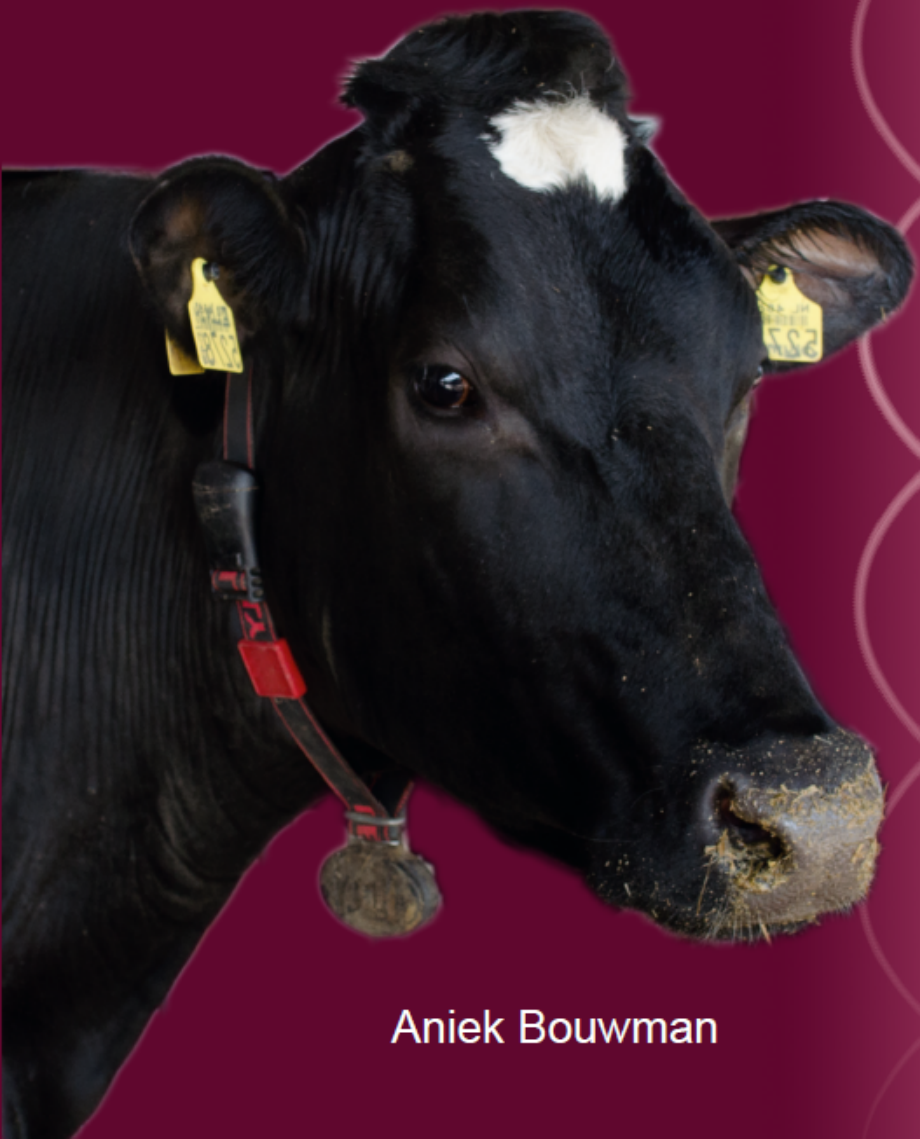
Unraveling the genetic background of bovine milk fat composition



Aniek Bouwman

Unraveling the genetic background of bovine milk fat composition

Aniek Bouwman

Thesis committee

Promotor
Prof. Dr J.A.M. van Arendonk
Professor of Animal Breeding and Genetics
Wageningen University

Co-promotors
Dr H. Bovenhuis
Associate professor, Animal Breeding and Genomics Centre
Wageningen University

Dr M.H.P.W. Visker
Researcher, Animal Breeding and Genomics Centre
Wageningen University

Other members
Prof. Dr F. A. van Eeuwijk, Wageningen University
Prof. Dr E.J.M. Feskens, Wageningen University
Prof. Dr J.B. German, University of California, Davis, USA
Prof. Dr D-J. de Koning, Swedish University of Agricultural Sciences, Uppsala, Sweden

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

Unraveling the genetic background of bovine milk fat composition

Aniek Bouwman

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 16 May 2014
at 4 p.m. in the Aula.

Bouwman, A.C.
Unraveling the genetic background of bovine milk fat composition,
176 pages.

PhD thesis, Wageningen University, Wageningen, NL (2014)
With references, with summaries in English and Dutch

ISBN 978-90-6173-906-3

Abstract

Bouwman, A.C. (2014). Unraveling the genetic background of bovine milk fat composition. PhD thesis, Wageningen University, the Netherlands

Identification of genomic regions, and preferably individual genes, responsible for genetic variation in bovine milk fat composition enhances the understanding of biological pathways involved in fatty acid synthesis and is expected to increase opportunities for changing bovine milk fat composition by means of selective breeding. This thesis aimed to unravel the genetic background of bovine milk fat composition by detection, confirmation and fine-mapping of quantitative trait loci (QTL) for milk fatty acids in Dutch Holstein Friesian cattle. In addition, causal relations between fatty acids were explored. For this study roughly 2,000 dairy cows were genotyped with 50,000 DNA markers and phenotyped for individual fatty acids in both winter and summer milk samples using gas chromatography. Genome-wide association studies (GWAS) showed that milk fat composition has a complex genetic background with three major QTL that explain a relatively large fraction of the genetic variation of several milk fatty acids, and many QTL that explain a relatively small fraction of the genetic variation. Results from the GWAS for summer milk fatty acids confirmed most associations that were detected in the winter milk samples. Moving from linkage analysis toward GWAS confirmed and refined the size of previously detected QTL regions and resulted in new QTL regions. Performing GWAS based on individual fatty acids resulted in additional QTL as compared to GWAS based on fat percentage or yield. This shows that refinement of complex phenotypes into underlying components results in better links between genes and phenotypes. By increasing the marker density, the QTL on BTA19 was refined to a linkage disequilibrium block that contained 2 genes: coiled-coil domain containing 57 and fatty acid synthase. A search for causal relations between fatty acids resulted in a pathway from C4:0 to C12:0, which resembled the de novo synthesis pathway. Causal relation between the QTL on BTA19 and de novo fatty acids showed that the QTL affects C4:0, C6:0, C8:0, C10:0 and C14:0 directly, while C12:0 was indirectly affected by the QTL through its effect on C10:0. The potential of GWAS based on MIR predicted fatty acids was explored but failed to detect some QTL and resulted in additional QTL that were not detected based on GC measurements. Therefore, MIR predicted phenotypes add complexity to the genotype-phenotype relationship, and renders MIR predicted phenotypes less appropriate to identify candidate genes and to infer the biological background of traits.

Contents

1

General Introduction

1.1 Bovine milk fat composition

Milk is produced by mammals in their mammary glands as the primary source of nutrition for their neonates. Milk contains many components such as fat, protein, lactose, minerals and vitamins. This thesis focuses on the fat portion of bovine milk. Milk fat is a nutritious source of energy, fat-soluble vitamins, and bioactive lipids. The fatty acid composition of milk fat determines the flavor and texture of dairy products, such as spreadability of butter. Also, certain fatty acids are claimed to have unfavorable effects on human health, while others are claimed to be beneficial for human health (German and Dillard, 2006). Studying fatty acids in bovine milk can provide valuable information about the synthesis of fatty acids, which could be useful for approaches aimed at changing the fatty acid composition of dairy products and ultimately at improving human health.

Bovine milk fat consists mainly (~98%) of triglycerides, and each triglyceride is composed of three fatty acids attached to a glycerol backbone (Figure 1.1). Generally, a fatty acid consists of a straight chain of a number of carbon atoms that is surrounded with hydrogen atoms and a carboxyl group (−COOH) at one end. If the carbon-to-carbon bonds are all single, the fatty acid is saturated; if any of the bonds is double, the fatty acid is unsaturated. The numerous fatty acids present in bovine milk differ from each other by carbon length, level (e.g. mono) and position (e.g. *cis*9) of unsaturation, and branching. Differences in these characteristics result in different physiological properties related to for instance the texture and flavor of dairy products. (MacGibbon and Taylor, 2006)

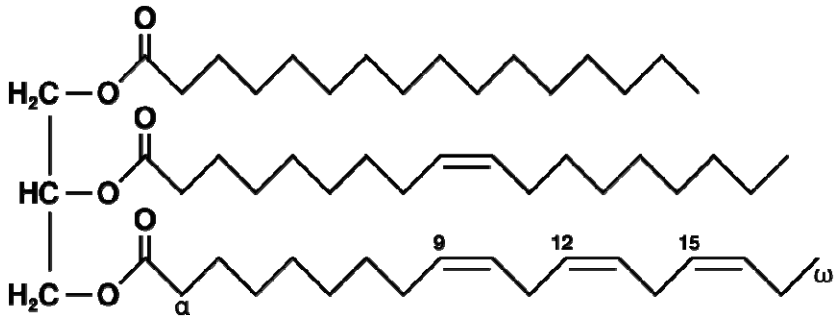


Figure 1.1 Triglyceride with on the left the glycerol backbone and at each of the three positions a fatty acid. At the first position a saturated fatty acid (i.e. palmitic acid (C16:0)) is depicted, at the second position a mono unsaturated fatty acid (i.e. oleic acid (C18:1*cis9*) with one double bond) and at the third position a poly unsaturated fatty acid (i.e. alpha-linolenic acid (C18:3n-3; ALA) with three double bonds).

In total 416 different fatty acids have been discovered in bovine milk (Jensen, 2002), of which a few are abundant and many are minor milk fatty acids. This thesis focuses on the saturated fatty acids C4:0, C6:0, C8:0, C10:0, C12:0, C14:0, C16:0, C18:0, the mono-unsaturated fatty acids C10:1*cis*9, C12:1*cis9*, C14:1*cis9*, C16:1*cis9*, C18:1*cis9*, and poly-unsaturated C18:2*cis9,trans11* (Conjugated linoleic acid; CLA).
Short- and medium-chain saturated fatty acids, i.e. C4:0, C6:0, C8:0, C10:0, C12:0, C14:0 and about half the C16:0 present in bovine milk, are produced de novo in the mammary gland by the multi-enzyme complex fatty acid synthase (FASN) (e.g., Neville and Picciano, 1997; Palmquist, 2006). In this metabolic pathway, the carbon chain is elongated in a sequential cyclic reaction from acetate and β-hydroxybutyrate until a C16:0 fatty acid is formed (e.g., Neville and Picciano, 1997; Palmquist, 2006). All the intermediate fatty acids can leave this elongation cycle by a chain termination mechanism (Smith, 1994) and, thus, end up in bovine milk. This termination mechanism differs between mammals resulting in different milk fat composition; for instance, human milk does not contain C4:0 (e.g., Zou et al., 2013).
A small fraction of the saturated fatty acids C10:0, C12:0, C14:0 and C16:0 is desaturated in the mammary gland into their *cis*9 mono unsaturated equivalent (C10:1-C16:1) by the enzyme stearoyl-CoA desaturase 1 (SCD1) (Palmquist, 2006).
Long-chain fatty acids, i.e. about half the C16:0 present in milk and all fatty acids with 18 or more carbons, are taken up by cows through the diet, but can also be released from adipose tissues. The dietary long-chain fatty acids are biohydrogenated into C18:0 and all kinds of intermediate products by micro-flora in the rumen before they are absorbed from the rumen into the bloodstream, and from the bloodstream into the mammary gland (Jenkins, 1993; Palmquist, 2006). From the diet, large quantities of C16:0, C18:0, and C18:1*cis*9, small quantities of C18:2, and limited amounts of other monoenoic and dienoic fatty acids end up in milk (MacGibbon and Taylor, 2006).
In the mammary gland fatty acids are attached to the glycerol backbone and excreted in milk. The glycerol backbone has three positions for the fatty acids and three enzymes are responsible for the attachment of the fatty acids onto these positions: glycerol-3-phosphate acyltransferase (GPAT), 1-acylglycerol-3-phosphate acyltransferase (AGPAT) and diacylglycerol acyltransferase (DGAT1) (Agarwal and Garg, 2003; Takeuchi and Reue, 2009). In theory numerous different triglycerides can be formed with all the different fatty acids present in milk (64×10^6); however, these enzymes appear to be somewhat specific for the type of fatty acid they attach onto the glycerol backbone, resulting in a limited number of specific triglycerides (Jensen, 2002).

1.2 Factors influencing bovine milk fat composition

Milk fat composition of cows differs from milk fat composition of other mammals; in fact, there is a lot of variation in milk fat composition between mammals. Ruminant milk contains for instance short-chain saturated fatty acids and trans fatty acids, which are almost absent in human milk (Zou et al., 2013). There is not only variation between mammals but also between cattle breeds. For example, Jersey, Brown Swiss and Guernsey have higher levels of saturated milk fatty acids compared to Holstein (Beaulieu and Palmquist, 1995; DePeters et al., 1995; Kelsey et al., 2003; Stull and Brown, 1964), and Jersey and Guernsey have lower levels of unsaturated milk fatty acids compared to Holstein (Beaulieu and Palmquist, 1995; DePeters et al., 1995; Stull and Brown, 1964). There is also variability in milk fat composition among local Dutch dairy breeds (Maurice-Van Eijndhoven et al., 2011) and among French dairy breeds (Lawless et al., 1999).

Variation in milk fat composition can even be found between individuals within a population or herd. The main sources of variation are feeding (e.g., Jensen, 2002; Palmquist, 2006; Palmquist et al., 1993) and genetics (Soyeurt et al., 2007; Stoop et al., 2008), but also factors like health status of the cow and lactation stage introduce variation (e.g., Bastin et al., 2011; German and Dillard, 2006; Karijord et al., 1982; Palmquist, 2006; Stoop et al., 2009a). Feeding directly influences the milk fat composition. For instance, pasture based cows have lower proportions of short- and medium-chain fatty acids, and higher proportions of long-chain fatty acids compared to indoor housed cows fed silage (reviewed in Jensen, 2002), and also fat supplements modify the milk fat composition (reviewed in Ashes et al., 1997). Genetics can be applied to change milk fat composition through breeding. Genetic variation in milk fat composition has been demonstrated, with heritabilities that range between 0.42 and 0.71 for de novo synthesized milk fatty acids and between 0.22 and 0.42 for long-chain milk fatty acids (Stoop et al., 2008).

1.3 Regions on the bovine genome associated with milk fatty acids

Given the substantial genetic variation in milk fat composition in dairy cattle, it is of interest to find the regions on the bovine genome responsible for this variation. This will increase our understanding of the biological processes involved in milk fat synthesis and provide possibilities for animal selection in breeding. To find such regions, also known as quantitative trait loci (QTL), the genome is screened using DNA markers, single nucleotide polymorphisms (SNP), with known location. The

aim of genome-wide association studies (GWAS) is to detect association between phenotypes and SNP.

Previously, 1,500 SNP have been used in family-based linkage studies resulting in large QTL regions for milk fatty acids (Schennink et al., 2009b; Stoop et al., 2009b). Linkage studies look within a family for significant differences in phenotypes between offspring-groups that inherited the opposite allele from the parent at a specific genomic region. Large chunks of DNA are transmitted from the parent to the offspring due to limited recombination within a family. This results in large regions that co-segregate within families and, therefore, linkage studies generally detect large QTL regions. Genome-wide association studies look for significant associations between phenotypes and genotypes of individuals within a population. The advantage of GWAS is that co-segregation within populations is smaller than co-segregation within families and, therefore, GWAS generally detect smaller QTL regions. However, more DNA markers are required. The development of SNP-chips with 50,000 SNP markers provided opportunities to perform GWAS for milk fatty acids.

A screen along the genome with 50,000 SNP markers provides insight into the genetic architecture of a trait: how many and which regions explain variation in the trait, and the distribution of effect sizes of these regions on the trait. It shows also whether there are QTL with major influence on the trait, like for instance diacylglycerol O-acyltransferase 1 (*DGAT1*) *K232A* has on milk production (Grisart et al., 2002), or whether the trait is mainly defined by many QTL with small effects or a combination of both. The power of GWAS to detect QTL with major effects on the phenotype is large; therefore, major QTL can be pursued for fine-mapping. The power to detect QTL with relatively small effects on the phenotype is lower and false positives may reside among the QTL with small effects, but at the same time many QTL with small effects may remain undetected due to false negatives. It is, thus, wise to confirm such QTL with relatively small effects in a second study before fine-mapping.

The ultimate goal is to find the actual causative DNA variant. With GWAS and a large number of SNP the QTL region can be reduced from hundreds of candidate genes in the region to a few and, if lucky, only one gene that is known to be functionally related to the trait. A candidate gene approach can then be applied to see which DNA variant in or directly around the gene causes the most variation. The advantage of GWAS is that it provides an unbiased search for candidate genes across the whole genome, rather than identifying candidate genes based on known biological function as done in traditional candidate gene studies. In any case, the discovery of causal variants remains rather challenging. One can end up with

multiple DNA variants in strong LD with each other as candidates (Mackay, 2001), and functional studies are required to validate the effect of the DNA variant on the phenotype.

1.4 Causal relationships among bovine milk fatty acids

In quantitative genetics, relationships among traits are often explored by correlations. Analyses involving phenotypic and genetic correlations between milk fatty acids (Karijord et al., 1982; Soyeurt et al., 2007; Stoop et al., 2008), clustering techniques (Heck et al., 2012; Massart-Leën and Massart, 1981), or principal component analysis (Fievez et al., 2003) reflect the biological pathways of de novo synthesis, biohydrogenation and desaturation of fatty acids. These studies suggest that certain fatty acids have a common origin, but are not able to distinguish direct from indirect relationships and, thus, do not imply causality. Causality is the relation between events, where one event is the direct consequence of the other event (the cause). Statistical causal inference aims to reason to the conclusion that something is, or is likely to be, the cause of something else. Visualizing these causal relationships among variables in a graph increases our understanding and interpretation of complex biological systems, while quantifying causal relations allows predicting outcomes of external interventions applied to such a causal network.

A search for causality can differentiate between direct and indirect relationships among variables. However, true causality is difficult to declare. A controlled experiment that isolates the effect of one variable on a system by holding constant all variables but the one under observation can declare true causality between two variables. Also a completely randomized experiment can declare true causality between two variables, where random assignment of each subject to different treatment groups coupled with random assignment of treatment level to each group results in averaging out potential sources of confounding effects. In traditional animal production data, variables that act as confounders are not averaged out but, once they are measured, they can be included in the model to correct for this confounding effect. Valente et al. (2010) developed an approach to deal with animal production data in a mixed model setting to search for causal relations between phenotypes. With this method partial correlations between milk fatty acids can be explored to determine causality between them. The resulting causal structure can then be used as condition for a structural equation model to estimate the magnitude of causal relationships among the fatty acids. Visualizing

causal relations between milk fatty acids may enhance our understanding of synthesis of milk fatty acids.

1.5 Aim and outline of this thesis

The research described in this thesis studied the genetic background of bovine milk fat composition and aimed to detect, confirm and fine-map QTL for individual milk fatty acids in Dutch Holstein Friesian cattle. In chapter 2 a GWAS for the most abundant fatty acids in winter milk samples was performed and detected 54 regions on 29 chromosomes that were significantly associated with one or more milk fatty acids. In chapter 3 a GWAS was performed on fatty acids from summer milk samples and detected 51 regions on 24 chromosomes that were significantly associated with one or more milk fatty acids. Associations detected in the GWAS of fatty acids based on summer milk samples was in agreement with most of the associations detected in the GWAS of fatty acids based on winter milk samples and, thus, confirmed these associations. Chapter 4 aimed to refine the location of the major QTL on BTA19 for bovine milk fat composition that was detected in chapter 2 and confirmed in chapter 3. The QTL region was narrowed down to a linkage disequilibrium block from 51,303,322 to 51,388,329 bp on BTA19 that contained 2 genes: coiled-coil domain containing 57 (*CCDC57*) and fatty acid synthase (*FASN*). Since many QTL regions were associated with multiple fatty acids chapter 5 aimed to provide more insight into the causal relations among the individual milk fatty acids. The general discussion (chapter 6) focused on what insights can be gained from this thesis and what more can be done to better understand the genetic background of milk fat composition. First, the methods used and results obtained in this thesis were discussed, followed by the importance of intermediate phenotypes to close the gap between QTL and complex phenotypes. Next, the inference of causal relations between QTL and phenotypes was explored. And finally, the potential of mid-infrared (MIR) predicted milk fatty acids instead of milk fatty acids measured by gas chromatography (GC) as phenotypes for GWAS was discussed.

2

Genome-wide association of milk fatty acids in Dutch dairy cattle

Aniek C. Bouwman, Henk Bovenhuis, Marleen H.P.W. Visker,
Johan A.M. van Arendonk

Animal Breeding and Genomics Centre, Wageningen University, P.O. Box 338,
6700 AH Wageningen, the Netherlands

BMC Genetics (2011) 12:43

Abstract

Background

Identifying genomic regions, and preferably individual genes, responsible for genetic variation in milk fat composition of bovine milk will enhance the understanding of biological pathways involved in fatty acid synthesis and may point to opportunities for changing milk fat composition via selective breeding. An association study of 50,000 single nucleotide polymorphisms (SNPs) was performed for even-chain saturated fatty acids (C4:0-C18:0), even-chain monounsaturated fatty acids (C10:1-C18:1), and the polyunsaturated C18:2*cis9,trans11* (CLA) to identify genomic regions associated with individual fatty acids in bovine milk.

Results

The two-step single SNP association analysis found a total of 54 regions on 29 chromosomes that were significantly associated with one or more fatty acids. *Bos taurus* autosomes (BTA) 14, 19, and 26 showed highly significant associations with seven to ten traits, explaining a relatively large percentage of the total additive genetic variation. Many additional regions were significantly associated with the fatty acids. Some of the regions harbor genes that are known to be involved in fat synthesis or were previously identified as underlying quantitative trait loci for fat yield or content, such as *ABCG2* and *PPARGC1A* on BTA 6; *ACSS2* on BTA 13; *DGAT1* on BTA 14; *ACLY*, *SREBF1*, *STAT5A*, *GH*, and *FASN* on BTA 19; *SCD1* on BTA26; and *AGPAT6* on BTA 27.

Conclusions

Medium chain and unsaturated fatty acids are strongly influenced by polymorphisms in *DGAT1* and *SCD1*. Other regions also showed significant associations with the fatty acids studied. These additional regions explain a relatively small percentage of the total additive genetic variance, but they are relevant to the total genetic merit of an individual and in unraveling the genetic background of milk fat composition. Regions identified in this study can be fine mapped to find causal mutations. The results also create opportunities for changing milk fat composition through breeding by selecting individuals based on their genetic merit for milk fat composition.

Key words: milk fatty acids, dairy, genome-wide association

2.1 Background

The biosynthesis of bovine milk fat is a complicated process regulated by many genes belonging to several pathways (Bionaz and Loor, 2008). Genetic analyses of bovine milk fatty acids have shown heritable variation. Short and medium chain fatty acids (C4:0 up to and including C16:0), which are synthesized de novo in the mammary gland, have moderate to high heritability (Soyeurt et al., 2007; Stoop et al., 2008). Long chain fatty acids (i.e. C16:0 and higher) are derived from blood lipids that originate mainly from the diet and endogenous lipids, nevertheless, they have low to moderate heritability (Mele et al., 2009; Soyeurt et al., 2007; Stoop et al., 2008). Identification of genomic regions, and preferably individual genes, responsible for genetic variation in milk fat composition will enhance the understanding of biological pathways involved in fatty acid synthesis and may point towards opportunities for changing milk fat composition via selective breeding. Candidate gene studies have shown that polymorphisms in diacylglycerol O-acyltransferase 1 (*DGAT1 K232A*) (Grisart et al., 2002) and stearoyl-CoA desaturase 1 (*SCD1 A293V*) (Taniguchi et al., 2004) have important effects on milk fat composition (Conte et al., 2010; Kgwatalala et al., 2009; Mele et al., 2007; Moioli et al., 2007; Schennink et al., 2007, 2008). Many genes are involved in the biosynthesis of milk fat, and analyzing these candidate genes one by one in a candidate gene approach is not an option; therefore, quantitative trait loci (QTL) studies try to identify regions associated with milk fat composition to identify candidate genes that are worth considering.

In order to identify genomic regions involved in the biosynthesis of milk fat, Schennink et al. (2009b) and Stoop et al. (2009b) performed genome-wide linkage analyses of milk fatty acids and detected genome-wise significant QTL and several suggestive QTL. Other linkage studies have been performed for single chromosomes (Morris et al., 2007) or the fat composition of adipose tissue in beef cattle (Abe et al., 2008; Alexander et al., 2007; Morris et al., 2007; Morris et al., 2010).

Recent developments in molecular genetics have made it possible to perform genome-wide association studies using thousands of single nucleotide polymorphism (SNP) markers to detect QTL. A genome-wide association study has higher power to detect QTL and provides more precise estimates of QTL locations compared to a linkage study. Some genome-wide associations for routinely evaluated traits in dairy cattle, such as milk production and fertility, have been published (Daetwyler et al., 2008; Lillehammer et al., 2009; Mai et al., 2010; Pryce

et al., 2010). To the best of our knowledge, no genome-wide association study of milk fatty acids has been reported.
The aim of this study was to perform a genome-wide association analysis using 50,000 SNP markers to identify QTL for individual fatty acids in bovine milk. Associations were studied for even-chain saturated fatty acids (C4:0-C18:0), even-chain monounsaturated fatty acids (C10:1-C18:1), and the polyunsaturated fatty acid C18:2*cis9,trans11* (CLA).

2.2 Methods

Phenotypes

The fat composition of winter milk samples from 1,905 first-lactation Dutch Holstein Friesian cows was available for this study. The cows were housed on 398 commercial farms throughout the Netherlands. At least three cows were sampled per farm. The cows were between 63 and 282 days in milk. The period of negative energy balance in early lactation was avoided by choosing cows over 63 days in lactation. The population consisted of five large paternal half-sib families from proven sires (200, 199, 195, 176, 101 daughters per sire) and 50 small paternal half-sib families from test-sires (10-24 daughters per sire), as well as 190 cows descending from 45 other proven sires (1-30 daughters per sire). The pedigree of the cows was supplied by CRV (Cooperative cattle improvement organization, Arnhem, the Netherlands) and consisted of 26,300 animals. Milk fat composition was measured by gas chromatography. Many fatty acids were measured, but only the major fatty acids are reported here: even-chain saturated fatty acids C4:0 to C18:0, even-chain (*cis9*) monounsaturated fatty acids C10:1 to C18:1, and the polyunsaturated fatty acid CLA. The fatty acids were expressed in terms of weight-proportion of total fat weight (w/w%). In total, these fatty acids made up 89% of the total fat. Table 2.1 presents the mean, phenotypic standard deviation, and intra-herd heritability for the fatty acids included in this study. More detailed information about the population and phenotypes can be found in Stoop et al. (2008).

Genotypes

Blood samples were collected from the cows for DNA isolation. The cows were genotyped using a custom Infinium Array (Illumina, San Diego, CA, USA) designed by CRV. In total, 1,810 cows were successfully (call rate > 90%) genotyped. The cows were genotyped for 50,855 technically successful SNPs. The assumed map positions of the SNPs were based on the bovine genome assembly BTAU 4.0 (Liu et al., 2009). From these 50,855 SNPs, a total of 776 SNPs could not be mapped to any

of the *Bos taurus* (BTA) chromosomes and were assigned to BTA 0. In addition, 591 of the SNPs were located on the X chromosome. The SNPs on BTA 0 and the X chromosome were included in the study. The average distance between SNPs was 52,452 bp. Monomorphic SNPs (n = 245), SNPs with a genotyping rate < 80% (n = 383), and SNPs with a genotype frequency < 0.006 (1-9 observations for one of the genotype classes, SNPs with two genotype classes instead of three were included in the final marker set; n = 5,494) were discarded from the original SNP set of 50,855 SNPs, resulting in the final marker set of 44,733 SNPs used for the association analysis. Table 2.2 provides an overview of the number of SNPs available for the association study per chromosome.

Table 2.1 Descriptive statistics of milk fatty acids. Mean (w/w%), phenotypic standard deviation (σ_P), and intra-herd heritability (h^2_{IH}) for the fatty acids of winter milk samples

Trait	Mean	σ_P [1]	h^2_{IH}
C4:0	3.50	0.24	0.44
C6:0	2.22	0.14	0.47
C8:0	1.37	0.12	0.61
C10:0	3.03	0.35	0.72
C12:0	4.11	0.46	0.64
C14:0	11.61	0.78	0.62
C16:0	32.59	2.15	0.43
C18:0	8.72	1.18	0.24
C10:1	0.37	0.06	0.34
C12:1	0.12	0.02	0.38
C14:1	1.36	0.23	0.34
C16:1	1.44	0.30	0.44
C18:1	18.18	1.57	0.26
CLA	0.39	0.07	0.42

[1] Phenotypic standard deviation after adjusting for systematic environmental effects: days in milk, age at first calving, season of calving, and herd.

Table 2.2 SNP information per *Bos Taurus* chromosome. Total number of SNPs, map length, average SNP interval, number of monomorphic SNPs, number of SNPs with a genotyping rate (genorate) < 80% and number of SNPs with a genotype frequency (freq) < 0.006 for all *Bos Taurus* (BTA) chromosomes.

BTA	SNP	Length (Mbp)	SNP interval (bp)	Monomorph	Genorate	Freq
0[1]	776	-	-	1	6	84
1	3,011	160.91	53,371	4	25	355
2	2,451	140.64	57,333	20	19	269
3	2,342	127.13	54,212	12	11	260
4	2,300	124.09	53,930	11	20	239
5	2,215	125.78	56,788	5	18	214
6	2,844	122.39	43,050	12	25	334
7	2,017	111.67	55,392	8	14	209
8	2,131	116.93	54,818	15	15	243
9	1,860	108.05	58,090	14	20	206
10	1,911	106.10	55,406	14	11	202
11	2,193	110.01	50,187	18	12	239
12	1,512	85.22	56,324	6	14	147
13	1,689	84.00	49,732	11	12	155
14	2,122	81.29	38,272	9	17	222
15	1,446	84.23	58,130	5	7	169
16	1,455	77.83	53,454	8	12	183
17	1,561	76.40	48,942	5	14	159
18	1,282	66.04	51,429	3	8	131
19	1,452	65.13	44,826	5	7	147
20	1,479	75.41	50,985	3	9	156
21	1,246	69.08	55,440	5	15	130
22	1,256	61.75	49,161	4	10	139
23	1,169	53.27	45,570	7	10	107
24	1,296	64.93	50,141	10	13	159
25	1,256	43.44	34,617	6	9	109
26	1,131	51.00	45,097	7	6	152
27	933	48.73	52,280	3	13	117
28	899	46.01	51,184	6	5	94
29	1,029	51.78	50,371	6	4	109
X	591	88.46	149,940	2	2	55
Total	50,855	2,628		245	383	5,494

[1] unmapped SNP

Statistical analysis

For the association study, both phenotype and genotype information was available for 1,706 individuals. A two-step single SNP association analysis was performed. In the first step, the genome was screened for interesting regions using a general linear model. In the second step, the interesting regions were verified using an animal model.

