G | GA | 0.21 | 5.2e-14 |
| C14:0 | 17 | rs384370770 | 51,231,279 | T | C | 0.09 | 3.67e-8 |
| C14:0 | 19 | rs109016955 | 51,381,233 | G | C | 0.04 | 4.07e-11 |

**Table 1.** Summary of the top variants on chromosomes 11, 13, 17 and 19 determined by single-marker association analyses of sequence variants for fatty acids C4:0 to C14:0. FA, fatty acid; BTA, *bos taurus* chromosome; rs, rs number; Top variant, position of the most significant markers in base pairs; Ref allele, reference allele; Alt allele, alternative allele; MAF, minor allele frequency.

decreasing significance for shorter (C6:0) and longer acids (C10:0–C14:0). Finally, the QTL on BTA19 becomes the most significant for the synthesis of the longest of the analysed acids; C14:0.

All major QTL regions spanned genes with an established function in milk fat biosynthesis. The QTL on BTA11 was detected close to the *associated endometrial protein* (*PAEP*) gene at 103.3 Mb. The QTL region on BTA13 was rather broad and covered at least two potential candidate genes; *nuclear receptor coactivator 6* (*NCOA6*) at 64.6 Mb and *acyl-CoA synthetase short-chain family member 2* (*ACSS2*) gene at 64.8 Mb. BTA17 displayed a QTL close to the *acetoacetyl-CoA synthetase* (*AACS*) gene at 53 Mb. Closer examinations of BTA19 revealed that the associations were located in two distinct regions; one at 37.4 Mb which is around 500 kb from *acyl-CoA synthetase family member 2, mitochondrial precursor* (*ACSF2*) at 36.9 Mb. The second QTL region was close to *fatty acid synthase* (*FASN*) at 51.4 Mb. However, analysis of sequence variants revealed that the significant markers detected around 36.9 Mb were not situated within or very close to *ACSF2*. Since no other convincing candidate gene was detected in this region, we chose not to follow up this QTL in further analyses.

**Fine-mapping of imputed sequence variants on selected chromosomes.** To characterize as much genetic variation as possible in and around the candidate genes we imputed SNPs and indels identified from whole genome sequence data resulting in a more than 20-fold increase in marker density after quality filtering in the regions 90–110 Mb on BTA11, 60–70 Mb on BTA13, 20–60 Mb on BTA17, and 45–55 Mb on BTA19. The quality of imputation relates most to marker allele frequencies. The Beagle software[18] reports an internally calculated parameter, allelic r-squared (AR$^2$), that is the estimated squared correlation between the most likely allele and the true allele for each marker[19]. The mean value of this parameter ranged from 0.84 for variants with minor allele frequency (MAF) below 0.05 to 0.94 for variants with MAF above 0.05. Imputed SNPs and indels, with AR$^2$ above 0.7, were included in a reanalysis of the QTLs for significant phenotype associations. Detailed information of the top significant variants on each chromosome for each FA tested, is shown in Table 1.

**Chromosome 11.** Results for BTA11 showed the strongest associations between C4:0 and a group of markers that were situated within and immediately outside of *PAEP* (Fig. 2). The most significant marker (rs109837926; p-value = 3.5e-9) was found at position 103,300,697 bp which is ≈800 bp upstream from *PAEP*'s transcription start site. The minor A allele (MAF = 0.34) was associated with a slight, but noteworthy increase in C4:0 levels (0.02 g/100 g milk fat). Closer examination of the haplotype containing the top ranked markers (all of which had p-values, effects and frequencies similar to rs109837926) revealed that the minor alleles for all markers were included in a single haplotype (frequency ≈ 0.3) that covered a region beginning 11 kb upstream from *PAEP* and extending into the neighbouring gene *glycosyltransferase 6 domain containing 1* (*GLT6D1*). The high level of linkage disequilibrium (LD) among these markers (Fig. 2) restricted our ability to pinpoint any one of them as causal. Among other top-ranking markers were two missense variants in *PAEP*, known to produce the A and B protein variants of beta-lactoglobulin (at 103,303,475 bp in exon 3 and 103,304,757 bp in exon 4), one splice region variant (at 103,304,656 bp in *PAEP* exon 3), and three SNPs in the 5′UTR of *PAEP* (103,301,561 bp, 103,301,690 bp, and 103,301,694 bp). The four top-ranked markers were grouped in a region ≈1,000 bp upstream of *PAEP* which

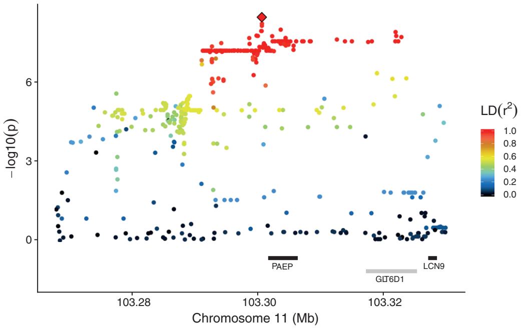


**Figure 2.** Results for BTA11 - C4:0 association analysis using imputed sequence variant data on BTA11, zoomed in on the region between 103.27 and 103.33 Mb. The ordinate provides $-\log_{10}$(p-value) for each marker – trait association, while the abscissa denotes marker position. The red diamond indicates the most significant marker for C4:0; rs109837926 at position 103,300,697 bp. Colouring indicates the level of LD (r$^2$) between each marker and rs109837926. Gene annotation information (Ensembl Bos taurus annotation release 86) is shown with grey and black bars reflecting positive and negative strand orientations respectively.

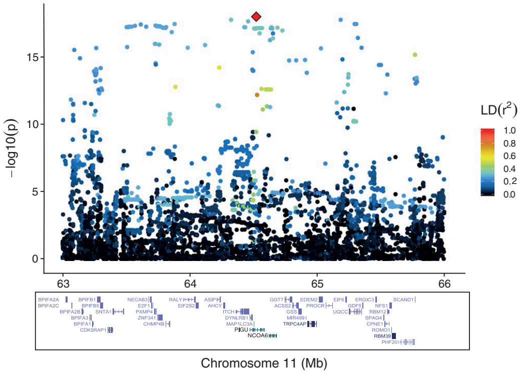


**Figure 3.** Results for BTA13 - C8:0 association analysis using imputed sequence variant data on BTA13, zoomed in on the region between 63 and 66 Mb. The ordinate provides $-\log_{10}$(p-value) for each marker – trait association, while the abscissa denotes marker position. The red diamond indicates the most significant marker for C8:0; rs381037433 at position 64,523,817 bp. Colouring indicates the level of LD (r$^2$) between each marker and rs381037433. Gene annotation information (Ensembl Bos taurus annotation release 86) is shown with blue bars reflecting position and exon structure.

can suggest a regulatory role of the QTL. Results for all tested markers and trait combinations are shown in Supplementary Fig. S6 and Supplementary Table S2.

### Chromosome 13.

On BTA13, the most significant results were found for C8:0, with decreasing significance levels for acids with shorter and longer chain lengths. For all traits, we detected similar p-values for a large number of markers in a region spanning from approximately 63.5 to 65.4 Mb (Fig. 3) that covered at least 39 characterised genes (NCBI Bos taurus Annotation Release 105, UMD 3.1.1) including the two genes regarded as most potent candidates; *NCOA6* and *ACSS2*. The most significant marker for C8:0, C10:0, C12:0 and C14:0 was rs381037433 at 64,523,817 bp (p-values = 1.08e-18, 7.6e-16, 4.8e-14 and 5.2e-14, respectively), which is an intronic insertion (G/GA) in *phosphatidylinositol glycan anchor biosynthesis class U* (*PIGU*). The insertion had a frequency of 0.21 and was associated with a reduction of C8:0 level of 0.02 g/100 g milk fat. C6:0 was most significantly associated to rs41700742 at 64,648,620 bp (p-value = 6.8e-16), which is a synonymous SNP in *NCOA6*. LD (r$^2$) between these two markers is 0.3.

Many other markers in the 63.5 to 65.4 Mb region displayed p-values and allele substitution effects similar to those of rs381037433. However, MAFs varied from 0.07 to 0.47. Haplotype analyses revealed that the least frequent allele of all these markers was present in one specific haplotype with a frequency of approximately 0.08 that spanned the entire 63.5 to 65.4 Mb region. For markers where the MAF was higher than 0.08, the least frequent allele was also found in other haplotypes. Hence, the LD (r$^2$) among the most significant markers were generally low.

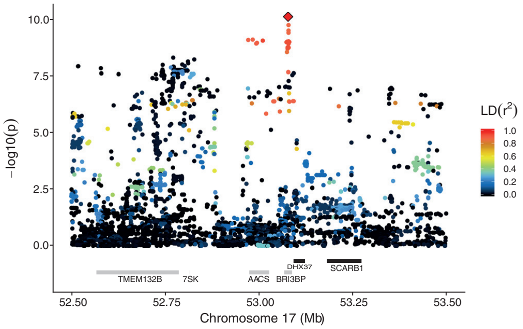


**Figure 4.** Results for BTA17 - C4:0 association analysis using imputed sequence variant data on BTA17, zoomed in on the region between 52.5 and 53.3 Mb. The ordinate provides $-\log_{10}$(p-value) for each marker – trait association, while the abscissa denotes marker position. The red diamond indicates the most significant marker for C4:0; rs477658921 at position 53,078,216 bp. Colouring indicates the level of LD ($r^2$) between each marker and rs477658921. Gene annotation information (Ensembl Bos taurus annotation release 86) is shown with grey and black bars reflecting positive and negative strand orientations respectively.

Only two non-synonymous SNPs were found among these top-ranking markers; rs383480158 in *peroxisomal membrane protein 4* (*PXMP4*) and rs446495267 in *PIGU*. Also, there were two 3′ UTR variants in *zinc finger protein 341* (*ZNF341*) and *ENSBTA00000000308*. However, neither of these genes have a function that can easily be related to milk fat synthesis. All other significant markers were either synonymous (i.e. not causing an amino acid shift) or positioned in non-coding regions such as introns and intergenic regions. This suggests a regulatory role also for this QTL. Results for all tested markers and trait combinations are shown in Supplementary Fig. S7 and Supplementary Table S3.

**Chromosome 17.** In agreement with the GWAS analysis, imputed sequence variants on BTA17 were found to have a main effect on short C4:0 and C6:0 fatty acids. The most significant association was found for C4:0 and rs477658921 at 53,078,216 bp (Fig. 4). This is a 7-bp indel (GAAAGTGA/G) where the minor G allele (MAF = 0.08) was associated with an increase of C4:0 level of 0.05 g/100 g milk fat. This variant showed a lower significance against C6:0 (p-value = 1.3e-9) and no significance with other longer acids.

The rs477658921 indel is located within intron 1 of *BRI3 binding protein* (*BRI3BP*), which does not appear to be an especially good functional candidate gene, but it is also in close proximity to *AACS* at 52.97–53.03 Mb which may be involved in utilizing ketone body for fatty acid-synthesis. LD among the top-ranking markers was high ($r^2$ higher than 0.84), indicating that the significance of the rs47765892 polymorphism on C4:0 could be a result of polymorphisms in or near *AACS*. As with BTA11, we found that the least frequent alleles of the most significant markers were contained within a haplotype with a frequency of 0.083 that spanned *AACS* and *BRI3BP*. All the top-ranking markers are either situated in introns of or outside these two genes and suggest a regulatory role of the QTL.

A second peak was detected at 51.49 Mb within the *zinc finger protein 280B* (*ZNF280B*). The LD between significant SNPs within this QTL region and the QTL embracing *AACS* and *BRI3BP* at 53.07 Mb is low which suggests that these are two different QTLs. The p-values of this second peak were approximately 1e-9 for all traits. Results for all tested markers and trait combinations are shown in Supplementary Fig. S8 and Supplementary Table S4.

**Chromosome 19.** The QTL on BTA19 was most strongly associated to C14:0 with significance levels decreasing for acids with shorter chain length and until it dropped below the significance threshold for C8:0 and shorter acids. The most significant marker for C14:0, C12:0 and C10:0 was rs109016955 at 51,381,233 bp (p-values = 4.1e-11, 7.5e-9 and 5.2e-9, respectively). This G/C SNP has a MAF of 0.04 where the minor C allele was associated with a reduction of C14:0 level of 0.14 g/100 g milk fat. It is situated ≈3,7 kb upstream of the transcription start site of *FASN* and annotated both as an upstream gene variant of *FASN* and as a 3′ untranslated region (3′ UTR) variant and a non-coding exon variant in various predicted transcript variants of *coiled-coil domain containing 57* (*CCDC57*). Similarly high significance levels were found for 24 variants situated either in introns of *CCDC57* and *FASN* or in the region between these two genes (Fig. 5). The MAF of all these markers were very low (0.03 to 0.05), and most of them showed only moderate LD with rs109016955 (Fig. 5). Results for all tested markers and trait combinations are shown in Supplementary Fig. S9 and Supplementary Table S5.

**Investigations of expression levels and transcripts.** To verify that the candidate genes we detected from GWAS are present in the udder during lactation we isolated mRNA from somatic milk cells and measured their level of expression. Specifically, all genes in the region between 63.5 to 65.4 Mb on BTA13 were tested, along with candidates and, where appropriate, neighbour candidates on BTA11, 17 and 19. Expression levels in the form of mean normalised gene count can be found as Supplementary Table S10. *PAEP* was the most abundantly expressed gene of those found to be significant in the association analyses, with a mean count of approximately

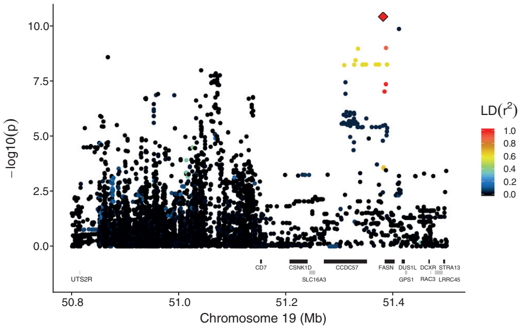


**Figure 5.** Results for BTA19 – C14:0 association analyses using imputed sequence variant data on BTA19, zoomed in on the region between 50.8 and 51.5 Mb. The ordinate provides $-\log_{10}$(p-value) for each marker – trait association, while the abscissa denotes marker position. The red diamond indicates the most significant marker for C14:0; rs109016955 at position 51,381,233 bp. Colouring indicates the level of LD ($r^2$) between each marker and rs381037433. Gene annotation information (Ensembl Bos taurus annotation release 86) is shown with grey and black bars reflecting positive and negative strand orientations respectively.

444,000 reads. No reads were found for its neighbour *GLT6D1*. The QTL region on BTA13 contains at least 39 characterised genes (NCBI *Bos taurus* Annotation Release 105, UMD 3.1.1) of which 28 genes were found to be expressed in somatic milk cells. Mean expression levels varied from two to 5,407 normalised counts for these genes, with the highest expression found for *ACSS2*. *NCOA6* showed the eight highest expression level of these genes. On BTA17, both *AACS* and *BRI3BP* were expressed in the SMC, but expression level of *AACS* was much higher than for *BRI3BP*. *FASN* on BTA19 showed the second highest expression level of the studied genes with a normalised read count of ~29,000, which was approximately 233 times higher than the expression level of the neighbour *CCDC57*.

## Discussion

Understanding the role of genetic variation on fatty acid composition in bovine milk may reveal opportunities to produce superior raw product, and at the very least will improve our understanding of the genetics of fatty acid synthesis. Uniquely, our study combined data from 4.6 million FTIR recordings (FA composition phenotypes) representing 640,000 cows, with combined high-density genotyping and whole genome sequencing representing 1,811 bulls. This analysis allowed us to reveal a number of genetic variants associated with the synthesis of short- and medium-chained FAs C4:0 to C14:0.