In the first step, a genome-wide association study was performed with a general linear model using the R package 'SNPassoc' (González et al., 2007). In this step, the analyzed phenotypes were pre-adjusted for systematic environmental effects, and the general linear model accounted for the SNP effect and the effect of sire. The general linear model used in the first step was:

$$y^{*}_{ij} = \mu + \text{sire}_{i} + \text{SNP}_{j} + e_{ij}\,, \tag{1}$$

where **y*** was the phenotype adjusted for the systematic environmental effects; sire was the fixed effect of sire; SNP was the fixed effect of SNP genotype; and **e** was the random residual. Sire effect was included in the SNPassoc model to account for paternal half-sib relations. Phenotypes were adjusted for days in milk, age at first calving, calving season, and herd. Adjusted phenotypes were obtained from the phenotypes of 1,905 cows using an animal model in ASReml (Gilmour et al., 2006):

$$\begin{aligned} y_{ijklmn} = {} & \mu + b_1 \times \text{dim}_i + b_2 \times e^{-0.05 \times \text{dim}_i} + b_3 \times \text{afc}_j + b_4 \times \text{afc}_j^2 + \\ & \text{season}_k + \text{scode}_l + \text{herd}_m + \text{animal}_n + e_{ijklmn}\,, \end{aligned} \tag{2}$$

where **y** was the (unadjusted) phenotype; μ was the overall mean; dim was the covariate describing the effect of days in milk; afc was the covariate describing the effect of age at first calving; season was the fixed effect of the class of calving season (June-Aug 2004, Sept-Nov 2004, or Dec 2004-Jan 2005); scode was the fixed effect accounting for differences in genetic level between groups of proven bull daughters, young bull daughters, and other bull daughters; herd was the random effect of herd, distributed as $N(0, \mathbf{I}\sigma^2_{\text{herd}})$, with identity matrix **I** and herd variance σ^2_{herd}; animal was the random additive genetic effect of the individual, distributed as $N(0, \mathbf{A}\sigma^2_a)$, with the additive genetic relationship matrix **A** and the additive genetic variance σ^2_a; and **e** was the random residual, distributed as $N(0, \mathbf{I}\sigma^2_e)$, with identity matrix **I** and residual variance σ^2_e.

The genome-wide false discovery rate (FDR) was controlled according to the method described by Storey and Tibshirani (Storey and Tibshirani, 2003), by separately calculating the genome-wide FDR based on the *P*-values from the general linear model for each trait using the R package 'qvalue'. Associations with a genome-wide FDR < 0.05 for the general linear model were considered significant.

The first step was performed to identify interesting regions, which were then further analyzed with an animal model to account for all relationships among individuals. Including a polygenic effect and accounting for genetic relationships would be more appropriate (Kennedy et al., 1992). The model including a polygenic effect is computationally demanding when analyzing many traits, SNPs, and animals; therefore, in the second step we only analyzed the regions that contained multiple SNPs that were significant in the first step.

A region started at the first significant SNP on a chromosome that was followed by an additional significant SNP within 10 Mbp; the region was extended as long as another significant SNP occurred within 10 Mbp from the previous one and ended at the last significant SNP that was not followed by another significant SNP within the next 10 Mbp. Thus, a region contained at least two significant SNPs. More than one region could be present on the same chromosome when there were groups of significant SNPs located within 10 Mbp from each other but further than 10 Mbp from the other region(s) on the chromosome. The 10 Mbp distance between significant SNPs is rather large, but it was chosen to prevent having many small regions on one chromosome, each containing a small number of significant SNPs.

In the second step, all SNPs in regions with significant effects were analyzed using animal model (2) extended with an SNP effect in ASReml (Gilmour et al., 2006). In this model the phenotypes were simultaneously adjusted for systematic environmental effects, for all genetic relationships among individuals, and for the SNP genotype. Associations with a $-\log_{10}(P\text{-value}) \geq 3$ were considered significant.

The genetic variance explained by an SNP was calculated from the estimated genotype effects from animal model (2) extended with an SNP effect and the observed genotype frequencies. The result was expressed as a percentage of the total additive genetic variance. These percentages can be overestimated, especially when the effect of an SNP is small, this is due to the so called Beavis effect (Beavis, 1998). The percentage of the total additive genetic variance explained by the most significant SNP per trait per region is reported. The most significant SNP can differ per trait for a region associated with multiple traits.

2.3 Results

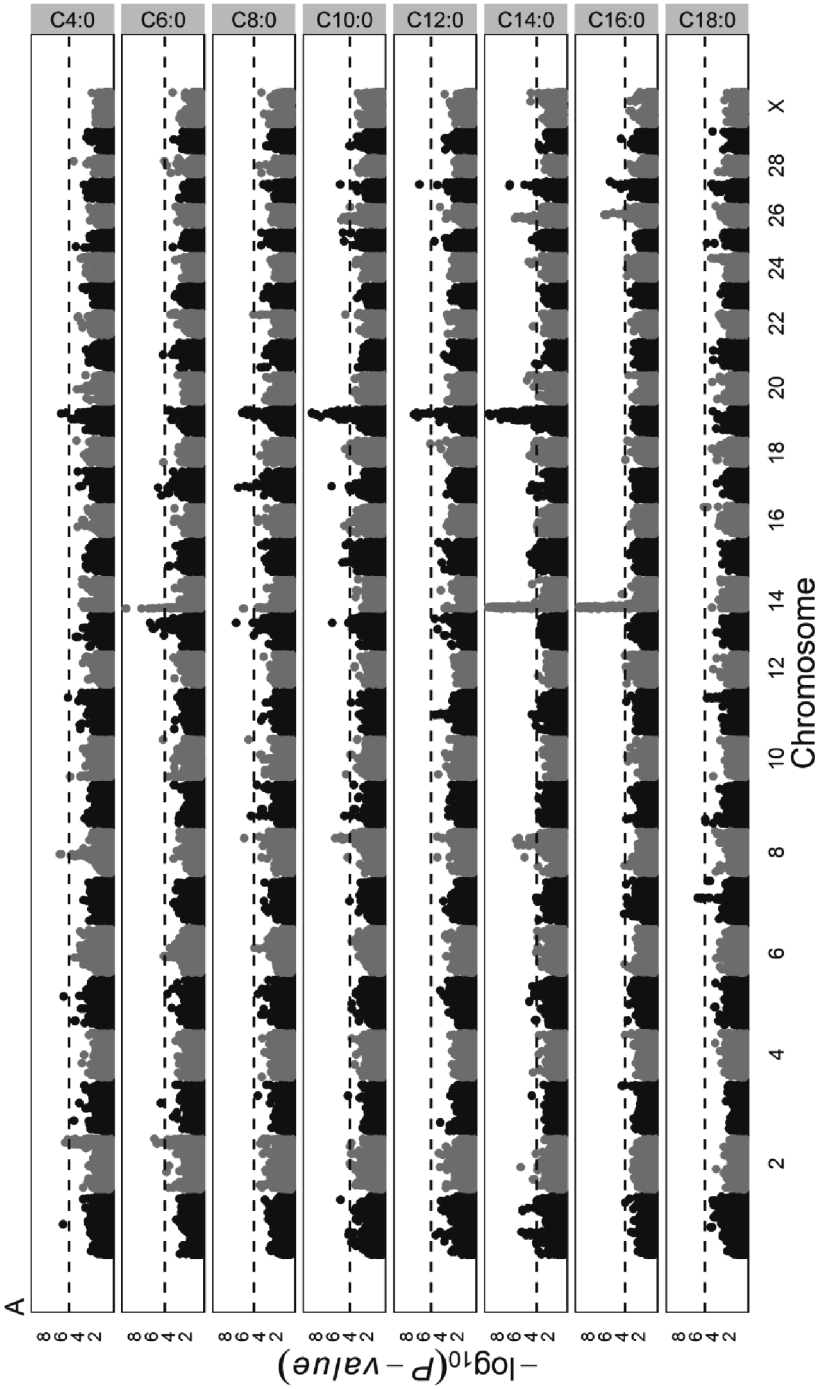
In the first step of the single SNP association study, all SNPs were analyzed using a general linear model. The analysis resulted in many significant (FDR < 0.05) associations between SNPs and the studied fatty acids. Figure 2.1 shows the genome-wide plots of -$\log_{10}$(*P*-values) for all of the studied fatty acids. All analyzed fatty acids showed significant associations with at least one genomic region.

From the results of the general linear model, a total of 64 significant regions were defined, affecting one or more (up to 10) traits. Significant associations with SNPs were detected on all chromosomes; however, on BTA 29 we identified only a single SNP associated with a trait instead of multiple SNPs in a region. Therefore, no regions were defined on BTA 29. These 64 regions were analyzed with an animal model and 54 remained significant at -$\log_{10}$(*P*-values) ≥ 3. Thus, most of the regions identified with the general linear model were confirmed with the animal model. The correlation between *P*-values in both analyses was 0.90. All results mentioned hereafter originate from the animal model and concern the 54 regions that remained significant at -$\log_{10}$(*P*-values) ≥ 3 with the animal model.

Significant associations were detected with several regions for most of the fatty acids, especially C14:0 (19 regions) and C16:1 (18 regions), whereas C18:0 and C12:0 were associated with only one region (Table 2.3). The X chromosome showed significant association with C14:0. Table 2.3 shows that several regions affect more than one trait. In particular, the medium chain unsaturated fatty acids were often associated with the same region.

Sixteen of the unmapped SNPs showed significant associations with one or more of the studied fatty acids. Comparing the sequence of these significant unmapped SNPs against genome assembly UMD 3.0 (Zimin et al., 2009) identified 14 of them as mapping to regions 14a, 19b, and 26, which were already identified as being associated with the traits. The other two unmapped SNPs (ULGR_BTA-38166 and ULGR_BTA-38169) were significantly associated with C10:0 and located at 24.2 Mbp on BTA 16 according to the UMD 3.0 genome assembly (Zimin et al., 2009).

For each region, we estimated the variance explained by the most significant SNP for each trait. Within a region, this most significant SNP can be a different SNP per trait. The percentage of total additive genetic variance explained by the most significant SNP per region ranged from 0.9 for C10:0 on BTA 1 to 61.9 for C18:1 on BTA 14 (Table 2.3). These percentages can be overestimated, especially when the effect of an SNP is small, this is due to the so called Beavis effect (Beavis, 1998). The QTL with large effects, such as region 14a and 26, are less likely overestimated.


A
C4:0
C6:0
C8:0
C10:0
C12:0
C14:0
C16:0
C18:0
8
6
4
2
−log10(P − value)
2
4
6
8
10
12
14
16
18
20
22
24
26
28
X
Chromosome

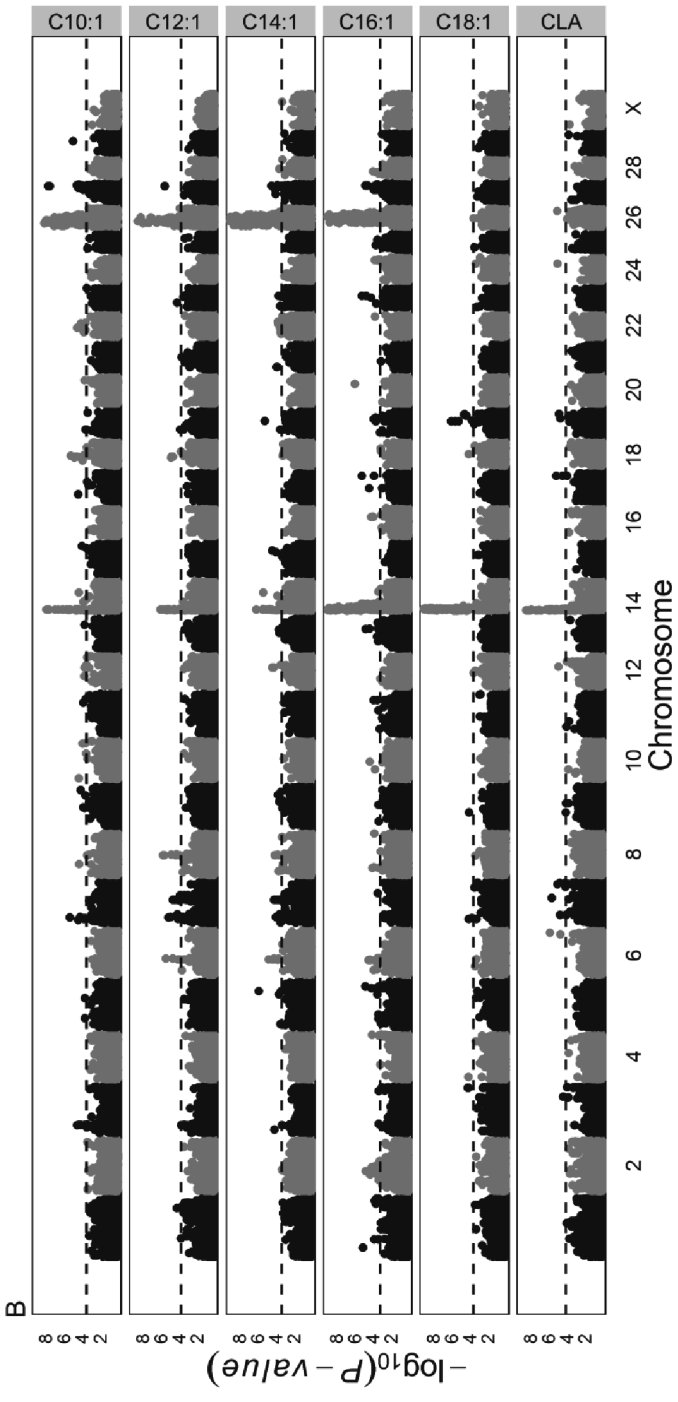


Figure 2.1 Genome-wide association plots for milk fatty acids. Genome-wide plots of -$\log_{10}$(*P*-values) (y-axis) for association of loci with saturated fatty acids (A) and unsaturated fatty acids (B) analyzed with a general linear model. The genomic position is represented along the x-axis and chromosome numbers are given on the x-axis. The dashed horizontal lines represent the 0.05 false discovery rate thresholds. The y-axis are cut off at -$\log_{10}$(*P*-values) of 9.

Table 2.3 Regions significantly (-$\log_{10}$(*P*-values) ≥ 3) associated with fatty acids and the percentage of the total additive genetic variance explained. Percentage of total additive genetic variance explained by the most significant SNP associated with the trait per region, analyzed with an animal model.

Region	Start (Mbp)	End (Mbp)	# SNP in reg	Trait													#Traits /Region
				C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA	
1a	25	25	2												3.3			1
1b	54	63	148				0.9		2.0									2
1c	122	130	120						3.0									1
1d	146	161	304												2.1			1
2a	28	28	3						1.8									1
2b	44	70	351												2.7			1
2c	126	140	257	4.0	3.3													2
3a	22	25	55									4.1						1
3b	125	127	39							2.9						5.3		2
4	122	124	30												2.2			1
5a	66	66	3						1.5									1
5b	82	100	275						2.7					4.8				2
5c	108	114	104						1.8						2.7			2
6a	31	45	630										4.3	1.3	3.1			3
6b	111	116	81														4.2	1
7a	11	24	206									4.1	3.9	3.3		4.3		4

Region	Start (Mbp)	End (Mbp)	# SNP in reg	Trait														#Traits /Region
				C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA	
7b	62	64	46						2.0		7.4		3.4	3.0				4
7c	100	111	166														3.1	1
8a	21	22	28												2.7			1
8b	53	58	104	4.5									4.2					2
8c	77	101	406				1.3		2.4									2
9a	21	22	2				1.1											1
9b	37	37	3													3.9		1
9c	49	50	25									3.9						1
9d	67	67	4											4.8				1
10a	8	8	3				1.2											1
10b	99	100	34									2.8						1
11	35	48	218						2.3									1
12	52	60	116									2.2		2.8				2
13	50	72	377		3.0	2.2	1.9							4.4	3.3			5
14a	0	26	1121		4.9	2.2			17.7	39.6		6.2	4.8	4.5	34.8	61.9	12.2	10
14b	40	41	7							3.7								1
15a	21	27	121				1.2		1.8									2
15b	61	65	63											3.4				1
15c	72	78	103									3.8						1

Region	Start (Mbp)	End (Mbp)	# SNP in reg	Trait														#Traits /Region
				C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA	
16	53	54	9												2.3			1
17a	31	34	43			2.9	1.9											2
17b	69	69	3												3.5		3.0	2
18a	14	30	305						1.8			3.1	3.5					3
18b	54	55	15						1.9									1
19a	6	6	3												2.3			1
19b	33	62	529	6.5		3.0	4.4	5.3	13.8						3.0	7.1	3.9	8
20	73	74	18						1.6									1
21	6	6	4											3.0				1
22	9	42	593									3.9		3.5				2
23	26	32	141												2.9			1
24a	40	49	142						1.3									1
24b	56	59	27												2.2			1
25	43	43	2				1.5											1
26	3	41	724				3.3		4.5	4.8		35.4	20.5	67.8	23.2			7
27	29	48	358						2.7	3.7		7.6	4.1	4.4	3.2			6
28a	7	13	116												2.8			1
28b	20	20	2											2.9				1
X	85	86	16						2.0									1

Region	Start (Mbp)	End (Mbp)	# SNP in reg	Trait														#Traits /Region
				C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA	
Sum			8,605	14.9	11.2	10.3	18.9	5.3	68.7	54.7	7.4	77.1	48.7	114.0	102.3	82.6	26.5	
All SNPs simultaneous in model[2]				14.9	10.2	10.9	16.3	5.3	41.8	53.9	7.4	62.0	45.2	97.4	79.2	78.6	24.1	
# Regions/trait				3	3	4	10	1	19	5	1	11	8	14	18	5	5	107

[1] A region starts at the first significant SNP and proceeds if the next significant SNP is positioned within the next 10 Mbp on the same chromosome, ending at the position of the last significant SNP matching this requirement, with a minimum of 2 significant SNP per trait in a region. The number of the region stands for the BTA number plus an a, b, c or d indicating different regions within a BTA.

[2] Percentage of the total additive genetic variance explained by all the most significant SNPs per region together. This was analyzed with the animal model and all the most significant SNPs per region simultaneous in the model.

Three chromosomes showed highly significant regions, which were associated with multiple fatty acids: region 14a showed significant associations with 10 traits, region 19b with 8 traits, and region 26 with 7 traits. The region 14a harbors *DGAT1* and region 26 harbors *SCD1*; these genes are known to affect fat composition (Conte et al., 2010; Kgwatalala et al., 2009; Mele et al., 2007; Moioli et al., 2007; Schennink et al., 2007, 2008). These regions explained a large portion of the total additive genetic variation. The results also showed that additional regions had an effect on single fatty acids or a certain group of fatty acids.

In this section the SNP showing the strongest association of the whole genome for each trait is reported. Strongest associations for C14:0 ($-\log_{10}$(*P*-values) = 36.6), C16:0 ($-\log_{10}$(*P*-values) = 51.3), and C16:1 ($-\log_{10}$(*P*-values) = 54.1) were found with two SNPs (ULGR_SNP_AJ318490_1b and ULGR_SNP_AJ318490_1c), located at 0.4 Mbp on BTA 14, that are responsible for the *DGAT1 K232A* polymorphism. For C6:0 ($-\log_{10}$(*P*-values) = 8.4), C18:1 ($-\log_{10}$(*P*-values) = 48.6), and CLA ($-\log_{10}$(*P*-values) = 15.8) the *DGAT1 K232A* SNPs were the most significant SNPs immediately after an unmapped SNP (ULGN_SNP_AJ318490_2). After comparing this unmapped SNP sequence against the sequence of *DGAT1*, we identified it as an SNP located in *DGAT1* that is in high LD (r^2=0.99) with the *DGAT1 K232A* SNPs. The strongest associations for C10:1 ($-\log_{10}$(*P*-values) = 41.1), C12:1 ($-\log_{10}$(*P*-values) = 22.8), and C14:1 ($-\log_{10}$(*P*-values) = 80.0) were found with an unmapped SNP (ULGR_SNP_SCD). Also, C16:1 showed strong association ($-\log_{10}$(*P*-values) = 34.8) with this unmapped SNP on BTA 26, but C16:1 showed the strongest association with the *DGAT1 K232A* SNPs on BTA 14 as mentioned above. After comparing this unmapped SNP sequence against the sequence of *SCD1*, we identified it as the SNP underlying the *SCD1 A293V* polymorphism. Although *SCD1* was not mapped on genome assembly BTAU 4.0, it was mapped on genome assembly UMD 3.0 at 21.1 Mbp on BTA 26 (Zimin et al., 2009). The strongest associations for C10:0 ($-\log_{10}$(*P*-values) = 11.0) and C12:0 ($-\log_{10}$(*P*-values) = 9.7) were found with an SNP (ULGR_MARC_10099_486) located at 52.5 Mbp on BTA 19. The strongest association for C4:0 ($-\log_{10}$(*P*-values) = 6.6) was found with an SNP (ULGR_BTA-45866) located at 54.6 Mbp on BTA 19, and the strongest association for C8:0 ($-\log_{10}$(*P*-values) = 6.2) was found with an SNP (ULGR_BTA-45847) located at 55.1 Mbp on BTA 19. Also, C14:0 showed strong association ($-\log_{10}$(*P*-values) = 22.0) with an SNP (ULGR_BTA-45758) on BTA 19 located at 52.1 Mbp, but C14:0 showed the strongest association with the *DGAT1 K232A* SNPs on BTA 14 as mentioned above. The strongest association for C18:0 ($-\log_{10}$(*P*-values) = 5.8) was found with an SNP (ULGR_BTA-28678) located at 64.2 Mbp on BTA 7.

2.4 Discussion

This study is the first to report a genome-wide association study of bovine milk fatty acids. A two-step single SNP association analysis was performed. In the first step, the genome was screened for interesting regions using a general linear model. In the second step, the interesting regions were verified using an animal model. The animal model accounted for all relationships between individuals in the pedigree, whereas the general linear model accounted only for the paternal half-sib structure of the phenotyped cows. Ignoring relationships between individuals can cause false positive associations (Goddard and Hayes, 2009); therefore, the animal model was applied to verify the results from the general linear model.

The results showed that *DGAT1* and *SCD1* are highly associated with several of the fatty acids. The results also showed that, for some traits, other regions were more significantly associated, such as BTA 19 for some short and medium chain saturated fatty acids. In addition, many other regions were associated with fatty acids, but with smaller effects.

Schennink et al. (2009b) and Stoop et al. (2009b) reported a linkage study of milk fat composition using some of the same data used in the present study. In the linkage study, 1,341 SNPs were genotyped in 849 cows and their seven sires. Schennink et al. (2009b) and Stoop et al. (2009b) detected genome-wide significant QTL on BTA 6, 14, 19, and 26 for fatty acids included in the present study. The QTL on BTA 14, 19, and 26 were confirmed in our study. Our results suggest a QTL on BTA 6 for C6:0 and C8:0 (Figure 2.1A), but for C6:0 only one SNP exceeded the FDR threshold, and for C8:0 none of the SNPs exceeded the FDR threshold. Therefore, this region was not included in the animal model analysis in our second step. Given the quite stringent threshold, this region is still likely to harbor a QTL for C6:0 and C8:0.

The suggestive QTL found by Schennink et al. (2009b) and Stoop et al. (2009b) on BTA 2, 13, 14, 17, 19, and 26 was also confirmed in our study. Other suggestive QTL reported in the studies could not be confirmed, though in some cases a QTL was indicated, but this did not pass the threshold. In addition, several novel regions significantly associated with fatty acids were detected in our study but not reported by Schennink et al. (2009b) and Stoop et al. (2009b).

Some studies of the fat composition of adipose tissue in beef cattle were confirmed by our findings regarding milk fat composition. This result indicates that these regions are not unique for milk fat composition. Our findings confirmed the QTL detected by Morris et al. (2010) in subcutaneous fat from beef cattle: C16:1 on BTA 1; C14:0, C16:1, and C18:1 on BTA 19; C14:0 and C14:1 on BTA 26; and C14:0 on

BTA 27. Our findings confirmed the QTL detected by Abe et al. (2008) in back fat, intermuscular fat, or intramuscular fat from beef cattle: C14:0 and C18:1 on BTA 19. Our findings confirmed the QTL detected by Alexander et al. (2007) in the longissimus dorsi of beef cattle: CLA on BTA 7. Our findings confirmed the QTL detected by Uemoto et al. (2011) in intramuscular adipose tissue from beef cattle: C18:1 on BTA 19.

Major regions

Three major regions were detected in this genome-wide association study (regions 14a, 19b, and 26), with significant effects on milk fat composition. These regions showed highly significant associations with several fatty acids. The regions on BTA 14 and 26 are regions that have been studied previously, and our results confirm the earlier findings (Conte et al., 2010; Grisart et al., 2002; Kgwatalala et al., 2009; Mele et al., 2007; Moioli et al., 2007; Schennink et al., 2007, 2008; Taniguchi et al., 2004). The region on BTA 19 has not been studied extensively in relation to milk fat composition, but it harbors a number of candidate genes involved in fatty acid synthesis.

The region on BTA 14 showing an association with C6:0, C8:0, C14:0, C16:0, C10:1, C12:1, C14:1, C16:1, C18:1, and CLA is the region harboring *DGAT1*, which is known to influence milk production traits (Grisart et al., 2002) and milk fat composition (Conte et al., 2010; Schennink et al., 2007, 2008). For all of these traits, except C10:1, C12:1, and C14:1, the two most significant SNPs on BTA 14 (located at 0.4 Mbp) were the two SNPs corresponding to the dinucleotide substitution of *DGAT1* resulting in a *K* to *A* amino acid substitution (*DGAT1 K232A*). The *K* allele is associated with larger fractions of C6:0, C8:0, C16:0, and C16:1, and with smaller fractions of C14:0, C18:1, and CLA.

For C10:1 and C14:1, the most significant SNP (ULGR_BTC-068225) was located at 3.0 Mbp, and for C12:1 the most significant SNP (ULGR_BTC-067423) was located at 3.7 Mbp. However, for C10:1 and C12:1, the *DGAT1 K232A* SNPs were also significant. After correcting the phenotypes for the effect of the *DGAT1 K232A* polymorphism, the $-\log_{10}(P\text{-values})$ of these most significant SNPs decreased from 6.50 to 2.30 for C10:1, from 5.44 to 2.52 for C12:1, and from 4.63 to 2.40 for C14:1. The LD between these most significant SNPs and the *DGAT1 K232A* SNPs was moderate ($r^2 = 0.26$ and 0.31). Although the associations of C10:1, C12:1, and C14:1 were just below the significance threshold after correcting for the *DGAT1 K232A* genotype, the findings still suggest that an additional QTL may be present for the medium chain unsaturated fatty acids on BTA 14.

The region on BTA 26 that showed an association with C10:0, C14:0, C16:0, C10:1, C12:1, C14:1, and C16:1 is the region harboring *SCD1*, which is known to be associated with the desaturation of fatty acids (Kgwatalala et al., 2009; Mele et al., 2007; Moioli et al., 2007; Schennink et al., 2008). For all of these traits, except C16:0, the most significant SNP on BTA 26 corresponded to the nucleotide substitution in *SCD1* that causes an *A* to *V* amino acid substitution (*SCD1 A293V*) at 21.1 Mbp. The *A* allele was associated with larger fractions of C10:1, C12:1, and C14:1, and with smaller fractions of C10:0, C14:0, and C16:1. Thus, the *A* allele resulted in more C10:1 and C14:1 at the cost of C10:0 and C14:0. A similar effect was found for C12:0 and C12:1, though C12:0 was not significantly associated with the SNP. The opposite effect was found for C16:0 and C16:1; the *A* allele resulted in less C16:1 and more C16:0, though C16:0 was not significantly associated with the SNP. The *SCD1* gene codes for the SCD enzyme, which desaturates saturated fatty acids to Δ9 unsaturated fatty acids in the mammary gland (Pereira et al., 2003). The association of this SNP with, and its effects on, the medium chain unsaturated fatty acids and their equivalent saturated fatty acids is, therefore, in agreement with the function of the enzyme. The associations we identified for the medium chain unsaturated fatty acids confirm previous studies on the effect of the *SCD1 A293V* polymorphism on milk fatty acids (Kgwatalala et al., 2009; Moioli et al., 2007; Schennink et al., 2008). The associations we identified for the medium chain saturated fatty acids confirm only the results of Schennink et al. (2008), who used the same population as the present study.

The *SCD1 A293V* SNP was not significant for C16:0 ($-\log_{10}$(*P*-value) = 1.09), which also confirms previous studies (Kgwatalala et al., 2009; Mele et al., 2007; Schennink et al., 2008). The most significant SNP for C16:0 on BTA 26 was located at 28.8 Mbp. This SNP was not in LD with the *SCD1 A293V* SNP (r^2 = 0.08), and correcting for *SCD1 A293V* had little effect on the significance of the SNP associated with C16:0 ($-\log_{10}$(*P*-value) decreased from 5.52 to 4.46). Also, one allele of this SNP is associated with larger fractions of C10:1, C12:1, C14:1, and C16:0, and with smaller fractions of C10:0, C12:0, C14:0, and C16:1, suggesting that it has something to do with desaturation, but it was only significantly associated with C16:0. We did not identify obvious candidate genes in this region.

The region on BTA 19, at 32.7-61.8 Mbp, showed associations with C4:0, C8:0, C10:0, C12:0, C14:0, C16:1, C18:1, and CLA, i.e. mainly with the short and medium chain saturated fatty acids and with the long chain unsaturated fatty acids. No significant effects were found for C6:0, but a QTL was indicated below the threshold in the region on BTA 19 (Figure 2.1A). Morris et al. (2007) performed a

linkage analysis of milk fatty acids on BTA 19, detecting QTL for C8:0, C10:0, C12:0, C14:0, C18:1, and C18:2, which was confirmed by our findings and suggested fatty acid synthase (*FASN*, at 52.2 Mbp) as a candidate gene responsible for the observed effect. In addition to *FASN*, several other genes located within the region on BTA 19 are involved in the biosynthesis of milk fat, including sterol regulatory element binding transcription factor 1 (*SREBF1*, at 35.7 Mbp), ATP citrate lyase (*ACLY*, at 43.4 Mbp), signal transducer and activator of transcription 5A (*STAT5A*, at 43.7 Mbp), and growth hormone (*GH*, at 49.7 Mbp). These genes are all candidate genes because SNPs in the whole region showed an association with the traits, perhaps in LD with mutations in genes not captured by our marker set. The strongest association was found near *FASN*, but also near some other candidate genes as discussed below.

The region on BTA 19 was strongly associated with C14:0 and explained a large portion of the total additive genetic variation of C14:0. The SNP most significant for C14:0 ($-\log_{10}$(*P*-value) = 22.04) was also the most significant SNP for C18:1 ($-\log_{10}$(*P*-value) = 4.52) and located at 52.1 Mbp on BTA 19 (ULGR_BTA-45758 in *LOC518878*), which is 71,862 bp from *FASN*. This SNP also showed significant effects on C4:0 ($-\log_{10}$(*P*-value) = 3.57), C10:0 ($-\log_{10}$(*P*-value) = 7.01), C12:0 ($-\log_{10}$(*P*-value) = 5.96), and CLA ($-\log_{10}$(*P*-value) = 3.67), whereas the association with C8:0 ($-\log_{10}$(*P*-value) = 2.96) was just below the threshold. The effects of SNPs on C8:0, C10:0, C12:0, and C14:0 were in opposite direction of the effects on C4:0, C16:1, C18:1, and CLA. Fatty acid synthase (encoded by *FASN*) is a multi-enzyme system involved in de novo fatty acid synthesis. The SNP effects suggest that less C4:0 and more C8:0, C10:0, C12:0, and C14:0 are produced by fatty acid synthesis, or vice versa, which is in line with the function of *FASN*. Three SNPs in *FASN* were included in the marker set used in our study; however, two of them ($FASN_{16009}$ and $FASN_{763}$) were monomorphic for our population. The third SNP, $FASN_{17924}$, showed association with C14:0 ($-\log_{10}$(*P*-value) = 3.05). Schennink et al. (2009a) also studied $FASN_{16024}$, finding a significant association with C14:0 for the same population as in our study. Morris et al. (2007) did not find significant associations between $FASN_{17924}$ and C14:0, but did find a significant association between C14:0 and $FASN_{15531}$ and $FASN_{15603}$.

A haplotype of five *FASN* SNPs has been shown to be significantly associated with C6:0, C8:0, C10:0, C12:0, C14:0, and C18:1 in Friesian-sired cows (Morris et al., 2007), which are almost the same traits for which we found an association in the region, though not specifically with SNPs in *FASN*. This finding suggests a QTL in this region with an effect on short and medium chain fatty acids and long chain

unsaturated fatty acids, but whether it is actually *FASN* that causes the association remains unclear. Our genome-wide scan showed that an SNP outside of *FASN* is the most significant SNP for C14:0 and was also associated with some of the other traits. This SNP showed very strong association with C14:0, whereas the association of C14:0 with the SNP in *FASN* was just barely significant. Candidate gene studies have shown that *FASN* is mainly associated with C14:0, but we found associations with additional traits, similar to the haplotype findings of Morris et al. (2007). Perhaps the causal mutation is located outside of *FASN* and is mainly the effect on C14:0 strong enough to be detected by SNPs in LD with this mutation.

This region on BTA 19 also harbors *GH*, and SNPs in this gene have been associated with milk production traits, including fat yield (Yao et al., 1996). One SNP in our marker set was located in exon 5 of *GH* (GH-D30713-299) and showed significant association with C18:1 ($-\log_{10}$(*P*-value) = 3.81). The neighboring SNP showed even more associations: with C8:0 ($-\log_{10}$(*P*-value) = 3.48), C10:0 ($-\log_{10}$(*P*-value) = 5.13), C12:0 ($-\log_{10}$(*P*-value) = 3.86), and C14:0 ($-\log_{10}$(*P*-value) = 7.41).

No SNPs were located in the other candidate genes on BTA 19. The SNP showing the strongest association with C16:1 was located at 43.8 Mbp on BTA 19 (BFGL-NGS-111365), which is 13,873 bp from *STAT5A*. The SNPs neighboring *ACLY* and *STAT5A* showed significant associations with several of these traits, especially C14:0. All SNPs in the regions seem to be detecting the same effect, which is strongest with C14:0. Although previous studies suggested *FASN* as the candidate gene for association, which of the candidate genes causes the effect remains debatable. The causal mutation might even be in a gene not considered here.

Additional regions

In addition to the three major regions mentioned above, many additional regions showed associations with fatty acids (see Figure 2.1). We will not discuss all of the regions in this paper, but what follows are some select regions that showed association with three or more traits.

On BTA 6, region 6a was associated with C12:1, C14:1, and C16:1 (Table 2.3). The SNP effects were in the same direction for all three fatty acids. This region harbors the genes ATP binding cassette, subfamily G, member 2 (*ABCG2*, 37.4 Mbp) and peroxysome proliferator-activated receptor-gamma coactivator-1alpha (*PPARGC1A*, 44.8 Mbp). The SNPs most significantly associated with C12:1 ($-\log_{10}$(*P*-value) = 5.03, at 44.3 Mbp), C14:1 ($-\log_{10}$(*P*-value) = 4.39, at 41.2 Mbp), and C16:1 ($-\log_{10}$(*P*-value) = 4.63, at 40.2 Mbp) were located between these candidate genes. *ABCG2* has been associated with milk fat yield and percentage (Cohen-Zinder et al.,

2005; Olsen et al., 2007). One SNP in our marker set was located in *ABCG2* and showed no significant associations with the studied fatty acids. *PPARGC1A* has been associated with milk fat yield in German Holsteins (Weikard et al., 2005), but this was not confirmed by Khatib et al. (2007) in two larger American Holstein populations. Up to 10 SNPs in our marker set were located in *PPARGC1A*, but none of these were significantly associated with C12:1, C14:1, or C16:1. However, one of the 10 SNPs in *PPARGC1A* showed an association with C16:1, which was just below the significance threshold ($-\log_{10}(P\text{-value}) = 2.86$). In a candidate gene study, Schennink et al. (2009a) found a significant association in the same population as in the present study for two SNPs in *PPARGC1A* that were not included in our marker set: $PPARGC1A_{1790+514}$ with C16:1 and $PPARGC1A_{1892+19}$ with C14:1. Our genome-wide scan indicates that not *PPARGC1A*, but another region has the strongest effect on the unsaturated medium chain fatty acids.

On BTA 7, two regions showed an association with several fatty acids: region 7a associated with C10:1, C12:1, C14:1, and C18:1; and region 7b associated with C14:0, C12:1, C14:1, and C18:0 (Table 2.3).

Region 7a showed an association with almost all unsaturated fatty acids. In general, the SNP effects on C18:1 were in the opposite direction of the SNP effects on C10:1, C12:1, and C14:1. SNPs in this region were either associated with C10:1, C12:1, and C14:1, or with C18:1, suggesting one QTL for the medium chain unsaturated fatty acids and another for C18:1. The SNP effects on C10:0, C12:0, and C14:0 were in the same direction as the effects on C10:1, C12:1, and C14:1, but they were not significant. The SNP effects on C18:0 were in the same direction as the effects on C18:1, but they were not significant. Although only unsaturated fatty acids were significantly associated with this region, the SNP effects suggested that this QTL has nothing to do with desaturation because the SNP effects on unsaturated fatty acids were in the same direction as the effects on their saturated equivalents.

Region 7b showed associations with C14:0, C18:0, C12:1, and C14:1. The most significant SNP for each trait was different, but they were located in the same neighborhood, around 64.0-64.1 Mbp, and were in high LD with one another (r^2 = 0.53-0.97). This finding indicates the likelihood of one QTL in this region with an effect on these four traits. The effects of these most significant SNPs were in the opposite direction for C18:0 compared to C14:0, C12:1, and C14:1. The SNP effects on C10:0 and C12:0 were in the same direction as the effects on C14:0, C12:1, and C14:1, but they were not significant. The SNP effects on C16:0, C18:1, and CLA were in the same direction as the effects on C18:0, but they were not significant. These

SNP effects suggest that this QTL has something to do with a trade-off between long chain fatty acids and de novo synthesis of medium chain fatty acids. No candidate genes were located in this region.

On BTA 13, region 13 was associated with C6:0, C8:0, C10:0, C14:1, and C16:1 (Table 2.3). This region confirms the QTL for C6:0, C14:1, and C16:1 detected by Stoop et al. (2009b), who also found that C8:0 and C10:0 showed a QTL around the same position as C6:0, but these QTL were just below the threshold and, therefore, not reported. For C6:0, C8:0, and C10:0, the same SNP, located at 64.8 Mbp (ULGR_SNP_BES11_Contig346_1209), was the most significant SNP in the region, and the SNP effect was in the same direction for all three traits. These traits have a high genetic correlation (Stoop et al., 2008), which supports one QTL affecting these three short chain fatty acids. This SNP is located in acyl-CoA synthetase short-chain family member 2 (*ACSS2*), which activates acetate for de novo fatty acid synthesis (Bionaz and Loor, 2008); thus, *ACSS2* is a good candidate gene for a QTL with an effect on C6:0, C8:0, and C10:0. Given that this particular significant SNP is located in an intron (between exon 16 and 17) of *ACSS2*, this SNP is not likely the causal mutation, but it can be in high LD with the causal mutation. In addition, the region also had an effect on C14:1 and C16:1. The SNP effects for C14:1 and C16:1 were in the same direction, but these effects were in opposite direction of SNP effects on C6:0-C10:0. The SNP in *ACSS2* had no significant association with C14:1 or C16:1, which indicates an additional QTL with an effect on these unsaturated fatty acids.

On BTA 27, region 27 was associated with C14:0, C16:0, C10:1, C12:1, C14:1, and C16:1 (Table 2.3). The region on BTA 27 includes 1-acylglycerol-3-phosphate O-acyltransferase 6 (*AGPAT6*), which is the most abundant isoform of all *AGPAT* mRNA (~60%) in the mammary gland and involved in positioning fatty acids on the second position of the triglyceride backbone (Bionaz and Loor, 2008). Given that 62.2% of C14:0 and 43.1% of C16:0 is located at the second position of the triglyceride backbone (Jensen, 2002), this gene might be a candidate for this association. The SNP effects on C14:0 and C16:0 were in opposite directions, which suggests competition between C14:0 and C16:0 for the second position of the triglyceride backbone as an explanation for this association. In *AGPAT6* knock-out mice, the composition of the triacylglycerol is altered and contains proportionally more polyunsaturated fatty acids at the expense of monounsaturated fatty acids (Vergnes et al., 2006), which might explain the effect of this region on the mono-unsaturated fatty acids. In general, the SNP effects on C10:1, C12:1, C14:1, and C16:1 were in the same direction.

Variance explained

Table 2.3 shows that regions 14a and 26 explained a large portion of the total additive genetic variation of C10:1, C12:1, C14:0, C14:1, C16:0, C16:1, C18:1, and CLA. This variation is caused by *DGAT1* and *SCD1*. For other traits, however, no regions had such large effects. The sum of the total additive genetic variance explained by the most significant SNP per region for C4:0-C12:0 and C18:0 was less than 20% of the total additive genetic variance. This finding suggests that these traits are influenced by many genes with small effects. Regions explaining roughly 1% of the total additive genetic variation or more were detected by the studied design (e.g., C10:0 on BTA1), but additional regions with either undetectable small effects or that were not dense enough in our marker set to detect the effect is likely.

Even though short chain fatty acids are produced by de novo synthesis, less than 20% of the total additive genetic variance was explained by the analyzed regions. *FASN* plays an important role in de novo synthesis, but less than 6.5% of the total additive genetic variance of short chain fatty acids is explained by the region harboring *FASN*. De novo synthesis of fatty acids requires several compounds in addition to *FASN* enzymes to elongate fatty acids. One of the compounds is acetate, which is activated by *ACSS2* for de novo synthesis, *ACSS2* was indicated as a candidate gene associated with C6:0-C10:0 in this study. More genes like this that assist *FASN* in de novo synthesis and, therefore, explain a portion of the total additive genetic variation is likely. Also, genes involved in transport and triacyl-glyceride production might explain some of the total additive genetic variation.

2.5 Conclusions

A genome-wide association study of 50,000 SNPs was performed for milk fatty acids, resulting in many QTL. All over the genome regions were associated with milk fatty acids, some regions with just one fatty acid and other regions with multiple fatty acids. Milk fat composition is strongly influenced by polymorphisms in *DGAT1* and *SCD1*, genes that have large effects on medium chain fatty acids and unsaturated fatty acids. Several regions showed associations with these milk fatty acids, but with smaller effects. The short chain fatty acids, C12:0 and C18:0, are not strongly affected by genes with large effects, but are influenced by regions with small effects. Some regions included candidate genes involved in milk fat synthesis pathways. On BTA 19, there were several genes involved in fat synthesis underlying the region associated with multiple fatty acids. Only in a few cases was an SNP associated with fatty acids actually located in a candidate gene. Regions identified

in this study can be fine mapped to find causal mutations. The results also create opportunities for changing milk fat composition through breeding by selecting individuals based on their genetic merit for milk fat composition, which can be retrieved from the estimated SNP effects and the individual's genotype.

Acknowledgement

This study is part of the Dutch Milk Genomics Initiative and the project 'Melk op Maat', funded by Wageningen University (the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), the cooperative cattle improvement organization CRV (Arnhem, the Netherlands), the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands). The authors thank the herd owners for their help in collecting the data.

3

Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples

Aniek C. Bouwman, Marleen H.P.W. Visker, Johan A.M. van Arendonk, Henk Bovenhuis

Animal Breeding and Genomics Centre, Wageningen University, P.O. Box 338, 6700 AH Wageningen, the Netherlands

BMC Genetics (2012) 13:93

Abstract

Background

In this study we perform a genome-wide association study (GWAS) for bovine milk fatty acids from summer milk samples. This study replicates a previous study where we performed a GWAS for bovine milk fatty acids based on winter milk samples from the same population. Fatty acids from summer and winter milk are genetically similar traits and we therefore compare the regions detected in summer milk to the regions previously detected in winter milk GWAS to discover regions that explain genetic variation in both summer and winter milk.

Results

The GWAS of summer milk samples resulted in 51 regions associated with one or more milk fatty acids. Results are in agreement with most associations that were previously detected in a GWAS of fatty acids from winter milk samples, including eight 'new' regions that were not considered in the individual studies. The high correlation between the $-\log_{10}(P\text{-values})$ and effects of SNPs that were found significant in both GWAS imply that the effects of the SNPs were similar on winter and summer milk fatty acids.

Conclusions

The GWAS of fatty acids based on summer milk samples was in agreement with most of the associations detected in the GWAS of fatty acids based on winter milk samples. Associations that were in agreement between both GWAS are more likely to be involved in fatty acid synthesis compared to regions detected in only one GWAS and are therefore worthwhile to pursue in fine-mapping studies.

Key words: milk fatty acids, dairy, genome-wide association

3.1 Background

Dairy producers are looking for ways to optimize bovine milk fat composition for human health, and to improve physical and functional properties of milk. Increasing the knowledge about the synthesis of milk fatty acids, by unraveling the genetic background of milk fatty acids, can aid in modifying bovine milk fat composition. Polymorphisms in diacylglycerol-O-acyltransferase 1 (*DGAT1*) and stearoyl-CoA desaturase 1 (*SCD1*) are known to have an effect on milk fatty acids, e.g. the *DGAT1 K232A* polymorphism explains 40% of the genetic variation of C16:0 (Schennink et al., 2007) and the *SCD1 A293V* polymorphism explains 23% of the genetic variation of C16:1 (Bouwman et al., 2011). However, there is still a considerable amount of genetic variation in milk fat composition that has not been assigned to polymorphisms.

Chromosomal regions associated with milk fatty acids can be detected by screening the whole genome in a genome-wide association study (GWAS). In GWAS studies many thousands of single nucleotide polymorphisms (SNPs) are being tested for associations. In general it is expected that only a small proportion of the SNPs will have a true association and only those that have an effect that is large enough will be significant. Setting a significance threshold is finding the balance between limiting the number of false positives and maintaining sufficient power. Replication of results in independent samples is a strategy to separate false positives from true associations (Chanock et al., 2007; van den Oord, 2008).

In a previous GWAS we identified interesting regions of the bovine genome associated with milk fatty acids from winter milk samples (Bouwman et al., 2011). To our knowledge, at present that is the only GWAS on bovine milk fatty acids. Not many populations that are large enough for GWAS have been phenotyped for milk fatty acids, because accurate measurement of milk fatty acids using gas chromatography is expensive and time consuming. In addition, genotyping a large number of individuals for a large numbers of SNPs is costly. The population that was used for our previous GWAS, based on milk samples taken in winter, has been phenotyped for milk fatty acids in a second milk sample which was taken in summer. Repeating the GWAS for milk fatty acids based on winter samples using the summer samples can confirm the previously detected associations and result in new associations.

In this study we performed a GWAS for fatty acids based on summer milk samples. This study repeats our previous GWAS for fatty acids based on winter milk samples from the same population and largely the same animals. Fatty acids from summer and winter milk are genetically similar traits and we therefore compared the

regions detected in summer milk to the regions previously detected in winter milk to confirm associations.

3.2 Methods

Phenotypes

The phenotypes of the winter milk samples were described earlier in Bouwman et al. (2011), the phenotypes of the summer milk samples will be described in detail, as well as the differences between winter and summer samples.

The fat composition of summer milk samples from 1,795 first-lactation Dutch Holstein Friesian cows was available for this study. The cows were housed on 383 commercial farms throughout the Netherlands. At least three cows were sampled per farm. The cows were between 97 and 335 days in lactation when the summer samples were taken. The pedigree of the cows was provided by CRV (Cooperative cattle improvement organization, Arnhem, the Netherlands) and consisted of 26,300 animals.

Milk fat composition was measured by gas chromatography as described in Stoop et al. (2008). Many fatty acids were measured, but only the major fatty acids are reported here: even-chain saturated fatty acids C4:0 to C18:0, even-chain (*cis9*) monounsaturated fatty acids C10:1 to C18:1, and the polyunsaturated fatty acid C18:2*cis9,trans11* (CLA). These fatty acids made up 88% of the total milk fat. The fatty acids are expressed in terms of weight-proportion of total fat weight (w/w%).

The winter and summer milk samples were taken from the same cows during the same lactation. The moment of sampling resulted in some differences between the two samples. Winter milk samples were taken from February to March 2005, when Dutch cows are mainly kept indoors and fed silage. Summer milk samples were taken from May to June 2005, when Dutch cows are often grazing for at least some part of the day. Some cows sampled in winter were not in lactation anymore during summer, therefore additional cows were sampled in summer to assure at least 3 cows per herd. The cows were on average 167 days (63-282) in lactation when the winter samples were taken, and 247 days (97-335) in lactation when the summer samples were taken.

Statistical analysis of phenotypes

Variance components, heritabilities, and correlations between the fatty acids from summer and winter milk samples were estimated using a bivariate animal model in ASReml (Gilmour et al., 2006):

$$y_{ijklmn} = \mu + b_1 \times \text{dim}_i + b_2 \times e^{-0.05 \times \text{dim}_i} + b_3 \times \text{afc}_j + b_4 \times \text{afc}_j^2 + \text{season}_k + \text{scode}_l + \text{herd}_m + \text{animal}_n + e_{ijklmn}, \quad (1)$$

where **y** was the phenotype; μ was the overall mean; dim was the covariate describing the effect of days in milk; afc was the covariate describing the effect of age at first calving; season was the fixed effect of the class of calving season (June-Aug 2004, Sept-Nov 2004, or Dec 2004-Jan 2005); scode was the fixed effect accounting for differences in genetic level between proven sire daughters and test-sire daughters, proven sires are selected and therefore their daughters might have a different genetic level than daughters of test sires; herd was the random effect of herd, distributed as N(0, $\mathbf{I}\sigma^2_{\text{herd}}$), with identity matrix **I** and herd variance σ^2_{herd}; animal was the random additive genetic effect, distributed as N(0, $\mathbf{A}\sigma^2_a$), with the additive genetic relationship matrix **A** and the additive genetic variance σ^2_a; and e was the random residual, distributed as N(0, $\mathbf{I}\sigma^2_e$), with identity matrix **I** and residual variance σ^2_e.

Genotypes

Blood samples were collected from the cows for DNA isolation. Of the 1,795 cows phenotyped for the summer milk sample 1,656 were successfully genotyped for 50,855 single nucleotide polymorphisms (SNPs) using a custom Infinium Array (Illumina, San Diego, CA, USA) designed by CRV. The assumed map positions of the SNPs were based on the bovine genome assembly BTAU 4.0 (Liu et al., 2009). The average distance between SNPs was 52,452 bp. Of the 50,855 SNPs, 591 SNPs were located on the X chromosome, and 776 SNPs could not be mapped to any of the *Bos taurus* (BTA) chromosomes and were assigned to BTA 0. The SNPs on BTA 0 and the X chromosome were included in the study. Single nucleotide polymorphisms with a genotyping rate < 80% (n=392), monomorphic SNPs (n=236), and SNPs with 1-9 observations for one of the genotype classes (SNPs with such low number of observations in one of the genotype classes were excluded from further analyses to reduce the number of spurious associations) (n=5,646) were discarded from the original SNP set, resulting in the final marker set of 44,581 SNPs used for the summer GWAS.

Ethical approval

Genomic DNA of the cows was isolated from whole blood samples of the cows. Blood samples were collected in accordance with the guidelines for the care and use of animals as approved by the ethical committee on animal experiments of Wageningen University (protocol: 200523.b).

Genome-wide association based on summer milk samples

For the GWAS based on the summer milk samples, both phenotype and genotype data were available for 1,656 individuals. A single SNP GWAS was performed using an univariate animal model in ASReml that was the same as model 1 but extended with a fixed effect for the SNP genotype.

The genome-wide FDR was based on the *P*-values from the animal model using the R package 'qvalue' (Storey and Tibshirani, 2003). A genome-wide FDR was calculated for each trait individually. Associations with a genome-wide FDR < 0.05 were considered significant.

SNPs significant for one trait located close to each other were termed a "region". All significant SNPs in a region might be associated with the same causal mutation. A region was defined as follows: it started at the first significant SNP on a chromosome that was followed by an additional significant SNP within 10 Mbp; the region was extended as long as another significant SNP occurred within 10 Mbp from the previous one and ended at the last significant SNP that was not followed by another significant SNP within the next 10 Mbp. Thus, each region contained at least two SNPs significant for the trait. More than one region could be present on the same chromosome when there were groups of significant SNPs located within 10 Mbp from each other but further than 10 Mbp from the other region(s) on that chromosome.

The genetic variance explained by a SNP was calculated from the estimated genotype effects from the statistical model and the observed genotype frequencies. The result was expressed as a percentage of the total additive genetic variance obtained from model 1. These percentages can be overestimated due to the so called Beavis effect, especially when the effect of a SNP is small (Beavis, 1998). For each trait, the proportion of genetic variance explained by the most significant SNP in a region was reported. Note that the most significant SNP in a region can differ between traits.

Comparison of summer and winter GWAS results

The GWAS results from the winter milk samples from Bouwman et al. (2011) and from the summer milk samples were compared to each other. A total of 1,564 cows were studied in both the summer and the winter GWAS, 92 cows were only studied in the summer GWAS, and 142 only in the winter GWAS.

In Bouwman et al. (2011) a two-step single SNP approach was used for the GWAS of fatty acids from winter milk samples. Due to computation time, in the study by Bouwman et al. (2011) only regions which showed significant associations in analyses using a general linear model were re-analysed using an animal model to account for all relations between the animals. Here we used a one-step approach using the animal model only. To make results from the previous study and the present study comparable, the GWAS based on winter milk samples was redone using an animal model for all SNPs. The Pearson correlation between the $-\log_{10}(P$-values) of the general linear model for winter milk samples and the animal model for winter milk samples was 0.95, which indicates that the general linear model used correctly identified the regions of interest and that the results for winter milk samples of Bouwman et al. (2011) are comparable with the results presented in the current study.

For the individual GWAS studies for winter and summer milk samples a FDR < 0.05 was used. A FDR < 0.05 is stringent, especially when looking for agreement of results. Therefore, a SNP was qualified as associated with both summer and winter milk fatty acids when the FDR threshold was smaller than 0.20 in both GWAS studies. We choose a FDR threshold of 20%, because the FDR of agreement between the two GWAS studies would then be 4% (20%*20%) if the studies were independent (Liu et al., 2008b). SNPs in agreement were reported when at least more than one SNP was in agreement between the summer and winter GWAS in a region (were region was defined as above) or when the SNP was in agreement between the summer and winter GWAS for more than one trait.

3.3 Results

Table 3.1 shows that the phenotypic variation of milk fatty acids was larger in summer than in winter. Summer milk samples contained more long chain fatty acids (C18:0, C18:1, and CLA) and less C16:0 than winter milk samples (Table 3.1). Phenotypic correlations between fatty acids of summer and winter milk samples ranged between 0.36 and 0.67 (Table 3.2). Genetic correlations between fatty acids of summer and winter milk samples ranged between 0.77 and 1.00 (Table 3.2). Genetic correlations between summer and winter samples of C4:0, C6:0, C12:0, C18:0, C10:1, C12:1, C14:1, C16:1, and C18:1 (ranging between 0.90-1) were not

significantly different from one (Table 3.2). The genetic correlations for C8:0, C10:0, C14:0, C16:0, and CLA were significantly different from one but showed strong positive correlations (0.77-0.94, Table 3.2). Herd correlations between fatty acids of summer and winter milk samples ranged between 0.16 and 0.54 (Table 3.2).

Table 3.1 Mean (in w/w%), phenotypic variance ($\sigma^2_P=\sigma^2_a+\sigma^2_{herd}+\sigma^2_e$), intra-herd heritability ($h^2_{IH}=\frac{\sigma^2_a}{\sigma^2_a+\sigma^2_e}$), and proportion of variance due to herd (Herd=$\frac{\sigma^2_{herd}}{\sigma^2_a+\sigma^2_{herd}+\sigma^2_e}$) for the fatty acids of summer and winter milk samples, with their standard errors in subscript.

Sample		Trait	Mean	σ^2_P	h^2_{IH}	Herd
Summer	n=1,795					
		C4:0	$3.52_{0.35}$	$0.1267_{0.01}$	$0.37_{0.09}$	$0.24_{0.02}$
		C6:0	$2.17_{0.21}$	$0.0435_{0.00}$	$0.41_{0.09}$	$0.18_{0.02}$
		C8:0	$1.32_{0.17}$	$0.0288_{0.00}$	$0.41_{0.09}$	$0.19_{0.02}$
		C10:0	$2.87_{0.46}$	$0.2240_{0.01}$	$0.56_{0.10}$	$0.19_{0.02}$
		C12:0	$3.78_{0.73}$	$0.5550_{0.03}$	$0.52_{0.10}$	$0.40_{0.03}$
		C14:0	$11.15_{1.06}$	$1.1560_{0.05}$	$0.51_{0.10}$	$0.34_{0.03}$
		C16:0	$29.17_{3.50}$	$12.4400_{0.60}$	$0.36_{0.10}$	$0.50_{0.03}$
		C18:0	$9.88_{1.77}$	$3.1540_{0.13}$	$0.18_{0.07}$	$0.30_{0.03}$
		C10:1	$0.35_{0.07}$	$0.0051_{0.00}$	$0.48_{0.10}$	$0.25_{0.03}$
		C12:1	$0.11_{0.03}$	$0.0010_{0.00}$	$0.47_{0.10}$	$0.30_{0.03}$
		C14:1	$1.38_{0.28}$	$0.0754_{0.00}$	$0.46_{0.09}$	$0.15_{0.02}$
		C16:1	$1.40_{0.30}$	$0.0938_{0.00}$	$0.39_{0.09}$	$0.09_{0.02}$
		C18:1	$20.56_{2.80}$	$7.8000_{0.34}$	$0.37_{0.10}$	$0.34_{0.03}$
		CLA	$0.56_{0.28}$	$0.0796_{0.00}$	$0.28_{0.09}$	$0.58_{0.02}$
Winter	n=1,905					
		C4:0	$3.50_{0.27}$	$0.0775_{0.00}$	$0.43_{0.09}$	$0.16_{0.02}$
		C6:0	$2.22_{0.17}$	$0.0278_{0.00}$	$0.48_{0.10}$	$0.16_{0.02}$
		C8:0	$1.37_{0.14}$	$0.0202_{0.00}$	$0.62_{0.11}$	$0.20_{0.02}$
		C10:0	$3.03_{0.43}$	$0.2009_{0.01}$	$0.74_{0.11}$	$0.23_{0.02}$
		C12:0	$4.11_{0.69}$	$0.5041_{0.02}$	$0.64_{0.11}$	$0.43_{0.03}$
		C14:0	$11.61_{0.92}$	$0.8953_{0.04}$	$0.58_{0.10}$	$0.17_{0.02}$
		C16:0	$32.59_{2.83}$	$8.2030_{0.35}$	$0.37_{0.10}$	$0.30_{0.03}$
		C18:0	$8.72_{1.42}$	$1.9770_{0.07}$	$0.24_{0.07}$	$0.19_{0.02}$
		C10:1	$0.37_{0.07}$	$0.0044_{0.00}$	$0.33_{0.08}$	$0.10_{0.02}$
		C12:1	$0.12_{0.03}$	$0.0008_{0.00}$	$0.37_{0.08}$	$0.21_{0.02}$
		C14:1	$1.36_{0.26}$	$0.0614_{0.00}$	$0.33_{0.08}$	$0.07_{0.02}$
		C16:1	$1.44_{0.32}$	$0.1047_{0.00}$	$0.42_{0.09}$	$0.07_{0.02}$
		C18:1	$18.18_{2.04}$	$4.1790_{0.17}$	$0.28_{0.09}$	$0.29_{0.03}$
		CLA	$0.39_{0.11}$	$0.0130_{0.00}$	$0.44_{0.10}$	$0.51_{0.02}$

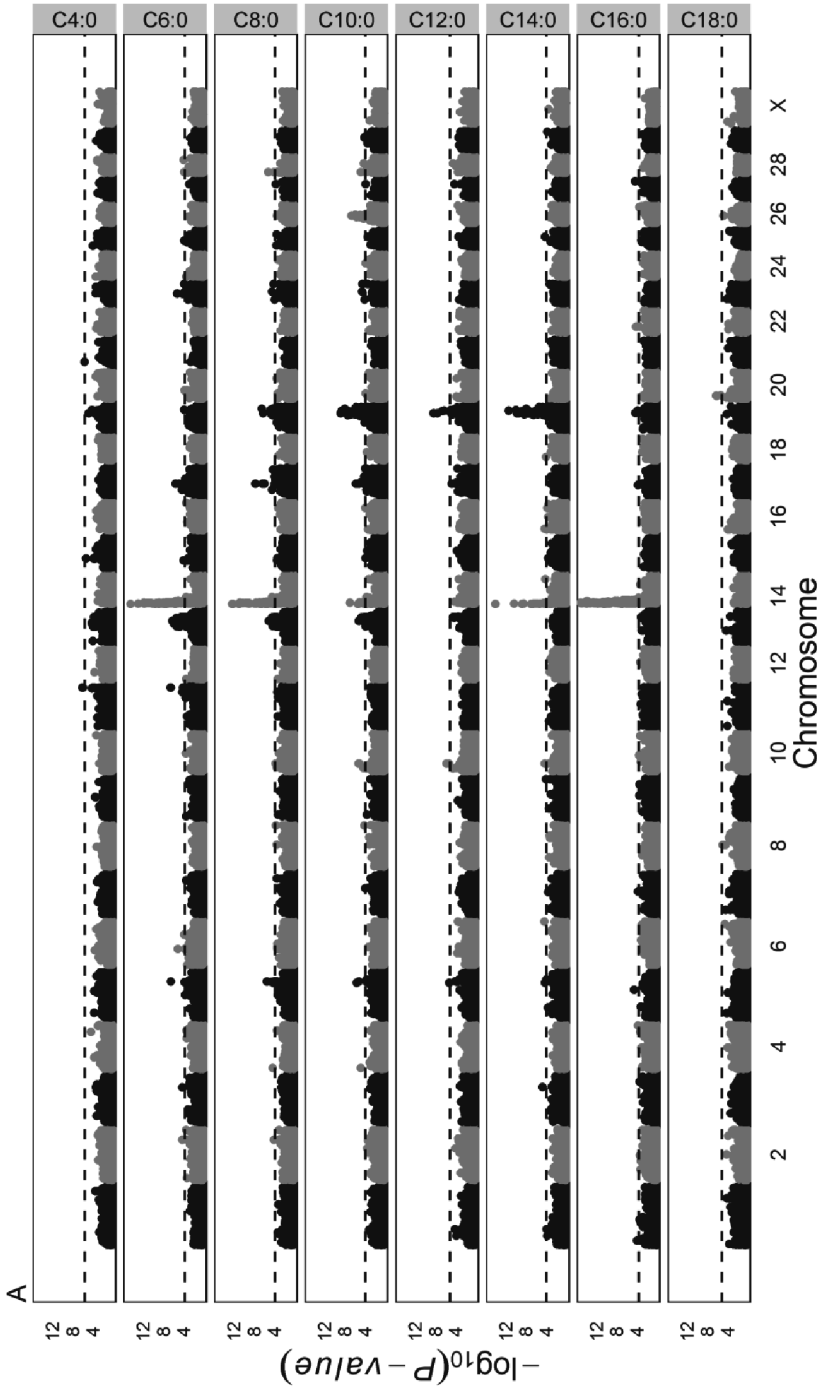
Table 3.2 Phenotypic (r_p), additive genetic (r_a), herd (r_{herd}), and residual correlation (r_e) between winter and summer milk samples, with their standard errors (se).

Trait	r_p [1]	se	r_a [2]	se	r_{herd}	se	r_e	se
C4:0	0.48	0.02	0.94^{ns}	0.06	0.31	0.08	0.25	0.09
C6:0	0.55	0.02	0.95^{ns}	0.05	0.42	0.08	0.29	0.09
C8:0	0.52	0.02	0.93^{*}	0.05	0.39	0.08	0.17	0.14
C10:0	0.56	0.02	0.94^{*}	0.04	0.41	0.07	-0.03	0.26
C12:0	0.54	0.02	0.98^{ns}	0.03	0.54	0.05	-0.06	0.21
C14:0	0.52	0.02	0.94^{*}	0.04	0.37	0.07	0.14	0.15
C16:0	0.42	0.03	0.77^{**}	0.11	0.20	0.06	0.47	0.07
C18:0	0.45	0.02	0.90^{ns}	0.10	0.26	0.08	0.41	0.05
C10:1	0.44	0.02	1.00^{ns}	0.03	0.31	0.10	0.15	0.10
C12:1	0.49	0.02	1.00^{ns}	0.03	0.37	0.07	0.21	0.10
C14:1	0.61	0.02	1.00^{ns}	0.02	0.16	0.14	0.46	0.06
C16:1	0.67	0.02	0.97^{ns}	0.03	0.19	0.17	0.53	0.06
C18:1	0.41	0.03	0.92^{ns}	0.08	0.19	0.07	0.33	0.07
CLA	0.36	0.03	0.81^{**}	0.11	0.30	0.06	0.24	0.08

[1] r_p is based on $\sigma^2_p=\sigma^2_a+\sigma^2_{herd}+\sigma^2_e$.
[2] Superscripts indicate whether the genetic correlation differs significantly from 0.995 (~1), where **P-value < 0.01, * P-value ≤ 0.05 and ns = non-significant, i.e., P > 0.05.