We identified one gene, *PAEP*, which is a novel candidate gene in the context of fatty acid content, and several previously known candidate genes. Our results revealed that concentration of C4:0 was most strongly affected by *PAEP* on BTA11 and *AACS* on BTA17. The QTL on BTA13, which is most likely caused by *NCOA6* or alternatively *ACSS2*, seems to be related to the generation of longer chain length acids, while the *de novo*-synthesis of the longest chain length acid, C14:0, is most strongly affected by a polymorphism in or around *FASN* on BTA19.

A key condition for using phenotype data (FA composition) predicted from FTIR profiles in an association study is that individual acids can be predicted with a high degree of confidence. The effectiveness of mid-infrared spectroscopy to predict bovine milk fatty acid composition have been thoroughly discussed in a number of papers[12,15–17,20–24]. Inaccurate predictions and correlations among acids or between acids and other milk components may reduce the ability to identify true QTLs and determine exactly which fatty acids that are affected. We have previously reported that FAs with a concentration of 1% or higher are predicted with acceptable accuracies[12]. This finding was also reflected in the current study, where all the tested *de novo*-synthesized acids (i.e., C4:0 to C14:0) were present in concentrations above 1% and had prediction accuracies (cross-validated squared Pearson product-moment correlation coefficients; $R^2$CV) ranging from 0.73 (C4:0) to 0.90 (C6:0 - C14:0) and hence were considered well predictable. An argument against using FTIR to predict FA composition is that the acids are correlated to total fat and the prediction merely reflects total fat rather than individual acids[20]. This correlation was accounted for in our two studies by presenting the fatty acid concentrations as percentages of total fat instead of as gram acid per unit of milk[21–23]. Soyeurt[21] suggested that predictions were due to real absorbance of the acids if the calibration correlations were higher than the correlations between the acids and total fat. As reported in our previous study[12], the squared correlation to total fat ranged between 0.001 and 0.012, indicating that the predicted concentrations are due to real absorbance values of these acids. A consequence of this normalization is that the prediction accuracies are expected to be lower than when FA concentrations are expressed as a quantity per unit of milk[21–23], however with the exception of C4:0, the accuracies were found to be comparable to those obtained by milk-based models[12,21–23]. Although C4:0 was predicted with lower accuracy than the other FAs included in our study, our analysis detected two candidate genes with functions judged relevant for C4:0 content. Separating FAs with similar chain lengths such as C4:0 and C6:0 using an FTIR approach can be challenging since their chemical structure is relatively similar, however, the technology allowed us to identify two clearly different QTL profiles for C4:0 and C6:0 which, if the phenotype measurements were severely confounded together, would not be as distinct as they appear to be.

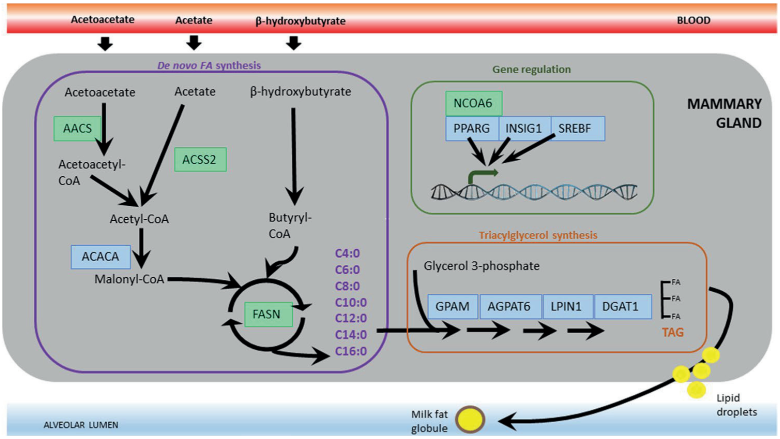


**Figure 6.** Illustration of the most relevant pathways and genes involved in *de novo* synthesis of short- and medium-chained fatty acids in the bovine mammary gland. Detected candidate genes are highlighted in green, whereas some additional well studied genes of high importance are shown in blue.

The genes highlighted as candidates for *de novo*-synthesis have, essentially, defined roles in bovine milk fat synthesis, and operate across the core pathways responsible for *de novo*-synthesis and triacylglycerol metabolism (Fig. 6). Early in *de novo*-synthesis, *ACSS2* facilitates the conversion of acetate to acetyl-CoA[25]. Alternatively, acetyl-CoA may be derived from acetoacetyl-CoA in the process beginning with the production of acetoacetate-CoA from acetoacetate by *AACS*[26]. Later, *FASN* oversees a process whereby palmitate (C16:0) is synthesised from acetyl-CoA and malonyl-CoA in a repeated, cyclic reaction. Importantly, intermediate length acids (C4:0 to C14:0) can leave the elongation cycle before the chain reaches full length. The entire lipid synthesis machine is regulated by a network of genes encoding transcription factors and nuclear receptors. One of these, *peroxisome proliferator-activated receptor gamma* (*PPARG*), is a well described transcriptional regulator affecting lipid storage[25,27], while *NCOA6* (which is a ligand for *PPARG* and *PPARA*[28,29]) is a transcriptional coactivator enhancing the activity of, among other things, *PPARG*.

*PAEP* encodes the milk protein beta-lactoglobulin (β-LG) which is the major whey protein in bovine milk. Although the effect of *PAEP* on milk production traits including total fat yield and fat percentage has been well documented[30,31], its influence on individual fatty acids is poorly understood. β-LG is found to bind both saturated and unsaturated FA *in vitro*[32], which might suggest a function as an intracellular transporter of FAs. The B variant of the β-LG protein is commonly known to be less abundant than the A variant[31,33], and it is unclear if the effect of *PAEP* on C4:0 found in our study is due to the polymorphisms causing the A and B β-LG variants or to regulatory sites affecting *PAEP* expression. Although the promoter region has been extensively studied, the causal polymorphism has not been identified due to an extensive level of LD between the two polymorphisms that produce the A and B protein variants and polymorphisms situtated within putative transcription factor binding sites[34–37]. The effect of this QTL on C4:0 could possibly be due to the combined influence of alterations in several sites simultaneously rather than to one specific SNP in one single site.

With regards to BTA13, previous studies have pointed towards *ACSS2*[38–40] and *NCOA6*[12] as positional and functional candidates. Due to high levels of LD among SNPs in the 2 Mb QTL region embracing these genes, our association analyses have expanded the BTA13 candidate list to include 39 characterised genes of which several have functions related to milk fat biosynthesis. This list also includes *E2F transcription factor 1* (*E2F1*) which is shown to regulate important genes involved in FA synthesis such as *ACSL1*, *FASN* and *PPARG*[41], and *agouti signalling protein* (*ASIP*) which might regulate lipid metabolism in adipocytes[42]. However, the most significant markers detected by the association analyses were either found in non-coding regions or genes without known relevant functions in fat synthesis such as *PIGU*. Expression analyses revealed that 28 of the 39 genes were expressed in the bovine mammary gland during lactation. While *ACSS2* was distinctly more expressed than *NCOA6* in all samples (Supplementary Table S10), variants within and near *ACSS2* also showed a weaker association to the traits. Furthermore, since *NCOA6* contained variants that were among the top-ranking SNPs, we consider this gene to be the most promising positional candidate gene.

Our finding of an association between C4:0 and markers near *AACS* at ~53 Mb on BTA17 has not been reported in other GWA studies as far as we know. Li *et al.*[40] and Duchemin *et al.*[43] reported associations to markers on BTA17 in Chinese Holstein and Dutch Holstein-Friesian, respectively, but in other regions than *AACS*. This discrepancy may be explained by differences in study design (direct FA measurements compared to our study using DYDs estimated from millions of spectra from 640,000 cows) or the use of different breeds.

For BTA19, several authors have reported significant associations within or near *FASN* and the neighbouring gene *CCDC57*[40,44]. *CCDC57* is poorly characterised, and its putative role in milk fat synthesis is unknown. Medrano *et al.*[45] reported that *CCDC57* was expressed in mammary tissues of a lactating cow with expression levels higher than that of *FASN*. This is in contrast with the results of the present paper, where *FASN* was expressed more than 200 times higher than *CCDC57*. *FASN* is an obvious candidate gene because of its known role in fat synthesis. *FASN* has been extensively studied in candidate gene studies for fat content in milk and adipose tissue[46–51], but similarly to the present study, this has not yet resulted in a clear identification of a causal polymorphism.

Most genome scans performed in other cattle breeds than Norwegian Red cattle have reported strong associations between milk fatty acids and the genes *diacylglycerol acyltransferase 1* (*DGAT1*) on BTA14 and *stearoyl-coenzyme A desaturase 1* (*SCD*) on BTA26. *DGAT1* encodes an enzyme that catalyses the final stage of triacylglycerol synthesis[52], while *SCD* on BTA26 is involved in the synthesis of monounsaturated FAs by introducing a double bond in the delta-9 position of C14:0, C16:0 and C18:0, primarily, thus producing the *cis*-9 variant of these acids[53]. No significant associations have been detected near these genes in any of our studies of Norwegian Red cattle. Resequencing of 147 widely used NR bulls revealed that all individuals were homozygous for the A variant of the DGAT1 K232A polymorphism (data not shown), suggesting that this variant is almost fixed in Norwegian Red cattle. The *SCD* polymorphism does segregate in our breed but was not significantly associated with any fatty acid neither in the present study nor in the previous study where a larger number of acids also including C14:1*cis*-9 and C16:1*cis*-9 were analysed[12]. However, these acids were poorly predicted by the FTIR approach[12], which hampers the possibility to detect significant associations for these traits.

Imputation from HD-density to sequence level is heavily dependent upon MAFs and number of animals in the reference dataset[54]. In this study, 153 animals were whole genome sequenced and used as the imputation reference. When performed within breed, imputation for high-density genotypes to sequence has previously been shown to work acceptably with reference dataset of about 130[55]. We did not perform a cross validation procedure to test the expected accuracy in our dataset, but Beagle outputs a measure (AR$^2$), defined as estimated squared correlation between most probable and true genotype, depending on the internally calculated uncertainty in the imputation model for each marker[56]. All markers with AR$^2$ below 0.7 was filtered from our marker list before association analysis, as values above this threshold has shown to be a good indicator for reliable imputation accuracies[55,57,58]. Overall, mean AR$^2$ was 0.92 for all sequence-level imputed variants, and 0.84 for variants with MAF below 0.05. AR$^2$ was close to 1 for all variants found significant by the association analyses of sequence-level variants.

## Conclusions

Understanding of genes and polymorphisms controlling milk fat synthesis may reveal opportunities to tailor the fatty acid content and thereby improve the nutritional value and quality of dairy products. In this study we identified a set of positional candidate genes within milk fat synthesis pathways by combining dense genotyping and whole genome sequencing with high-throughput phenotypes for *de novo* synthesis of milk fatty acids. These genes were *PAEP* (on BTA11), *AACS* (BTA17), *NCOA6* or *ACSS2* (BTA13) and *FASN* (BTA19). Their roles in fatty acid synthesis were further supported by their expression levels in milk.

## Methods

**Ethics statement.** All animals included in the study were Norwegian Red cattle, and experiments were conducted in accordance with the rules and guidelines outlined in the Norwegian Animal Welfare Act 2009, issued by the Norwegian Ministry of Agriculture and Food. Most data were generated as part of routine commercial activities outside the scope of that requiring formal committee assessment and ethical approval (as defined by the above guidelines).

**Estimation of bovine milk fat composition from FTIR spectroscopy data.** Milk fat composition was estimated from FTIR spectroscopy data as described in Olsen *et al.*[12], with some adjustments to number of spectra and animals used. In brief, 224 milk samples obtained from a feeding experiment and 659 samples from field sampling were analysed in parallel by FTIR and gas chromatography with flame ionization detector (GC-FID) reference analysis. FTIR spectra (regressors) were subsequently calibrated against GC-FID reference values (regressands) by using powered partial least squares regression (PPLSR[59]). Regressands were presented as percentages of GC-FID fatty acid values to total fat to reduce to a minimum value the correlation between the FA and total fat in milk samples. The calibration model was applied to a total of 4,619,737 infrared spectra from 640,304 cows sampled in the periods February to November 2007 and July 2008 to June 2014. The traits that were calibrated for in this study were C4:0, C6:0, C8:0, C10:0, C12:0 and C14:0.

A detailed description of the estimation of heritabilities and DYDs is given in in Olsen *et al.*[12]. In short, the estimation of heritabilities were performed on a reduced dataset of 2,209,486 profiles from 426,505 cattle with a pedigree of 716,753 animals using the DMU software version 6 release 5.1[60]. The data were analysed with the following mixed linear animal repeatability model:

$$Y = RYM_i + RPL_j + htd_k + pe_l + a_m + e_{ijklm} \tag{1}$$

where RYM is the fixed effect of region (9 regions) by year and month of the test-day, with i ranging from 1 to 740; RPL is the fixed effect of region by lactation number by 10-day period in lactation of the test-day, with j ranging from 1 to 1,116; htd is the random effect of herd by test-day, with k ranging from 1 to 168,483; pe is the random permanent environmental effect of the cow on her repeated records, with l ranging from 1 to 426,505; a is a random additive genetic effect of the animal, with m ranging from 1 to 716,753; and e is a random residual effect.

DYDs for the GWAS were then estimated using the 4,619,737 spectra for the full dataset of 640,304 cows with a pedigree of 999,470 animals as the sire averages of daughters' predicted FA compositions, which were each corrected for her fixed effects, non-genetic random effects and half of her dam's genetic effect[12].

The concentration of each fatty acid together with the accuracy of prediction (in the form of cross-validated squared Pearson product-moment correlation coefficients; $R^2CV$) and heritabilities were as reported in Olsen *et al.*[12]. Mean concentration ranged from 1.48% of total fat for C8:0 to 11.21% of total fat for C14:0. $R^2$ ranged from 0.73 for C4:0 to 0.91 for C8:0, C10:0 and C12:0, while heritabilities ranged from 0.11 for C14:0 to 0.35 for C4:0.

**Construction of a dense SNP dataset.** Genotypes for the studied animals were available from other projects and the routine genotyping performed by Geno Breeding and AI Association. DNA was extracted from semen samples of bulls and from blood samples of cows using standard phenol-chloroform-based protocols. The bulls were genotyped on at least one of four different platforms in order to make a genome-wide high-density SNP dataset for the association analyses; the Affymetrix 25 K SNP array (Affymetrix, Santa Clara), a custom Affymetrix 50 K SNP array, the Illumina 54 K BovineSNP50 BeadChip (Illumina, San Diego) and the 777 K Illumina BovineHD Genotyping BeadChip (Illumina, San Diego).

Imputation was done in a step-wise manner, were the 25 K Affymetrix dataset was imputed to the custom 50 K Affymetrix density, and then the combined Illumina 54 K and Affymetrix 50 K dataset were imputed to 777 K. The Affymetrix 50 K reference counted 5,009 NR animals and the Illumina 777 K reference consisted of 750 widely used AI bulls. Imputation was done using Beagle version 4.1[18], with effective population size (Ne) set to 200 and number of phasing iterations (niterations) set to 20. Remaining parameters were set to default. Map positions were based on the UMD 3.1 reference assembly[61].