Genome-wide association based on summer milk samples

Figure 3.1 shows the genome-wide plots of -$\log_{10}$(*P*-values) of the GWAS of the fatty acids based on the summer milk samples. In total, 51 regions were associated with one or more fatty acids. Table 3.3 gives all detected regions and the percentage of the total additive genetic variation explained by the most significant SNP in that region for each of the fatty acids (for more detailed information about those most significant SNPs see Additional table 3.1). The most significant SNPs per region explained 2.2% up to 50.1% of the total additive genetic variation (Table 3.3). When all these most significant SNPs per region per fatty acid were analysed simultaneously they explained between 5.5% for C4:0 and 92.5% for C16:1 (Table 3.3). Three regions with major effects were associated with multiple fatty acids: BTA 14, BTA 19, and BTA 26. First we will describe the results for these three regions with major effects and then for the other regions associated with more than one fatty acid. Regions associated with only one fatty acid are given in Table 3.3 but will not be described here.


A
C4:0
C6:0
C8:0
C10:0
C12:0
C14:0
C16:0
C18:0
12
8
4
2
4
6
8
10
12
14
16
18
20
22
24
26
28
X
Chromosome
−log10(P − value)

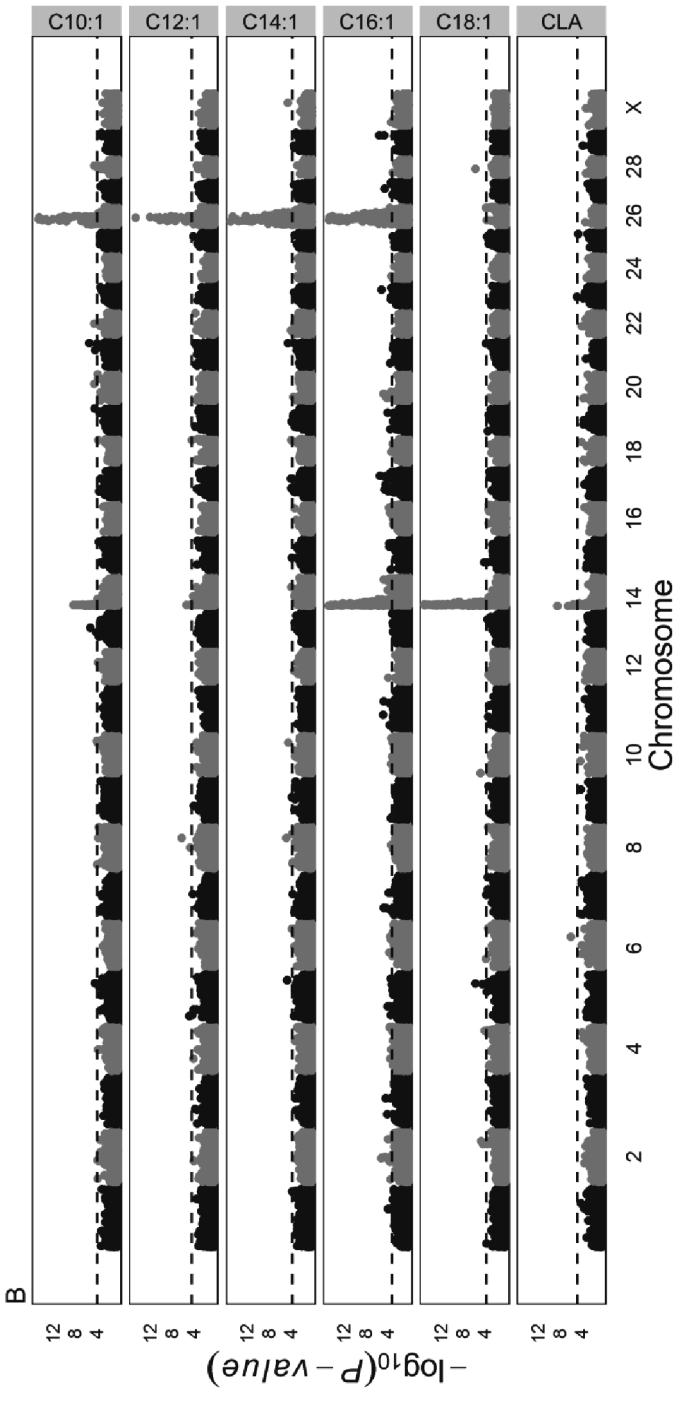


Figure 3.1 Genome-wide association plots for bovine milk fatty acids of summer milk samples. Genome-wide plots of -$\log_{10}$ (*P*-values) (y-axis) for association of SNPs with saturated fatty acids (A) and unsaturated fatty acids (B). The genomic position is represented along the x-axis and chromosome numbers are given on the x-axis. The dashed horizontal lines represent the 0.05 false discovery rate thresholds. The y-axis are cut off at -$\log_{10}$(*P*-value) of 15.

Table 3.3 Regions significantly associated with fatty acids of the summer milk samples and the percentage of the total additive genetic variance explained by the most significant SNP in that region.

Region[1]	Start (Mbp)	End (Mbp)	Trait														#Traits/ region
			C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA	
1a	105	106												3.3			1
1b	154	155											2.5				1
2a	56	70												4.0			1
2b	106	114													5.0		1
2c	119	119												3.0			1
3	73	73												3.5			1
4a	60	60												2.8			1
4b	122	124												3.2			1
5a	9	13										4.0					1
5b	36	36												3.4			1
5c	96	109		5.3	4.9	3.8		2.8			3.2		3.2	2.7	6.5		8
6a	41	41												2.9			1
6b	45	45		4.0													1
6c	77	85												2.8			1
6d	85	85		3.0													1
6e	105	106											2.9				1
7a	15	22												3.7			1
7b	64	64												3.0			1
10a	10	11				2.9											1

Region[1]	Start (Mbp)	End (Mbp)	Trait													#Traits/ region
			C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA	
10b	22	22				3.0	3.9										2
10c	86	90									3.1						1
11a	70	74												3.6			1
11b	95	107	5.5	5.2													2
13	41	68		5.5	4.6	3.2											3
14a	0	19		14.1	11.0	4.2		8.1	47.4		7.1	4.6		31.1	50.1	10.5	10
14b	44	51												3.6			1
14c	69	77												4.1			1
15a	21	21													3.6		1
15b	64	73												3.0			1
16a	2	4						2.2									1
16b	51	68												3.0			1
17a	15	24		3.3	3.9									3.7			3
17b	25	25											3.0				1
17c	29	34		5.3	8.2	4.1											3
17d	50	69											3.3	4.4			2
17e	74	74				2.2											1
19a	6	6												2.9			1
19b	37	62			6.0	5.7	6.3	12.3	4.3								5
20	9	26		2.8						15.6				4.0			3
21	64	65									4.2						1
22a	12	12											2.9				1

Region[1]	Start (Mbp)	End (Mbp)	Trait														#Traits/ region
			C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA	
22b	16	16							3.7								1
23a	16	24												2.7			1
23b	26	33		4.3													1
23c	43	49												4.1			1
26	1	39				4.5					20.6	15.0	46.4	30.7			5
27	43	48												2.6			1
28	3	3			4.9	3.1											2
29a	33	33				2.6											1
29b	44	44												3.5			1
X	63	63											3.9				1
Sum			5.5	52.8	43.6	39.3	10.2	25.4	55.4	15.6	38.3	23.6	68.0	145.3	65.2	10.5	
All SNPs in model[2]			5.5	28.4	35.6	29.3	6.5	21.4	51.2	15.6	37.1	22.0	61.8	92.5	63.6	10.5	
# Regions/trait			1	10	7	11	2	4	3	1	5	3	8	27	4	1	87

[1] A region starts at the first significant SNP and proceeds if the next significant SNP is positioned within the next 10 Mbp on the same chromosome, ending at the position of the last significant SNP matching this requirement, with a minimum of 2 significant SNPs per trait in a region. The number of the region stands for the BTA number plus an a, b, c, d or e indicating different regions within a BTA.

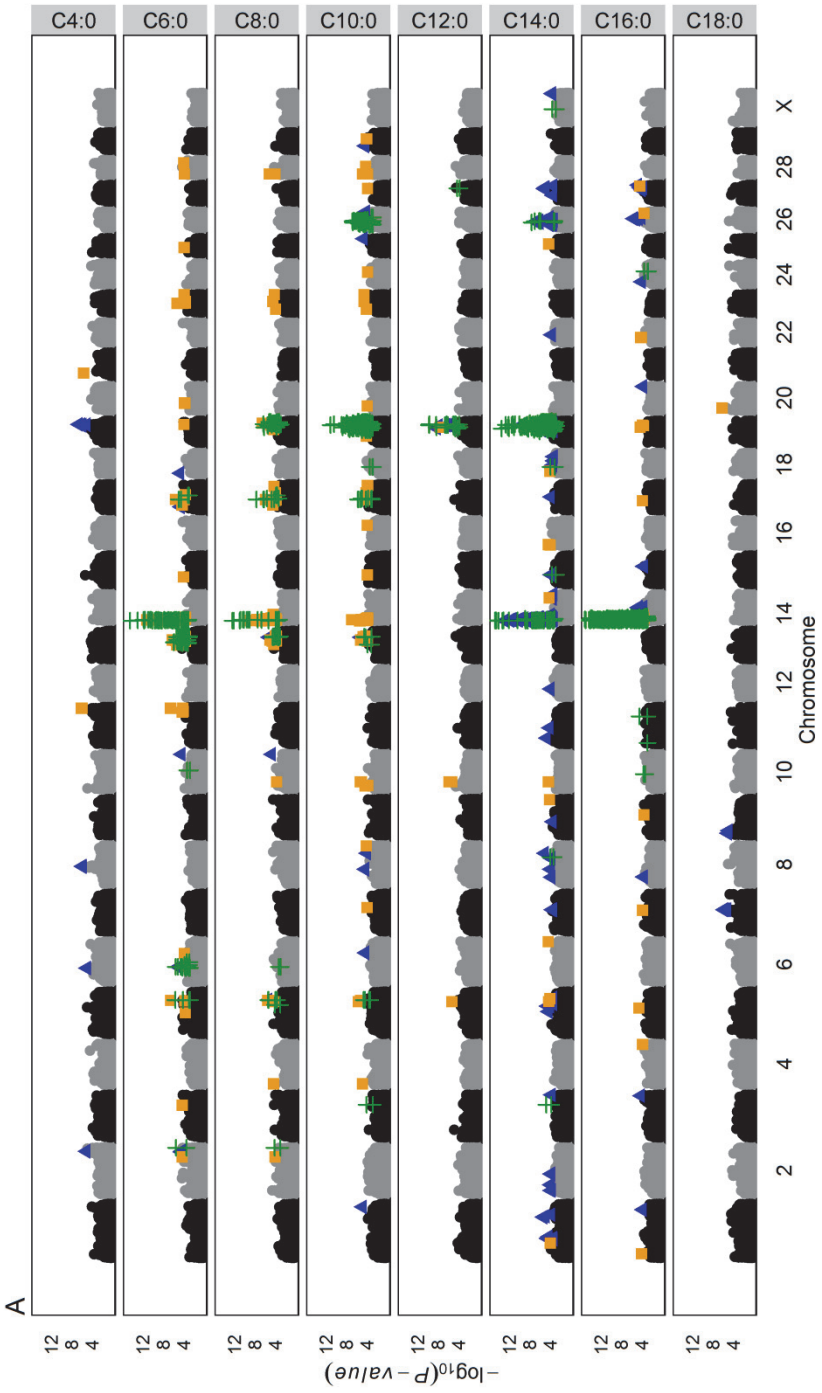
[2] Percentage of the total additive genetic variance explained by all the most significant SNPs per region together. This was analyzed with the animal model and all the most significant SNPs per region simultaneous in the model.

The association detected on BTA 14, between 0 and 18.9 Mbp (region 14a), with C6:0, C8:0, C14:0, C16:0, C16:1, C18:1, and CLA was most significant for three SNPs located in the *DGAT1* gene at 0.4 Mbp (including the two SNPs underlying the *DGAT1 K232A* dinucleotide polymorphism). The *DGAT1* SNPs explained 8.1-50.1% of the total additive genetic variation of these fatty acids (Table 3.3). The association detected on BTA 14 with C10:0 and C10:1 was also significant for the SNPs in the *DGAT1* gene. However, for these fatty acids the SNPs in the *DGAT1* gene were not the most significant ones in this region on BTA 14. The most significant SNPs were located at 3.6 Mbp for C10:0 and 3.0 Mbp for C10:1. The association detected on BTA 14 with C12:1 was not significant for the SNPs in the *DGAT1* gene ($-\log_{10}$(*P*-value) = 1.89). The association detected on BTA 14 with C12:1 was most significant for a SNP located at 3.2 Mbp. After correcting C10:0, C10:1, and C12:1 for the effect of the *DGAT1 K232A* polymorphism these most significant SNPs remained significant. The linkage disequilibrium (LD) between these SNPs and the *DGAT1 K232A* SNPs was moderate (r^2 = 0.14-0.34).

The association detected on BTA 19, between 37.3-62.3 Mbp (region 19b), with C8:0, C10:0, C12:0, C14:0, and C16:0 was most significant around 46 Mbp for C10:0, C12:0, and C16:0; around 52 Mbp for C14:0; and around 58 Mbp for C8:0. The most significant SNPs explained 4.3-12.3% of the total additive genetic variation of these fatty acids (Table 3.3).

The association detected on BTA 26, between 1.4 and 39.0 Mbp, with C10:0, C10:1, C12:1, C14:1, and C16:1 was most significant near the *SCD1* gene. The *SCD1* gene is not mapped on the BTAU 4.0 (Liu et al., 2009), but is mapped at 21 Mbp on BTA 26 according to the UMD 3.0 map (Zimin et al., 2009). Two SNPs located in the *SCD1* gene (including the *SCD1 A293V* polymorphism) were the most significant SNPs for the traits associated with the region on BTA 26. The *SCD1* SNPs explained 4.5% of the genetic variation of C10:0 and 15-46.4% of the total additive genetic variation of the medium chain unsaturated fatty acids (Table 3.3).

Beside these three regions with major effects, nine additional regions were associated with more than one fatty acid (Table 3.3). Region 5c was associated with C6:0, C8:0, C10:0, C14:0, C10:1, C14:1, C16:1, and C18:1. Region 10b was associated with C10:0, and C12:0. Region 11b was associated with C4:0, and C6:0. The region on BTA 13 was associated with C6:0, C8:0, and C10:0. On BTA 17 there were three regions associated with multiple traits: region 17a between 15.0 and 23.9 Mbp was associated with C6:0, C8:0, and C16:1; region 17c between 28.6 and 34.3 Mbp was associated with C6:0, C8:0, and C10:0; region 17d between 49.6 and 68.7 Mbp was associated with C14:1, and C16:1. The region on BTA 20 was associated with C6:0, C18:0, and C16:1. The region on BTA 28 was associated with C8:0, and C10:0.


A
C4:0
C6:0
C8:0
C10:0
C12:0
C14:0
C16:0
C18:0
12
8
4
−log10(P − value)
2
4
6
8
10
12
14
16
18
20
22
24
26
28
X
Chromosome

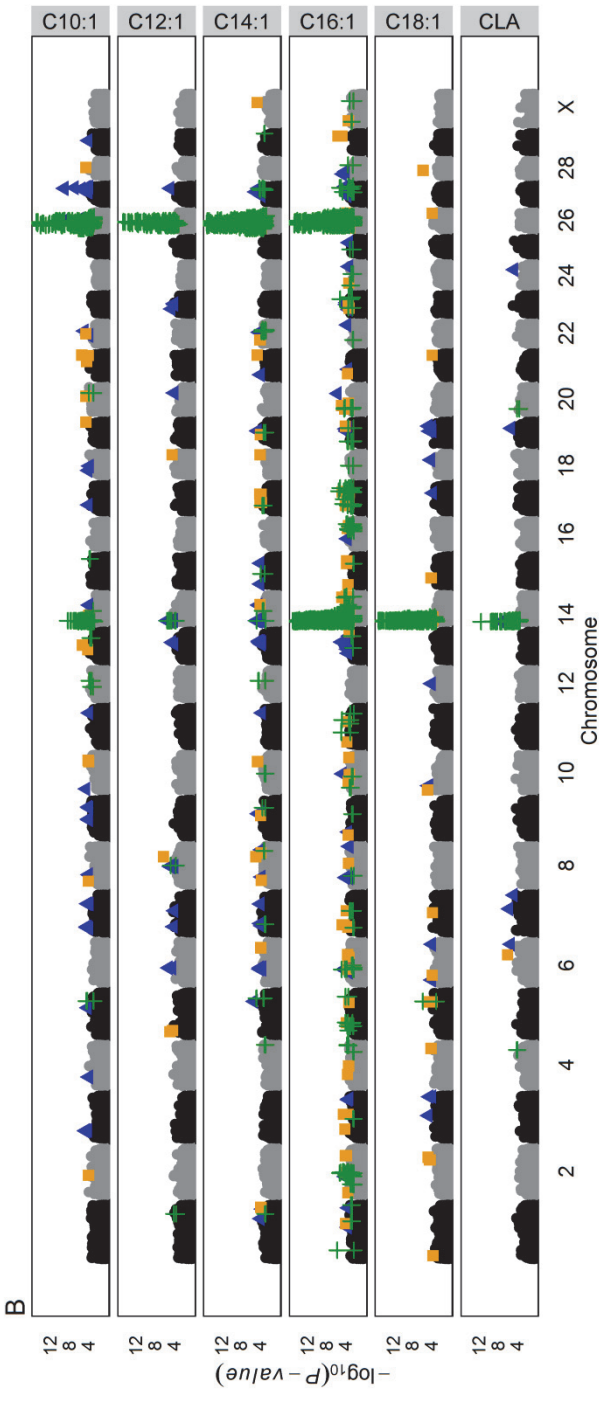


Figure 3.2 Results from winter and summer genome-wide association studies for bovine milk fatty acids combined in Manhattan-plots. Genome-wide plots of -$\log_{10}$ (*P*-values) (y-axis) for association of SNPs with saturated fatty acids (A) and unsaturated fatty acids (B). The genomic position is represented along the x-axis and chromosome numbers are given on the x-axis. The orange squares represent SNPs only significant for the summer milk samples (FDR_{summer} < 0.05, FDR_{winter} ≥ 0.20). Blue triangles represent SNPs only significant for the winter milk samples (FDR_{winter} < 0.05, FDR_{summer} ≥ 0.20). Green addition signs represent SNPs that were detected significant in both milk samples (FDR < 0.20 in both studies). Note that each SNP is represented twice in this figure, once at the -$\log_{10}$(*P*-value) for the summer GWAS and once at the -$\log_{10}$(*P*-value) for the winter GWAS. The y-axis are cut off at -$\log_{10}$(*P*-value) of 15.

Some SNPs located on BTA 0 were also significant. Blasting these SNPs against the UMD 3.0 map showed that they were mainly located in regions that already showed significant effects, such as region 14a, 19b and 26.

Comparison of summer and winter GWAS results

Figure 3.2 shows two -$\log_{10}$(*P*-values) for each SNP, one for the summer (significant SNPs are orange squares in Figure 3.2) and one for the winter (significant SNPs are blue triangles in Figure 3.2) GWAS. The -$\log_{10}$(*P*-values) of SNPs that had a FDR < 0.20 in both the winter and summer GWAS are indicated with green addition signs and show the regions that were found for both samples. Table 3.4 gives an overview of the regions associated with the summer milk fatty acids that were in agreement with the previous study of winter milk fatty acids. Only the regions that showed agreement between the summer and winter GWAS for more than one SNP or more than one trait are reported in table 3.4, resulting in 34 regions.

Three regions with major effects, BTA 14, 19, and 26, were found for both summer and winter milk fatty acids. These regions were highly significant in the GWAS based on winter milk samples and were therefore expected to be found for the summer milk samples too. More interesting are the additional regions that were found in both GWAS studies and especially the eight regions (1, 2a, 3, 5a, 10, 14b, 17c, and 24 (Table 3.4)) that were not reported for the individual studies based on winter or on summer milk samples because their FDR was between 0.05 and 0.20. In some regions agreement between the summer and winter GWAS was based on a single SNP but in other regions agreement was based on multiple SNPs. Also, some regions were associated with multiple fatty acids. We will report here the regions found to be associated with more than two fatty acids in both GWAS studies: regions 5c, 6, 13, 17b, and 27 (Table 3.4).

On BTA 5 the associations in region 5c with C6:0, C8:0, C10:0, C10:1, and C18:1 were in agreement with the winter GWAS (Bouwman et al., 2011). There are no obvious candidate genes located in this region. In both GWAS studies this region was also associated with C14:0, but different SNPs were significant in the two studies (see Figure 3.2). The agreement between summer and winter GWAS for C14:1 seems to be in a separate region, region 5d at 108.8 Mbp.

On BTA 6 the associations with C6:0, C8:0, and C16:1 (Table 3.4) were in agreement with the winter GWAS (Bouwman et al., 2011). This region contains the candidate gene peroxisome proliferator-activated receptor gamma, coactivator 1 alpha (*PPARGC1A*). Our SNP set contained 10 SNPs located in *PPARGC1A*, however, none of these SNPs were significant in winter nor in summer, but SNPs around (73,865 bp before and 797,923 bp after) the gene showed association in both the summer

and winter GWAS. In the winter GWAS the region was also associated with C12:1 and C14:1, but this was not the case in the summer GWAS.

On BTA 13 the associations with C6:0, C8:0, C10:0, and C10:1 (Table 3.4) were in agreement with the winter GWAS (Bouwman et al., 2011). The region on BTA 13 was associated with short chain fatty acids. A possible candidate gene in this region is acyl-CoA synthetase short-chain family member 2, which activates acetate for de novo fatty acid synthesis (Bionaz and Loor, 2008). In the winter GWAS the region was also associated with C14:1 and C16:1, but this was not the case in the summer GWAS.

On BTA 17 the associations in region 17b with C6:0, C8:0, C10:0, and C16:1 (Table 3.4) were in agreement with the winter GWAS (Bouwman et al., 2011). There are no obvious candidate genes located in this region.

On BTA 27 the associations with C12:0, C14:1, and C16:1 (Table 3.4) were in agreement with the winter GWAS (Bouwman et al., 2011). This region contains the candidate gene 1-acylglycerol-3-phosphate O-acyltransferase 6, which is involved in attaching fatty acids on the second position of the triglyceride backbone. In the winter GWAS this region was also associated with C14:0, C16:0, C10:1, and C12:1, but this was not the case in the summer GWAS.

Besides agreement of association for both summer and winter milk fatty acids it is also interesting to see how well the significance levels and effects of SNPs from the summer and winter GWAS correlate. High correlation implies that the effect of the QTL is similar in winter as in summer and, thus, that there is no genotype by season interaction for the regions that were found in both GWAS studies. Winter and summer $-\log_{10}(P\text{-values})$ of SNPs that had a FDR < 0.20 in both GWAS studies are plotted in Figure 3.3A and showed a correlation of 0.89. Winter and summer additive SNP effects of SNPs that had a FDR < 0.20 in both GWAS studies, expressed in phenotypic standard deviation, are plotted in Figure 3.3B and show a correlation of 0.97.

Table 3.4 Regions associated (FDR < 0.20) with milk fatty acids in both the summer and winter GWAS. The number of SNPs in agreement between summer and winter GWAS in the region are given and only regions that had more than one SNP or more than one trait in agreement between summer and winter GWAS are reported here.

Region[1]	Start	End	Trait													
	(Mbp)	(Mbp)	C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA
1	134.0	134.0										1	1			
2a	31.9	32.5												3		
2b	56.3	69.5												19		
2c	139.5	139.5		1	1											
3	97.9	97.9				1		1								
4	121.8	123.7											1	2		
5a	24.4	25.3												2		
5b	33.9	36.1												3		
5c	100.0	101.1		2	2	2					2				1	
5d	108.8	108.8											2			
6	40.0	59.2		11	1									6		
7	64.1	64.1												4		
9	66.2	66.2												2		
10	7.4	7.4												3		
12	52.4	52.5									3		3			
13	46.1	68.4		28	10	3					1					
14a	0.0	18.9		36	15			45	256		42	12	2	375	233	50
14b	32.7	32.7									1		1			

Region[1]	Start	End	Trait													
	(Mbp)	(Mbp)	C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA
14c	40.8	50.9												4		
14d	73.0	76.5												2		
16	46.7	68.0												15		
17a	13.5	21.0											1	6		
17b	31.4	34.7		1	4	5								1		
17c	43.4	43.4		1	1											
17d	56.4	69.8												14		
18	22.8	22.8				2		2								
19a	6.1	7.6												4		
19b	32.9	64.9			20	64	11	103					2	2		
20	8.6	12.3												3		1
22	36.1	42.4											5			
23	42.7	48.5												3		
24	39.2	39.2							2							
26	2.5	40.8				29		6		1	133	61	260	202		
27	30.2	47.6					1						4	7		
Total # SNPs in agreement			0	80	54	106	12	157	258	1	182	74	282	682	234	51

[1] Region number corresponds to the BTA number plus an a, b, c, or d indicating different regions within a BTA.

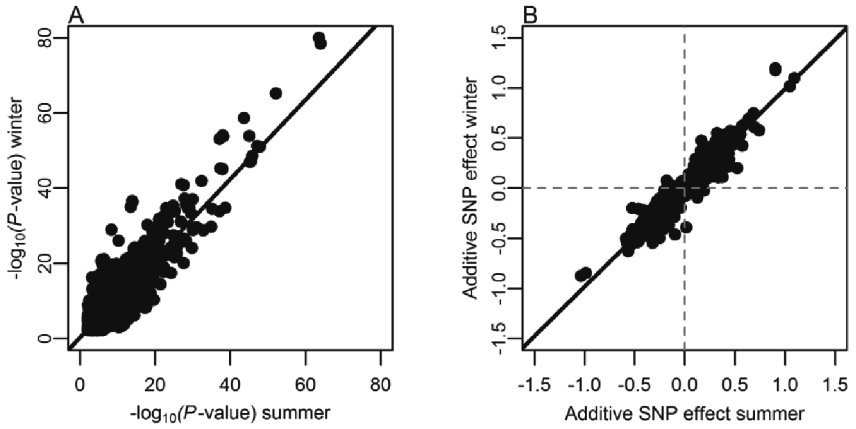


Figure 3.3 Significance level (A) and additive SNP effects expressed in phenotypic standard deviation ($\sigma_p=\sqrt{\sigma_a^2+\sigma_e^2}$) (B) of the SNPs that were found significant in both the summer and winter GWAS (FDR < 0.20).

3.4 Discussion

The aim of this study was to perform a GWAS of bovine fatty acids based on summer milk samples and to compare them to previous results of a GWAS of fatty acids based on winter milk samples (Bouwman et al., 2011). For this GWAS we used different milk samples from largely the same set of cows with the same genotypes at a different stage of the same lactation. The main difference between the two seasons was the herd management, including feeding. In winter, all herds were kept indoors and fed silage, while in summer about half of the herds were grazing outside for at least part of the day and a few other herds were fed fresh grass. Diets of cows including fresh grass are known to alter milk fat composition (e.g. Fievez et al., 2003; Smith et al., 2003). This was reflected in our results: summer milk contained more long chain fatty acids and less C16:0 compared to winter milk. The phenotypic correlations for the different fatty acids between summer and winter samples ranged between 0.36-0.67 (Table 3.2), indicating that the summer samples provide additional information compared to the winter samples. Genetic correlations between summer and winter samples of C4:0, C6:0, C12:0, C18:0, C10:1, C12:1, C14:1, C16:1, and C18:1 (ranging between 0.90-1, Table 3.2) showed that these fatty acids are genetically the same trait in summer and winter (Duchemin et al., 2013). The genetic correlations for C8:0, C10:0, C14:0, C16:0, and CLA were significantly different from one (Duchemin et al., 2013) but showed strong positive correlations (0.77-0.94, Table 3.2), suggesting that also for these

fatty acids summer and winter samples have most genetic variation in common. It is important here to consider that strong positive genetic correlations are required to ensure that the traits have the same genetic background. However, for the summer milk sample to provide additional information to our previous GWAS phenotypic correlations should be weak.

This GWAS of fatty acids based on summer milk samples shows agreement with most associations detected in our previous GWAS of fatty acids based on winter milk samples (Bouwman et al., 2011). Three regions with major effects detected in the winter GWAS (Bouwman et al., 2011) were also found in the summer GWAS. On BTA 14 a dinucleotide polymorphism in *DGAT1* is causing the major effect. The *DGAT1 K232A* polymorphism is known to be associated with fat content and composition, so our results are in line with other studies (Conte et al., 2010; Grisart et al., 2002; Schennink et al., 2007). The rather large region associated with the short and medium chain saturated milk fatty acids on BTA 19 confirm previous linkage studies (Morris et al., 2007; Schennink et al., 2009b; Stoop et al., 2009b). Several candidate genes related to fat synthesis are located in this region, e.g. ATP citrate lyase, sterol regulatory element-binding transcription factor 1, signal transducer and activator of transcription 5A, growth hormone, and fatty acid synthase. There might be more than one QTL in this region given the size of this region and the many candidate genes, but the actual polymorphism(s) causing the effect(s) has not yet been identified. On BTA 26 a polymorphism in *SCD1* is causing the major effect. The gene *SCD1* is known to be associated with medium-chain unsaturated fatty acids, so our results are in line with other studies (Conte et al., 2010; Kgwatalala et al., 2009; Mele et al., 2007; Moioli et al., 2007; Schennink et al., 2008).

Our results from both GWAS studies also suggest that there are additional QTL on BTA 14 besides *DGAT1* that were associated with fatty acids. These additional QTL on BTA 14 were located at 3.0-3.8 Mbp, 32.7 Mbp, 40.8-50.9 Mbp, and 73.0-76.5 Mbp, and confirm detected QTL for milk production traits in linkage analyses (reviewed in Wibowo et al., 2008). Candidate genes for these regions might be corticotropin releasing hormone at 30.5 Mbp, fatty acid binding protein 5 at 41.9 Mbp, and fatty acid binding protein 4 at 42.0 Mbp.

The three highly significant regions with major effects mentioned above were expected to be found in both GWAS studies. More interesting are the additional regions that were found in both GWAS studies such as the regions associated with more than two fatty acids: region 5c, 6, 13, 17b, and 27. Also worthwhile to mention are the 'new' regions that had a suggestive FDR between 5-20% and were not considered in the individual studies based on winter or on summer milk

samples only, but became of interest because they were found in both studies. There are eight 'new' regions like this: region 1, 2a, 3, 5a, 10, 14b, 17c, and 24 (Table 3.4).

There are two different confirmation strategies regarding GWAS: replication and validation. Replication studies are meant to confirm that the actual association is a true association and should therefore be based on samples from the same population with minimal systematic differences (Igl et al., 2009; König, 2011). Validation studies are meant to see if the association can be generalized over different populations and should therefore be based on samples from a different population, where this population can be different concerning genetic background, phenotype definition, sampling strategy, and time point of investigation (Igl et al., 2009; König, 2011). A correctly performed replication is more likely to be successful in finding the same association again than validation, however, when an association is validated the associated SNP is probably closer to the actual polymorphism. In literature, replication and validation are often used interchangeably which complicates the interpretation of the results, especially when a study is called replication study but population, phenotype or study design are too different from the original study to be a replication study. Our GWAS has elements of a replication as well as of a validation study; it met criteria for a replication such as sufficient sample size, phenotypes were measured using the same method, same set of markers and a very similar population was used. It also met criteria for validation because the phenotypes were measured at different time points. The difference in season of measuring the phenotype provides additional information. Ideally an independent population of cows should have been sampled, but this was practically not feasible.

Replication of the study with largely the same set of animals and the same genotypes led to minor differences in LD and allele frequencies between studies, therefore it is more likely to confirm previously detected results (Liu et al., 2008b). However, spurious associations due to population structure or genotyping errors are more likely to be detected twice using the same set of animals and genotypes.

Agreement between the two GWAS studies was based on a FDR threshold of 20% in each study. If the winter and summer GWAS would be independent, the FDR of a region found in both studies would be 4% (20%*20%) (Liu et al., 2008b). A FDR of 4% gives enough reason to investigate such a region further. Lowering the threshold from 5% to 20% FDR resulted in the eight 'new' regions mentioned above, besides the regions that were already discovered in one of the individual studies and were in agreement with the other.