For each imputation step, the following quality control of the markers was applied: Variants with MAF less than 0.01 and Hardy-Weinberg Equilibrium p-values less than 1e-7 were filtered. Animals with more than 10% Mendelian errors were removed from the dataset, and all remaining genotypes with Mendelian errors were set to missing and later imputed. Markers and animals with a call rate below 95% were removed. Markers on sex chromosomes were discarded. For each step, the imputation quality was tested using 5-fold cross validation. Markers with discordance between true and imputed genotypes above 10% were removed, as these markers are likely to be misplaced in the reference assembly[62]. SNPs on unplaced scaffolds and sex chromosomes were also discarded from the dataset.

A total of 2,434 genotyped AI bulls were considered for the initial 777 k GWAS analysis. After filtering bulls with less than 20 daughters, the dataset contained 1,811 bulls with imputed genotypes for the 777 K Illumina BovineHD BeadChip. Of the 1,811 bulls, 57 bulls had genotypes imputed from the Affymetrix 25 K array, 237 were imputed from the custom Affymetrix 50 K SNP array, 1,113 animals from the Illumina 54 K BeadChip and 404 were already genotyped on the 777 k Illumina BovineHD BeadChip. The resulting dataset consisted of 1,811 bulls with trait data in the form of DYDs based on 20 or more daughters for the relevant FAs and with genotypes for 609,361 SNPs distributed on all 29 autosomes.

**Whole-genome sequencing and variant calling.** Whole-genome sequencing data were obtained from 153 animals (132 AI bulls and 31 cows) as described in Olsen *et al.*[63]. The AI bulls were selected based on maximum number of daughters in production and by ensuring an even contribution to the population structure of Norwegian Red cattle, by manually examining the recorded pedigree. Animals were sequenced to an average coverage of 9 × using Illumina sequencing (Illumina, San Diego). All reads were aligned against UMD 3.1 using BWA-mem version 0.7.10[64]. Variant calling was done with FreeBayes version 1.0.2[65]. Missing genotypes in the resulting Variant Call Format (VCF) file were imputed and phased using Beagle version 4.1[18]. This phased dataset was used as a reference panel for imputing the 1,811 animal high-density panel to full sequence with the Beagle software using the same imputation parameters as described before except that expected allele miscall rate (err) were set to 0.01. In a final filtering step, variants with minor allele frequency above 0.02 were kept. Also, variants with Beagle's reported allelic $R^2$ ($AR^2$) below 0.7 were filtered, as this has been shown to be a robust and reliable threshold for filtering of imputed sequence variants[56–58]. The raw, unfiltered VCF-file were kept for future reference.

**Genotyping of cows.** The 36 cows used for the RNA sequencing were also genotyped on the Illumina BovineSNP50 BeadChip (54 K, Illumina, San Diego). Blood samples were collected by certified personnel, and DNA extraction and genotyping on the Illumina BovineSNP50 BeadChip (54 K, Illumina, San Diego) were performed according to the manufacturer's protocol. Genotypes were quality checked and imputed to sequence density as described above.

**Genome-wide association studies.** A single-marker genome-wide association analysis was performed for the fatty acids C4:0 to C14:0, and 609,361 genome-wide distributed SNPs. This analysis was conducted with the GCTA software[66] for computational feasibility. A mixed linear model association analysis was performed with the –mlma-loco option of GCTA. The model fitted to the performance information for each trait and each SNP was:

$$DYD = a + bx + g^{-} + e \tag{2}$$

were DYD is the performance of the bull, a is the mean term, b is the fixed additive effect of the candidate SNP to be tested for association, x is the SNP genotype indicator variable coded as 0, 1 or 2, $g^-$ is the random polygenic effect, i.e. the accumulated effect of all SNPs except those on the chromosome where the candidate SNP is located, and e is the residual. The var($g^-$) will be re-estimated each time when a chromosome is excluded from calculating the genomic relationship matrix. The chromosome-wide significance level was set at p = 1e-5 which is a default value from qqman[67]. The genome-wide significance level was set at (0.05/609,361*6) = 1.37e-8, corresponding to a nominal type I error rate of 0.05 and Bonferroni correction for 609,361 markers and 6 traits.

**Re-analyses of selected regions on sequence-level variants.** All sequence-level polymorphisms that passed quality control and were situated in the QTL regions detected by the GWAS were analysed using the ASReml package version 2.0[68]. ASReml were selected for this step since it allowed us to weight the DYDs by number of daughters as well as to use genotype dosage data as input.

Analysed regions and traits were 100 to 107 Mb on BTA11 (C4:0), 60 to 70 Mb on BTA13 (C6:0 to C14:0), 20 to 60 Mb on BTA17 (C4:0 and C6:0) and 45 to 55 Mb on BTA19 (C10:0 to C14:0).

The model that was fitted to the information on performance for each trait – marker combination was:

$$\mathbf{DYD} = \mathbf{1}\mu + \mathbf{X}\mathrm{b} + \mathbf{Za} + \mathbf{e} \tag{3}$$

where **DYD** is the vector of bull performances weighed by the number of daughters, **1** is a vector of ones, $\mu$ is the overall mean, **X** is a vector of marker genotypes coded as a decimal number between 0 and 2 depending on the estimated dosage of the alternate allele (as reported by Beagle 4.1), b is the fixed effect of the marker, **Z** is an incidence matrix relating phenotypes to the corresponding random polygenic effects, **a** is a vector of random polygenic effects, and **e** is a vector of residual effects. Genetic and residual variances were estimated from the data. **a** was assumed to follow a normal distribution $\sim N(\mathbf{0}, \mathbf{A}\sigma_{\mathbf{A}}^2)$ where **A** is the relationship matrix derived from the pedigree, and $\sigma_{\mathbf{A}}^2$ is the additive genetic variance. **e** was assumed to follow a normal distribution $\sim N(\mathbf{0}, \mathbf{I}\sigma_{\mathbf{e}}^2)$ where $\sigma_e^2$ is the residual variance. Association analysis was performed for each individual marker. Since ASReml does not output p-values for the marker effect automatically, these were calculated from the F statistics for the conditional sum of squares, the numerator degrees of freedom and the denominator degrees of freedom with the R function pf() from the stats package version 3.4.0.

**Haplotype analyses.** Pairwise LD measurements ($r^2$) were estimated and haplotypes were identified for the top ranking markers within the relevant QTL regions using the Haploview 4.2 software[69] on phased genotypes. Haplotypes were defined by Haploview according to the confidence intervals strategy[70] or the four gamete rule[71].

**RNA isolation, sequencing and read mapping.** Gene expression levels were obtained using read counts from mRNA isolated from somatic milk cells (SMC) of 36 cows from the research facilities at the Norwegian University of Life Sciences, Aas, Norway. Pedigree information was used to avoid selection of close relatives. The cows were part of a research herd at our University. All milk samples were collected approximately 50 days (range 47 to 55) after calving. This sampling period was chosen since it coincides approximately with peak expression of several relevant genes involved in bovine milk fat synthesis including *FASN*[25] and also with the peak of synthesis and import of FAs in bovine milk[25] and the top of the lactation curve of Norwegian Red cows[72]. The cows were in different parities due to the limited size of the research herd. All animals were fed an equal standard diet.

Milk is excreted by the mammary epithelial cells (MEC) lining the inside of the udder, which are subject to turnover and shed into the milk and therefore represent a proportion of the somatic cells found in milk[73]. Cánovas *et al.*[74] found that compared to other sources (e.g. mammary gland tissue, laser dissected MEC), the quality of the total RNA extracted from the SMC was high. Moreover, the expression of genes investigated in SMC derived material was highly correlated with the expression observed in laser-dissected MEC. Several studies have confirmed the usefulness of this method[73,75,76].

Milk samples were collected manually 2–3 hours after milking to maximise the amount of viable cells present in the milk. Teats were cleaned with water followed by 70% ethanol before milking by hand, and 2 × 50 ml milk from each animal was collected in Falcon tubes. Samples were stored on ice immediately after collection and centrifuged at 4 °C for 10 min at 2,300 g within 1.5 hours to collect cells in the bottom of tubes. After centrifugation, most of the fat layer was removed with a clean pipette tip and supernatant decanted. Each pellet was dissolved in 4 ml 1xPBS by pipetting up and down. The liquid was transferred to a new 50 ml Falcon tube. Samples were centrifuged at 4 °C for 10 min at 2,300 g and supernatant decanted. Cell pellets were dissolved in 1 ml Trizol (Qiagen), and cells were lysed by pipetting up and down. Samples were stored in −80 °C until RNA extraction with Qiagen RNeasy Plus Universal Tissue Mini Kit (Qiagen) according to the manufacturer's protocol. RNA concentrations and quality were measured with a NanoDrop8000 spectrophotometer (Thermo Fisher Scientific) and Agilent RNA 6000 assay on Agilent BioAnalyzer 2100 (Agilent Technologies), respectively. All samples had an RNA integrity number (RIN) between 6.6 and 9.2. Samples were prepared for paired-end sequencing (2 × 150 bp) using the Illumina® TruSeq® stranded mRNA library preparation kits and sequenced by the Norwegian Sequencing Centre (www.sequencing.uio.no) using an Illumina HiSeq. 3000 platform.

Before mapping, raw read quality were assessed using fastQC version 0.11.5 https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Illumina adaptors were removed, and the sequences were quality-trimmed using cutadapt[77]. Cutadapt was set to cut adaptors with a minimum overlap length of 8 and low-quality 3′ ends were removed by setting a quality threshold of 20 (phred quality + 33). An index of the UMD 3.1 reference genome was built, and reads were aligned to the reference using STAR version 2.3.1[78]. Sorting, indexing and conversion to the BAM file format (the compressed binary version of a SAM file) of the resulting SAM files were completed using SAMtools version 1.3[79]. The code for the described RNAseq mapping method is available as part of a bash-script pipeline (version 1.1.0) found at https://gitlab.com/fabian.grammes/RNAseq-analysis/.

**Variant annotations.** All variants were annotated using the web version of Ensembl Variant Effect Predictor[80] based on the Ensembl Bos taurus annotation release 86.

**Availability of data.** DNA and RNA sequence data will be submitted to the European Nucleotide Archive, http://www.ebi.ac.uk/ena. Phenotype and genotype data are available only upon agreement with Geno Breeding and AI Organization (http://www.geno.no).

## References

1. Mensink, R. P., Zock, P. L., Kester, A. D. & Katan, M. B. Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: a meta-analysis of 60 controlled trials. *Am J Clin Nutr* **77**, 1146–55 (2003).
2. Mandal, M., Olson, D. J., Sharma, T., Vadlamudi, R. K. & Kumar, R. Butyric acid induces apoptosis by up-regulating Bax expression via stimulation of the c-Jun N-terminal kinase/activation protein-1 pathway in human colon cancer cells. *Gastroenterology* **120**, 71–8 (2001).
3. Sun, C. Q., O'Connor, C. J. & Roberton, A. M. Antibacterial actions of fatty acids and monoglycerides against Helicobacter pylori. *FEMS Immunol Med Microbiol* **36**, 9–17 (2003).
4. Thormar, H., Isaacs, C. E., Kim, K. S. & Brown, H. R. Inactivation of visna virus and other enveloped viruses by free fatty acids and monoglycerides. *Ann N Y Acad Sci* **724**, 465–71 (1994).
5. Bobe, G., Hammond, E. G., Freeman, A. E., Lindberg, G. L. & Beitz, D. C. Texture of butter from cows with different milk fatty acid compositions. *J Dairy Sci* **86**, 3122–7 (2003).
6. Coppa, M. *et al*. Milk fatty acid composition and cheese texture and appearance from cows fed hay or different grazing systems on upland pastures. *J Dairy Sci* **94**, 1132–45 (2011).
7. MacLeod, I. M. *et al*. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* **17**, 144 (2016).
8. Bauman, D. E. & Griinari, J. M. Nutritional regulation of milk fat synthesis. *Annu Rev Nutr* **23**, 203–27 (2003).
9. Stoop, W. M., van Arendonk, J. A., Heck, J. M., van Valenberg, H. J. & Bovenhuis, H. Genetic parameters for major milk fatty acids and milk production traits of Dutch Holstein-Friesians. *J Dairy Sci* **91**, 385–94 (2008).
10. Bastin, C., Soyeurt, H. & Gengler, N. Genetic parameters of milk production traits and fatty acid contents in milk for Holstein cows in parity 1-3. *J Anim Breed Genet* **130**, 118–27 (2013).
11. Lopez-Villalobos, N. *et al*. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in New Zealand dairy cattle. *J Dairy Res* **81**, 340–9 (2014).
12. Olsen, H. G. *et al*. 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* **49**, 20 (2017).
13. Neville, M. C. & Picciano, M. F. Regulation of milk lipid secretion and composition. *Annu Rev Nutr* **17**, 159–83 (1997).
14. Smith, S. The animal fatty acid synthase: one gene, one polypeptide, seven enzymes. *Faseb j* **8**, 1248–59 (1994).
15. Afseth, N. K. *et al*. Predicting the fatty acid composition of milk: a comparison of two Fourier transform infrared sampling techniques. *Appl Spectrosc* **64**, 700–7 (2010).
16. Soyeurt, H. *et al*. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *J Dairy Sci* **94**, 1657–67 (2011).
17. Maurice-Van Eijndhoven, M. H., Soyeurt, H., Dehareng, F. & Calus, M. P. Validation of fatty acid predictions in milk using mid-infrared spectrometry across cattle breeds. *Animal* **7**, 348–54 (2013).
18. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* **98**, 116–26 (2016).
19. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499–511 (2010).
20. Eskildsen, C. E. *et al*. Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: understanding predictions of highly collinear reference variables. *J Dairy Sci* **97**, 7940–51 (2014).
21. Soyeurt, H. *et al*. Estimating fatty acid content in cow milk using mid-infrared spectrometry. *J Dairy Sci* **89**, 3690–5 (2006).
22. Rutten, M. J., Bovenhuis, H., Hettinga, K. A., van Valenberg, H. J. & van Arendonk, J. A. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. *J Dairy Sci* **92**, 6202–9 (2009).
23. De Marchi, M. *et al*. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. *Animal* **5**, 1653–8 (2011).
24. Bonfatti, V., Degano, L., Menegoz, A. & Carnier, P. Short communication: Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. *J Dairy Sci* **99**, 8216–21 (2016).
25. Bionaz, M. & Loor, J. J. Gene networks driving bovine milk fat synthesis during the lactation cycle. *BMC Genomics* **9**, 366 (2008).
26. Buckley, B. M. & Williamson, D. H. Acetoacetyl-CoA synthetase; a lipogenic enzyme in rat tissues. *FEBS Lett* **60**, 7–10 (1975).
27. Liu, L. *et al*. Regulation of peroxisome proliferator-activated receptor gamma on milk fat synthesis in dairy cow mammary epithelial cells. *In Vitro Cell Dev Biol Anim* **52**, 1044–1059 (2016).
28. Caira, F., Antonson, P., Pelto-Huikko, M., Treuter, E. & Gustafsson, J. A. Cloning and characterization of RAP250, a novel nuclear receptor coactivator. *J Biol Chem* **275**, 5308–17 (2000).
29. Lemay, D. G., Neville, M. C., Rudolph, M. C., Pollard, K. S. & German, J. B. Gene regulatory networks in lactation: identification of global principles using bioinformatics. *BMC Syst Biol* **1**, 56 (2007).
30. Tsiaras, A. M., Bargouli, G. G., Banos, G. & Boscos, C. M. Effect of kappa-casein and beta-lactoglobulin loci on milk production traits and reproductive performance of Holstein cows. *J Dairy Sci* **88**, 327–34 (2005).
31. Berry, S. D. *et al*. Mapping a quantitative trait locus for the concentration of beta-lactoglobulin in milk, and the effect of beta-lactoglobulin genetic variants on the composition of milk from Holstein-Friesian x Jersey crossbred cows. *N Z Vet J* **58**, 1–5 (2010).
32. Le Maux, S., Bouhallab, S., Giblin, L., Brodkorb, A. & Croguennec, T. Bovine beta-lactoglobulin/fatty acid complexes: binding, structural, and biological properties. *Dairy Sci Technol* **94**, 409–426 (2014).
33. Folch, J. M., Dovc, P. & Medrano, J. F. Differential expression of bovine beta-lactoglobulin A and B promoter variants in transiently transfected HC11 cells. *J Dairy Res* **66**, 537–44 (1999).
34. Wagner, V. A., Schild, T. A. & Geldermann, H. DNA variants within the 5′-flanking region of milk-protein-encoding genes II. The beta-lactoglobulin-encoding gene. *Theor Appl Genet* **89**, 121–6 (1994).
35. Lum, L. S., Dovc, P. & Medrano, J. F. Polymorphisms of bovine beta-lactoglobulin promoter and differences in the binding affinity of activator protein-2 transcription factor. *J Dairy Sci* **80**, 1389–97 (1997).
36. Braunschweig, M. H. & Leeb, T. Aberrant low expression level of bovine beta-lactoglobulin is associated with a C to A transversion in the BLG promoter region. *J Dairy Sci* **89**, 4414–9 (2006).
37. Ganai, N. A., Bovenhuis, H., van Arendonk, J. A. & Visker, M. H. Novel polymorphisms in the bovine beta-lactoglobulin gene and their effects on beta-lactoglobulin protein concentration in milk. *Anim Genet* **40**, 127–33 (2009).
38. Bouwman, A. C., Bovenhuis, H., Visker, M. H. & van Arendonk, J. A. Genome-wide association of milk fatty acids in Dutch dairy cattle. *BMC Genet* **12**, 43 (2011).
39. Buitenhuis, B. *et al*. Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. *BMC Genomics* **15**, 1112 (2014).
40. Li, C. *et al*. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. *PLoS One* **9**, e96186 (2014).
41. Denechaud, P. D. *et al*. E2F1 mediates sustained lipogenesis and contributes to hepatic steatosis. *J Clin Invest* **126**, 137–50 (2016).
42. Albrecht, E., Komolka, K., Kuzinski, J. & Maak, S. Agouti revisited: transcript quantification of the ASIP gene in bovine tissues related to protein expression and localization. *PLoS One* **7**, e35282 (2012).
43. Duchemin, S. I., Visker, M. H., Van Arendonk, J. A. & Bovenhuis, H. 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**, 7276–85 (2014).
44. Bouwman, A. C., Visker, M. H., van Arendonk, J. M. & Bovenhuis, H. Fine mapping of a quantitative trait locus for bovine milk fat composition on Bos taurus autosome 19. *J Dairy Sci* **97**, 1139–49 (2014).