Even though we used largely the same set of animals, same genotypes, same phenotype measurement and a lower threshold for agreement between summer and winter GWAS not all regions were found in both summer and winter GWAS. This can be due to genotype by season interaction, due to lack of power (false negative QTL) or because these QTL were false positive QTL. It is not possible to determine which of these three reasons apply. However, the genetic correlations indicated that fatty acids are genetically similar traits in summer and winter, which suggests that genotype by season interaction may have only a small effect on the results. So lack of agreement between summer and winter GWAS is either due to lack of power or false positives.

3.5 Conclusions

This GWAS of fatty acids based on summer milk samples is in agreement with most associations that were previously detected in a GWAS of fatty acids based on winter milk samples. Lowering the FDR threshold from 5% in individual studies to 20% for agreement between both the summer and winter GWAS led to eight 'new' regions that were not considered in the individual studies, but had a suggestive FDR between 5-20% in both studies. It is more likely that genomic regions are involved in fatty acid synthesis when associations are found in both summer and winter GWAS compared to regions detected in only one GWAS. Detected associations that were in agreement between summer and winter GWAS are therefore worthwhile to pursue in fine-mapping studies.

Acknowledgement

This study is part of the Dutch Milk Genomics Initiative and the project 'Melk op Maat', funded by Wageningen University (the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), the cooperative cattle improvement organization CRV (Arnhem, the Netherlands), the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands).

Additional files

Additional Table 3.1 Most significant SNP per trait for each region significantly associated with fatty acids of the summer milk sample (corresponding to table 3.3), SNP position, significance level and the percentage of total additive genetic variance explained by the SNP.

Reg	Trait	SNP	Chr	Position BTAU4	$-Log_{10}$ (*P*-value)	% of σ_a^2 explained by SNP
1a	C16:1	ULGR_BTA-39422	1	106298619	3.84	3.31
1b	C14:1	ULGR_BTA-58250	1	154751350	3.63	2.52
2a	C16:1	ARS-BFGL-NGS-101408	2	65960161	5.03	4.03
2b	C18:1	ULGR_rs29018764	2	113931092	4.54	5.02
2c	C16:1	ULGR_BTA-93268	2	118914271	3.84	2.98
3	C16:1	ULGR_rs29027883	3	72584745	4.25	3.49
4a	C16:1	ULGR_rs29015971	4	59974296	3.22	2.81
4b	C16:1	ULGR_rs29024031	4	121815498	3.53	3.18
5a	C12:1	ULGR_BTA-74162	5	9144954	4.69	4.02
5b	C16:1	ULGR_rs29024155	5	35966378	3.88	3.37
5c	C6:0	ULGR_AAFC03122217_7089	5	99946095	6.43	5.28
5c	C8:0	ULGR_AAFC03122217_7089	5	99946095	5.36	4.91
5c	C10:0	ULGR_BTA-61859	5	96448996	5.56	3.77
5c	C14:0	ULGR_BTA-61859	5	96448996	4.16	2.78
5c	C10:1	ULGR_rs29016908	5	101090417	4.24	3.22
5c	C14:1	ULGR_BTA-74571	5	101084555	2.85	1.95
5c	C16:1	ULGR_BTA-93285	5	96297427	3.15	2.65
5c	C18:1	ULGR_AAFC03122217_7089	5	99946095	5.60	6.47
6a	C16:1	ULGR_BTC-050897	6	40617870	3.57	2.89
6b	C6:0	ULGR_BTC-038642	6	44772742	5.04	4.03
6c	C16:1	BTA-76959-no-rs	6	84955939	3.39	2.80
6d	C6:0	ULGR_rs29012416	6	85288859	3.78	2.96
6e	C14:1	ULGR_BTA-77644	6	106088917	3.66	2.86
7a	C16:1	BTB-02031452	7	21946038	4.58	3.72
7b	C16:1	ULGR_BTA-28678	7	64163271	3.69	2.96
10a	C10:0	ULGR_BTA-105496	10	9818142	4.21	2.86
10b	C10:0	ARS-BFGL-NGS-28483	10	22024690	5.16	2.99
10b	C12:0	ARS-BFGL-NGS-28483	10	22024690	5.73	3.86
10c	C10:1	ULGR_BTA-15583	10	89892045	3.85	3.13

Reg	Trait	SNP	Chr	Position BTAU4	-Log_{10} (*P*-value)	% of σ_a^2 explained by SNP
11a	C16:1	ULGR_AAFC03072692_59348	11	74307761	4.58	3.61
11b	C4:0	ULGR_SNP_X14710_1740	11	107166278	6.00	5.46
11b	C6:0	ULGR_SNP_X14710_1740	11	107166278	6.50	5.20
13	C6:0	ULGR_rs29027599	13	57042707	6.13	5.49
13	C8:0	ULGR_BTA-33016	13	57022323	4.96	4.64
13	C10:0	ULGR_BTA-33016	13	57022323	5.11	3.21
14a	C6:0	ULGN_SNP_AJ318490_2	0	0	17.66	14.05
14a	C8:0	ULGN_SNP_AJ318490_2	0	0	12.32	10.98
14a	C10:0	ULGR_BTC-067569	14	3618555	6.84	4.24
14a	C14:0	ULGR_SNP_AJ318490_1c	14	445086	13.96	8.09
14a	C16:0	ULGN_SNP_AJ318490_2	0	0	47.81	47.41
14a	C10:1	ULGR_BTC-068221	14	3011407	8.17	7.05
14a	C12:1	ULGR_BTC-067762	14	3190704	5.21	4.55
14a	C16:1	ULGR_SNP_AJ318490_1b	14	445087	38.18	31.09
14a	C18:1	ULGN_SNP_AJ318490_2	0	0	45.82	50.13
14a	CLA	ULGR_SNP_AJ318490_1b	14	445087	8.35	10.54
14b	C16:1	ULGR_AAFC03000860_10938	14	45169820	3.90	3.57
14c	C16:1	ULGR_AAFC03063557_87858	14	74282593	4.59	4.10
15a	C18:1	ULGR_BTA-121008	15	20854340	3.91	3.61
15b	C16:1	ULGR_BTA-37283	15	64123483	3.61	3.00
16a	C14:0	ULGR_rs29019632	16	3732521	4.21	2.16
16b	C16:1	ULGR_BTA-40002	16	68010217	3.63	3.03
17a	C6:0	ULGR_BTA-19253	17	15030516	4.25	3.34
17a	C8:0	ULGR_BTA-19275	17	15039540	4.36	3.90
17a	C16:1	ULGR_BTA-40634	17	21040244	4.40	3.65
17b	C14:1	ULGR_BTA-88832	17	24727021	3.76	3.00
17c	C6:0	ULGR_BTA-40805	17	31496257	5.47	5.32
17c	C8:0	ULGR_BTA-40805	17	31496257	7.61	8.18
17c	C10:0	ULGR_BTA-40805	17	31496257	5.67	4.12
17d	C14:1	ULGR_BTA-41023	17	50852615	3.84	3.26
17d	C16:1	ULGR_BTA-41264	17	58681668	5.26	4.35
17e	C10:0	ULGR_rs41255340	17	73990637	3.81	2.21
19a	C16:1	ULGR_BTA-44680	19	6087775	3.43	2.86

Reg	Trait	SNP	Chr	Position BTAU4	-Log_{10} (*P*-value)	% of σ_a^2 explained by SNP
19b	C8:0	ULGR_rs41257373	19	58223746	6.40	6.03
19b	C10:0	ARS-BFGL-NGS-24479	19	45901284	8.68	5.65
19b	C12:0	ARS-BFGL-NGS-24479	19	45901284	8.23	6.29
19b	C14:0	ULGR_BTA-45758	19	52099860	16.69	12.30
19b	C16:0	ARS-BFGL-NGS-31468	19	46499482	4.19	4.28
20	C6:0	ULGR_BTA-50053	20	24052804	3.72	2.81
20	C18:0	BTB-00771394	20	9181457	6.13	15.59
20	C16:1	ULGR_AAFC03097520_1842	20	19545988	4.65	4.03
21	C10:1	ULGR_BTA-53024	21	65176950	5.19	4.24
22a	C14:1	ULGR_BTA-55267	22	11843506	3.84	2.94
22b	C16:0	ULGR_BTA-114990	22	16202758	4.10	3.68
23a	C16:1	ULGR_BTA-55534	23	16383738	3.27	2.73
23b	C6:0	ULGR_BTA-56106	23	27096170	5.16	4.34
23c	C16:1	ULGR_AAFC03029112_9488	23	42701448	5.00	4.13
26	C10:0	ULGR_rs41255702	0	0	7.62	4.53
26	C10:1	ULGR_rs41255702	0	0	27.58	20.65
26	C12:1	ULGR_rs41255702	0	0	18.63	14.98
26	C14:1	ULGR_rs41255702	0	0	63.99	46.35
26	C16:1	ULGR_SNP_SCD	0	0	38.66	30.74
27	C16:1	BTB-00968596	27	47560218	3.28	2.63
28	C8:0	ULGR_BTA-107346	28	3124529	5.09	4.95
28	C10:0	ULGR_BTA-107346	28	3124529	4.72	3.09
29a	C10:0	ULGR_BTA-22806	29	32659799	4.02	2.63
29b	C16:1	ULGR_BTA-65824	29	44267475	5.42	3.55
X	C14:1	ULGR_BTA-30449	30	63061338	4.40	3.89

4

Fine mapping of a quantitative trait locus for bovine milk fat composition on *Bos taurus* autosome 19

Aniek C. Bouwman, Marleen H.P.W. Visker, Johan A.M. van Arendonk, Henk Bovenhuis

Animal Breeding and Genomics Centre, Wageningen University, P.O. Box 338, 6700 AH Wageningen, the Netherlands

Journal of Dairy Science (2014) 97: 1139-1149

Abstract

A major quantitative trait locus (QTL) for milk fat content and fatty acids in both milk and adipose tissue has been detected on *Bos taurus* autosome 19 (BTA19) in several cattle breeds. The objective of this study was to refine the location of the QTL on BTA19 for bovine milk fat composition using a denser set of markers. Opportunities for fine mapping were provided by imputation from 50,000 genotyped single nucleotide polymorphisms (SNP) toward a high-density SNP panel with up to 777,000 SNP. The QTL region was narrowed down to a linkage disequilibrium block formed by 22 SNP covering 85,007 bp, from 51,303,322 bp to 51,388,329 bp on BTA19. This linkage disequilibrium block contained 2 genes: coiled-coil domain containing 57 (*CCDC57*) and fatty acid synthase (*FASN*). The gene *CCDC57* is minimally characterized and has not been associated with bovine milk fat previously, but is expressed in the mammary gland. The gene *FASN* has been associated with bovine milk fat and fat in adipose tissue before. This gene is a likely candidate for the QTL on BTA19 because of its involvement in de novo fat synthesis. Future studies using sequence data of both *CCDC57* and *FASN*, and eventually functional studies, will have to be pursued to assign the causal variant(s).

Key words: *Bos taurus* autosome 19, fine mapping, milk fatty acid, quantitative trait loci

4.1 Introduction

Many linkage and genome-wide association studies (GWAS) have been performed to identify QTL in cattle. These studies have detected numerous chromosomal regions affecting traits of interest (e.g., http://www.animalgenome.org/cgi-bin/QTLdb/index, Hu et al., 2010; Khatkar et al., 2004). Typically, the location of the QTL is estimated inaccurately and the confidence interval contains several candidate genes. To fine-map QTL, researchers have genotyped additional markers on the same animals, genotyped additional animals, and sometimes sequenced candidate genes (e.g., Blott et al., 2003; Cohen-Zinder et al., 2005; Druet et al., 2008; Gautier et al., 2006; Grisart et al., 2002; Karim et al., 2011; Kim et al., 2009; Meuwissen et al., 2002). Currently, opportunities for fine-mapping QTL in cattle are provided by the availability of high density SNP panels with up to 777,000 SNP. Genotyping at higher density, or imputation of genotypes to higher densities, increases the power of GWAS, gives a more detailed view of associated regions, and increases the chance of one of the SNP being in strong linkage disequilibrium (LD) with the causal variant of the QTL (Marchini and Howie, 2010; Marchini et al., 2007; Spencer et al., 2009).

For milk fatty acids, a few genome-wide linkage studies and GWAS have been performed. Those studies showed that 3 regions exist with major effects on milk fatty acids, located on BTA14, BTA19 and BTA26 (Bouwman et al., 2011, 2012; Schennink et al., 2009b; Stoop et al., 2009b). The region on BTA14 has been studied extensively and a dinucleotide polymorphism in diacylglycerol-O-acyltransferase 1 (*DGAT1*) has been suggested as the causal variant (Grisart et al., 2002; Schennink et al., 2007). For the region on BTA26, a polymorphism in stearoyl-CoA desaturase 1 (*SCD1*) has been suggested as causal variant (Schennink et al., 2008; Taniguchi et al., 2004). For the region on BTA19, the causal variant has not been identified.

The QTL on BTA19 shows association with several fatty acids in both milk and adipose tissue (Bouwman et al., 2011, 2012; Ishii et al., 2013; Morris et al., 2007). Fatty Acid Synthase (*FASN*) has been suggested as a candidate gene and several SNP in *FASN* are significantly associated with fatty acids in both beef and dairy cattle (Abe et al., 2009; Morris et al., 2007; Oh et al., 2012; Ordovás et al., 2008; Roy et al., 2006; Schennink et al., 2009a; Zhang et al., 2008), but the QTL region is rather large and shows much higher significance levels than those observed in the candidate gene studies (Bouwman et al., 2011). Therefore, the objective of this study was to refine the location of the QTL on BTA19 for bovine milk fat composition previously reported by Bouwman et al. (2011), using a denser set of markers.

4.2 Materials and methods

Population

Detailed fat composition was measured in milk samples from 1,905 first lactation Dutch Holstein Friesian cows, which were housed on 398 commercial farms throughout the Netherlands. At least 3 cows were sampled per herd. Milk samples were taken in winter (February to March 2005), when Dutch cows are mainly kept indoors and fed silage. The cows were between 63 and 282 DIM at the day of sampling. About half of the sampled cows were descendants from 5 proven sires (101-200 daughters per sire); the other half of the sampled cows descended from 50 test sires or 45 other sires (1-30 daughters per sire). The pedigree of the cows was provided by the Cooperative Cattle Improvement Organization (CRV, Arnhem, the Netherlands) and consisted of 26,300 animals.

Phenotypes

Milk fatty acids were measured by gas chromatography and were expressed in terms of weight-proportion of total milk fat weight (w/w%). More details about the phenotypes can be found in Stoop et al. (2008).

This study focuses on C14:0, because in previous studies, this milk fatty acid showed the strongest association with the region on BTA19 (Bouwman et al., 2011; Morris et al., 2007; Stoop et al., 2009b). Milk fat of the 1,905 cows contained, on average, 11.61% C14:0, the phenotypic standard deviation of C14:0 was 0.78, the heritability of C14:0 was 0.62, and the QTL on BTA19 explained approximately 13.8% of the genetic variation in C14:0 (Bouwman et al., 2011).

Genotypes

Initially, 1,810 cows and 55 of their sires (all proven and test sires) were genotyped with a custom 50,000 (50K)-SNP array (Illumina Inc., San Diego, CA) designed by CRV. The 55 sires were regenotyped using the BovineHD BeadChip (Illumina Inc.) with 777,000 (777K) SNP. The high-density (HD) genotypes of these 55 sires were combined with HD genotypes of other Dutch Holstein Friesians available at CRV to form a reference population for imputation of, in total, 1,333 HD genotyped Dutch Holstein Friesian animals.

Animals with pedigree inconsistencies (71 cows) were removed before imputation. Pedigree inconsistencies were assumed when more than 0.5% of the 50K SNP for which both sire and daughter were homozygous, were homozygous for the opposite allele. The software BEAGLE 3.3 (Browning and Browning, 2009) was used to phase and impute missing genotypes for the HD reference animals. These phased genotypes were then used to impute the 50K genotypes of the cows to HD

genotypes. The assumed map positions of the SNP were based on the bovine genome assembly UMD 3.1 (Zimin et al., 2009).

In the 50K data, 1,454 SNP were located on BTA19, with an average distance of 44,888 bp. The number of SNP on BTA19 increased to 18,893, using the HD imputed data, with an average distance between SNP of 3,386 bp. We found 998 SNP overlapping between the 50K SNP panel and the HD SNP panel. A total of 1,572 monomorphic SNP in the HD-imputed data were excluded, and in addition 2,659 SNP were excluded because they had a low genotype frequency (i.e., 1-9 individuals within 1 of the 3 genotype classes), resulting in 14,662 SNP used in this study.

Single SNP association analysis

For 1,640 cows with both C14:0 phenotypes and HD-imputed genotypes, a single SNP analysis was performed for BTA19 using the following mixed model in ASReml software (VSN International Ltd., Hemel Hempstead, UK):

$$y_{ijklmno} = \mu + b_1 \times \dim_i + b_2 \times e^{-0.05 \times \dim_i} + b_3 \times afc_j + b_4 \times afc_j^2 + season_k + scode_l + herd_m + genotype_n + animal_o + e_{ijklmno} \quad (1)$$

where $y_{ijklmno}$ was the phenotype; μ was the overall mean; b_1 to b_4 were regression coefficients of corresponding covariates; $\dim_i$ was the covariate describing the effect of DIM; afc_j was the covariate describing the effect of age at first calving; $season_k$ was the class variable accounting for calving season (June-August 2004, September-November 2004, or December 2004-January 2005); $scode_l$ was the class variable accounting for differences in genetic level between proven-sire daughters and test-sire daughters; $herd_m$ was the random effect of herd, distributed as $N(0, \mathbf{I}\sigma^2_{herd})$, with identity matrix **I** and herd variance σ^2_{herd}; $genotype_n$ was the class variable accounting for the genotype of the SNP; $animal_o$ was the random additive genetic effect, distributed as $N(0, \mathbf{A}\sigma^2_a)$, with additive genetic relationship matrix **A** based on the full pedigree and additive genetic variance σ^2_a; and $e_{ijklmno}$ was the random residual, distributed as $N(0, \mathbf{I}\sigma^2_e)$, with identity matrix **I** and residual variance σ^2_e. To speed up the single SNP association analysis of all 14,662 SNP on BTA19, the genetic and herd variances were fixed to the variances estimated using model 1 without the genotype effect.

A significance threshold was calculated using a Bonferroni-like correction for multiple testing according to the method of Šidák (1967): $1-(1-\alpha)^{1/n}$, where n is the number of SNP (here, 14,662) and α is the significance level (here, 0.1%). This rather conservative threshold was chosen because the main interest of this study

was reducing false positives, whereas missing false-negative associations was of less concern because we have previously shown highly significant evidence for the presence of a QTL on BTA19 (Bouwman et al., 2011, 2012), which has also been confirmed in other cattle populations (Ishii et al., 2013; Morris et al., 2007).
The genetic variance explained by a SNP was calculated from the estimated genotype effects and the observed genotype frequencies. The result was expressed as a percentage of the total additive genetic variance obtained from model 1 without the genotype effect.
The R package biomaRt (Durinck and Huber, 2012) was used on all 14,662 SNP to determine if SNP were located in genes and, if so, in which genes they were located. For the SNP mentioned in the tables, the Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to compare the SNP sequences with the bovine genome sequence to confirm the location of the SNP in the genes and to infer the functional consequences of the SNP.

Haplotype analysis

The imputation in BEAGLE resulted in phased genotypes for all individuals, which provided the opportunity to study haplotypes. Haplotypes better characterize a chromosomal segment than any single SNP. This is relevant when the causal variant is not one of the typed SNP or in full LD with any of the typed SNP. In such a case, a haplotype might capture more LD with the causal variant and, consequently, provide a better estimate of the effect of the causal variant than any of the typed SNP. When the haplotype characterizes the chromosomal segment that contains the causal variant and the causal variant is not one of the typed SNP or in full LD with any of the typed SNP, the effect of the haplotype should be larger than the effect of any of the typed SNP. Two different types of haplotype analysis were performed and are described in more detail in the following section: (1) sliding window of 2 consecutive SNP and (2) LD block analysis.

Sliding window

For the 1,640 cows (with both genotypes and phenotypes), phased genotypes of 2 neighboring SNP were used to create 2-SNP haplotypes, and the 2 haplotypes per cow were combined into a genotype. The genotypes of the 2-SNP haplotypes were included as a class variable in model 1 instead of the single SNP genotype. This association analysis was used to screen the whole chromosome 19 with a sliding window of 2 consecutive SNP, shifting 1 SNP at a time.

LD block

Haploview 4.2 software (Barrett et al., 2005) was used to define LD blocks for the region between 51.2 and 51.5 Mbp on BTA19 that contained the most significant SNP from both the single SNP and the sliding-window analyses. The LD blocks were defined using all 84 SNP in the region and all 1,640 cows (with both phenotypes and genotypes). The LD blocks were based on the D' measure of LD (default Haploview options).

Six LD blocks were defined in the region, of which LD block 3 contained the most significantly associated SNP in both the single SNP and the sliding-window analyses. Therefore, LD block 3 was studied in more detail. Nine haplotypes within LD block 3 had reasonable frequencies (>1%) and were tested for association with C14:0. For 1 of the 9 haplotypes at a time, the SNP genotype in model 1 was replaced by the number of copies of the haplotype tested (0, 1 or 2) as a covariable. This association analysis was repeated for each of the 9 haplotypes.

Correction for most significant SNP and 2-SNP window

To test whether multiple QTL were located on BTA19, the single SNP association analysis was repeated for C14:0 phenotypes (**y**) corrected for the genotype effects of (a) the most significant SNP of the single SNP analysis and (b) the most significant 2-SNP window of the sliding-window analysis. In both cases, BTA19 was screened using a single SNP analysis as described above, using model 1 on the precorrected phenotypes ($\mathbf{y}^*$), where $\mathbf{y}^* = \mathbf{y} - \mathbf{X}_{\mathbf{geno}}\mathbf{b}$, with $\mathbf{X}_{\mathbf{geno}}$ representing an incidence matrix for the genotype of (a) the most significant SNP of the single SNP analysis or (b) the most significant 2-SNP window of the sliding-window analysis and **b** representing a vector with the estimated effect of each genotype.

4.3 Results

Single SNP in *CCDC57* most significantly associated with C14:0

Figure 4.1 shows the significance of association between C14:0 and 14,662 SNP on BTA19, resulting from a chromosome-wide single SNP association analysis. Imputed SNP showed a stronger signal than empirically genotyped SNP, but the most significant empirically genotyped SNP was located in the same region as the most significant imputed SNP (Figure 4.1). Based on the Bonferroni threshold with α = 0.1% ($-\log_{10}$(*P*-value) = 7.17), there were 284 significant SNP, which were located between 40,228,233 and 58,072,578 bp. Within this part of the chromosome, 3 regions were most pronounced: around 42.4 Mbp, around 51.3 Mbp, and around 54.1 Mbp. Focus for the next section will be on the region around 51.3 Mbp because this region contained the most significant SNP.

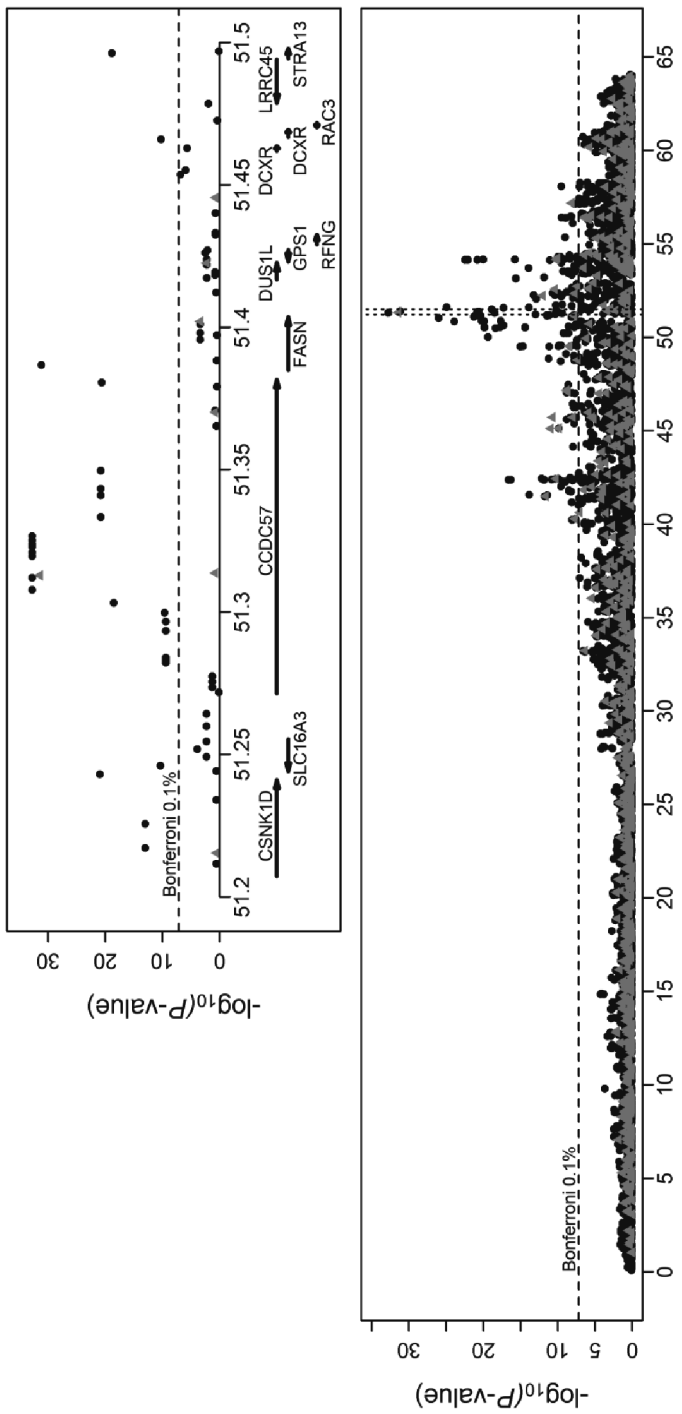


Figure 4.1 Manhattan plot of the single SNP association analysis of C14:0 on BTA19, with a zoom view of the region between 51.2 and 51.5 Mbp. Gray triangles represent genotyped SNP, whereas black dots represent imputed SNP. All genes located in the region between 51.2 and 51.5 Mbp are represented by the arrows in the zoom view. The dotted vertical lines indicate the zoom view area.

Table 4.1 Details of the 10 most significant SNP from a single SNP association analysis of C14:0 on BTA19

SNP Name[1]	Position (bp)	MAF[2]	-Log_{10} (*P*-value)[3]	Allele substitution effect[4]	Gene[5]	Functional consequence
14348	51,307,827	0.45	32.8	0.47	*CCDC57*	Intron
14349	51,312,107	0.45	32.8	0.47	*CCDC57*	Intron
14354	51,319,695	0.45	32.8	0.47	*CCDC57*	Intron
14355	51,320,976	0.45	32.8	0.47	*CCDC57*	Intron
14356	51,322,876	0.45	32.8	0.47	*CCDC57*	Intron
14357	51,323,849	0.45	32.8	0.47	*CCDC57*	Intron
14358	51,325,153	0.45	32.8	0.47	*CCDC57*	Intron
39328[6]	51,326,752	0.45	32.8	0.47	*CCDC57*	Intron
14350	51,312,889	0.44	31.4	0.47	*CCDC57*	Intron
14372	51,386,738	0.34	31.2	-0.42	*FASN*	Intron

[1] The SNP names begin with BovineHD19000 before the 5-digit number given, unless otherwise indicated.
[2] Minor allele frequency.
[3] The $-\log_{10}$ of the *P*-value from the association of the SNP with C14:0.
[4] The allele substitution effect calculated using model 1 with a regression on the number of copies (0, 1, or 2) of the minor allele. Standard errors ranged between 0.03 and 0.04.
[5] *CCDC57* = coiled-coil domain containing 57; *FASN* = fatty acid synthase.
[6] The SNP name begins with ARS-BFGL-NGS-.

Focusing on the most significant region around 51.3 Mbp shows that 8 intronic SNP were located in the coiled-coil domain containing 57 (*CCDC57*) gene that were in perfect LD (r^2 = 1) with each other and were the most significant SNP ($-\log_{10}$(*P*-value) = 32.8; Figure 4.1 and Table 4.1). These SNP showed an allele substitution effect of 0.47 wt/wt% (Table 4.1) and explained 21.6% of the total additive genetic variation in C14:0. The next most significant SNP (BovineHD1900014350; $-\log_{10}$(*P*-value) = 31.4; Table 4.1) was also located in *CCDC57*, and the r^2 between this SNP and the 8 most significant SNP was 0.98. This SNP was actually empirically genotyped and was the most significant SNP in previous GWAS (Bouwman et al., 2011, 2012). The next most significant SNP (BovineHD1900014372; $-\log_{10}$(*P*-value) = 31.2; Table 4.1) was located in the *FASN* gene, which is adjacent to *CCDC57*, and the r^2 between this SNP and the 8 most significant SNP was 0.40.

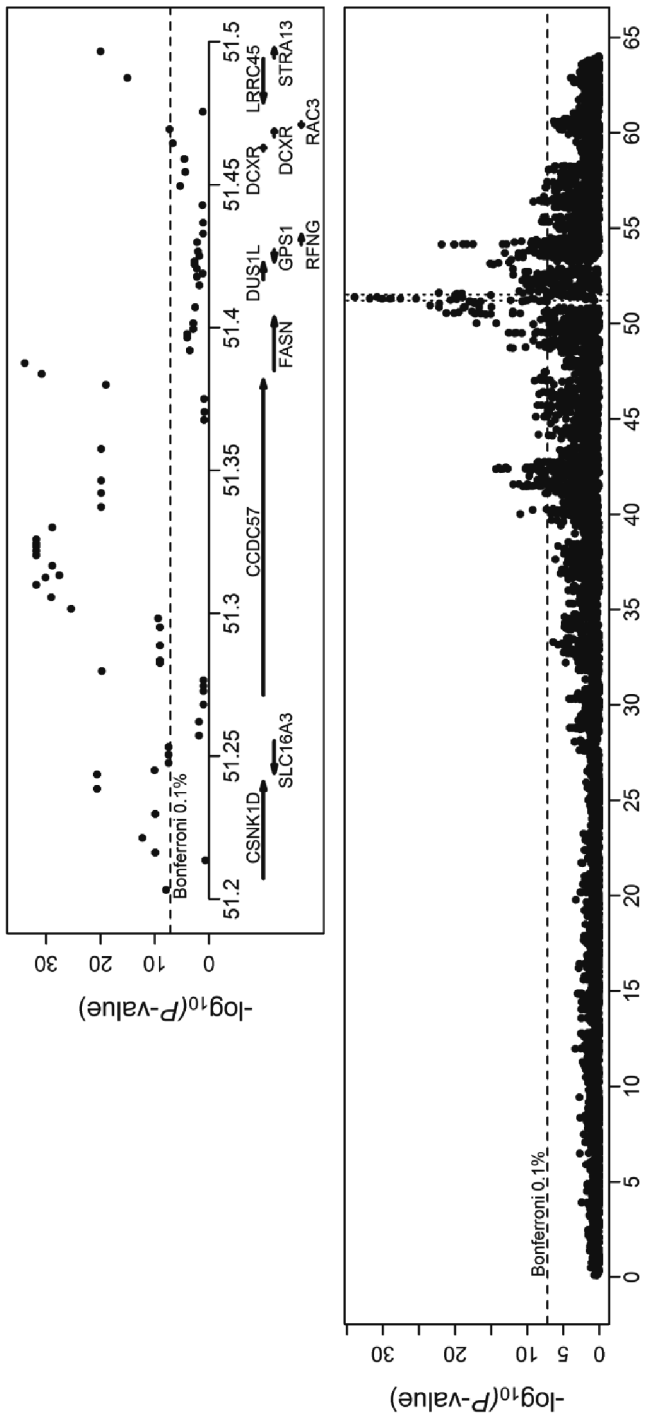


Figure 4.2 Manhattan plot of association analysis of C14:0 on BTA19 using a sliding window of 2 phased SNP, with a zoom view of the region between 51.2 and 51.5 Mbp. The marker positions in this figure were based on the average positions of the 2 SNP in each window. All genes located in the region between 51.2 and 51.5 Mbp are represented by the arrows in the zoom view. The dotted vertical lines indicate the zoom area.