45. Medrano, J., Rincon, G. & Islas-Trejo, A. Comparative analysis of bovine milk and mammary gland transcriptome using RNA-seq. *Proc. 9th World Congr. Genet. Appl. Livest. Prod*. Leipzig, Germany (2010).
46. Roy, R. *et al*. Association of polymorphisms in the bovine FASN gene with milk-fat content. *Anim Genet* **37**, 215–8 (2006).
47. Zhang, S., Knight, T. J., Reecy, J. M. & Beitz, D. C. DNA polymorphisms in bovine fatty acid synthase are associated with beef fatty acid composition. *Anim Genet* **39**, 62–70 (2008).
48. Abe, T. *et al*. Novel mutations of the FASN gene and their effect on fatty acid composition in Japanese Black beef. *Biochem Genet* **47**, 397–411 (2009).
49. Schennink, A., Bovenhuis, H., Leon-Kloosterziel, K. M., van Arendonk, J. A. & Visker, M. H. Effect of polymorphisms in the FASN, OLR1, PPARGC1A, PRL and STAT5A genes on bovine milk-fat composition. *Anim Genet* **40**, 909–16 (2009).
50. Li, C., Aldai, N., Vinsky, M., Dugan, M. E. & McAllister, T. A. Association analyses of single nucleotide polymorphisms in bovine stearoyl-CoA desaturase and fatty acid synthase genes with fatty acid composition in commercial cross-bred beef steers. *Anim Genet* **43**, 93–7 (2012).
51. Oh, D. *et al*. Fatty acid composition of beef is associated with exonic nucleotide variants of the gene encoding FASN. *Mol Biol Rep* **39**, 4083–90 (2012).
52. Cases, S. *et al*. Identification of a gene encoding an acyl CoA:diacylglycerol acyltransferase, a key enzyme in triacylglycerol synthesis. *Proc Natl Acad Sci USA* **95**, 13018–23 (1998).
53. Ntambi, J. M. & Miyazaki, M. Recent insights into stearoyl-CoA desaturase-1. *Curr Opin Lipidol* **14**, 255–61 (2003).
54. Pausch, H. *et al*. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genet Sel Evol* **49**, 24 (2017).
55. Daetwyler, H. D. *et al*. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**, 858–65 (2014).
56. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210–23 (2009).
57. van Binsbergen, R. *et al*. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol* **46**, 41 (2014).
58. Littlejohn, M. D. *et al*. Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. *Sci Rep* **6**, 25376 (2016).
59. Indahl, U. A twist to partial least squares regression. *Journal of Chemometrics* **19**, 32–44 (2005).
60. Madsen, P. J. J. DMU: a user's guide. A package for analysing multivariate mixed models. Version 6, release 5.1 edn (Danish Institute of Agricultural Sciences, Foulum, Denmark, 2012).
61. Zimin, A. V. *et al*. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol* **10**, R42 (2009).
62. Erbe, M. *et al*. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci* **95**, 4114–29 (2012).
63. Olsen, H. G. *et al*. Fine mapping of a QTL on bovine chromosome 6 using imputed full sequence data suggests a key role for the group-specific component (GC) gene in clinical mastitis and milk production. *Genet Sel Evol* **48**, 79 (2016).
64. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv:1303.3997v2 [q-bio.GN] (2013).
65. Garrison, E. M. G. Haplotype-based variant detection from short-read sequencing. Preprint at arXiv:1207.3907v2 [q-bio.GN] (2012).
66. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).
67. Turner, S. D. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. Preprint at *bioRxiv* https://doi.org/10.1101/005165 (2014).
68. Gilmour, A., Gogel, B., Cullis, M. & Thompson, R. ASReml User Guide Release 2.0. VSN International Ltd., Hemel Hempstead, UK (2006).
69. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–5 (2005).
70. Gabriel, S. B. *et al*. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–9 (2002).
71. Wang, N., Akey, J. M., Zhang, K., Chakraborty, R. & Jin, L. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* **71**, 1227–34 (2002).
72. Andersen, F., Osteras, O., Reksen, O. & Grohn, Y. T. Mastitis and the shape of the lactation curve in Norwegian dairy cows. *J Dairy Res* **78**, 23–31 (2011).
73. Boutinaud, M. & Jammes, H. Potential uses of milk epithelial cells: a review. *Reprod Nutr Dev* **42**, 133–47 (2002).
74. Canovas, A. *et al*. Comparison of five different RNA sources to examine the lactating bovine mammary gland transcriptome using RNA-Sequencing. *Sci Rep* **4**, 5297 (2014).
75. Feng, S., Salter, A. M., Parr, T. & Garnsworthy, P. C. Extraction and quantitative analysis of stearoyl-coenzyme A desaturase mRNA from dairy cow milk somatic cells. *J Dairy Sci* **90**, 4128–36 (2007).
76. Boutinaud, M., Rulquin, H., Keisler, D. H., Djiane, J. & Jammes, H. Use of somatic cells from goat milk for dynamic studies of gene expression in the mammary gland. *J Anim Sci* **80**, 1258–69 (2002).
77. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
78. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
79. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
80. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

## Acknowledgements

We would like to acknowledge Tore Bendos and staff at the national establishment for experiments and education in livestock production (Aas, Norway), and Kristil Sundsaasen with staff at the CIGENE genomics lab (NMBU, Norway) for facilitating sampling and isolation of RNA from milk samples of lactating animals. The authors would also like to thank Torfinn Nome for his valuable advice on bioinformatics analyses. The project was funded by the Norwegian Research Council (Project Number 225173) with additional funding from GENO Breeding and AI Organization and Tine Norwegian Dairy Company.

## Author Contributions

T.M.K., H.G.O. and S.L. conceived the experiments. V.T. and A.K. did the calibration of GC/FTIR data. M.S. calculated variance components and heritabilities. M.P.K. planned and facilitated the sequencing of whole genome and RNAseq data. T.M.K. did sequence alignment, variant calling and imputation. T.M.K. and H.G.O. did data analysis. T.M.K. and H.G.O. wrote the main manuscript text with input and critical evaluation from all other authors.

## Additional Information

**Supplementary information** accompanies this paper at https://doi.org/10.1038/s41598-018-20476-0.

**Competing Interests:** Patent application (PCT/EP2017/065798) includes results from the current study and has been jointly submitted by TINE SA, Geno Breeding and AI Organization; S.L., H.G.O., T.M.K. and A.K. are listed as inventors. M.S. is an employee of GENO AS and AI Organization which supplies bovine germplasm. All other authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# Paper III

# Genetic variants associated with fatty acid composition offer new opportunities to breed for healthier milk.

**Short title:** Genetic variants associated with fatty acid composition in bovine milk

Tim Martin Knutsen[1]*, Hanne Gro Olsen[1], Isaya Appelesy Ketto[2], Kristil Kindem Sundsaasen[1], Achim Kohler[3], Valeria Tafintseva[3], Morten Svendsen[4], Matthew Peter Kent[1], Sigbjørn Lien[1].

[1] Centre for Integrative Genetics, Department of Animal and Aquacultural Sciences, Faculty of Life Sciences, Norwegian University of Life Sciences, Ås, Norway

[2] Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences Ås, Norway.

[3] Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

[4] Geno Breeding and AI Association, Ås, Norway

*Corresponding author

E-mail: tim.knutsen@nmbu.no (TMK)

# Abstract

While bovine milk is regarded as healthy and nutritious, its high content of saturated fatty acids (FA) may be harmful to cardiovascular health. Palmitic acid (C16:0) is the predominant saturated FA in milk whose adverse health effects might be countered by substituting it with higher levels of unsaturated FA; such as oleic acid (C18:1*cis*-9). In this work, we performed genome-wide association analyses using high-density SNP-array and whole genome sequence data to detect genetic variants underlying levels of C16:0 and C18:1*cis*-9 and investigate positional candidate genes by transcript profiling and protein level analyses. Genome-wise significant associations were detected for C16:0 on *Bos taurus* autosomes (BTA) 11, 16 and 27, and for C18:1*cis*-9 on BTA 5, 13 and 19. Closer examination of a significant loci on BTA11 identified *PAEP*, which encodes the milk protein *β*-lactoglobulin, as a particularly attractive positional candidate gene. We discovered a tightly linked cluster of genetic variants in coding and regulatory sequences that had opposing effects on levels of C16:0 and C18:1*cis*-9. The favourable haplotype, linked to reduced levels of C16:0 and increased C18:1*cis*-9 was also associated with a marked reduction in *PAEP* expression and *β*-lactoglobulin levels. *β*-lactoglobulin is an abundant milk protein whose levels in milk affect important dairy production parameters such as cheese yield. The genetic variants detected in this study could be used in breeding to promote milk with an improved FA health-profile and enhanced cheese making properties.

# Introduction

Bovine milk is a staple food in billions of people's diet, where it serves as an important source of proteins, fat, minerals and vitamins. Nonetheless, the positive effects of cow milk on human health has been debated, primarily due to its high content of saturated fatty acids (FAs) as compared to the level of unsaturated acids (Mensink et al., 2003; Lindmark Månsson, 2008). Palmitic (C16:0) and oleic (C18:1*cis*-9) acids are the dominant saturated and unsaturated milk FAs respectively, and together they represent 40 - 50% of the total milk fat content (Jensen, 2002). Replacing dietary saturated with unsaturated fat has been shown to reduce the risk of cardiovascular diseases (Mensink et al., 2003; Hooper et al., 2015), and might also reduce the risk of insulin resistance and type-2 diabetes (Kennedy et al., 2008).

Both C16:0 and C18:1*cis*9 show moderate heritability across a range of 0.1 - 0.3 in the extensively studied Holstein-Friesian breed (Stoop et al., 2008; Krag et al., 2013; Lopez-Villalobos et al., 2014). In Norwegian Red cattle, the heritability estimates are 0.13 and 0.14 for C18:1*cis*9 and C16:0 respectively (Olsen et al., 2017), which raises the possibility of using selective breeding to improve the FA profile of cow's milk.

Detection of causal polymorphisms and implementation of genome information in selection typically requires phenotypic data from thousands of individuals. Traditionally, characterisation of milk fat composition has been performed using gas chromatography (GC), but this becomes costly when thousands of samples must be analysed. An alternative is to predict milk fat composition using Fourier transform infrared spectroscopy (FTIR) (Afseth et al., 2010; Soyeurt et al., 2006b; Rutten et al., 2009; Maurice-Van Eijndhoven et al., 2013; Olsen et al., 2017; Knutsen et al., 2018), which produces fast, cheap and detailed phenotypes.

Compared to the widely used single nucleotide polymorphism (SNP) panels, the use of whole genome sequence data has the potential to detect causative variants underlying a given trait, or at least genetic variants in very close linkage disequilibrium (LD) to the causative variants. Once identified, such variants can be used to develop cost-effective genotyping panels for improved quantitative trait loci (QTL) discovery and genomic predictions that persists across diverse genetic backgrounds and multiple generations (Druet et al., 2014; van den Berg et al., 2016). Moreover, coordinated international actions to generate genome-wide maps of functional elements for animal genomes will provide valuable knowledge to understand the context where these variants operate and might eventually pin down the variants and candidate genes underlying the genetic basis of complex traits (Andersson et al., 2015).

The current study seeks to identify and improve our understanding of the genetic variants underlying content of C16:0 and C18:1*cis*-9 using a combination of imputed sequence data and mRNA- and protein-expression profiling. Initially, FTIR-predicted phenotypes were combined with array-based SNP genotypes in a genome-wide association study (GWAS) to identify QTLs with impact on the concentration of the two FAs. Next, a candidate gene region was re-analysed using the imputed sequence variants (SNPs and indels). Finally, gene expression data from mammary epithelial cells and milk protein measurements were used to validate our analysis.

# Results

## Genome-wide association analyses on a high-density SNP dataset

To identify chromosomal regions with a major impact on C16:0 and C18:1*cis*-9 levels, we performed an initial GWAS using 1811 animals genotyped for 609,391 SNPs. As shown in Fig 1, genome-wise significant associations (p-value < 4.1e-8) were detected for C16:0 on *Bos taurus* autosomes (BTA)11, 16 and 27, and for C18:1*cis*-9 on BTA5, 13 and 19. Suggestive findings (p < 1e-5) were detected on BTA1, 4, 5, 6, 8, 17 and 18 for C16:0 and on BTA2, 7, 11, 14, 16, 22 and 26 for C18:1*cis*-9 (Fig 1). Results for all significant marker and trait combinations are provided in Supplementary Table S1.

The most significant associations were found between C16:0 and five SNPs spanning a 24-kb region located at 103.3 Mb on BTA11. This region included the *progestagen-associated endometrial protein* (*PAEP*) gene encoding *β*-lactoglobulin (*β-LG*) and the *glycosyltransferase 6 domain containing 1* (*GLT6D1)* gene encoding a protein of the same name. The two top SNPs for C16:0 had equal p-values and frequencies (p-value = 3.34e-14, MAF = 0.34). The first (rs110186753; A/G) is situated in *PAEP* intron 1 at 103,302,351 bp, while rs109087963 (G/A) is located 1,940 kb downstream of *PAEP* at 103,308,330 bp. These SNPs also showed an association with C18:1*cis*-9 (p-value 1.91e-6), with alleles having opposing effects. That is, the G and A alleles of rs110186753 and rs109087963, respectively, were associated with elevated levels of C16:0 and reduced levels of C18:1*cis*-9. The proportion of daughter yield deviation (DYD) variance explained by each of these SNPs was 3.4 % for C16:0 (allele substitution effect: 0.18g/100g milk fat) and 1.4 % for C18:1*cis*-9 (allele substitution effect: -0.12g/100g milk fat).

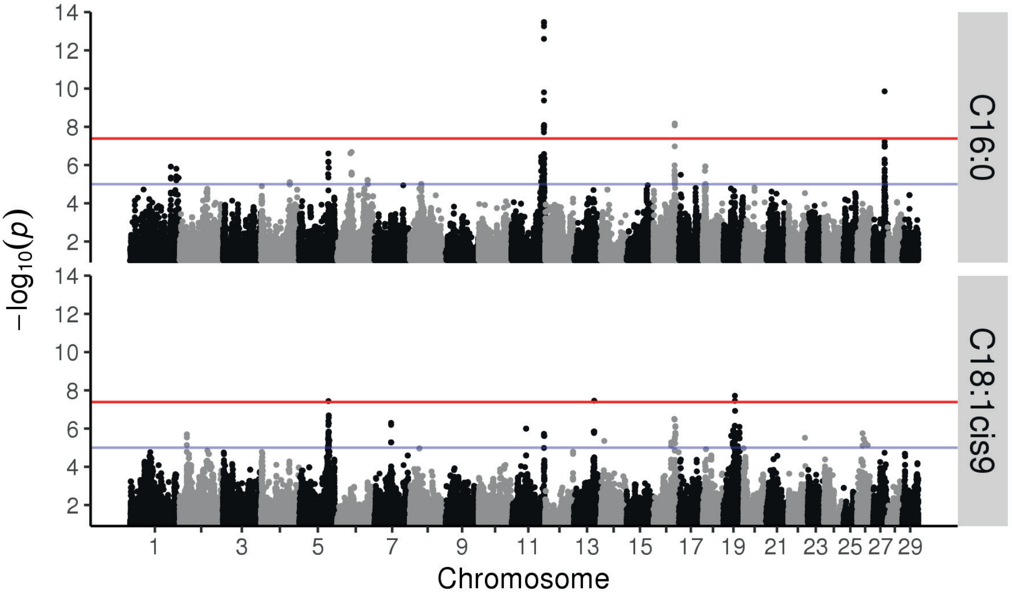


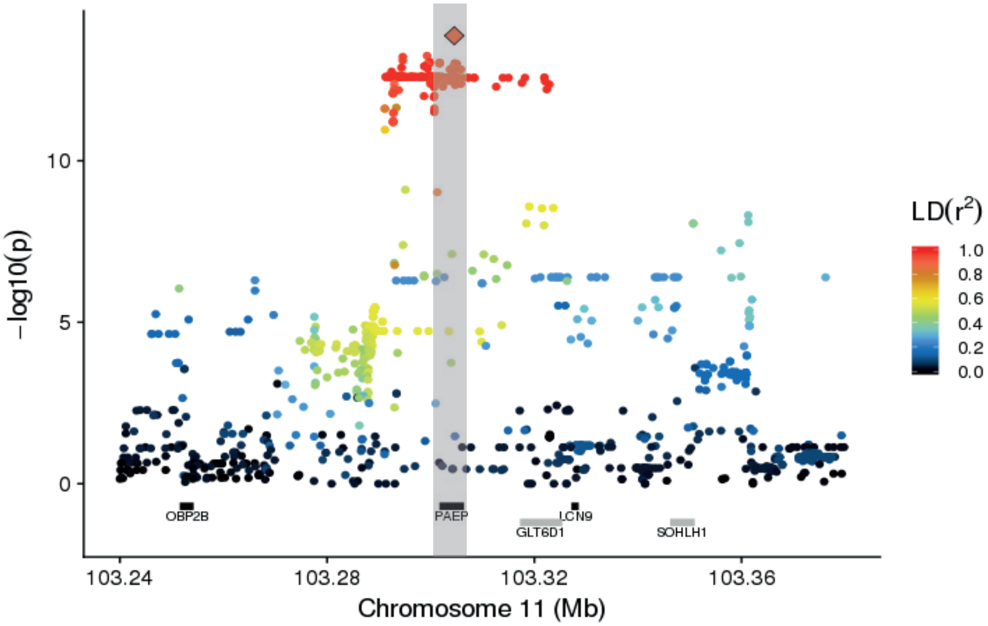
**Fig. 1. Manhattan plots of GWAS results for C16:0 (top) and C18:1*cis*-9 (bottom).** Chromosomes and marker order are represented on the x-axis, with the significance of association (–$\log_{10}$ p-value) between each marker and trait shown on the y-axis. The red line represents the genome-wise significance level (p-value < 4.1e-8), while the blue line represents the suggestive significance level (p-value < 1e-5).

### Fine-mapping of the QTL region on BTA11

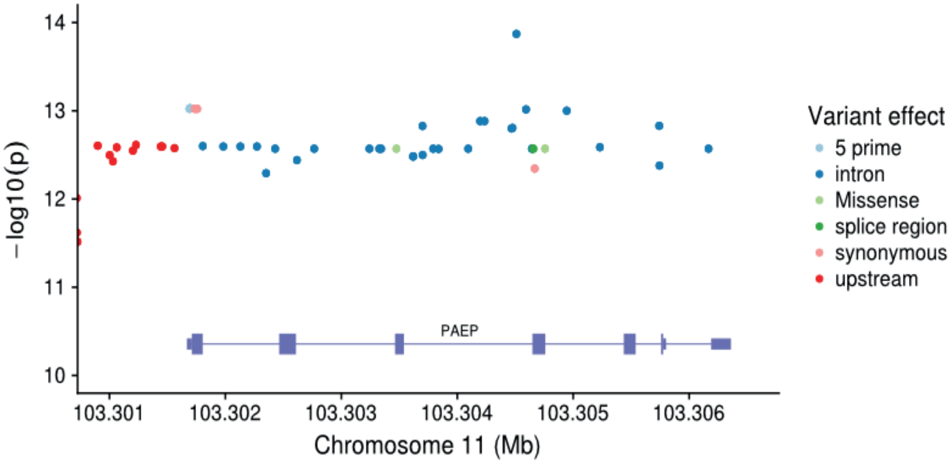
To fine map the QTL on BTA11 and possibly identify underlying causal variants, we re-analysed phenotype data for C16:0 and C18:1*cis*-9 using 109,401 imputed sequence variants spanning a region from 90 to 107 Mb. The results revealed a cluster of 174 variants associated with both C16:0 and C18:1*cis*-9 with largely similar p-values, MAFs and allele substitution effects (Fig. 2). Alleles associated with increased concentration of C18:1*cis*-9 were linked to reduced C16:0 concentration and vice versa. Results for all significant marker and trait combinations are provided in Supplementary Table S2.

Closer examination of pairwise (linkage disequilibrium) LD measurements ($r^2$) between variants in the region, revealed that all 174 variants were in almost perfect LD with each other and could be combined into two major haplotypes extending from ≈10.5 kb upstream of the *PAEP* transcription start site, through *PAEP* and into the neighbouring gene *GLT6D1* (Fig 2). Two predominant haplotypes had frequencies of 0.29 and 0.54, while less frequent haplotypes, differing from the two major haplotypes only by two and three SNPs, were found with frequencies of 0.04 and 0.06. Two missense variants (rs110066229 in exon 3 and rs109625649 in exon 4) code for to the A and B variants of the *PAEP* protein *β-LG* (Caroli et al., 2009), and were present in the identified haplotype block. Accordingly, our two major haplotypes were denoted A and B. The more frequent B haplotype includes alleles associated with reduced levels of C16:0 (allele substitution effect: -0.2g/100g milk fat) and increased levels of C18:1*cis*-9 (allele substitution effect: 0.14g/100g milk fat), i.e. the desirable FA ratio. Supplementary Table S3 provides a more detailed description of the 174 markers assembling the haplotype block, including the haplotype A and B alleles and variant effect predictions.

The haplotype included variants in both the coding and regulatory regions of *PAEP*. After variants annotation, a polymorphism in exon 3 (rs109990218 at 103,304,656 bp) was found to potentially affect alternative splicing of exons into different transcripts (Fig 2), but no transcript splice variants (freq. > 0.05) were found. The most significant SNP for C16:0 was situated in *PAEP* intron 3 (rs110992345; 103,304,509 bp, p = 1.35e-14), while the top-ranking marker for C18:1*cis*-9 was found 2 kb upstream of *PAEP* (rs110920335; 103,300,718 bp, p-value = 1.35e-8), but no obvious causal function could be assigned to either of these SNPs. Tightly linked to these top SNPs, and highly significant, were the two known missense variants determining the *β-LG* A and B variants. Lastly, the haplotype block contained two variants in the 5' untranslated region of *PAEP*, a region that might influence gene expression (rs41255685; position 103,301,690 bp, and rs41255686, position 103,301,694 bp, both with a p-value of 9.5e-14).



−log10(p)
10
5
0
LD(r²)
1.0
0.8
0.6
0.4
0.2
0.0
OBP2B
PAEP
LCN9
GLT6D1
SOHLH1
103.24
103.28
103.32
103.36
Chromosome 11 (Mb)



−log10(p)
14
13
12
11
10
PAEP
103.301
103.302
103.303
103.304
103.305
103.306
Chromosome 11 (Mb)
Variant effect
5 prime
intron
Missense
splice region
synonymous
upstream

**Fig 2. Analysis of C16:0 using sequencing data.** (Top). Association analysis of C16:0 in the region between 103.2 and 103.4Mb on BTA11 using variants imputed from sequence data. The zoomed region showed in the bottom figure, is indicated with a vertical grey bar. The y-axis shows $-\log_{10}$(p-value) for each marker-trait association, while the x-axis denotes marker position. The red diamond indicates the most significant marker for C16:0; rs110992345 at 103,304,509 bp. Colouring indicates the level of LD ($r^2$) between each marker and rs110992345. Gene annotation information according to the Ensembl annotation release 88 is shown with grey and black bars reflecting positive and negative strand orientations respectively. (Bottom). An expanded plot showing variants and their effect relative to the position in the *PAEP* gene structure. The y-axis shows $-\log_{10}$(p-value) for each marker-trait association, while the x-axis denotes marker position. Point colour indicates variant effect class according to the Ensemble annotation release 88.

## Gene expression analyses

To investigate whether any of the significant variants within the two haplotypes were associated with differential gene expression of the two genes spanned by the haplotype block (i.e. generate a cis expression QTL effect; cis eQTL), mRNA was isolated from somatic milk cells and sequenced to quantify expression of the genes. Despite being present in the QTL region, *GLT6D1* was not found to be expressed in any sample. In contrast, *PAEP* was found highly expressed in all samples. Therefore, subsequent analyses were directed towards this gene.

SNPs significant at the genome-wise level, and/or situated within a region extending 5kb up- and downstream from *PAEP*, were tested for their association to the expression level of *PAEP* adjusted by total read count of all measurable milk protein mRNAs (see Methods section). The analysis showed that all 93 tested polymorphisms were significantly (p-value < 0.03) associated with *PAEP* expression (Supplementary Table S4). Their association (p-values) were relatively similar, reflecting the similarity in allele frequency and LD between the tested variants. To illustrate, *PAEP* expression levels relative to genotypes for rs110992345, the marker most significantly associated with C16:0, is shown in Fig 3a. In this Figure, the T allele of rs110992345 which is present in the frequent and favourable B haplotype and hence associated with lower *PAEP* expression is compared to the C allele found in the A haplotype.

To validate the apparent difference in allele-dependent expression levels, we also tested for allele specific expression (ASE) in the 15 animals that were heterozygous for the seven variants located in exons and UTRs of *PAEP*. Concordant with the results of the eQTL analysis, we found that in 98 out of 105 tests for ASE, the alleles present in the B haplotype was expressed at a significantly (adjusted p-value < 0.05) lower level than the alleles present in the A haplotype (Fig 3b). Fifty of the ASE-tests showed extremely low adjusted p-values (< 5.3e-50), with the most significant having 6,598 reads from the A haplotype and 2,635 reads from B haplotype (Supplementary Table S5).

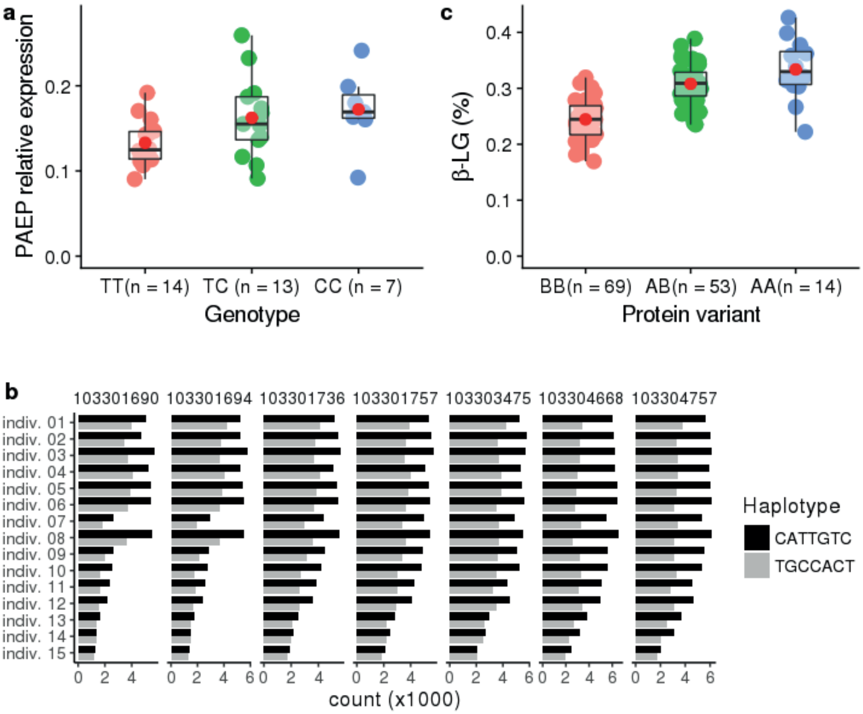


**Fig. 3. Effects of the top associated variants on expression of the *PAEP* locus. a)** The relationship between cow genotypes (n = 34) of the top associated variant (Chr11_103304509_*T_C;* rs110992345) and the expression of *PAEP*. The Y-axis denotes the expression of *PAEP* relative to the sum of expression of the five other milk protein genes. The red dot represents the mean expression value within each group. **b)** ASE for 15 cows heterozygous for seven exonic SNPs (position shown in bp on BTA11) within *PAEP*. The X-axis shows mean normalised counts (x1,000) per haplotype allele. Haplotype A is coloured black, and haplotype B is coloured grey. **c)** The relationship between the two *β-LG* protein variants and the percentage of *β-LG* measured in 136 milk samples. The red dot represents the mean expression value within each group.

**Protein analyses**

Finally, *β-LG* protein levels were quantified to test whether the haplotypes associated with differences in FA and *PAEP* expression levels also reflect differences in protein concentration level. One-hundred and thirty-six cows were genotyped for the two SNPs determining the A and B *β-LG* variants tagging the A and B haplotypes, respectively. The results showed that animals homozygous for the B variant of *β-LG* (i.e. alleles of haplotype B) had on average 35% less *β-LG* than cows homozygous for the haplotype tagged by the A variant (Fig 3c).