Table 4.2 Details of the 10 most significant 2-SNP windows associated with C14:0 on BTA19

SNP 1		SNP 2			
SNP Name[1]	Position (bp)	SNP Name[1]	Position (bp)	-Log_{10} (*P*-value)[2]	Gene[3]
14372	51,386,738	14373	51,388,329	33.9	*FASN*
14348	51,307,827	14349	51,312,107	31.8	*CCDC57*
14354	51,319,695	14355	51,320,976	31.8	*CCDC57*
14355	51,320,976	14356	51,322,876	31.8	*CCDC57*
14356	51,322,876	14357	51,323,849	31.8	*CCDC57*
14357	51,323,849	14358	51,325,153	31.8	*CCDC57*
14358	51,325,153	39328[4]	51,326,752	31.8	*CCDC57*
14371	51,380,688	14372	51,386,738	30.8	*CCDC57/FASN*
14349	51,312,107	14350	51,312,889	30.1	*CCDC57*
14346	51,303,322	14348	51,307,827	29.0	*CCDC57*

[1] The SNP names begin with BovineHD19000 before the 5-digit number given, unless otherwise indicated.
[2] The $-\log_{10}$ of the *P*-value from the association of the 2-SNP window with C14:0.
[3] *FASN* = fatty acid synthase; *CCDC57* = coiled-coil domain containing 57.
[4] The SNP name begins with ARS-BFGL-NGS-.

2-SNP window in *FASN* most significantly associated with C14:0

Screening BTA19 with a sliding window of 2 consecutive phased SNP, shifting 1 SNP at a time, resulted in the same chromosomal region showing the strongest signal as in the single SNP analysis (Figure 4.2), but focusing on the region around 51.3 Mbp showed some small differences. The most significant window consisted of 2 SNP located in *FASN* (Table 4.2) instead of SNP in *CCDC57*. The other highly significant windows given in Table 4.2 consisted of 2 SNP located in *CCDC57*, except 1 window that consisted of the last SNP in *CCDC57* and the first SNP in *FASN*.

For the most significant window with both SNP located in *FASN* (BovineHD1900014372-BovineHD1900014373), a regression on the number of copies of each haplotype was performed. Haplotype G-G had a frequency of 0.48 and an allele substitution effect of 0.45 (±0.04); haplotype A-G had a frequency of 0.34 and an effect of -0.42 (±0.03); haplotype G-A had a frequency of 0.18 and an effect of 0.05 (±0.04; haplotype A-A was not present in the population studied). The allele substitution effects of the A-G and G-G haplotypes were of similar size compared with the allele substitution effect of BovineHD1900014372 from the single SNP analysis (Table 4.1).

Table 4.3 Haplotypes for linkage disequilibrium (LD) block 3 (51,303,322-51,388,329 bp on BTA19) [1]

	SNP[2]																					
	CCDC57																				*FASN*	
Haplotype	14346	14348	14349	14350	14273[3]	14354	14355	14356	14357	14358	39328[4]	14360	14361	14363	14364	14367	14274[2]	14368	14370	14371	14372	14373
-Log_{10} (*P*-value)[5]	19	33	33	31	1	33	33	33	33	33	33	21	21	21	21	1	1	1	0	21	31	1
HAPLO1	**C**	G	A	G	**G**	A	A	A	C	A	A	A	A	A	A	**A**	**A**	**A**	**G**	A	**G**	**G**
HAPLO2	**C**	**A**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	**A**	**A**	**A**	**G**	**C**	A	**G**
HAPLO3	A	G	A	G	**G**	A	A	A	C	A	A	A	A	A	A	**A**	**A**	**A**	**G**	A	**G**	**G**
HAPLO4	**C**	**A**	**G**	**A**	A	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	G	G	G	A	**C**	**G**	A
HAPLO5	**C**	**A**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	**A**	**A**	**A**	**G**	**C**	**G**	A
HAPLO6	**C**	**A**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	G	**A**	**A**	A	**C**	**G**	A
HAPLO7	**C**	**A**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	A	A	A	A	**A**	**A**	**A**	**G**	A	**G**	**G**
HAPLO8	**C**	**A**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	A	A	A	A	**A**	**A**	**A**	**G**	A	A	**G**
HAPLO9	**C**	**A**	**G**	**A**	**G**	**G**	**G**	**G**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	G	**A**	**A**	A	**C**	A	**G**

[1]Differences in haplotype sequence are indicated in bold for the SNP alleles more common in the different haplotypes.
[2]The SNP names begin with BovineHD19000 before the 5-digit number given, unless otherwise indicated. *CCDC57* = coiled-coil domain containing 57; *FASN* = fatty acid synthase.
[3]The SNP name begins with BovineHD41000.
[4]The SNP name begins with ARS-BFGL-NGS-.
[5]–Log_{10} (*P*-value) for individual SNP as obtained from the single SNP analysis.

LD block contains both *CCDC57* and *FASN*

Haploview analysis of the 84 SNP in the region between 51.2 and 51.5 Mbp on BTA19 revealed 6 LD blocks. Linkage disequilibrium block 3 (51,303,322-51,388,329 bp) contained 22 SNP, including the *CCDC57* SNP that were most significant in the single SNP analysis and the 2 *FASN* SNP that were most significant in the sliding-window analysis. Therefore, LD block 3 was investigated in more detail. Within LD block 3, 9 different haplotypes had a frequency >1% in the population (Table 4.3); HAPLO3 was the most frequent haplotype (0.368), followed by HAPLO2 (0.287; Table 4.4).

Association between C14:0 and the number of copies of the haplotype present per individual were analyzed for each of the 9 haplotypes. The haplotype HAPLO2 was most significantly associated with C14:0 ($-\log_{10}$(*P*-value) = 23.84), followed by HAPLO3 ($-\log_{10}$(*P*-value) = 18.67; Table 4.4). The effects of HAPLO2 and HAPLO3 on C14:0 were similar but in opposite directions: the regression coefficient of HAPLO2 was -0.37 (±0.03), whereas the regression coefficient of HAPLO3 was 0.35 (±0.04; Table 4.4). This agreed with the fact that HAPLO2 and HAPLO3 differed from each other at the SNP in LD block 3 that were significant in the single SNP analysis (Table 4.3). These haplotype allele substitution effects were a little smaller compared with the allele substitution effect of the most significant SNP in the single SNP analysis and the allele substitution effect of the most significant 2-SNP window in the sliding-window analysis. Therefore, the LD block haplotypes did not seem to capture more LD with the causal variant of the QTL compared with the single SNP and sliding-window analyses.

2-SNP window captures most of the QTL variance

Figure 4.3A shows the Manhattan plot of BTA19 after correction for the most significant SNP of the single-SNP association analysis (Bovine1900014354). The original single SNP analysis showed 3 significant regions: around 42.4, 51.3, and 54.1 Mbp. After correction for the most significant SNP located in *CCDC57* (51.3 Mbp), the region around 42.2 Mbp dropped below the significance threshold, whereas the regions around 51.3 and 54.1 Mbp remained significant. The most significant SNP after correction (BovineHD1900018628) was located at 51,612,335 bp in the gene Aly/REF THO complex 4 (*ALYREF*). The LD between this SNP in *ALYREF* and the most significant SNP from the original single SNP analysis was 0.17. The next 5 most significant SNP after correction were located between 54,144,924 and 54,166,828 bp in the gene ubiquitin specific peptidase 36 (*USP36*), of which one (BovineHD1900015164) was a missense mutation, causing an alanine to aspartic acid amino acid change. The LD between the most significant SNP in *USP36*

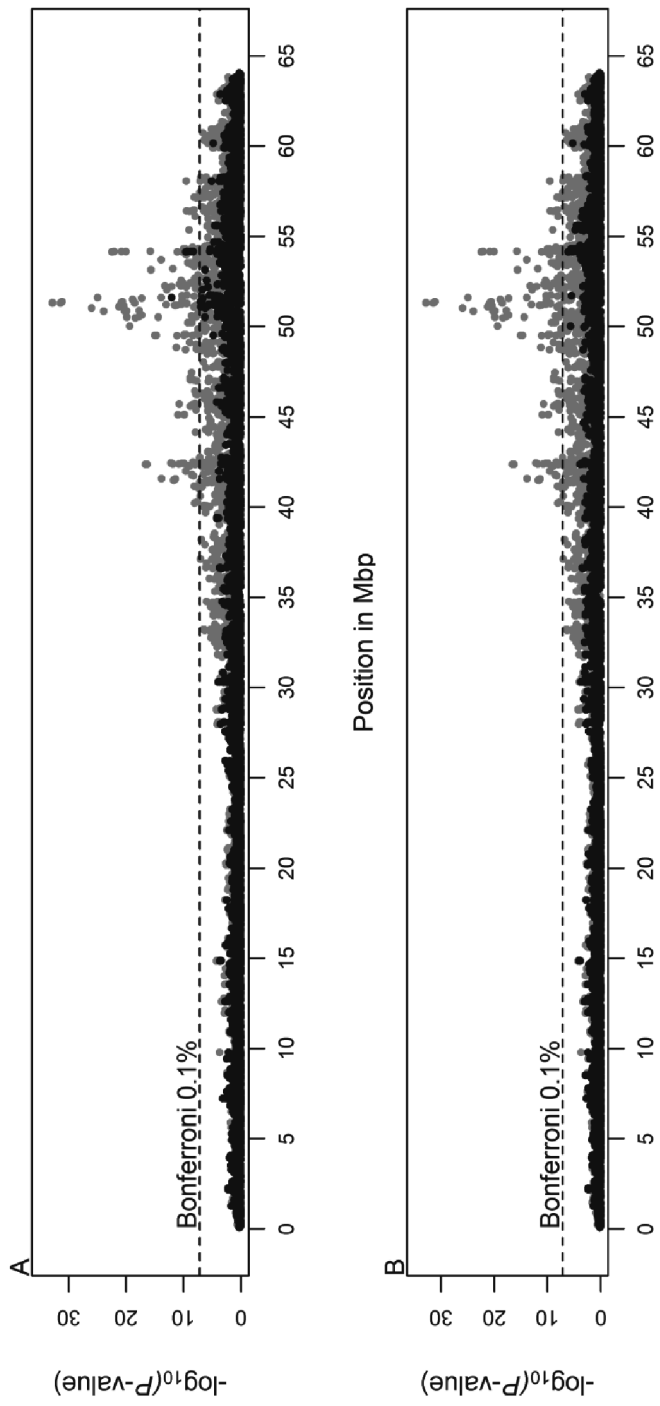


Figure 4.3 Manhattan plot of single SNP association analysis of C14:0 on BTA19 after correction for the effect of the most significant SNP of the single SNP analysis (A) and after correction for the most significant 2-SNP window of the sliding window analysis (B). In gray are the original single SNP analysis results.

Table 4.4 Frequencies, significance levels, and allele substitution effects of haplotypes for linkage disequilibrium block 3 (51,303,322-51,388,329 bp on BTA19) associated with C14:0

Haplotypes	Frequency	-Log_{10} (P-value)	Allele substitution effect	SE
HAPLO1	0.073	5.27	0.28	0.06
HAPLO2	0.287	23.84	-0.37	0.03
HAPLO3	0.368	18.67	0.35	0.04
HAPLO4	0.083	1.18	-0.11	0.06
HAPLO5	0.051	2.14	0.21	0.08
HAPLO6	0.049	1.01	0.12	0.07
HAPLO7	0.035	0.29	-0.06	0.08
HAPLO8	0.023	4.96	-0.45	0.10
HAPLO9	0.019	0.48	-0.11	0.11

and the most significant SNP from the original single SNP analysis was 0.19, whereas the LD between this *USP36* SNP and the SNP in *ALYREF* was 0.43. The significance level of all other SNP significant in the original single SNP analysis dropped below the Bonferroni 0.1% threshold after correction for the most significant SNP.

Figure 4.3B shows the Manhattan plot of BTA19 after correction for the most significant 2-SNP window (Bovine1900014372-Bovine1900014373) of the sliding-window analysis. After correction for the most significant 2-SNP window located in *FASN* (51.3 Mbp), all SNP significant in the original single SNP analysis dropped below the Bonferroni 0.1% threshold. This indicates that the 2-SNP window in *FASN* captured most of the QTL variance.

4.4 Discussion

This study aimed to fine map a QTL for C14:0 content in bovine milk fat on BTA19. The previously identified QTL spanned a rather large region of almost half the chromosome (Bouwman et al., 2011). Using 10 times more SNP in the present study gave a more detailed view of the associated region and showed that the most significant SNP were located in an LD block that contained 2 genes: *CCDC57* and *FASN*.

Fine mapping using imputed SNP data

The QTL on BTA19 was fine mapped using genotypes imputed from a 50K SNP panel to a 777K SNP panel. Fine mapping the QTL on BTA19 using genotypes imputed from the 50K SNP panel to the 777K SNP panel recovered the same region as being most significant compared with a previous 50K genome-wide analysis (Bouwman et al., 2011). However, in the previous study, it could be suggested that SNP in *CCDC57* picked up the effect of *FASN*, but now due to the higher density of SNP, we see that SNP in both genes are significantly associated with C14:0 and LD between SNP in the 2 genes is not high.

High-density SNP panels and sequence data will enhance QTL fine mapping and detection of causal variants. The 50K SNP panel has a useful density to screen the genome for such QTL, whereas the higher-density SNP panels are very useful to fine map the regions detected using the 50K SNP panels. Fine mapping using sequence data could even lead to detection of candidate causal variants because the causal variant should be present as a SNP in the sequence data of a population that is segregating for the QTL. However, functional studies are required to declare a candidate SNP as causal variant.

The numbers of animals with HD genotypes and especially with sequence data are still limited. Imputation of genotypes can, therefore, be a useful tool to increase the density of genotypes. Also, imputation is useful to facilitate combining genotype data generated with different genotyping arrays. In human studies, it has been shown that the power of GWAS increases when imputation is used to increase the number of genotypes (Guan and Stephens, 2008; Marchini and Howie, 2010; Marchini et al., 2007; Spencer et al., 2009). Small changes in imputation accuracy lead to small changes in power of GWAS; also, poor imputation accuracy can still improve power compared with no imputation (Guan and Stephens, 2008).

On the other hand Almeida et al. (2011) showed that the type-I error of association is higher for imputed SNP compared with empirically genotyped SNP; especially SNP with minor allele frequency close to 0.5 were false-positively associated. In the current study, one of the highly significant SNP in the studied LD block was genotyped (BovineHD1900014350), and that SNP was also the most significant SNP for C14:0 on BTA19 in Bouwman et al. (2011, 2012). In addition, a GWAS (50K) in beef cattle showed that 1 of the 8 most significant SNP (ARS-BFGL-NGS-39328) was most significantly associated with C14:0 in adipose tissue (Ishii et al., 2013), indicating that it is unlikely that the detected associations with SNP in the haplotype block were false-positive associations.

Imputation accuracy depends on many factors, such as LD between SNP, allele frequency, number of animals genotyped with high density, and relationships

between reference population and target population. Schrooten et al. (2012) showed that only 0.55 to 0.76% imputation errors were made using BEAGLE to impute individuals genotyped with 50K SNP panels to HD panels using 488 HD genotyped individuals. Because the HD genotyped reference population was much larger in the current study (1,333 individuals) and included the 55 sires of the imputed cows, it can be assumed that imputation errors were even lower.
Imputation accuracy also depends on the quality of the genome build. Erbe et al. (2012) and Pausch et al. (2013) showed that certain regions on the genome have poor imputation accuracy. Erbe et al. (2012) suggested that this poor imputation accuracy could be due to mapping errors. Imputation accuracy improved after remapping problem regions based on LD (Erbe et al., 2012). According to the studies of Erbe et al. (2012) and Pausch et al. (2013) no indication existed of low imputation accuracy or mapping errors on BTA19 in Holstein and Fleckvieh, respectively.

CCDC57

Fine-mapping the QTL for C14:0 content in bovine milk fat on BTA19 gave a more detailed view of the associated region and showed that the most significant SNP were located in an LD block that contained 2 genes: *CCDC57* and *FASN*. The gene *CCDC57* is minimally characterized and has not been associated with bovine milk fat previously; it is transcribed into a coiled-coil domain containing protein. Coiled-coil domains are structural motifs in proteins, with many involved in important biological functions such as DNA binding and regulation of gene expression. The gene *CCDC57* is located next to *FASN*, which is a more pronounced candidate gene because of its known biological relation to fat synthesis. However, the candidacy of *CCDC57* is supported by Medrano et al. (2010), who showed that *CCDC57* was expressed in mammary tissue of a second-lactation cow and that the expression level of *CCDC57* was higher than that of *FASN*.

FASN

In addition to *CCDC57*, the current study also suggested *FASN* as a candidate gene, because, besides SNP in *CCDC57*, a 2-SNP window located in *FASN* also was highly significantly associated with C14:0 content. The gene *FASN* encodes a multi-enzyme system that catalyzes de novo fatty acid synthesis. Even though *FASN* has been studied extensively in candidate gene studies for fat content, milk fatty acids and fatty acids in adipose tissue (Abe et al., 2009; Li et al., 2012; Oh et al., 2012; Roy et al., 2006; Schennink et al., 2009a; Zhang et al., 2008), this has not yet resulted in identification of the causal variant for the QTL on BTA19.

Three SNP on the HD SNP panel located in *FASN* have been associated with milk fatty acids in dairy cattle and with fatty acids in adipose tissue in beef cattle: BovineHD1900014375 (g.13126C>T; position relative to the sequence of *FASN* with accession AF285607), BovineHD1900014377 (g.16907T>C), and BovineHD1900014275 (g.17924A>G; Li et al., 2012; Morris et al., 2007; Oh et al., 2012; Schennink et al., 2009a; Zhang et al., 2008). A fourth SNP on the HD SNP panel located in *FASN*, BovineHD1900014376 (g.13965C>T), was detected by Abe et al., (2009) but not studied in detail. The current study showed that associations with these 4 SNP were relatively weak ($-\log_{10}$(*P*-values) were 0.4, 3.4, 21.3, 3.4, respectively) compared with the considerably stronger association that was found with another *FASN* SNP BovineHD1900014372 ($-\log_{10}$(*P*-value) = 31.2).

Other fat synthesis-related genes

In addition to *CCDC57* and *FASN*, additional candidate genes known to be involved in milk fat synthesis underlie the QTL on BTA19, such as ATP citrate lyase (*ACLY*), sterol regulatory element-binding transcription factor 1 (*SREBF1*), signal transducer and activator of transcription 5A (*STAT5A*), and growth hormone 1 (*GH1*). Of the 8 SNP in *ACLY*, 1 SNP ($-\log_{10}$(*P*-value) = 7.18) just exceeded the Bonferroni threshold and 3 SNP ($-\log_{10}$(*P*-value) = 7.13) were just below that threshold, but they reduced considerably in significance when corrected for the effect of the most significant SNP or 2-SNP window (Figure 4.3). The 2 SNP in our HD SNP panel located in *SREBF1* and 4 SNP located in *STAT5A* did not show any association with C14:0. The only SNP located in *GH1* was below the Bonferroni threshold ($-\log_{10}$(*P*-value) = 6.82), and also reduced considerably in significance when corrected for the effect of the most significant SNP or 2-SNP window (Figure 4.3).

4.5 Conclusions

We reduced the QTL for C14:0 on BTA19 to an LD block formed by 22 SNP covering 85,007 bp (51,303,322-51,388,329 bp). This LD block contained 2 genes: *CCDC57* and *FASN*. The gene *CCDC57* is minimally characterized and has not been associated with bovine milk fat previously, but is expressed in the mammary gland. In addition to SNP located in *CCDC57*, a 2-SNP window located in *FASN* also was highly significantly associated with C14:0 content. The gene *FASN* is involved in de novo fat synthesis and has been studied in candidate gene studies. Future studies using sequence data of both *CCDC57* and *FASN* and, eventually, functional studies will have to be pursued to assign the causal variant(s).

Acknowledgement

Chris Schrooten (Cooperative Cattle Improvement Organization (CRV), Arnhem, the Netherlands) is acknowledged for the imputation of the genotypes. This study is part of the Dutch Milk Genomics Initiative and the project “Melk op Maat”, funded by Wageningen University (Wageningen, the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), CRV, the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands).

5

Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context

Aniek C. Bouwman[1], Bruno D. Valente[2], Luc L.G. Janss[3], Henk Bovenhuis[1], Guilherme J.M. Rosa[2,4]

[1] Animal Breeding and Genomics Centre, Wageningen University, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [2] Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA; [3] Faculty of Science and Technology, Department of Molecular Biology and Genetics, University of Aarhus, DK-8830 Tjele, Denmark; [4] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53706, USA

Genetics Selection Evolution (2014) 46:2

Abstract

Background

Knowledge regarding causal relationships among traits is important to understand complex biological systems. Structural equation models (SEM) can be used to quantify the causal relations between traits, which allow prediction of outcomes to interventions applied to such a network. Such models are fitted conditionally on a causal structure among traits, represented by a directed acyclic graph and an Inductive Causation (IC) algorithm can be used to search for causal structures. The aim of this study was to explore the space of causal structures involving bovine milk fatty acids and to select a network supported by data as the structure of a SEM.

Results

The IC algorithm adapted to mixed models settings was applied to study 14 correlated bovine milk fatty acids, resulting in an undirected network. The undirected pathway from C4:0 to C12:0 resembled the de novo synthesis pathway of short and medium chain saturated fatty acids. By using prior knowledge, directions were assigned to that part of the network and the resulting structure was used to fit a SEM that led to structural coefficients ranging from 0.85 to 1.05. The deviance information criterion indicated that the SEM was more plausible than the multi-trait model.

Conclusions

The IC algorithm output pointed towards causal relations between the studied traits. This changed the focus from marginal associations between traits to direct relationships, thus towards relationships that may result in changes when external interventions are applied. The causal structure can give more insight into underlying mechanisms and the SEM can predict conditional changes due to such interventions.

Key words: dairy, inductive causation, milk fatty acids, structural equation model

5.1 Background

In animal breeding and genetics, relationships between traits are traditionally studied using multi-trait mixed models (Henderson and Quaas, 1976). Such models do not allow for recursive relationships between traits that are generally present in biological systems. Structural equation modelling (SEM) is a statistical technique for testing and estimating such recursive relationships (Haavelmo, 1943; Pearl, 2009; Wright, 1921). Gianola and Sorensen (2004) described SEM in a quantitative genetics context in order to account for possible feedback or recursive relations among traits in multi-trait mixed models settings. In most applications of SEM in animal breeding and genetics, only few hypothesized networks are typically tested and compared, and those that best fit the data are declared as most plausible (de los Campos et al., 2006; de Maturana et al., 2009; Jamrozik et al., 2010; König et al., 2008; Wu et al., 2007). Although such an approach avoids the computational challenges involved in testing every possible network, it does not explore the full space of possible networks. However, data driven exploration of the space is possible using the Inductive Causation (IC) algorithm (Verma and Pearl, 1990).

The IC algorithm is based on conditional independencies tests, such that under multivariate normality, it can be implemented by using partial correlations tests. When all partial correlations between a pair of traits are non-null for each conditioning subset of traits (i.e., they are dependent conditionally on all possible sets of other traits), then a direct causal relation between this pair of traits is declared. When a partial correlation between two traits is null (i.e., they are independent conditionally on at least one set of other traits), then there is no direct causal relation between this pair of traits. Therefore, partial correlations can be explored to study how a set of traits is causally related and this can be qualitatively represented by a graph or network (Pearl, 2009). If the resulting network is completely directed, it can be used as a causal structure of a SEM, and the magnitude of causal relationships among traits (represented by structural coefficients) can be estimated by fitting such a model. Furthermore, visualization of the causal relationships among variables on a graph could help understand and interpret complex biological systems, while their quantification allows prediction of outcomes of external interventions applied to such a causal network.

The inferred structural coefficients associated with connections between traits in a network only carry a causal interpretation under specific causal assumptions. For example, structural coefficients inferred from a SEM with an acyclic causal structure and independent residuals only keep their causal meaning under the assumption that there are no hidden causal effects that have a direct influence on

two or more traits in the network. In livestock, removing such confounding effects can be achieved by performing randomized experiments. However, most livestock data come from non-randomized field studies and are prone to the influence of several sources of systematic variation. When measured, the confounding generated by these systematic sources of variation can be controlled by correcting for them in a model. One example of hidden factors that may affect two or more traits in the network is correlated genetic effects. Thus, the genetic covariances are background sources of phenotypic covariances among traits that confound not only the inference of causal effects between pairs of traits, but also the search for causal structures, because algorithms may interpret such covariances as due to causal relations among phenotypes. Therefore, Valente et al. (2010) proposed to use the inferred residual (co)variance matrix of a standard multi-trait mixed model (which represents the covariance matrix among traits conditionally on the genetic confounders) as input for the IC algorithm, instead of the observed data, when searching for causal structures in mixed effects settings. Valente et al. (2010, 2011) used simulated data to show that applying the IC algorithm to the posterior distribution of the residual (co)variance matrix of a multi-trait mixed model recovered the correct network, and Valente et al. (2011) used the methodology on real data from quails to study causal networks involving five traits.

Here, we applied the same approach to a set of 14 highly correlated milk fatty acids to analyze their causal relations. Fatty acids are important components in human diets with either beneficial or unfavorable effects on human health, depending on the fatty acid. Studying causal relations between bovine fatty acids in milk can provide valuable information about the synthesis of fatty acids, which could be useful for approaches aimed at changing the fatty acid composition of dairy products and ultimately at improving human health. Since a considerable amount of knowledge about the synthesis of fatty acids is available, the network obtained from the adapted IC algorithm can be compared to known biological pathways. However, the network may also reveal new relations that could confirm existing hypotheses or create new ones. The known biological pathways include de novo synthesis, biohydrogenation and desaturation of milk fatty acids. Most of these pathways are reflected in the results of analyses that involve phenotypic and genetic correlations between milk fatty acids (Karijord et al., 1982; Soyeurt et al., 2007; Stoop et al., 2008), clustering techniques (Heck et al., 2012; Massart-Leën and Massart, 1981), or principal component analysis (Fievez et al., 2003). These studies suggest that certain fatty acids have a common origin, but they cannot distinguish between direct and indirect relationships.

Our aim was to explore causal networks between milk fatty acids by applying for the first time the adapted IC algorithm as presented by Valente et al. (2010) to 14 highly correlated traits. In addition, the selected network was used as the causal structure of a SEM to quantify the relationships between the milk fatty acids.

5.2 Methods

Data

Data on the fat composition of winter milk samples from 1,902 first-lactation Dutch Holstein Friesian cows were used. The cows were housed on 397 commercial farms throughout the Netherlands. At least three cows between 63 and 282 days in milk were sampled per farm. The pedigree of the cows was supplied by CRV (Cooperative cattle improvement organization, Arnhem, the Netherlands) and included information from the last four generations (4,676 animals).

Table 5.1 Mean and phenotypic standard deviation[1] for bovine milk fatty acids (in g/kg milk)

Trait	Mean	σ_p [1]
C4:0	1.53	0.26
C6:0	0.97	0.17
C8:0	0.60	0.11
C10:0	1.32	0.28
C12:0	1.79	0.37
C14:0	5.05	0.77
C16:0	14.27	2.84
C18:0	3.80	0.84
C10:1	0.16	0.04
C12:1	0.05	0.01
C14:1	0.59	0.13
C16:1	0.63	0.19
C18:1	7.87	1.20
CLA	0.17	0.04

[1] $\sigma_p=\sqrt{\sigma_a^2+\sigma_e^2}$.

Milk fat composition was measured by gas chromatography (details about the phenotyping are in Stoop et al. (2008)). Fourteen fatty acids with the highest concentration in milk fat were considered: even-chain saturated fatty acids C4:0, C6:0, C8:0, C10:0, C12:0, C14:0, C16:0, C18:0, even-chain (*cis9*) monounsaturated fatty acids C10:1, C12:1, C14:1, C16:1, C18:1, and the polyunsaturated fatty acid CLA (conjugated linoleic acid, C18:2*cis9*,*trans11*). Gas chromatography was

performed on fat samples and provided relative amounts of fatty acids expressed on a fat basis in g/100g fat. However, these relative amounts do not properly represent the biological relationships among fatty acids; therefore the fatty acids were expressed on a milk basis in g/kg milk. Table 5.1 presents the mean and adjusted phenotypic standard deviation for the fatty acids included in this study.

Multi-trait analysis

Genetic and residual (co)variances among traits were estimated by fitting a Bayesian multi-trait mixed model that uses latent variables to fit (co)variance structures and a random walk Metropolis-Hastings algorithm to obtain Markov chain Monte Carlo (MCMC) samples for variance components, similar to the latent variable models to estimate genomic (co)variances in Sørensen et al. (2012). Latent variables were used to fit (co)variance structures because most of the milk fatty acids were strongly correlated, both genetically and residually. Fitting a standard multi-trait model for 14 milk fatty acids resulted in convergence issues, but using latent variables to reduce the dimensionality of the data improved convergence of the Bayesian multi-trait mixed model.

Phenotypes were standardised to traits with a mean of 0 and a standard deviation of 1 to reduce scale differences between the milk fatty acids in the multi-trait mixed model. The following multi-trait model was fitted:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e},$$

with the joint distribution of vectors **u** and **e** as:

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim N\left\{\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{G_0} \otimes \mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{R_0} \otimes \mathbf{I}\end{bmatrix}\right\},$$

where **y** is a vector of phenotypes; **β** is a vector for systematic effects, for each trait the same systematic effects were included: a covariate for days in milk modelled with a Wilmink curve (Wilmink, 1987), a covariate for age at first calving, a covariate for age at first calving squared, a fixed effect for calving season (June-Aug 2004, Sept-Nov 2004, or Dec 2004-Jan 2005), a fixed effect for sire code (accounting for differences in the genetic level between proven sire daughters and test-sire daughters), and a fixed effect for herd; **X** is a known incidence matrix of **β** on **y**; **u** is a vector of random additive genetic effects; **Z** is a known incidence matrix of **u** on **y**; and **e** is a vector of random residuals. $\mathbf{G_0}$ is the additive genetic (co)variance matrix; **A** is the additive genetic relationship matrix; $\mathbf{R_0}$ is the residual (co)variance matrix; **I** is an identity matrix. The (co)variances between genetic

effects and between residuals were modelled employing *k* latent vectors $\mathbf{v}_k$ to model residual (co)variances, and *k* latent vectors $\mathbf{w}_k$ to model genetic (co)variances, such that $\mathbf{e}_i \sim N(\sum_k r_{k,i} v_k , \tau_{e_i}^2 \mathbf{I})$ and $\mathbf{u}_i \sim N(\sum_k s_{k,i} w_k , \tau_{u_i}^2 \mathbf{A})$, with $\mathbf{v}_k \sim N(0, \mathbf{I})$ and $\mathbf{w}_k \sim N(0, \mathbf{A})$ as standard Normal latent vectors, $r_{k,i}$ and $s_{k,i}$ as regressions or "loadings" on the latent vectors with uniform priors [-∞, ∞], and $\tau_{e_i}^2$ and $\tau_{u_i}^2$ as the independent remaining variances for residuals and genetic effects per trait *i*. From the latent variable model, the residual variance for trait *i* is $\sum_k r_{k,i}^2 + \tau_{e_i}^2$, and the residual covariance between traits *i* and *j* is $\sum_k r_{k,i} r_{k,j}$. In a similar manner, the variances ($\sum_k s_{k,i}^2 + \tau_{u_i}^2$) and covariances ($\sum_k s_{k,i}\, s_{k,j}$) were obtained for the additive genetic effects.

In order to maintain mixing in the MCMC sampling algorithm, the remaining independent variances $\tau_{e_i}^2$ and $\tau_{u_i}^2$ must remain well above 0. Initially, this large set of highly correlated traits resulted in residual and polygenic variances $\tau_{e_i}^2$ and $\tau_{u_i}^2$ that were close to 0, thus it was necessary to set a minimum value for them and uniform priors [0.02, ∞] were used on these parameters to achieve this. Because standardised traits were used, these bounds imply that at least 2% of the residual variance for each trait was not explained by residual covariances with other traits, and likewise at least 2% of the genetic variance for each trait was not explained by genetic covariances with other traits. All fixed and random effects (including latent variables) and the regression loadings were conditionally normal, and conditional distributions for variance parameters were scaled inverse Chi square in the MCMC implementation.