# Discussion

C16:0 and C18:1*cis*-9 are the most abundant FAs in bovine milk, but may have opposing effects on human health (Mensink et al., 2003; Kennedy et al., 2008; Hooper et al., 2015), and genome-based selection strategies increasing the ratio of C18:1*cis*-9 to C16:0 in milk may offer ways to improve fat composition. In the current study, we combined milk FA composition phenotypes with high-density SNP information and whole genome sequence data, followed by gene expression and protein level analyses to reveal genetic variants influencing levels of these two acids in milk from Norwegian Red cattle.

The results revealed genome-wise significant QTLs for C16:0 on BTA11, 16 and 27, and for C18:1*cis*-9 on BTA 5, 13 and 19. Subsequent analyses focused on the QTL on BTA11 since it was the most significant and showed antagonistic effects on levels of C16:0 and C18:1*cis*-9. This analysis revealed a haplotype block spanning multiple variants in regulatory and coding regions of *PAEP*, including the two SNPs coding for the A and B variants of the *PAEP* gene product *β-LG*. The most frequent haplotype in the block (haplotype B, encoding the B protein variant) was associated with (i) a more favourable C16:0 to C18:1*cis*-9 ratio, (ii) lower *PAEP* expression and (iii) lower *β-LG* levels as compared to haplotype A.

*β-LG* is one of the most abundant proteins in bovine milk (Ng-Kwai-Hang and Kim, 1996). The two major protein isoforms, variant A and B, differ at mRNA positions 64 and 118 leading to ASP>GLY and VAL>ALA substitutions, respectively (Caroli et al., 2009). The association between *PAEP* allelic variants and milk production traits such as protein percentage, total fat yield and fat percentage in cows has been well documented (Tsiaras et al., 2005; Berry et al., 2010). Previous studies have shown that *β-LG* can bind both saturated and unsaturated FAs, especially C16:0, *in vitro* (Le Maux et al., 2014). In dairy sheep, *β-LG* variants were shown to affect the concentration of C16:0 along with other FAs (Mele et al., 2007). Furthermore, the B protein variant associated with reduced C16:0 levels has been linked to favourable chemical composition and technological parameters such as shorter coagulation time, a lower concentration of whey proteins together with higher casein levels and higher cheese yield (Puppel et al., 2016; Ketto et al., 2017).

Still, the mechanism for how different *β-LG* variants or the *β-LG* protein concentration in milk could influence individual FAs is not well understood. But given the strong C16:0 binding capacity of *β-LG*, we the QTL effect on the C16:0 to C18:1*cis*-9 ratio may well be caused by differences in the affinity for the FAs between the protein variants, a change in the concentration of *β-LG* due to differential expression of *PAEP*, or a combination of these effects.

We found evidence for differential expression of the two protein variants but believe that this is more likely related to linked polymorphisms within regulatory regions rather than the protein variants themselves (Lum et al., 1997; Folch et al., 1999). *PAEP* expression in lactating mammals is reported to be regulated by *signal transducer and activator of transcription 5* (*STAT5*, also known as *milk protein binding factor*) and *activator proteins 1* and *2* (Qian and Zhao, 2014). Several polymorphisms located in putative binding sites for these transcription factors have been identified (e.g. Wagner et al., 1994; Lum et al., 1997, Braunschweig & Leeb, 2006; Ganai et al., 2009), but the extensive levels of LD in the region hamper our ability to pinpoint one specific variant as the underlying causal factor. However, several of our top-ranked variants were situated in these binding sites. We therefore hypothesize that the effect on gene expression can be due to the combined impact of alterations at several regulatory sites within the haplotypes, rather than to one specific SNP.

In addition to *PAEP*, our GWAS highlights several other genes with functions related to milk FA composition. For example, the QTL on BTA5 at 93.9 Mb affected both C16:0 and C18:1*cis*-9 in opposite directions, with the most significant SNP for C18:1*cis*-9 being situated in the first intron of *microsomal glutathione S-transferase 1* (*MGST1*). Although the role of this gene in milk fat synthesis is unclear, it is known to be strongly associated with levels of milk fat, protein, and milk yield (Littlejohn et al., 2016; Raven et al., 2016; Xiang et al., 2017).

BTA13 harbour a QTL for C18:1*cis*-9 in a region that also affects *de novo*-synthesis of short- and medium-chained saturated acids (especially C8:0) in our population (Olsen et al., 2017; Knutsen et al., 2018). This QTL region contains at least two functional candidate genes, *nuclear receptor coactivator 6* (*NCOA6*) at 64.6 Mb and *acyl-CoA synthetase short-chain family member 2* (*ACSS2*) gene at 64.8 Mb. *ACSS2* facilitates the conversion of acetate to acetyl-CoA early in the *de novo* synthesis of FAs (Bionaz and Loor, 2008b), while *NCOA6* is a transcriptional coactivator enhancing, among other things, the activity of the *peroxisome proliferator-activated receptor gamma* (*PPARG),* which is a well-described transcriptional regulator affecting lipid storage (Lemay et al., 2007; Bionaz and Loor, 2008b; Liu et al., 2016).

Two distinct QTLs were found for C18:1*cis*-9 on BTA19, of which the one at 51.38 Mb was located to *fatty acid synthase* (*FASN*), a multifunctional enzyme that catalyses *de novo* synthesis of milk FAs (Bionaz and Loor, 2008b).

We also detected chromosome-wise significant associations between C18:1*cis*-9 and markers situated near the *stearoyl-coenzyme A desaturase 1* (*SCD*) on BTA26. *SCD* is involved in the synthesis of monounsaturated FAs by introducing a double bond in the delta-9 position of C14:0, C16:0 and C18:0, primarily, thus producing the *cis*-9 variant of these acids (Ntambi and Miyazaki, 2003).

The QTL affecting C18:1*cis*-9 at 36.2 Mb on BTA27 spans the gene *glycerol-3-phosphate acyltransferase 4* (*GPAT4*) which encodes the rate-limiting enzyme in the triacylglycerol biosynthesis pathway and plays a crucial role in milk fat biosynthesis (Bionaz and Loor, 2008a).

An essential requirement when using phenotype data (FA composition) from FTIR profiles is that individual acids are predicted with high confidence. The prediction accuracy of mid-infrared spectroscopy has been demonstrated (Soyeurt et al., 2006a; Rutten et al., 2009;

Afseth et al., 2010; De Marchi et al., 2011; Soyeurt et al., 2011; Maurice-Van Eijndhoven et al., 2013; Bonfatti et al., 2016; Olsen et al., 2017). However, since FA are correlated to total fat, a possible concern is that the prediction values reflect total fat rather than individual acids (Eskildsen et al., 2014). To address this, we assess FA concentrations as percentages of total fat instead of gram-acid-per-unit-of-milk (Olsen et al., 2017), which has led to a prediction accuracy (in the form of cross-validated squared Pearson product-moment correlation coefficients) of 0.77 for C16:0 and 0.94 for C18:1*cis*-9. Soyeurt et al. (2006) suggested that the predicted concentrations were due to real absorbance values specific to the FAs if the calibration correlations were higher than the correlations between total fat and FA. As reported in Olsen et al.(2017), the C16:0 and C18:1*cis*-9 squared correlation to total fat was 0.19 and 0.03, respectively, which is markedly lower than the cross-validated squared Pearson product-moment correlation coefficients. A consequence of correcting for total fat is that the prediction accuracies are expected to be lower than when FA concentrations are expressed as a quantity per unit of milk (Soyeurt et al., 2006a; Rutten et al., 2009; De Marchi et al., 2011). This was the case for C16:0, while the prediction accuracy of C18:1*cis*-9 was found to be comparable to those obtained by milk-based models (Rutten et al., 2009; De Marchi et al., 2011; Olsen et al., 2017).

In recent years, methods exploring ways to apply imputed sequence variants in GWAS and genomic predictions in dairy cattle has emerged (van den Berg et al., 2016; Goddard, 2017; VanRaden et al., 2017). The current study utilised sequence imputation to fine map several QTL regions associated with 16:0 and C18:1*cis*9 levels in milk. With sequence density genotypes, we expect to have the causal variants present in the data for the direct estimation of their GWA p-value, and hence also their effect on the trait. While GWAS analysis with imputed sequence data has previously confirmed causal loci in cattle (MacLeod et al., 2016), imperfect imputation, extensive LD and sampling error may result in the causal polymorphism not being identified as the most highly associated variant. However, using non-linear prediction models were most variant effects are set to zero and some to have larger effects seem promising (Erbe et al., 2012; MacLeod et al., 2016). Others have shown improved genomic prediction reliabilities when including selected sequence variants from GWA in the prediction (van den Berg et al., 2016; VanRaden et al., 2017). Both these strategies could be used with the results from the current paper. Nonetheless, further research to discover functional variants in the genome, and improvements to the computational and statistical methodology of GWA and genomic prediction strategies is critical to realising the full potential of the sequence data approach.

# Conclusions

The current study revealed a haplotype block with two major haplotypes spanning both coding and regulatory sequences of *PAEP*, including the polymorphisms underlying the A and B variants of the *β-LG* protein. The most frequent haplotype B was associated with a favourable C16:0 to C18:1*cis*-9 ratio and a marked reduction in *PAEP* expression and *β-LG* levels, which suggests a regulatory role of causal variants underlying the QTL. Furthermore, the B variant is considered beneficial for milk production traits. Our results may, therefore, be applied in breeding to produce milk with healthier FA profile and more favourable cheese-making properties.

# Materials and methods

## Ethics statement

All animals included in the study were Norwegian Red cattle, and experiments were conducted in accordance with the rules and guidelines outlined in the Norwegian Animal Welfare Act 2009, issued by the Norwegian Ministry of Agriculture and Food. Data generated as part of routine commercial activities are considered outside the scope of that requiring formal committee assessment and ethical approval.

## Estimation of bovine milk fat composition from FTIR spectroscopy data

Milk fat composition was estimated from FTIR spectroscopy data as described in Olsen et al. (2017) with some adjustments to the number of spectra and animals used. In brief, 224 milk samples obtained from a previous feeding experiment and 659 samples from field sampling were analysed in parallel by FTIR and GC with flame ionisation detector (GC-FID) reference analysis. FTIR spectra (regressors) were subsequently calibrated against GC-FID reference values (regressands) by using powered partial least squares regression (PPLSR; Indahl, 2005). Regressands were presented as percentages of GC-FID FA values to total fat to reduce to a minimum value the correlation between the FA and total fat in milk samples. The calibration model was applied to a total of 4,619,737 infrared spectra from 640,304 cows sampled in two periods; February to November 2007 and July 2008 to June 2014. The traits that were utilised in this study were C16:0 and C18:1*cis*-9.

A detailed description of the estimation of heritabilities and DYDs is given in in Olsen et al. (2017). In short, the heritability estimates were performed on a reduced dataset of 2,209,486

FA profiles from 426,505 cows with a pedigree of 716,753 animals using the DMU software version 6 release 5.1 (Madsen and Jensen, 2008). The data were analysed with the following mixed linear animal repeatability model:

$$Y = RYM_i + RPL_j + htd_k + pe_l + a_m + e_{ijklm} \qquad (1)$$

where RYM is the fixed effect of region (9 regions) by year and month of the test-day, with i ranging from 1 to 740; RPL is the fixed effect of region by lactation number by 10-day period in lactation of the test-day, with j ranging from1 to 1,116; htd is the random effect of herd by test-day, with k ranging from 1 to 168,483; pe is the random permanent environmental effect of the cow on her repeated records, with l ranging from 1 to 426,505; a is a random additive genetic effect of the animal, with m ranging from 1 to 716,753; and e is a random residual effect.

GWAS DYDs were estimated using the 4,619,737 spectra for the full dataset of 640,304 cows with a pedigree of 999,470 animals as the sire averages of daughters' predicted FA compositions, which were each corrected for her fixed effects, non-genetic random effects and half of her dam's genetic effect (Olsen et al., 2017).

The concentration of the two acids together with the accuracy of prediction (in the form of cross-validated squared Pearson product-moment correlation coefficients; $R^2$CV) and heritabilities were as reported in Olsen et al. (2017). Mean concentrations were 25.25 and 21.4 % of total fat for C16:0 and C18:1*cis*-9, respectively. $R^2$CV was 0.75 and 0.94, and heritabilities 16.06 and 14.34 for C16:0 and C18:1*cis*-9, respectively.

## Construction of a dense SNP dataset

Genotypes for the studied animals were available from other projects and the routine genotyping performed by Geno Breeding and AI Association. DNA was extracted from semen samples of artificial insemination (AI) bulls, and from blood samples of cows using standard phenol-chloroform-based protocols. The bulls were genotyped on at least one of four different platforms in order to make a genome-wide high-density SNP dataset for the association analyses; the Affymetrix 25K SNP array (Affymetrix, Santa Clara), a custom Affymetrix 50K SNP array, the Illumina 54K BovineSNP50 BeadChip (Illumina, San Diego) and the 777K Illumina BovineHD Genotyping BeadChip (Illumina, San Diego).

Imputation was done step-wise, with the 25K Affymetrix dataset first imputed to the custom 50K Affymetrix density, and then the combined Illumina 54K and Affymetrix 50K dataset imputed to 777K. The Affymetrix 50K reference counted 5,009 animals and the Illumina 777K reference consisted of 750 widely used AI bulls. Imputation was done using Beagle version 4.1 (Browning and Browning, 2016), with effective population size (Ne) set to 200 and number of phasing iterations (niterations) set to 20. Remaining parameters were set to default. Map positions were based on the UMD 3.1 reference assembly (Zimin et al., 2009).

For each imputation step, several genotype quality control steps were applied: 1) Variants with MAF less than 0.01 and Hardy-Weinberg Equilibrium p-values less than 1e-7 were filtered. 2) Animals with more than 10 % Mendelian errors were removed from the dataset, and all remaining genotypes with Mendelian errors were set to missing and later imputed. 3) Markers and animals with a call rate below 95% and markers on sex chromosomes were discarded. 4) For each step, the imputation quality was tested using 5-fold cross-validation.

Markers with discordance between true and imputed genotypes above 10% were removed, as these markers are likely to be misplaced in the reference assembly (Erbe et al., 2012). SNPs on unplaced scaffolds and sex chromosomes were also discarded from the dataset.

A total of 2,434 genotyped AI bulls were considered for the initial 777k GWAS analysis. After filtering bulls with less than 20 daughters, the dataset contained 1,811 animals with imputed genotypes for the 777K Illumina BovineHD BeadChip. Of the 1,811 bulls, 57 bulls had genotypes imputed from the Affymetrix 25K array, 237 were imputed from the custom Affymetrix 50K SNP array, 1,113 animals from the Illumina 54K BeadChip and 404 were already genotyped on the 777K Illumina BovineHD BeadChip. The resulting dataset consisted of 1,811 bulls with trait data in the form of DYDs based on 20 or more daughters for the relevant FAs and with genotypes for 609,361 SNPs distributed on all 29 autosomes.

**Whole-genome sequencing, variant calling and sequence imputation.**

Whole-genome sequencing data were obtained from 153 animals (132 AI bulls and 31 cows) as described in Olsen et al. (2016). The AI bulls were selected based their number of daughters in production and by ensuring an even genetic contribution to the population structure of Norwegian Red cattle, by examining the recorded pedigree. Animals were sequenced to an average coverage of 9 x using Illumina sequencing (Illumina, San Diego). All reads were aligned against UMD 3.1 using BWA MEM version 0.7.10. Variant calling was done with FreeBayes version 1.0.2 (Garrison and Marth, 2012). Missing genotypes in the resulting Variant Call Format (VCF)-file were imputed and phased using Beagle version 4.1 (Browning and Browning, 2016). This phased dataset was used as a reference panel for imputing the 1,811 animal high-density panel to full sequence with Beagle using the same imputation parameters as described before, except that allele miscall rate (err) was set to 0.01. In a final filtering step, variants with minor allele frequency above 0.02 were retained. Also, variants with Beagle's reported allelic $R^2$ ($AR^2$) below 0.7 were filtered, as this has been shown to be a robust and reliable threshold for filtering of imputed sequence variants (Littlejohn et al., 2016; Browning and Browning, 2008; van Binsbergen et al., 2014).

### Genotyping of cows

The 36 cows used for the RNA sequencing were also genotyped on the Illumina BovineSNP50 BeadChip (54K, Illumina, San Diego). Blood samples were collected by certified personnel, and DNA extraction and genotyping on the Illumina BovineSNP50 BeadChip (54K, Illumina, San Diego) were performed according to the manufacturer's protocol. Genotypes were quality checked and imputed to sequence density as previously described.

### Genome-wide association study

This study was initiated by conducting a single marker genome-wide association study for C16:0 and C18:1*cis*-9 concentration with genotypes for 609,361 genome-wide distributed SNPs and phenotypes in the form of DYD from 1,811 elite AI bulls, with follow-up analyses of selected regions imputed to sequence level density. The initial GWAS was conducted with the GCTA software (Yang et al., 2014) for computational feasibility, while the follow-up analyses of selected regions were analysed using ASReml package version 3.0 (Gilmour et al., 2009) to be able to weight the analysis by number of daughters for each DYD and to be able to use genotype dosage data in the model.

A mixed linear model single-marker association analysis was performed with the –mlma-loco option of GCTA. The model fitted to the performance information for each trait and each SNP was:

$$\mathrm{DYD} = a + bx + g^{-} + e \qquad (2)$$

were DYD is the performance of the bull, a is the mean term, b is the fixed additive effect of the candidate SNP to be tested for association, x is the SNP genotype indicator variable coded as 0, 1 or 2, $g^{-}$ is the random polygenic effect, i.e. the accumulated effect of all SNPs except those on the chromosome where the candidate SNP is located, and e is the residual. The var($g^{-}$) will be re-estimated each time when a chromosome is excluded from calculating the genomic relationship matrix. The suggestive significance level was set at p = 1e-5, which is a default setting in the R-package qqman used for producing manhattan plots (Turner, 2014). The genome-wise significance level was set at (0.05/609,361*2) = 4.1e-8, corresponding to a nominal type I error rate of 0.05 and Bonferroni correction for 609,361 markers and two traits.

**Re-analyses of the candidate gene region on BTA11 using sequence-level variants**

All sequence-level polymorphisms situated between 90 and 107 Mb on BTA11 that passed quality control were analysed for association to C16:0 and C18:1*cis*-9 using ASReml. The model that was fitted to the information on performance for each trait – marker combination was:

$$\mathbf{DYD} = \mathbf{1}\mu + \mathbf{X}\mathrm{b} + \mathbf{Za} + \mathbf{e}, \qquad (3)$$

where $\mathbf{DYD}$ is the vector of bull performances weighed by the number of daughters, $\mathbf{1}$ is a vector of ones, $\mu$ is the overall mean, $\mathbf{X}$ is a vector of marker genotypes coded as a decimal number between 0 and 2 depending on the estimated dosage of the alternate allele (as reported by Beagle 4.1), b is the fixed effect of the marker, $\mathbf{Z}$ is an incidence matrix relating phenotypes to the corresponding random polygenic effects, $\mathbf{a}$ is a vector of random polygenic effects, and $\mathbf{e}$ is a vector of residual effects. Genetic and residual variances were estimated from the data. $\mathbf{a}$ was assumed to follow a normal distribution $\sim N(\mathbf{0}, \mathbf{A}\sigma_{\mathbf{A}}^{2})$ where $\mathbf{A}$ is the relationship matrix derived from the pedigree, and $\sigma_{\mathbf{A}}^{2}$ is the additive genetic variance. $\mathbf{e}$ was assumed to follow a normal distribution $\sim N(\mathbf{0}, \mathbf{I}\sigma_{\mathbf{e}}^{2})$ where $\sigma_{\mathbf{e}}^{2}$ is the residual variance. Association analysis was performed for each individual marker. Since ASReml does not output p-values for the marker effect automatically, these were calculated from the F statistics for the conditional sum of squares, the numerator degrees of freedom and the denominator degrees of freedom  with the R-function pf() from the stats package version 3.4.0 (R Core Team, 2017).

The fraction of genetic and phenotypic DYD variance explained by each SNP for each phenotype was calculated as 2p(1-p)α2, divided by the additive genetic variance and phenotypic variance, respectively (Falconer and Mackay, 1996). Here p is the frequency of one allele of a biallelic SNP, and α is the allele substitution effect.

**Haplotype analyses**

Pairwise LD measurements ($r^2$) were estimated and haplotypes were identified for the top-ranking markers within the QTL region using the Haploview 4.2 software (Barrett et al.,

2005) on phased genotypes. Haplotypes were defined by Haploview according to the confidence intervals strategy (Gabriel, 2002).

### RNA isolation, sequencing and read mapping

Gene expression levels were obtained using read counts from mRNA isolated from somatic milk cells (SMC) of 36 cows from the research herd at the Norwegian University of Life Sciences, Ås, Norway. The animal pedigree was used to avoid selection of close relatives. All milk samples were collected approximately 50 days (range 47 to 55) after calving. This sampling period was chosen since it roughly coincides with peak expression of several relevant genes involved in bovine milk fat synthesis, including *FASN* (Bionaz and Loor, 2008b) and with the top of the lactation curve of Norwegian Red cows (Andersen et al., 2011)· The cows were in different parities due to the limited size of the research herd. All cows were fed the same diet.

In our study, mRNA was isolated from somatic milk cells. Most common is the use of mammary tissue from biopsies, which is invasive and represent technical challenges and management issues in the recovery of the animals. Contrary to this, milk is excreted by the mammary epithelial cells (MEC) lining the inside of the udder, which is subject to turnover and shed into the milk and therefore represents a proportion of the somatic cells found in milk (Boutinaud and Jammes, 2002). Cánovas et al. (2014) found that compared to other sources (e.g. mammary gland tissue, laser dissected MEC), the quality of the total RNA extracted from the SMC was high. Moreover, the expression profile of genes investigated in SMC derived material was highly correlated with the expression observed in laser-dissected

MEC. Several studies have confirmed the usefulness of this method (Boutinaud and Jammes, 2002; Boutinaud et al., 2002; Feng et al., 2007).

Milk samples were collected manually 2-3 hours after milking to maximise the number of viable cells present in the milk. Teats were cleaned with water followed by 70% ethanol before milking by hand, and 2 x 50 ml milk from each animal was collected in Falcon tubes. Samples were stored on ice immediately after collection and centrifuged at 4°C for 10 min at 2,300g within 1.5 hours to collect cells in the bottom of tubes. After centrifugation, most of the fat layer was removed with a clean pipette tip and supernatant decanted. Each pellet was dissolved in 4 ml 1xPBS by pipetting up and down. The liquid was transferred to a new 50 ml Falcon tube. Samples were centrifuged at 4°C for 10 min at 2,300g and supernatant decanted. Cell pellets were dissolved in 1 ml Trizol (Qiagen), and cells were lysed by pipetting up and down. Samples were stored in -80 °C until RNA extraction with Qiagen RNeasy Plus Universal Tissue Mini Kit (Qiagen) according to the manufacturer's protocol. RNA concentrations and quality were measured with a NanoDrop8000 spectrophotometer (Thermo Fisher Scientific) and Agilent RNA 6000 assay on Agilent BioAnalyzer 2100 (Agilent Technologies), respectively. All samples had an RNA integrity number (RIN) between 6.6 and 9.2.  Samples were prepared for paired-end sequencing (2x150 bp) using the Illumina® TruSeq® stranded mRNA library preparation kits and sequenced by the Norwegian Sequencing Centre (www.sequencing.uio.no) using the Illumina HiSeq 3000 platform.

Before mapping, raw read quality was assessed using fastQC version 0.11.5 https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Illumina adaptors were removed, and the sequences were quality-trimmed using cutadapt (Martin, 2011). Cutadapt was set to cut adaptors with a minimum overlap length of 8 and low-quality 3′ ends were removed by setting a quality threshold of 20 (phred quality + 33). An index of the UMD 3.1 reference genome was built, and reads were aligned to the reference using STAR version 2.3.1 (Dobin et al., 2013). Sorting and indexing of the resulting BAM files were completed using SAMtools version 1.3 (Li and Durbin, 2009). The code for the described RNAseq mapping method is available as part of a bash-script pipeline found at https://gitlab.com/fabian.grammes/RNAseq-analysis/ (version 1.1.0). To look for novel splice variants of candidate genes, the BAM-files were assembled into transcripts using stringtie version 1.3.3 (Pertea et al., 2015). Isoform fraction was set to 5 %. All other settings were set to default.

**Effect of genotype on gene expression.**

Detection of cis-acting eQTLs was performed using the linear eQTL method implemented in the R package Matrix eQTL version 2.1.1 (Shabalin, 2012). Cis distance (CisDist) was set to 5 kb so that all variants within and ±5 kb of the tested gene are included for association with the expression level of that gene. A weakness we identified in using somatic milk cells as the basis for RNAseq analysis was that the expression levels of FA metabolism genes varied remarkably between the sampled cows. Even after accounting for sequence library size, there was an approximately 100-fold difference in the expression level of key FA metabolism genes (such as *FABP3*, *SCD1* and *DGAT1*) between samples with highest and lowest levels of expression. Given that we collected the samples from cows eating the same feed, in the same environment, at the same lactation stage, we believe this is due to variation in the proportion of mammary epithelial cells compared to white blood cells (immune cells) in each sample. To adjust for this effect, we included an effect of the total expression level of the other five major milk protein genes (gene names: CSN1S1, CSN2, CSN1S2, CSN3 and LALBA) as a covariate in the linear model run by Matrix eQTL. Use of this covariate will be an indirect way of adjusting for the sample MEC to white blood cell fraction.

The percentage *PAEP* expression variance explained by the top-SNP genotype was calculated by modelling the expression as a function of the animal genotype using the lm function in R.

**Allele-specific expression**

ASE analysis was accomplished using the tool ASEReadCounter from the Genome Analysis Toolkit (McKenna et al., 2010) with default settings. Before running the tool, duplicated reads were removed using markdup from Sambamba (Tarasov et al., 2015). ASEReadcounter produces a table with separate read counts for every heterozygous bi-allelic variant in the provided BAM files. To test for significant levels of ASE, we used a two-sided Exact Binomial Test using the R-function *binom.test* with the number of trials equal to total read counts at each locus. The test gives a p-value for the hypothesis that the number of reads for each allele at heterozygous loci will be approximately equal when sequenced (Castel et al., 2015). The p-values were adjusted using the p.adjust R-function with method = "bonferroni".

**Protein analysis.**

The relative concentration of *β-LG* was determined by using an Agilent capillary electrophoresis (CE) system (G1600AX), installed with Agilent ChemStation software (Agilent Technologies, Germany) as described in Ketto et al. (2017). Composition of *β-LG* was determined by adjusting the relative concentration of *β-LG* with the total protein content determined by MilkoScan FT1 (Foss Electric A/S, Hillerød, Denmark) as described in Jorgensen et al. (2016). The effects of milk protein genotypes on the *β-LG* concentration of milk were analysed using the lme4 R package (Bates et al., 2014), where the effect of cow was treated as a random effect. Effects of parity, selection line and stage of lactation were not found to be significant and therefore excluded from the statistical analysis.

### Variant annotations

All variants were annotated using the Ensembl Variant Effect Predictor web tool (McLaren et al., 2016), based on the Ensembl *Bos taurus* annotation release 88.

### Availability of data

DNA and RNA sequence data will be submitted to the European Nucleotide Archive, http://www.ebi.ac.uk/ena. Phenotype and genotype data are available only upon agreement with Geno Breeding and AI Organization (http://www.geno.no).

## Acknowledgements

We would like to acknowledge Tore Bendos and staff at the Centre for livestock production (Aas, Norway), and the CIGENE genomics lab staff for sampling and isolation of RNA from milk samples of lactating animals. The authors would also like to thank Fabian Grammes for development of his RNA-seq mapping and SNP-calling pipeline, and Torfinn Nome for his valuable advice on bioinformatics analyses. The project was funded by the Norwegian Research Council (Project Number 225173) with additional funding from GENO Breeding and AI Organization and Tine Norwegian Dairy Company.

## References

Afseth, N.K., H. Martens, Å. Randby, L. Gidskehaug, B. Narum, K. Jørgensen, S. Lien, and A. Kohler. 2010. Predicting the Fatty Acid Composition of Milk: A Comparison of Two Fourier Transform Infrared Sampling Techniques. *Appl. Spectrosc.* 64:700–707. doi:10.1366/000370210791666200.

Andersen, F., O. Østers, O. Reksen, and Y.T. Gröhn. 2011. Mastitis and the shape of the lactation curve in Norwegian dairy cows. *J. Dairy Res.* 78:23–31. doi:10.1017/S0022029910000749.

Andersson, L., A.L. Archibald, C.D. Bottema, R. Brauning, S.C. Burgess, D.W. Burt, E. Casas, H.H. Cheng, L. Clarke, C. Couldrey, B.P. Dalrymple, C.G. Elsik, S. Foissac, E. Giuffra, M.A. Groenen, B.J. Hayes, L.S.S. Huang, H. Khatib, J.W. Kijas, H. Kim, J.K. Lunney, F.M. McCarthy, J.C. McEwan, S. Moore, B. Nanduri, C. Notredame, Y. Palti, G.S. Plastow, J.M. Reecy, G.A. Rohrer, E. Sarropoulou, C.J. Schmidt, J. Silverstein, R.L. Tellam, M. Tixier-Boichard, G. Tosser-Klopp, C.K. Tuggle, J. Vilkki, S.N. White, S. Zhao, and H. Zhou. 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* 16:57. doi:10.1186/s13059-015-0622-4.

Barrett, J.C., B. Fry, J. Maller, and M.J. Daly. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*. 21:263–265. doi:10.1093/bioinformatics/bth457.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2014. Fitting Linear Mixed-Effects Models using lme4. *J. Stat. Software; Vol 1, Issue 1* . doi:10.18637/jss.v067.i01.

van den Berg, I., D. Boichard, and M.S. Lund. 2016. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. *Genet. Sel. Evol.* 48:1–18. doi:10.1186/s12711-016-0259-0.

Berry, S.D., N. Lopez-Villalobos, E.M. Beattie, S.R. Davis, L.F. Adams, N.L. Thomas, A.E. Ankersmit-Udy, A.M. Stanfield, K. Lehnert, H.E. Ward, J.A. Arias, R.J. Spelman, and R.G. Snell. 2010. Mapping a quantitative trait locus for the concentration of beta-lactoglobulin in milk, and the effect of beta-lactoglobulin genetic variants on the composition of milk from Holstein-Friesian x Jersey crossbred cows. *N. Z. Vet. J.* 58:1–5. doi:10.1080/00480169.2010.65053.

van Binsbergen, R., M.C. Bink, M.P. Calus, F.A. van Eeuwijk, B.J. Hayes, I. Hulsegge, and R.F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* 46:41. doi:10.1186/1297-9686-46-41.