The dimension of latent variables *k* is to be pre-set but good indications for this dimension can be obtained by a principal component analysis on the traits analysed, which gives information on the number of latent variables suitable to model the joint (co)variance structure. In order to limit the constraints on the covariance structure, the number of principal components was chosen such that together they explained 90% of the variance. Principal component analysis of the 14 fatty acids showed that the first four principal components explained ~90% of the variance; therefore four latent factors were chosen.

The MCMC software Bayz 2.1 (Janss, 2010) was used for parameter inference. Eight chains of 1 million iterations each were run, with a burn-in of 100,000 for each chain, and a thinning of 1,000 iterations. Convergence was checked by visual inspection of the sample trace plots, of posterior density plots and by determining effective sample size using the Coda package in R (Plummer et al., 2006).

Inductive causation (IC) algorithm

By fitting the multi-trait mixed model described above, the data can be corrected for systematic effects and for genetic (co)variances and thus, inferences regarding the joint distribution of the traits conditionally on genetic and systematic effects can be made. This is important to search for the causal structure using the IC algorithm, because correlated genetic effects are confounding factors, since they are sources of phenotypic correlation due to the genetic background but not due to recursive relations among traits (Valente et al., 2010). The relevant information to be used in a causal structure search is in the residual (co)variance matrix that results from a multi-trait mixed model. Therefore Valente et al. (2010) proposed to use this matrix as input for the IC algorithm to search for causal networks, instead of using the observed data.

The IC algorithm performs a series of statistical decisions based on partial correlations between traits. The posterior distributions of partial correlations were obtained using the posterior samples of residual (co)variance matrices from the multi-trait analysis and these were then used to test for non-null partial correlations. A partial correlation was declared non-null whenever the highest posterior density (HPD) interval did not include zero. The expected output for the IC algorithm is a partially oriented graph that represents a set of statistically equivalent causal structures.

The IC algorithm consisted of three steps (Pearl, 2009):

Step 1

Partial correlations were used to search for edges that connect adjacent variables (two vertices that are endpoints of an edge) to obtain an undirected graph (e.g., $Y_1 - Y_2$). If all partial correlations of two traits conditional on each possible set of other traits were different from zero, an edge was placed between the traits.

Step 2

Partial correlations were also used to search for unshielded colliders (three connected variables in a path directed as $Y_1 \rightarrow Y_2 \leftarrow Y_3$) to orient some edges of the undirected graph provided by step 1. If partial correlations of two non-adjacent traits (e.g., Y_1 and Y_3) that have a common adjacent trait (Y_2) in such an undirected graph are dependent conditional on any possible set that includes the adjacent trait (Y_2), the edges should be oriented towards the common adjacent trait (Y_2), such as in $Y_1 \rightarrow Y_2 \leftarrow Y_3$.

Step 3
When possible, remaining undirected edges were oriented in a way that introduced no new unshielded colliders or cycles. This step could only be performed when the graph obtained in step 2 contained unshielded colliders and the orientation followed unambiguously from the graph.

Structural equation model
Relationships represented by the causal network obtained from the IC algorithm were quantified using a SEM, as in Gianola and Sorensen (2004). The SEM was fitted using Bayesian methods that fit a multi-trait mixed model in software Bayz 2.1 (Janss, 2010), where causal parents (e.g., Y_1 is causal parent of Y_2 in $Y_1 \rightarrow Y_2$) of a given trait were considered as covariates in the equations assigned to this trait, and a diagonal residual (co)variance matrix was imposed. Therefore, the following model was fitted:

$$\mathbf{y} = (\mathbf{\Lambda} \otimes \mathbf{I})\mathbf{y} + \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\mathbf{u}^* + \mathbf{e}^*,$$

with the joint distribution of vectors **u** and **e** as:

$$\begin{bmatrix}\mathbf{u}^* \\ \mathbf{e}^*\end{bmatrix} \sim N\left\{\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{G}_\mathbf{0}^* \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_\mathbf{0} \otimes \mathbf{I}\end{bmatrix}\right\},$$

where the model was similar to the multi-trait model as described above but with the addition of $(\mathbf{\Lambda} \otimes \mathbf{I})\mathbf{y}$, where **Λ** is a $t \times t$ (with t equal to the number of traits) matrix with 0's on the diagonal and with structural coefficients or 0's on the off-diagonals. The causal structure defines which of the off-diagonal entries of **Λ** must be estimated and which ones are set to 0. $\mathbf{G}_\mathbf{0}^*$ is the SEM additive genetic (co)variance matrix and $\mathbf{\Psi}_\mathbf{0}$ is a diagonal matrix with the SEM residual variances. The residual covariances between the traits in the SEM were assumed to be 0, which confers identifiability to the structural coefficients in the likelihood function. The priors used for the SEM were the same as those used for the multi-trait model.
The SEM was compared with the multi-trait model using the deviance information criterion (DIC) (Spiegelhalter et al., 2002). The DIC takes the trade-off between model goodness-of-fit and corresponding complexity of model into account. Models with smaller DIC are better supported by the data.

5.2 Results

Multi-trait analysis

Eight independent MCMC chains of the Bayesian multi-trait animal model for the 14 bovine milk fatty acids converged to similar estimates of the variance components, which was confirmed by trace and density plots. The effective sample size for heritabilities, correlations and (co)variance components ranged from 391 to 2,431 samples. Posterior means of the heritabilities, genetic correlations and residual correlations between milk fatty acids are shown in Table 5.2. Fatty acids that are consecutively synthesized de novo (e.g., C4:0 and C6:0, C6:0 and C8:0, etc.) generally showed strong positive correlations, both genetically and residually. Residual correlations between medium chain unsaturated fatty acids (C10:1, C12:1, C14:1) and long chain fatty acids (C18:0, C18:1, CLA), and between CLA and C8:0, C10:0, and C12:0 were weak and showed large standard deviations. There were no strong negative correlations between fatty acids.

Table 5.2 Multi-trait genetic parameters[1] for bovine milk fatty acids[2,3]

	C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0	C18:0	C10:1	C12:1	C14:1	C16:1	C18:1	CLA
C4:0	**0.53**	0.91	0.83	0.77	0.71	0.87	0.89	0.66	0.50	0.39	0.48	0.50	0.54	0.28
C6:0	0.91	**0.49**	0.94	0.90	0.86	0.93	0.87	0.63	0.63	0.54	0.54	0.49	0.43	0.16
C8:0	0.78	0.91	**0.48**	0.95	0.92	0.93	0.81	0.59	0.67	0.61	0.53	0.45	0.35	0.08
C10:0	0.56	0.76	0.89	**0.43**	0.94	0.92	0.76	0.58	0.64	0.62	0.50	0.40	0.30	0.02
C12:0	0.47	0.66	0.81	0.88	**0.42**	0.89	0.71	0.54	0.62	0.63	0.48	0.37	0.26	−0.01
C14:0	0.56	0.74	0.87	0.91	0.90	**0.39**	0.88	0.65	0.60	0.57	0.56	0.53	0.46	0.16
C16:0	0.87	0.82	0.69	0.49	0.47	0.53	**0.33**	0.55	0.61	0.55	0.69	0.73	0.53	0.31
C18:0	0.80	0.75	0.63	0.44	0.30	0.41	0.69	**0.37**	−0.01	−0.06	0.02	0.13	0.70	0.23
C10:1	0.63	0.70	0.74	0.65	0.70	0.73	0.62	0.41	**0.61**	0.88	0.81	0.62	0.00	0.04
C12:1	0.39	0.48	0.57	0.57	0.71	0.68	0.47	0.16	0.88	**0.63**	0.82	0.63	−0.03	0.00
C14:1	0.45	0.49	0.53	0.47	0.60	0.60	0.50	0.26	0.90	0.93	**0.67**	0.83	0.22	0.23
C16:1	0.61	0.61	0.57	0.48	0.54	0.52	0.71	0.31	0.54	0.53	0.46	**0.49**	0.43	0.37
C18:1	0.82	0.86	0.84	0.72	0.71	0.76	0.79	0.60	0.79	0.66	0.66	0.72	**0.52**	0.48
CLA	0.39	0.49	0.58	0.61	0.69	0.66	0.42	0.09	0.57	0.62	0.50	0.66	0.66	**0.56**

[1] Heritabilities are shown in bold on the diagonal, genetic correlations below the diagonal and residual correlations above diagonal.
[2] In g/kg milk.
[3] Time-series standard errors for the variance components and correlations ranged from 0.0007 to 0.0091 and posterior standard deviations for the variance components and correlations ranged between 0.018 and 0.211.

Inductive causation (IC)

The IC algorithm based on the 95% HPD interval retrieved the undirected network presented in Figure 5.1 (solid edges). Consecutive fatty acids C4:0, C6:0, C8:0, C10:0 and C12:0 formed a path of connected nodes. The fatty acids C10:1 and C12:1, as well as C14:1 and C16:1, were also connected to each other. The HPD interval content was reduced to see if there were additional less strong connections between the fatty acids, which may give better results if posterior distributions are not very sharp (Shipley, 2002; Valente et al., 2011). Reducing the HPD to a probability of 90% resulted in the same network as the HPD interval of 95% (solid edges in Figure 5.1). Reducing the HPD interval to 85% resulted in four additional edges: between C4:0 and C16:0, between C6:0 and C14:0, between C8:0 and C12:0, and between C18:0 and C18:1 (dashed edges in Figure 5.1). Reducing the interval further to 80% resulted in two additional edges: between C8:0 and C14:0 and between C18:1 and CLA (dotted edges in Figure 5.1).

No unshielded colliders were recovered from the data in step 2 of the IC algorithm. Therefore, step 3 of the IC algorithm did not result in any additional edge orienting and the resulting network remained undirected.

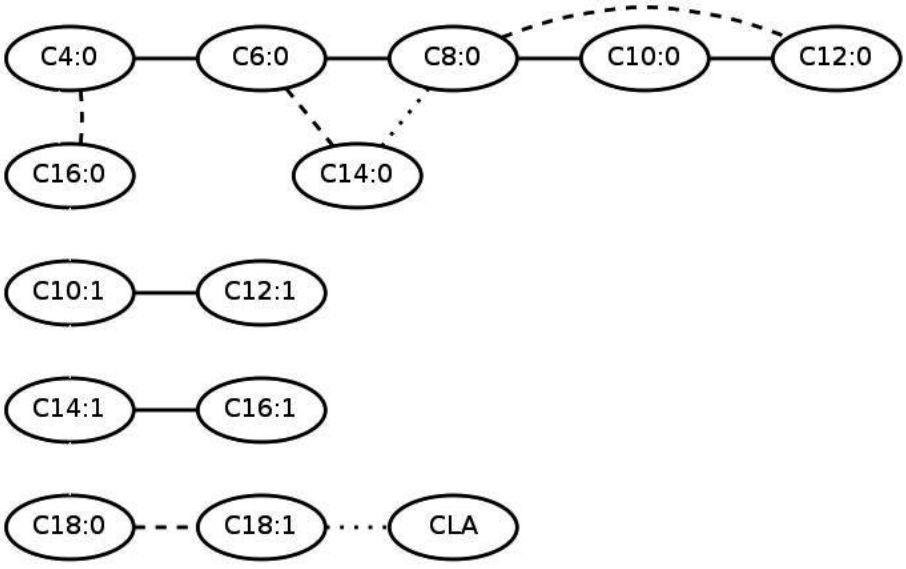


Figure 5.1 Network obtained from the inductive causation (IC) algorithm with different highest posterior density (HPD) intervals. The connections obtained with a HPD interval of 95% and 90% are given in solid lines, with a HPD interval of 85% in dashed lines, and with a HPD interval of 80% in dotted lines.

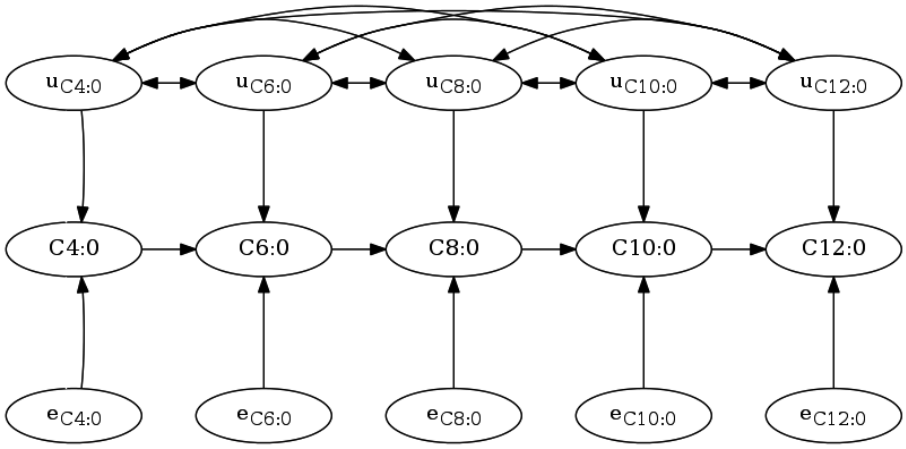


Figure 5.2 The fitted causal structure of the structural equation model. The edges in the fitted structure represent the causal relations for the observed variables (C4:0-C12:0), with independent residuals ($e_{C4:0}$-$e_{C12:0}$) and correlated additive genetic effects ($u_{C4:0}$-$u_{C12:0}$).

Structural equation model (SEM)

A SEM was used to quantify the causal relationships between the milk fatty acids based on a causal structure that was chosen based on the outputs of the IC algorithm. Since a fully oriented structure is required to specify a SEM, the undirected network obtained with the 95% HPD interval (Figure 5.1, solid edges) was oriented according to prior biological knowledge about the sequence in which the fatty acids are synthesized in the mammary gland. In this sense, the path C4:0—C6:0—C8:0—C10:0—C12:0 agreed with the de novo synthesis of milk fatty acids. According to the de novo synthesis, C4:0 should precede C6:0, C6:0 should precede C8:0, and so on. On this basis, the path C4:0—C6:0—C8:0—C10:0—C12:0 could be directed from C4:0 to C12:0, that is C4:0 → C6:0 → C8:0 → C10:0 → C12:0. The five traits involved in this path were analyzed with both a multi-trait model and a SEM. Both models were compared in terms of fit and parameter inferences.

The causal network chosen for the SEM shown in Figure 5.2 resulted in the following structure for the **Λ**-matrix:

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \lambda_{C6:0,C4:0} & 0 & 0 & 0 & 0 \\ 0 & \lambda_{C8:0,C6:0} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{C10:0,C8:0} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{C12:0,C10:0} & 0 \end{bmatrix}.$$

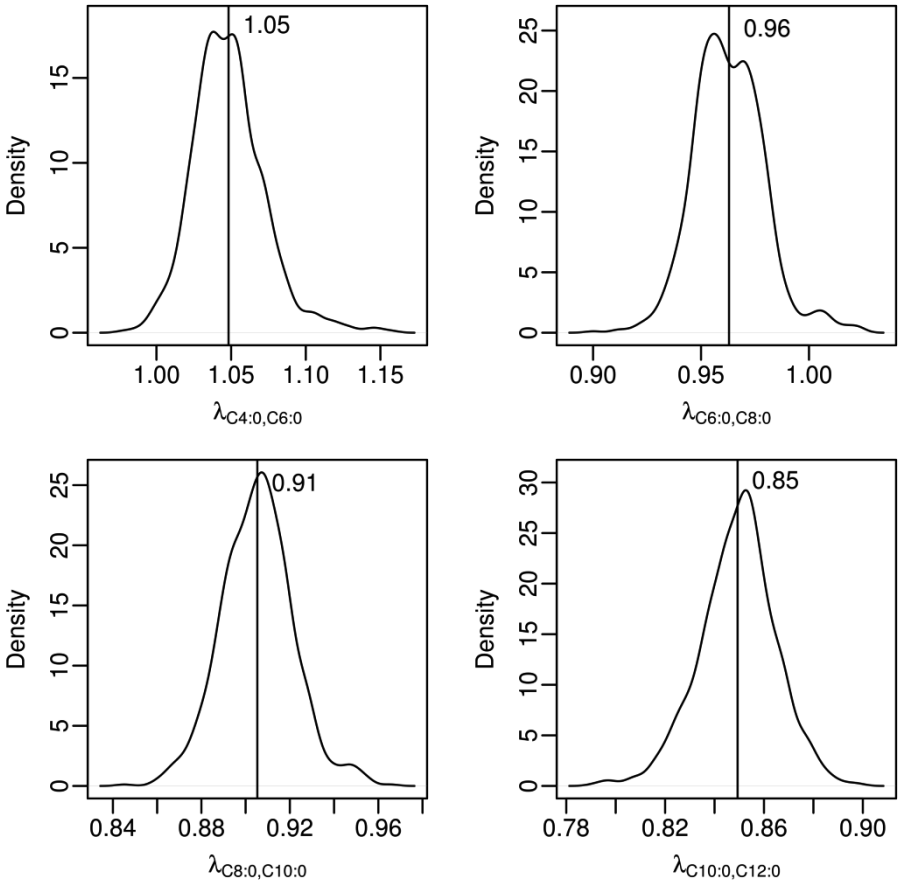


Figure 5.3 Posterior densities of structural coefficients for the fitted causal structure of the structural equation model

The posterior densities of the structural coefficients that resulted from the SEM are in Figure 5.3. The posterior means of these parameters ranged from 0.85 to 1.05. Table 5.3 shows the posterior means for the parameters from both the multi-trait model and the SEM for C4:0, C6:0, C8:0, C10:0 and C12:0. As pointed out by Valente et al. (2013), genetic effects from multi-trait models and SEM have different meanings: the latter represent direct genetic effects (i.e., genetic effects that are not mediated by other traits in the causal network), while the former represent overall genetic effects (i.e., a combination of all direct and indirect genetic effects on each trait). Model specific genetic (co)variances refer to the

Table 5.3 Posterior means of the variance components for the multi-trait and the structural equation model of C4:0 to C12:0

Variance component[1]	Multi-Trait			SEM		
	Mean	SD[2]	Time-series SE[3]	Mean	SD[2]	Time-series SE[3]
σ^2_e C4:0	0.549	0.108	0.003	0.455	0.091	0.002
σ^2_e C6:0	0.606	0.102	0.004	0.003	0.002	0.000
σ^2_e C8:0	0.599	0.100	0.004	0.000	0.000	0.000
σ^2_e C10:0	0.560	0.102	0.004	0.006	0.002	0.000
σ^2_e C12:0	0.459	0.087	0.003	0.059	0.004	0.000
r_e C4:0,C6:0	0.938	0.019	0.001	.	.	.
r_e C4:0,C8:0	0.885	0.046	0.001	.	.	.
r_e C4:0,C10:0	0.808	0.084	0.002	.	.	.
r_e C4:0,C12:0	0.754	0.101	0.003	.	.	.
r_e C6:0,C8:0	0.950	0.014	0.000	.	.	.
r_e C6:0,C10:0	0.906	0.036	0.001	.	.	.
r_e C6:0,C12:0	0.859	0.053	0.002	.	.	.
r_e C8:0,C10:0	0.950	0.014	0.000	.	.	.
r_e C8:0,C12:0	0.911	0.028	0.001	.	.	.
r_e C10:0,C12:0	0.934	0.017	0.001	.	.	.
σ^2_g C4:0	0.360	0.151	0.005	0.460	0.122	0.002
σ^2_g C6:0	0.325	0.143	0.005	0.114	0.023	0.001
σ^2_g C8:0	0.310	0.140	0.005	0.073	0.009	0.000
σ^2_g C10:0	0.319	0.141	0.005	0.066	0.008	0.000
σ^2_g C12:0	0.276	0.121	0.004	0.026	0.005	0.000
r_g C4:0,C6:0	0.855	0.074	0.002	−0.440	0.123	0.004
r_g C4:0,C8:0	0.675	0.157	0.005	−0.417	0.116	0.004
r_g C4:0,C10:0	0.424	0.237	0.007	−0.400	0.109	0.003
r_g C4:0,C12:0	0.331	0.255	0.008	−0.084	0.089	0.002
r_g C6:0,C8:0	0.863	0.069	0.002	0.761	0.033	0.001
r_g C6:0,C10:0	0.697	0.148	0.004	0.730	0.036	0.001
r_g C6:0,C12:0	0.617	0.179	0.006	0.160	0.154	0.004
r_g C8:0,C10:0	0.862	0.071	0.002	0.692	0.036	0.001
r_g C8:0,C12:0	0.805	0.102	0.003	0.152	0.147	0.004
r_g C10:0,C12:0	0.899	0.052	0.002	0.148	0.142	0.004

[1] σ^2_e is residual variance, σ^2_g is genetic variance, r_e is residual correlation, r_g is genetic correlation.
[2] SD is the posterior standard deviations of the component.
[3] Time-series SE is the time-series standard error of the component.

(co)dispersion of the genetic effects of each model, and therefore have distinct meanings as well. The posterior means of the genetic variances of the multi-trait model for C4:0, C6:0, C8:0, C10:0 and C12:0 were fairly similar to each other (i.e., between 0.360 for C4:0 and 0.276 for C12:0), while the posterior means of the SEM genetic variances for C4:0, C6:0, C8:0, C10:0 and C12:0 showed a gradual decrease (i.e., 0.460 for C4:0, 0.114 for C6:0, 0.073 for C8:0, 0.066 for C10:0 and 0.004 for C12:0), indicating that indirect genetic effects from upstream traits were gradually explaining a larger portion of genetic variability. Such reduction was even stronger for the SEM residual variance. Statistically, this result was expected because conditioning on the strongly correlated traits in the SEM removed a large proportion of the observed variance. On the basis of the given causal structure, this indicates that the variability of each of these fatty acids can be almost fully explained by the conditioning (parent) fatty acid. The posterior means of the genetic and residual variances of C4:0 for the SEM were similar to those for the multi-trait model, because C4:0 was not conditioned on any of the other traits. The posterior means of the genetic correlations from the SEM refer to the genetic covariance that is left after conditioning on the appropriate fatty acids, i.e., it expresses the correlation between direct genetic effects for each trait. For that reason, the SEM genetic correlations were different from the correlations estimated with the multi-trait model.

The DIC for the multi-trait model for C4:0, C6:0, C8:0, C10:0 and C12:0 was -21,083, while the DIC for the SEM using the structure depicted in Figure 5.2 was -32,406, indicating that the studied structure is plausible (Spiegelhalter et al., 2002). This lower DIC for the SEM is partly due to a lower penalty for model complexity in the DIC for the SEM. Although the SEM introduces sources of covariance from the causal associations, the residuals of the SEM were assumed to be uncorrelated, which resulted in a model that was more parsimonious than the multi-trait model. This lower penalty for model complexity was reflected by a lower effective number of parameters (5,152 for the SEM and 6,829 for the multi-trait model).

5.4 Discussion

The aim of this study was to explore causal networks of milk fatty acids by applying the IC algorithm in a mixed model context. Undirected acyclic graphs were obtained for several HPD intervals. A subset of five fatty acids formed a structure that could be directed based on prior knowledge and this structure was then used in a SEM to quantify the relationships between them.

Direction of network based on prior knowledge

The networks obtained for the 14 fatty acids were undirected. Based on the known sequence of the synthesis of fatty acids, edges could be directed without creating cycles or unshielded colliders that were not supported by the data. Fatty acid C4:0 precedes C6:0, which in turn precedes C8:0 and so on in the de novo synthesis, which led us to suggest that the path containing C4:0, C6:0, C8:0, C10:0 and C12:0 is directed from C4:0 to C12:0. This means that the final network is not completely data-driven. However, the structure for this subset of fatty acids that is plausible based on biological knowledge does not have colliders, so the fact that the algorithm could not detect directions was expected. Therefore, not finding any unshielded colliders among these fatty acids supports the hypothesis of a path directed from C4:0 to C12:0.

Linearity

The search space does not contain cyclic structures and non-linear relations are also not considered in the specific application presented here. Instead, as in most studies, it was assumed that relationships between traits were linear but in reality they could be non-linear. In contrast to the assumptions of the adapted IC algorithm applied here, SEM can be extended to include, for instance, interactions, feedback mechanisms (cyclic relations), quadratic terms or polynomials to determine which model fits the data best (Schumacker and Marcoulides, 1998; Spirtes, 1993). In addition, the search algorithm could make decisions based on alternative tests for conditional independence instead of on partial correlations (Pearl, 2009).

Causal sufficiency assumption

Connections between variables are often referred to as causal relations, but the only widely accepted method for declaring causation between two variables is a randomised experiment. This involves random assignment of each subject to different treatment groups, coupled with random assignment of treatment level to each group, and results in averaging out potential sources of confounding effects. In the analysed data, variables that act as confounders are not averaged out, but when they are measured, they can be included in the model to correct for this confounding effect. Based on model assumptions, causation can be inferred, but because of the impossibility of declaring with absolute certainty that there are no additional unmeasured causal variables, these assumptions cannot be guaranteed. The IC algorithm is based on the assumption that there are no hidden variables that affect more that one of the variables considered in the model, which is called the

causal sufficiency assumption (Spirtes et al., 2000). If this assumption does not hold, there may be direct connections between variables that are not causal relations but that are due to other sources, such as common hidden causes. Although a SEM does not require this assumption, it is commonly applied for the sake of model identifiability.

Comparison between the network obtained and known biological networks

Metabolic pathways involved in the synthesis of milk fatty acids, such as de novo synthesis, desaturation and biohydrogenation, could be reflected in the structure provided by the IC algorithm. In the following, the network obtained with the IC algorithm will be compared with known metabolic pathways of milk fat synthesis. For this comparison, two aspects should be noted. First, the variables studied here are fatty acids excreted in the milk, which are not necessarily the same variables as the corresponding fatty acids involved in the milk fat synthesis pathways, e.g., C6:0 measured in milk is not the same as a C6:0 in the elongation cycle of the de novo synthesis being transformed into C8:0. This is especially important considering that the SEM expresses the causal effect between fatty acids excreted in the milk, which are the recorded phenotypes. These causal effects reflect expected results of (ideal) external interventions. However, the expected consequences of modifying a fatty acid that is excreted in the milk on other fatty acids may not be the same as the consequences of manipulating the amount of a specific fatty acid during synthesis in the mammary gland.

The second aspect is that the proposition that an object B originates from an object A does not necessarily imply that causal effects between measurements a and b made respectively on A and B must be directed as $a \rightarrow b$. Therefore, if fatty acid B originates from fatty acid A in the synthesis process in the mammary gland, measurements of the concentration of these fatty acids in the milk (a and b) are not necessarily directed as $a \rightarrow b$ if they are causally connected. So it is possible that edges may actually have alternative directions, and that is not a strict contradiction of known biochemical paths. For example, inoculating C8:0 in the mammary gland could affect the amount of C6:0 released in the milk, which would be an effect that is opposite to the description of how C8:0 originates from C6:0, but does not deny that C8:0 originates from C6:0. Although one could defend such an alternative structure (and other statistically equivalent ones), the structure chosen to fit the model is credible given its expected intervention outcome. For instance, the chosen structure expresses that if C8:0 is inoculated in the mammary gland, then C4:0 and C6:0 would remain the same, but such intervention would

affect C10:0 and also, indirectly, C12:0. This is compatible with a scenario in which C8:0 is inoculated: C4:0, and C6:0 would be normally produced since their synthesis occurs earlier in the cycle, and less C8:0 would be released in the milk, since its concentration is already high due to the inoculation (in case there is some regulation of fatty acids production by the concentration of free fatty acids). This would leave more "substrate" remaining within the cycle for the subsequent fatty acids and would result in increasing C10:0, and so forth. This is compatible with the causal meaning of the chosen structure (and the inferred structural coefficients, if they are positive). It should be noted that in this case, the meaning of the graph C4:0 → C6:0 → C8:0 → C10:0 → C12:0 depends on whether it is interpreted as a biochemical pathway that shows how fatty acids are originated or as a SEM that involves the concentrations of such fatty acids, although both interpretations could be represented with the same nodes and directed connections. For the structure of the SEM fitted (C4:0 → C6:0 → C8:0 → C10:0 → C12:0), directions were chosen that mirror the de novo pathway, because it is plausible (although not necessary) based on how the fatty acids are generated and on that basis, if the underlying causal structure indeed reflected the metabolic pathway, the expected output of the search algorithm would be exactly C4:0—C6:0—C8:0—C10:0—C12:0.

De novo synthesis

Short and medium chain saturated fatty acids (C4:0-C14:0 and about half of the C16:0 present in milk) are produced in the de novo synthesis pathway. In this metabolic pathway, the carbon chain is elongated in a sequential cyclic reaction from acetate and β-hydroxybutyrate until a C16:0 fatty acid is formed by fatty acid synthase in the mammary gland (e.g., Neville and Picciano, 1997; Palmquist, 2006). In the bovine, all intermediate fatty acids can leave the elongation cycle by a chain termination mechanism (Smith, 1994) and thus end up in bovine milk. The path from C4:0 to C12:0 that was obtained from the IC algorithm with a HPD interval of 95% (solid edges in Figure 5.1) mirrored this de novo synthesis. One could argue that the path obtained from the IC algorithm should also include C14:0 and C16:0 but part of C14:0 and C16:0 originate from the cows' diet, which might have reduced the degree of association with the remaining pathway, thus leading the search algorithm to declare them disconnected from the remaining variables, i.e. excluding them from the pathway. The structural coefficients that were estimated using the SEM with the causal structure C4:0 → C6:0 → C8:0 → C10:0 → C12:0 indicate that if C4:0 increases 1 g/kg milk, then C6:0 would respond by increasing 1.05 g/kg milk (Figure 5.3). However, the molar mass of C6:0 is 1.32 times the molar mass of C4:0, so although the relationship is nearly one to one unit-wise, is

less than one based on molar mass. The structural coefficients $\lambda_{C10:0,C8:0}$ and $\lambda_{C12:0,C10:0}$ were slightly lower than $\lambda_{C6:0,C4:0}$ and $\lambda_{C8:0,C6:0}$, possibly because a small part of C10:0 and C12:0 is desaturated into C10:1 and C12:1 in the mammary gland. These structural coefficients suggest that an intervention that increases the amount of C4:0 secreted in milk would result in an increase in C6:0 secreted in milk and that would in turn result in an increase in C8:0, C10:0 and C12:0 secreted in milk.

Desaturation

Medium chain saturated fatty acids (C10:0-C16:0) are desaturated by coenzyme A desaturase 1 (*SCD1*) into their equivalent mono-unsaturated fatty acids (C10:1-C16:1) in the mammary gland (Palmquist, 2006; Taniguchi et al., 2004). Structures that mirror this desaturation pathway (e.g., C10:0 → C10:1) were not recovered by the IC algorithm (Figure 5.1). The obtained structures (C10:1—C12:1 and C14:1—C16:1) showed that the amount of mono-unsaturated medium chain fatty acids measured in milk are not causally associated with the amount of their equivalent saturated fatty acid, but suggest that the mono-unsaturated medium chain fatty acids may have a common hidden causal variable among them.

Biohydrogenation

Long chain fatty acids (half of the C16:0 present in milk and all fatty acids with 18 or more carbons) originate from the diet fed to cows and are biohydrogenated by micro-flora in the rumen into C18:0 and multiple intermediate products (Palmquist, 2006). Some edges were recovered between the long chain fatty acids, e.g. between C18:0 and C18:1, and between C18:1 and CLA, which likely represent this biohydrogenation process. These edges were recovered when the HPD interval was relaxed to 80-85%, which indicates weak evidence for these edges (Figure 5.1).

Reducing the HPD interval resulted in additional edges. The edges that involve long chain fatty acids might be plausible associations but the edges between C4:0 and C16:0, C6:0 and C14:0, C8:0 and C12:0, C8:0 and C14:0 appear to be false positive associations due to the lowered threshold.

To conclude, although the fatty acids were measured when secreted in milk and not during their synthesis in the mammary gland, concentrations of fatty acids in milk mirror some of the metabolic pathways, and resemblance with the de novo synthesis pathway obtained most evidence.

Convergence issues of the multi-trait model

The search for causal structures among a set of variables makes sense if associations exist between them. However, if many traits have strong correlations with each other, fitting multi-trait mixed models may encounter convergence issues, which was the case in the current study. Most milk fatty acids were strongly correlated with each other, both genetically and residually. Fitting a standard multi-trait model for 14 milk fatty acids resulted in slow MCMC convergence, strong auto lag correlations in the chain and thus in a small number of effective samples.