Bionaz, M., and J.J. Loor. 2008a. ACSL1, AGPAT6, FABP3, LPIN1, and SLC27A6 are the most abundant isoforms in bovine mammary tissue and their expression is affected by stage of lactation. *J. Nutr.* 138:1019–1024. doi:138/6/1019 [pii].

Bionaz, M., and J.J. Loor. 2008b. Gene networks driving bovine milk fat synthesis during the lactation cycle. *BMC Genomics*. 9:366. doi:10.1186/1471-2164-9-366.

Bonfatti, V., L. Degano, A. Menegoz, and P. Carnier. 2016. Short communication: Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. *J. Dairy Sci.* 99:8216–8221. doi:10.3168/jds.2016-10953.

Boutinaud, M., and H. Jammes. 2002. Potential uses of milk epithelial cells: a review. *Reprod. Nutr. Dev.* 42:133–47. doi:10.1051/rnd.

Boutinaud, M., H. Rulquin, D.H. Keisler, J. Djiane, and H. Jammes. 2002. Use of somatic cells from goat milk for dynamic studies of gene expression in the mammary gland. *J. Anim. Sci.* 80:1258–1269. doi:10.2527/2002.8051258x.

Browning, B.L., and S.R. Browning. 2008. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84:210–223. doi:10.1016/j.ajhg.2009.01.005.

Browning, B.L., and S.R. Browning. 2016. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* 98:116–126. doi:10.1016/j.ajhg.2015.11.020.

Cánovas, A., G. Rincón, C. Bevilacqua, A. Islas-Trejo, P. Brenaut, R.C. Hovey, M. Boutinaud, C. Morgenthaler, M.K. VanKlompenberg, P. Martin, J.F. Medrano, A. Canovas, G. Rincon, C. Bevilacqua, A. Islas-Trejo, P. Brenaut, R.C. Hovey, M. Boutinaud, C. Morgenthaler, M.K. VanKlompenberg, P. Martin, and J.F. Medrano. 2014. Comparison of five different RNA sources to examine the lactating bovine mammary gland transcriptome using RNA-Sequencing. *Sci. Rep.* 4:5297. doi:10.1038/srep05297.

Caroli, A.M., S. Chessa, and G.J. Erhardt. 2009. Invited review: milk protein polymorphisms in cattle: effect on animal breeding and human nutrition. *J. Dairy Sci.* 92:5335–5352. doi:10.3168/jds.2009-2461.

Castel, S.E., A. Levy-Moonshine, P. Mohammadi, E. Banks, and T. Lappalainen. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* 16:195. doi:10.1186/s13059-015-0762-6.

Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T.R. Gingeras. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 29:15–21. doi:10.1093/bioinformatics/bts635.

Druet, T., I.M. Macleod, and B.J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity (Edinb).* 112:39–47. doi:10.1038/hdy.2013.13.

Erbe, M., B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, B.A. Mason, and M.E. Goddard. 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.

Eskildsen, C.E., M.A. Rasmussen, S.B. Engelsen, L.B. Larsen, N.A. Poulsen, and T. Skov. 2014. Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: understanding predictions of highly collinear reference variables. *J. Dairy Sci.* 97:7940–7951. doi:10.3168/jds.2014-8337.

FALCONER, D.S., and T.F.. MACKAY. 1996. Introduction to quantitative genetics. Introduction to quantitative genetics. Pearson Education India. 463 pp.

Feng, S., A.M. Salter, T. Parr, and P.C. Garnsworthy. 2007. Extraction and Quantitative Analysis of Stearoyl-Coenzyme A Desaturase mRNA from Dairy Cow Milk Somatic Cells. *J. Dairy Sci.* 90:4128–4136. doi:10.3168/jds.2006-830.

Folch, J.M., P. Dovc, and J.F. Medrano. 1999. Differential expression of bovine beta-lactoglobulin A and B promoter variants in transiently transfected HC11 cells. *J. Dairy Res.* 66:537–544.

Gabriel, S.B. 2002. The Structure of Haplotype Blocks in the Human Genome. *Science (80-. ).* 296:2225–2229. doi:10.1126/science.1069424.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. doi:arXiv:1207.3907 [q-bio.GN].

Gilmour, a R., B.J. Gogel, B.R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0. *VSN Int. Ltd.* 275. doi:10.1017/CBO9781107415324.004.

Goddard, M.E. 2017. Can we make genomic selection 100% accurate? *J. Anim. Breed. Genet.* 134:287–288. doi:10.1111/jbg.12281.

Hooper, L., N. Martin, A. Abdelhamid, and G. Davey Smith. 2015. Reduction in saturated fat intake for cardiovascular disease. *Cochrane database Syst. Rev.* CD011737. doi:10.1002/14651858.CD011737.

Jensen, R.G. 2002. The composition of bovine milk lipids: January 1995 to December 2000. *J. Dairy Sci.* 85:295–350. doi:10.3168/jds.S0022-0302(02)74079-4.

Kennedy, A., K. Martinez, C.-C. Chuang, K. LaPoint, and M. McIntosh. 2008. Saturated Fatty Acid-Mediated Inflammation and Insulin Resistance in Adipose Tissue: Mechanisms of Action and Implications. *J. Nutr.* 139:1–4. doi:10.3945/jn.108.098269.

Ketto, I.A., T.M. Knutsen, J. Øyaas, B. Heringstad, T. Ådnøy, T.G. Devold, and S.B. Skeie.

2017. Effects of milk protein polymorphism and composition, casein micelle size and salt distribution on the milk coagulation properties in Norwegian Red cattle. *Int. Dairy J.* 70. doi:10.1016/j.idairyj.2016.10.010.

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

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

Lemay, D.G., M.C. Neville, M.C. Rudolph, K.S. Pollard, and J. German. 2007. Gene regulatory networks in lactation: identification of global principles using bioinformatics. *BMC Syst. Biol.* 1:56. doi:10.1186/1752-0509-1-56.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754–1760. doi:10.1093/bioinformatics/btp324.

Lindmark Månsson, H. 2008. Fatty acids in bovine milk fat. *Food Nutr. Res.* 52:1821. doi:10.3402/fnr.v52i0.1821.

Littlejohn, M.D., K. Tiplady, T.A. Fink, K. Lehnert, T. Lopdell, T. Johnson, C. Couldrey, M. Keehan, R.G. Sherlock, C. Harland, A. Scott, R.G. Snell, S.R. Davis, and R.J. Spelman. 2016. Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. *Sci. Rep.* 6:25376. doi:10.1038/srep25376.

Liu, L., Y. Lin, L. Liu, L. Wang, Y. Bian, X. Gao, and Q. Li. 2016. Regulation of peroxisome proliferator-activated receptor gamma on milk fat synthesis in dairy cow mammary epithelial cells. *Vitr. Cell. Dev. Biol. - Anim.* 52:1044–1059. doi:10.1007/s11626-016-0059-4.

Lopez-Villalobos, N., R.J. Spelman, J. Melis, S.R. Davis, S.D. Berry, K. Lehnert, S.E. Holroyd, A.K.H. MacGibbon, and R.G. Snell. 2014. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in New Zealand dairy cattle. *J. Dairy Res.* 81:340–9. doi:10.1017/S0022029914000272.

Lum, L.S., P. Dovč, and J.F. Medrano. 1997. Polymorphisms of Bovine β-Lactoglobulin Promoter and Differences in the Binding Affinity of Activator Protein-2 Transcription Factor. *J. Dairy Sci.* 80:1389–1397. doi:10.3168/jds.S0022-0302(97)76068-5.

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

Madsen, P., and J. Jensen. 2008. An user's guide to DMU. A package for analysing multivariate mixed models. Version 6, release 4.7. 1–33.

De Marchi, M., M. Penasa, A. Cecchinato, M. Mele, P. Secchiari, and G. Bittante. 2011. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. *Animal*. 5:1653–1658. doi:10.1017/S1751731111000747.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 17:10. doi:10.14806/ej.17.1.200.

Maurice-Van Eijndhoven, M.H.T., H. Soyeurt, F. Dehareng, and M.P.L. Calus. 2013. Validation of fatty acid predictions in milk using mid-infrared spectrometry across cattle breeds. *Animal*. 7:348–354. doi:10.1017/S1751731112001218.

Le Maux, S., S. Bouhallab, L. Giblin, A. Brodkorb, and T. Croguennec. 2014. Bovine β-lactoglobulin/fatty acid complexes: binding, structural, and biological properties. *Dairy Sci. Technol.* 94:409–426. doi:10.1007/s13594-014-0160-y.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303. doi:10.1101/gr.107524.110.

McLaren, W., L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. The Ensembl Variant Effect Predictor. *Genome Biol.* 17:122. doi:10.1186/s13059-016-0974-4.

Mele, M., G. Conte, A. Serra, A. Buccioni, and P. Secchiari. 2007. Relationship between beta-lactoglobulin polymorphism and milk fatty acid composition in milk of Massese dairy ewes. *Small Rumin. Res.* 73:37–44. doi:10.1016/j.smallrumres.2006.10.021.

Mensink, R.P., P.L. Zock, A.D.M. Kester, and M.B. Katan. 2003. Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: A meta-analysis of 60 controlled trials. *Am. J. Clin. Nutr.*

77:1146–1155. doi:10.1161/CIRCULATIONAHA.114.010236.

Ng-Kwai-Hang, K.F., and S. Kim. 1996. Different amounts of β-lactoglobulin A and B in milk from heterozygous AB cows. *Int. Dairy J.* 6:689–695. doi:10.1016/0958-6946(95)00069-0.

Ntambi, J.M., and M. Miyazaki. 2003. Recent insights into stearoyl-CoA desaturase-1. *Curr. Opin. Lipidol.* 14:255–61. doi:10.1097/01.mol.0000073502.41685.c7.

Olsen, H.G., T.M. Knutsen, A. Kohler, M. Svendsen, L. Gidskehaug, H. Grove, T. Nome, M. Sodeland, K.K. Sundsaasen, M.P. Kent, H. Martens, and S. Lien. 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.* 49:20. doi:10.1186/s12711-017-0294-5.

Olsen, H.G., T.M. Knutsen, A.M. Lewandowska-Sabat, H. Grove, T. Nome, M. Svendsen, M. Arnyasi, M. Sodeland, K.K. Sundsaasen, S.R. Dahl, B. Heringstad, H.H. Hansen, I. Olsaker, M.P. Kent, and S. Lien. 2016. Fine mapping of a QTL on bovine chromosome 6 using imputed full sequence data suggests a key role for the group-specific component (GC) gene in clinical mastitis and milk production. *Genet. Sel. Evol.* 48:79. doi:10.1186/s12711-016-0257-2.

Pertea, M., G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell, and S.L. Salzberg. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33:290–295. doi:10.1038/nbt.3122.

Puppel, K., B. Kuczy??ska, T. Na??ecz-Tarwacka, M. Go??ebiewski, T. Sakowski, A.

Kapusta, A. Budzi??ski, and M. Balcerak. 2016. Effect of supplementation of cows diet with linseed and fish oil and different variants of ??-lactoglobulin on fatty acid composition and antioxidant capacity of milk. *J. Sci. Food Agric.* 96:2240–2248. doi:10.1002/jsfa.7341.

Qian, X., and F.-Q. Zhao. 2014. Current major advances in the regulation of milk protein gene expression. *Crit. Rev. Eukaryot. Gene Expr.* 24:357–378.

R Core Team. 2017. R: A language and environment for statistical computing.

Raven, L.A., B.G. Cocks, K.E. Kemper, A.J. Chamberlain, C.J. Vander Jagt, M.E. Goddard, and B.J. Hayes. 2016. Targeted imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. *Mamm. Genome*. 27:81–97. doi:10.1007/s00335-015-9613-8.

Rutten, M.J.M., H. Bovenhuis, K.A. Hettinga, H.J.F. van Valenberg, and J.A.M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. *J. Dairy Sci.* 92:6202–6209. doi:10.3168/jds.2009-2456.

Shabalin, A.A. 2012. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 28:1353–1358. doi:10.1093/bioinformatics/bts163.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006a. Estimating Fatty Acid Content in Cow Milk Using Mid-Infrared Spectrometry. *J. Dairy Sci.* 89:3690–3695. doi:10.3168/jds.S0022-

0302(06)72409-2.

Soyeurt, H., P. Dardenne, A. Gillon, C. Croquet, S. Vanderick, P. Mayeres, C. Bertozzi, and N. Gengler. 2006b. Variation in fatty acid contents of milk and milk fat within and across breeds. *J Dairy Sci*. 89. doi:10.3168/jds.S0022-0302(06)72534-6.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D.P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *J. Dairy Sci.* 94:1657–1667. doi:10.3168/jds.2010-3408.

Stoop, W.M., J.A.M. van Arendonk, J.M.L. Heck, H.J.F. van Valenberg, and H. Bovenhuis. 2008. Genetic parameters for major milk fatty acids and milk production traits of Dutch Holstein-Friesians. *J. Dairy Sci.* 91:385–394. doi:10.3168/jds.2007-0181.

Tarasov, A., A.J. Vilella, E. Cuppen, I.J. Nijman, and P. Prins. 2015. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*. 31:2032–2034. doi:10.1093/bioinformatics/btv098.

Tsiaras, A.M., G.G. Bargouli, G. Banos, and C.M. Boscos. 2005. Effect of kappa-casein and beta-lactoglobulin loci on milk production traits and reproductive performance of Holstein cows. *J. Dairy Sci.* 88:327–334. doi:10.3168/jds.S0022-0302(05)72692-8.

Turner, S.D. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*.

VanRaden, P.M., M.E. Tooker, J.R. O’Connell, J.B. Cole, and D.M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel.*

*Evol.* 49:32. doi:10.1186/s12711-017-0307-4.

Xiang, R., I.M. MacLeod, S. Bolormaa, and M.E. Goddard. 2017. Genome-wide comparative analyses of correlated and uncorrelated phenotypes identify major pleiotropic variants in dairy cattle. *Sci. Rep.* 7:9248. doi:10.1038/s41598-017-09788-9.

Yang, J., N.A. Zaitlen, M.E. Goddard, P.M. Visscher, and A.L. Price. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* 46:100–106. doi:10.1038/ng.2876.

Zimin, A. V, A.L. Delcher, L. Florea, D.R. Kelley, M.C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C.P. Van Tassell, T.S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J.A. Yorke, and S.L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol.* 10:R42. doi:10.1186/gb-2009-10-4-r42.

# Supporting information

Supporting information is available online from Figshare via the following link: https://figshare.com/s/8e244ac5a487997a3b38. DOI: 10.6084/m9.figshare.5981899

**Supplementary Table S1.** Table showing GWAS results for C16:0 and C18:1*cis*-9. All significant (p<1e-5) marker – trait combinations from the GWA analysis.

**Supplementary Table S2.** Results for single-marker association analyses (p<1e-5) of C16:0 and C18:1*cis*-9 on imputed sequence data in the region between 100 and 107 Mb on BTA11.

**Supplementary Table S3**. Detailed information of the 174 markers included in the haplotype block with antagonistic effects on C16:0 and C18:1*cis*-9, with haplotype alleles and variant effect predictions from Ensembl.

**Supplementary Table S4**. Results from the eQTL analyses showing 93 significant variants with p-values for the GWAS and eQTL linear model shown.

**Supplementary Table S5**. Results from the 105 binomial tests for ASE (statistical significance of deviations from the theoretically expected distribution of reads originating from the two alleles of a heterozygous SNP) conducted using 15 animals were heterozygous for the seven variants located in exons and UTRs of *PAEP*

ISBN: 978-82-575-1499-0
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no