Running the MCMC Bayz 2.1 (Janss, 2010) program using latent variables to reduce the dimensionality of the data improved convergence of the Bayesian multi-trait mixed model. A principal component analysis showed that using four latent variables was reasonable for the multi-trait model with 14 fatty acids. Using latent variables has some effect on the modelled (co)variance structures; because the latent variable model uses less parameters than the full (co)variance matrix, the (co)variance structure is somewhat restricted, similar to using only the main principal components in a principal component analysis or frequentist factor analytic model (e.g., Meyer and Kirkpatrick, 2008). In this case, the latent variable model used 70 parameters [$(4\ latent\ variables + 1) \times 14\ traits$] for each of the environmental and genetic (co)variance structures, whereas the full (co)variance matrix has 105 parameters. For the multi-trait model for C4:0, C6:0, C8:0, C10:0 and C12:0, two latent variables were used, resulting in 15 parameters and thus no restrictions on the (co)variance matrix. The multi-trait model for C4:0 to C12:0 resulted in the same pathway as the model with 14 traits, suggesting that the restriction in parameters due to latent variables did not influence this particular pathway.

A final measure to improve convergence was to set minimum bounds on the remaining independent variances $\tau^2_{e_i}$ and $\tau^2_{u_i}$ for residuals and genetic effects through the prior distributions. These minimum bounds were set at 0.02 (on standardized phenotypes), which implies that heritabilities were constrained to be between 2 and 98%, and that all correlations were forced to remain slightly below 1. These adaptations were required for the model to converge such that this dataset could be explored for causal networks.

Computation time of the adapted Inductive Causation (IC) algorithm

The approach suggested by Valente et al. (2010) is more complex and computationally demanding than the standard use of the IC algorithm and other similar methods that simply work with unconditional point estimates of covariance matrices, not requiring prior model fitting. Although this is appealing in the context of mixed effects SEM, there is a compelling reason to follow the approach of Valente et al. (2010) because mixed effects SEM allow direct genetic covariances, which are extra genetic sources of associations among traits, aside from causal effects. Assuming these genetic associations to be absent would be more difficult to accept, since genetics most likely affects multiple traits of a set in a way that is not mediated by other traits in the set. Using the IC algorithm on raw data assumes that these correlated direct genetic effects do not exist and, therefore, requires assumptions that are more difficult to accept. Furthermore, using the output from such an IC analysis in a mixed effects SEM with unstructured genetic covariances implies inconsistency of assumptions in the different analysis steps.

The computation time of the IC algorithm increases rapidly with an increasing number of analyzed traits, because of the increasing number of partial correlations to be tested. The IC algorithm required testing the partial correlations between each pair of fatty acids conditional on all possible subsets of the remaining fatty acids. With 14 traits there are 91 distinct pairs of traits [$n \times (n-1)/2$] and 4,096 possible conditioning sets (2^{n-2}), leading to 372,736 partial correlations to be calculated for each posterior sample of the residual (co)variance matrix (i.e., $2^{n-1} \times n \times (n-1)/2$). In addition, the size of the posterior sample also affects computation time. Additional thinning of the MCMC speeds up computation time for the adapted IC algorithm. Parallel computing would be a promising strategy to reduce the computation time of the algorithm. However, other refinements to the method used here will be needed when the number of variables increases strongly, for instance with high-throughput gene expression data, such as microarray or RNA-seq.

Possibilities

Correlations between traits play a role in livestock management practices. These correlations can result from different causal relationships, such as direct or indirect causal effects between traits, or from a common causal parent, or even from a combination of these. The concentrations of fatty acids in milk are clearly correlated, but the partial correlations indicate that only a few are directly connected in the network. Even an undirected structure is informative and reveals direct and indirect associations between variables. Nonetheless, prior knowledge

may be used to orient additional edges, and resulting causal inferences can then be confirmed with additional data and studies. Representing the associations between traits with networks may provide better insights into the underlying biological mechanisms and offer opportunities for management tools to focus on pathways instead of correlations. Response to interventions applied to a biological system can be predicted using SEM. Shifting the focus from correlation matrices to causal diagrams might result in faster and better understanding of responses to interventions. The principles of the IC algorithm and SEM can also be used to investigate gene regulatory networks in gene expression studies (de la Fuente et al., 2004; Liu et al., 2008a; Schadt et al., 2005). Understanding the relationships between genes can, for instance, identify targets for intervention that could contribute to the development of therapies for certain diseases.

5.5 Conclusions

Application of the adapted IC algorithm proposed by Valente et al. (2010) resulted in an undirected network for the 14 milk fatty acids studied. The pathway from C4:0 to C12:0 reflected the de novo synthesis pathway of short and medium chain saturated fatty acids. By using prior biological knowledge, directions were assigned to that part of the network and the resulting structure was used to fit an SEM. The edges between C10:1 and C12:1 and between C14:1 and C16:1 did not correspond to associations reported in the literature, which might be due to a common hidden causal variable. Other expected relations based on biological knowledge were not found or were detected only when the HPD interval was relaxed.

The output of the IC algorithm suggested causal relations between the studied traits. This changes the focus from marginal associations between traits to direct relationships that may result in changes when external interventions are applied. The causal structure can give more insight into underlying mechanisms and the SEM can predict conditional changes due to such interventions.

6

General Discussion

The aim of this thesis was to unravel the genetic background of bovine milk fat composition. Genome-wide association studies (GWAS) using 50,000 single nucleotide polymorphisms (SNP) detected quantitative trait loci (QTL) for bovine milk fatty acids and show that milk fat composition has a complex genetic background with three major QTL that explain a relatively large fraction of the genetic variation of several milk fatty acids, and many QTL that explain a relatively small fraction of the genetic variation (chapters 2 and 3). In chapter 4 the major QTL on BTA19 was fine-mapped. The QTL region was reduced to a linkage disequilibrium (LD) block of 22 SNP that covers 85,007 bp and contains two genes: coiled-coil domain containing 57 (*CCDC57*) and fatty acid synthase (*FASN*). In chapter 5 causal relations between milk fatty acids were inferred to gain insight in the biological mechanisms involved in milk fat synthesis.
In this general discussion, I will first discuss the insights that were gained from GWAS for milk fatty acids. Subsequently, I will discuss the importance of intermediate phenotypes to close the gap between QTL and complex phenotypes. Next, I will discuss how combining causal relations with QTL can help to better understand biological mechanisms. And finally, I will explore the potential of GWAS using milk fatty acids based on mid-infrared instead of gas chromatography.

6.1 Scanning the whole genome

The field of QTL detection has gained a lot of attention over the years and many QTL have been discovered since (Goddard and Hayes, 2009). Linkage studies performed on part of the Dutch milk genomics data (with 1,500 SNP for 5 large sire-families) showed that several QTL for milk fatty acids were segregating in the population (Schennink et al., 2009b; Stoop et al., 2009b). However, the regions detected were rather large and some detected regions needed confirmation. Technological development of high throughput 50k SNP panels enabled GWAS. Proceeding from linkage analysis to GWAS confirmed and refined QTL locations and resulted in new candidate QTL regions (chapters 2 and 3).
The success of GWAS depends strongly on the study design (e.g., Spencer et al., 2009). Power calculations provide clear expectations in terms of QTL detection and limits of the study. The Dutch milk genomics dataset used in this thesis was designed before large SNP panels were available, and intended for two main purposes: estimation of genetic parameters and linkage analysis. Linkage analysis requires large half-sib families; therefore, data was recorded on five sire-families of about 100 to 200 daughters each. Estimation of genetic parameters, such as heritabilities and genetic correlations, is preferably done in a population of

unrelated individuals. Therefore, data was recorded also on 50 small sire-families. These small families made the Dutch milk genomics population better suited for GWAS. Moreover, the pedigree was taken into account in the statistical analyses to deal with population stratification resulting from the family structure. With phenotypes and genotypes on approximately 1,700 cows, the power of the GWAS applied in this thesis was suited to detect variants that explain 2% or more of the genetic variation on milk fatty acids, which is in agreement with the QTL effect sizes detected in chapters 2 and 3. In chapter 2, GWAS for winter milk samples resulted in 64 significant regions, each of which explained between 0.9 and 67.8% of the additive genetic variance of a fatty acid. In chapter 3, GWAS for summer milk samples resulted in 51 significant regions, each of which explained between 2.2 and 50.1% of the additive genetic variance of a fatty acid. Of these regions, 34 were overlapping between winter and summer samples. All detected QTL together explain between 5.3% of the total additive genetic variation for C12:0 and 97.4% of the total additive genetic variation for C14:1 in winter milk and between 5.5% of the total additive genetic variation for C4:0 and 92.5% of the total additive genetic variation of C16:1 in summer milk (chapters 2 and 3). These percentages of explained additive genetic variation are probably overestimated due to the Beavis effect (Beavis, 1998) and may include some false positive QTL because of the false discovery rate (FDR) of 5%. Nonetheless, a considerable fraction of the total genetic variation of several milk fatty acids can be attributed to QTL detected in the GWAS presented in this thesis. The remaining, unexplained genetic variation might be caused by QTL with effects that are too small to be detected in the current study (below 2% of genetic variation), or for instance by epistasis or structural variants, among other possibilities (e.g., Manolio et al., 2009; van der Sluis et al., 2010).

The efforts to detect QTL using GWAS have often been questioned because inferring precise location is difficult and GWAS have rarely resulted in detection of the actual causal variants (Ron and Weller, 2007; Weller and Ron, 2011). It should be noted that GWAS is only the first step in the process of identifying causal variants. Genome-wide association studies provide a search for regions associated with the trait across the whole genome without requiring prior knowledge on location or gene function (Hirschhorn and Daly, 2005). Follow-up studies with appropriate designs have to be conducted to confirm QTL, fine-map QTL, and eventually, identify causal variants. In livestock, confirmation of QTL in an independent sample prior to publication is not a requirement as is the case in human studies. For major QTL like the ones detected in this thesis on BTA14, 19 and 26 confirmation might not be necessary, however, for QTL with less evidence a follow-up study to confirm the QTL in a different population is needed. In livestock

a limited number of detected QTL (mainly major QTL) have been followed up by fine-mapping and even fewer by functional studies. For animal breeding it is not essential to know the causal variants, while in human genetics causal variants are essential for treatment of genetic defects. Human QTL studies show a substantial number of validated QTL with strong candidates in over half of them (Visscher 2008; Weedon and Frayling 2008), but also in human studies identification of causal variants is lacking in most cases. The identification of causal variants for complex traits has been a difficult task, but technological development and reduced cost of next generation sequencing will enhance some of the problems associated with the detection of causal variants.

A complicating factor in livestock is that several traits have been under long term artificial selection and, as a consequence, polymorphisms with large effects on traits under selection have been fixed on the favorable allele (Weller and Ron, 2011). However, based on QTL mapping and GWAS we now know that some traits in Holstein cattle are influenced by major QTL, such as fat content and coat color (Hayes et al., 2010), as well as milk fat composition (this thesis). The most evident example of a major QTL in livestock is the QTL on BTA14 for milk yield and composition. Milk yield and composition have been under long term artificial selection, yet this QTL with major effects on these traits is still segregating in the current dairy cattle population. This major QTL is not fixed due to its opposing effects on milk yield and on fat and protein content, and, consequently, small effects on fat and protein yield. Simultaneous selection on milk yield and composition has, thus, retained this major QTL. Several linkage studies that detected the major QTL on BTA14 (Coppieters et al., 1998; Heyen et al., 1999) have been followed up by fine-mapping (Farnir et al., 2002; Riquet et al., 1999), and eventually Grisart et al. (2002, 2004) proposed diacylglycerol O-acyltransferase 1 (*DGAT1*) *K232A* as causal variant. The di-nucleotide polymorphism *DGAT1 K232A* is one of the few causal variants in dairy cattle that were successfully detected based on QTL mapping, followed by fine-mapping and functional studies. The proposed causal variant for this QTL on BTA14 has been shown to be associated with milk fatty acids (Schennink et al., 2007, 2008), and was one of the three major QTL detected in this thesis (chapters 2 and 3).

A different strategy to find causal variants is through a traditional candidate gene approach, where genes are selected based on prior knowledge regarding their biological function in relation to the phenotype, and then sequenced to find associated polymorphisms. Taniguchi et al. (2004) proposed stearoyl-CoA desaturase 1 (*SCD1*) *A293V* as a possible causal variant for unsaturation of fatty acids in beef cattle using such a candidate gene approach. This polymorphism was

also associated with milk fatty acid unsaturation in dairy cattle (Schennink et al., 2007, 2008), and is one of the three major QTL detected in this thesis.

The third major QTL detected in this thesis encompassed a rather large area on BTA19, which harbors quite a number of genes known to be related to fat synthesis. In chapter 4 the position of this QTL was refined by increasing the number of markers. The refined region contains two candidate genes: *CCDC57* and *FASN*. Fatty acid synthase has often been studied in traditional candidate gene studies, but significance levels reported in these studies (Abe et al., 2009; Li et al., 2012; Oh et al., 2012; Roy et al., 2006; Schennink et al., 2009a; Zhang et al., 2008) did not reach the significance reported in GWAS for the same chromosomal region described in this thesis. Although *FASN* is a good candidate gene for associations detected in this region, results from fine mapping suggest that not only *FASN*, but also its adjacent gene *CCDC57*, as well as the region in between these genes should be considered in the search for the causal mutation(s). Other genes located on BTA19 and related to fat synthesis, such as ATP citrate lyase (*ACLY*), sterol regulatory element-binding transcription factor 1 (*SREBF1*), signal transducer and activator of transcription 5A (*STAT5A*), and growth hormone 1 (*GH1*), had no significant effects on C14:0 after accounting for the most significant SNP on BTA19 located in *CCDC57* (chapter 4). Thus, these genes are not likely to contain the causal variant for the QTL on BTA19. Some of these fat synthesis related genes have been studied as candidate genes by others (Nafikov et al., 2013; Schennink et al., 2009a). Results from our fine-mapping study suggest that these candidate gene studies have picked up the effect of the QTL at 51Mbp through LD, rather than the effect of an actual causal variant in one of these genes. This confirms that candidate gene studies reveal only part of the picture and therefore might lead to wrong conclusions with respect to causal variants, whereas GWAS show a more complete, unbiased picture by scanning the whole genome.

Several other candidate genes related to fat synthesis have been proposed based on the GWAS in chapters 2 and 3, e.g. acyl-CoA synthetase short-chain family member 2 (*ACSS2*) on BTA13 and 1-acylglycerol-3-phosphate O-acyltransferase 6 (*AGPAT6*) on BTA27. The latter is supported by a recent study of Wang et al. (2012) who found a QTL for milk fat percentage on BTA27 and used sequence data to fine-map the QTL to the promoter region of *AGPAT6*. Chapters 2 and 3 revealed additional regions associated with milk fatty acids, for which no genes with known functional relation to milk fat could be assigned. Identifying candidate genes in such regions is rather challenging, first, because milk fat synthesis is a complex process that can be influenced by many gene regulated factors, some of them might influence milk fat synthesis indirectly, and second, because not all genes and

regulatory regions have been (well) characterized in the bovine genome. The bovine genome consortium has done a great job to provide researchers with a reference genome for cattle, however, it is incomplete and improvements are needed. Furthermore, efforts to improve the annotation of genes and regulatory regions will increase the potential for identification of candidate genes.

6.2 Intermediate phenotypes

Complex phenotypes can often be decomposed into underlying components. These components, or intermediate phenotypes, may arise from the same biological mechanism and show overlap in genetic associations. They may also arise from different biological mechanisms that interact and, to some extent, show different genetic associations. In this thesis, milk fat was decomposed into individual fatty acids. Several QTL have been detected previously for fat percentage and fat yield (Ashwell et al., 2004; Heyen et al., 1999; Wang et al., 2012). Some of these were also detected for individual fatty acids (e.g., on BTA5, 14 and 27), but many additional QTL were detected based on GWAS of individual fatty acids, such as the major QTL on BTA19 for de novo synthesized fatty acids and BTA26 for unsaturated fatty acids. These QTL do not significantly contribute to genetic variability in fat percentage or yield, but they do provide valuable information about milk fat synthesis.

Another example of information obtained from decomposed phenotypes relates to the QTL on BTA26 caused by *SCD1*. The enzyme produced by *SCD1* desaturates several saturated fatty acids into their unsaturated equivalents (e.g. C10:0 into C10:1). Desaturation indices of individual fatty acids C10, C12, C14, C16, C18 and CLA are associated with the QTL on BTA26. However, this QTL is not detected when studying the desaturation index of these fatty acids when combined into one group (Schennink et al., 2008). Apparently, opposite allelic effects of *SCD1* on the desaturation indices of C10, C12 and C14 compared to the indices of C16, C18 and CLA cancel each other out (Schennink et al., 2008). Consequently, the QTL on BTA26 may not be relevant for changing the overall amount of unsaturated fatty acids by genetic selection; however, it is of crucial importance to understand the mechanism of milk fat synthesis, in this specific example the unsaturation of milk fat.

Decomposing complex phenotypes can be part of system genetics. System genetics aims to fill the so called genotype-phenotype gap by unraveling how information flows from DNA to phenotype (Civelek and Lusis, 2014; Houle et al., 2010). More detailed characterization of complex phenotypes is a way to fill a piece of this gap,

as shown in this thesis. This can be further enhanced by studying transcripts, proteins, metabolites, and interactions among and between them (Civelek and Lusis, 2014; Houle et al., 2010). In system genetics, hypotheses are generated by phenotyping traits, assess their correlations, search for (expression)QTL associated with these phenotypes, group traits that appear to share characteristics and combine all information obtained in a network (Civelek and Lusis, 2014). One way to obtain such a network is to look at causal relationships. Ultimately, causal networks are built over all layers of information; however, less complex starting points can be causal networks among measured phenotypes as studied in chapter 5, or between phenotypes and QTL, which will be explored in the next section of this discussion.

6.3 Causal relations between milk fatty acids and QTL

Linking phenotypic data and QTL to build causal networks can increase our understanding of complex biological systems. Such a network can show the impact of specific QTL on the whole system. Schadt et al. (2005) tested causality between gene-expressions associated with a phenotype by including QTL information. A QTL associated with both the trait and the gene-expression can help to infer causality, i.e. to define whether the relationship between gene-expression (GExp) and phenotype (P) is causal (QTL→GExp→P), reactive (QTL→P→GExp) or independent (GExp←QTL→P). Similar causal relationships can be inferred for a QTL associated with two traits (i.e. QTL→y_1→y_2, QTL→y_2→y_1, y_1←QTL→y_2).

The results of this thesis were used to test whether linking phenotypic data and QTL in a causal network can further improve our knowledge on milk fatty acids. The GWAS in this thesis (chapters 2 and 3) have shown that milk fatty acids are influenced by more than one locus, and that several milk fatty acids have one or more loci in common. This is in agreement with known genetic correlations between fatty acids and current knowledge on fatty acid metabolism. This thesis has also shown that some non-genetic associations are indirectly mediated through other milk fatty acids, while only a few are direct associations between fatty acids (chapter 5). The same can be true for the associations found in GWAS: a QTL can have a direct effect on multiple milk fatty acids or have a direct effect on one fatty acid that influences other milk fatty acids indirectly through a pathway.

True causality can only be inferred in fully controlled experiments in the laboratory or by completely randomized experiments. Li et al. (2006) indicated that there is an analogy between QTL mapping and randomized experiments, as the genotype of the QTL can be seen as treatment groups and meiosis randomly allocates QTL

genotypes to individuals. The causal relationship between a QTL and phenotype is always directed from QTL towards phenotype (genes come before phenotypes): therefore, QTL added to the phenotype network can help to direct the network (Chaibub Neto et al., 2010).

QTL on BTA19

In chapter 5 of this thesis, causal relationships between winter milk fatty acids were inferred. Figure 6.1A shows the pathway recovered between C4:0, C6:0, C8:0, C10:0 and C12:0 that resembles the de novo synthesis of short and medium-chain saturated fatty acids. The de novo synthesized fatty acids C14:0 and C16:0 were not part of this pathway. Often the same QTL regions were detected for the de novo milk fatty acids. This was also the case for the major QTL on BTA19, which showed associations with C4:0, C8:0, C10:0, C12:0 and C14:0 in the winter milk samples (figure 6.1B). At first sight it seems that the QTL affects almost all traits in the pathway from C4:0 to C12:0, as well as C14:0. However, these QTL associations could also be indirectly mediated through this very pathway.

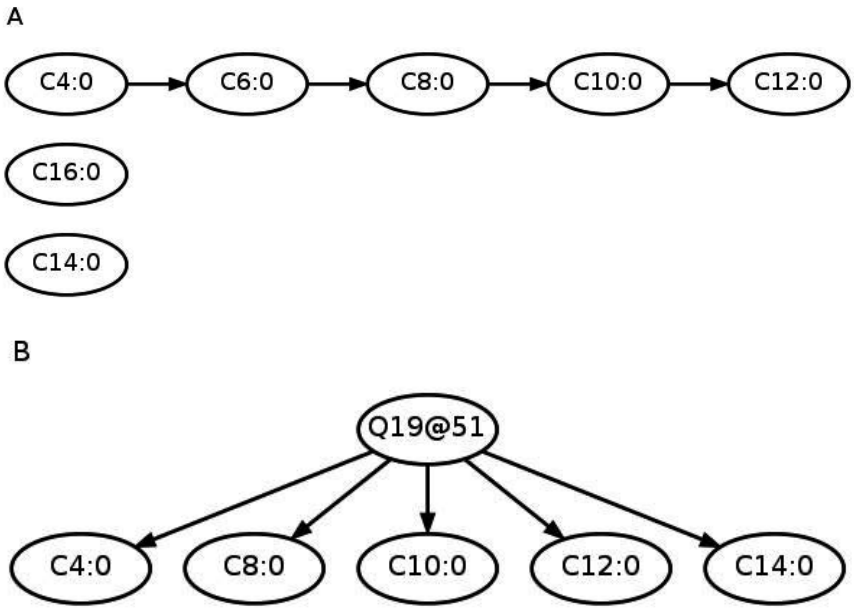


Figure 6.1 Graphical representation of A) the causal relations between de novo synthesized milk fatty acids (modified from chapter 5) and B) the associations between de novo synthesized fatty acids and the QTL on BTA19 at 51 Mbp (Q19@51) detected in chapter 2.

QTL on BTA19 conditioned for de novo synthesis

A graph for the QTL on BTA19 conditioned for the de novo synthesis pathway can be obtained by testing the models given in table 6.1 for two consecutive fatty acids (y_1 and y_2) associated with the same QTL. Models A and B in table 6.1 indicate whether the traits are associated with the QTL. These models have already been performed in the GWAS described in chapter 2. They were repeated here with the SNP (BovineHD1900014354) at 51.3 Mbp on BTA19 that showed the strongest association with C14:0 in the fine-mapping study of the QTL on BTA19 in chapter 4. Model C in table 6.1 will indicate whether the QTL has a direct effect on trait y_2 ($\beta_4 \neq 0$) or whether the effect is indirectly mediated through trait y_1 via the pathway QTL$\rightarrow y_1 \rightarrow y_2$ ($\beta_4 = 0$). In model C for y_2, the trait y_1 (the preceding fatty acid in the pathway) is added to the model as a covariable to condition on while testing for association with the SNP. These models were analyzed for each de novo synthesized fatty acid and the QTL on BTA19. The fatty acids C14:0 and C16:0 were treated as if they were also part of the pathway, even though they were not associated with other fatty acids in chapter 5. In the statistical analysis other systematic effects were accounted for as well as the polygenic background by using a mixed model similar to the models used throughout this thesis.

Table 6.1 Tests for direct QTL effects on two correlated traits (y_1 and y_2).

	Model	Test	Result
A	$y_1=\mu_1+\beta_1 QTL$	$\beta_1 \neq 0$	QTL detected for y_1
B	$y_2=\mu_2+\beta_2 QTL$	$\beta_2 \neq 0$	QTL detected for y_2
C	$y_2=\mu_1+\beta_3 y_1+ \beta_4 QTL$	$\beta_4 \neq 0$	Direct effect of QTL on y_2

Table 6.2 Significance levels ($-\log_{10}$(*P*-value)) of SNP BovineHD1900014354 in both the standard model and the conditional model in which the preceding fatty acid is included as a covariate.

	C4:0	C6:0	C8:0	C10:0	C12:0	C14:0	C16:0
Standard model	4.42	1.43	8.19	16.39	12.42	33.68	2.20
Conditional model	-	9.64	11.79	11.50	0.12	24.18	0.28

Table 6.2 (standard model) shows that the SNP on BTA19 was significantly ($-\log_{10}$(*P*-value) > 3) associated with all the de novo synthesized fatty acids except C6:0 and C16:0. The effect of the alleles on the de novo milk fatty acids was in the same direction for these de novo fatty acids except for C4:0 and C16:0. After conditioning on the preceding fatty acid (e.g. C6:0 conditioned on C4:0), all de novo synthesized fatty acids were significantly associated with the SNP except C12:0 and C16:0 (Table 6.2). Apparently the high genetic correlations between consecutive fatty acids do

not account for the associations of the SNP on BTA19 with multiple de novo synthesized fatty acids. In fact, the fatty acid included as a covariable in the model was highly significant, but the SNP still explained a significant part of the remaining phenotypic variation. Remarkably, C6:0 became significantly associated with the SNP after conditioning on C4:0. This is probably because the allele substitution effects of the SNP were opposite for C4:0 and C6:0. Milk fatty acid C12:0 was no longer associated with the SNP after conditioning on C10:0. The reduction in significance of C12:0 indicates that the variance in C12:0 explained by the SNP is part of the covariance between C10:0 and C12:0. This would mean that the SNP indirectly affects C12:0 through the path via C10:0. For C4:0, C6:0, C8:0, C10:0 and C14:0 there is evidence for a direct effect of the SNP on each phenotype. Furthermore, the SNP has no effect on C16:0, which is probably because approximately only 50% of C16:0 is produced de novo, while the other 50% is derived from the diet. These results lead to the updated causal graph depicted in figure 6.2.

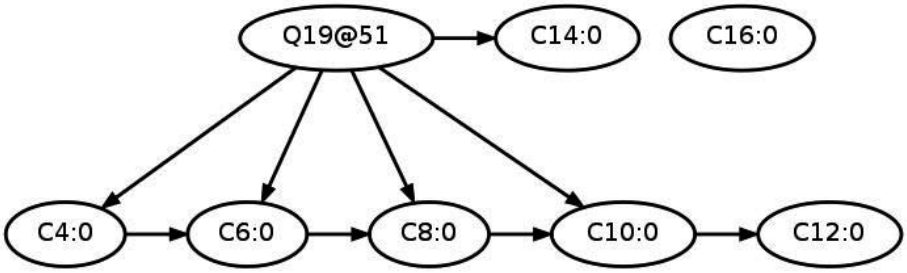


Figure 6.2 Graphical representation of the QTL on BTA 19 at 51 Mbp (Q19@51) and the causal relations between milk fatty acids after conditioning on the de novo synthesis pathway.

Figure 6.2 gives a slightly different view of the possible causal relations between the QTL and the fatty acids than figure 6.1. One of the candidate genes underlying the QTL at 51 Mbp on BTA19 is *FASN*. This gene is transcribed into a multi-enzyme complex that synthesizes de novo milk fatty acids by carbon chain elongation (Smith, 1994). Before studying the casual relations between the QTL and these fatty acids it could be hypothesized that the QTL has direct effects on multiple fatty acids, or that it influences only one fatty acid which influences the other fatty acids through the pathway. The latter would indicate that the QTL influences only a specific part (e.g. initiation) of the complex chain elongation by the *FASN* enzymes. However, here I showed that the QTL influences C4:0, C6:0, C8:0, C10:0, as well as C14:0, directly. This suggests that the QTL is involved in the chain elongation mechanism. How this exactly works and why C12:0 is not influenced directly by the

QTL remains unclear. However, Heck et al. (2012) showed that not all C12:0 is de novo synthesized; a part of the C12:0 in milk is derived from feed ingredients rich in C12:0. This suggests that the QTL has no significant effect on that part, but only on the amount of C12:0 that is produced by elongating its precursor C10:0.
Causal inference is a statistical way to generate hypotheses that has been used on the results of this thesis to demonstrate that the QTL on BTA19 influences several fatty acids directly. The statistical models applied here are very simple and not all possible causal networks have been explored; therefore, it is difficult to draw final conclusions on true causality. A model comparison test like Vuong et al. (1989) is preferred over the method applied here. However, model comparison tests require testing all possible models, which is computationally unfeasible when many variables are involved. Currently available and computationally efficient software to compare causal structures on the basis of model fitting are developed for experimental crosses and not suitable for outbred populations (Chaibub Neto et al., 2010). Application of these kinds of approaches on large scale livestock data would require development of efficient methods for outbred populations. Although the method applied here was rather straightforward, it has demonstrated that genetic information can aid in directing relationships between variables and in determining whether relationships are causal, reactive or independent. The genotype-phenotype gap can be reduced further by including different levels of intermediate phenotypes in such causal searches for information flows from DNA to phenotypes.

6.4 GWAS based on mid-infrared predicted milk fatty acids

The phenotypes used in this thesis were milk fatty acids measured by gas chromatography (GC). Gas chromatography is the so called ‘gold standard’ method that provides highly accurate measurements of milk fatty acids. But, GC is also expensive and time consuming. Therefore, routine measurement of fatty acids with GC is not feasible and populations phenotyped for milk fatty acids with this technique are limited in number and usually of small size. Prediction of fatty acids in milk based on mid-infrared spectra (MIR) (Rutten et al., 2009; Soyeurt et al., 2006) has made routine measurement of milk fatty acids possible. The MIR spectra are already used in routine milk recording to quantify fat, protein, lactose, and urea content. The recently developed prediction equations for milk fatty acids enable routine phenotyping of large populations at low cost. This facilitates breeding for desired fat composition, but also GWAS can benefit from large resource populations. This raises the question whether MIR predicted milk fatty acids would be suitable for unraveling the genetic background of milk fat composition with

GWAS. Both Bastin et al. (2012) and Govignon-Gion et al. (2012) performed GWAS based on MIR predicted milk fatty acids and detected many associated regions, which included regions known to be associated with fat composition. However, there were also discrepancies between regions found in our study based on GC data and those studies using MIR data. It is not clear whether these discrepancies are caused by differences between populations or differences between MIR predicted and GC measured milk fatty acids. The data used in this thesis is suited to address the implications of using MIR predicted milk fatty acids in GWAS studies, because both GC and MIR spectra were available on the same milk samples, and Rutten et al. (2009) developed prediction equations for milk fatty acids using this data.

GWAS results

Here I will compare GWAS results for both GC measured and MIR predicted milk fatty acids on the same milk samples to address the implications of using predicted milk fatty acids in GWAS studies. Four fatty acids were chosen for comparison: C14:0, C16:0, C18:1*cis*9 (C18:1), and C18:2*cis*9,*trans*11 (Conjugated Linoleic Acid, CLA). The equations used to predict milk fatty acids from the MIR spectra were built as described by Rutten et al. (2009) and based on weight percentages, i.e. g/100g fat. According to Rutten et al. (2009), the squared correlation of GC measured records with MIR predicted records (here referred to as prediction accuracy) was 0.73 for C14:0, 0.71 for C16:0, 0.84 for C18:1, and 0.58 for CLA (Table 6.3). The fatty acid C14:0 was chosen because it is a de novo synthesized fatty acid for which many QTL were detected (chapters 2 and 3); C16:0 because it is the most abundant fatty acid in milk; and C18:1 and CLA because they have a relatively high and low prediction accuracy, respectively. Both GC measurements and MIR predictions were available on the same milk samples of 1,614 cows with 50K genotype data. The genetic correlations between the GC measured records and the MIR predicted records for the four fatty acids ranged from 0.83 to 0.97 (Table 6.3).

Table 6.3 Prediction accuracy (r^2) of MIR predicted records[1], genetic correlation (r_g) between GC measured and MIR predicted records, and the correlation between allele substitution effects of all SNP from GWAS with GC data and GWAS with MIR predicted data (r_{ase}).

Trait	r^2	r_g (s.e.)	r_{ase}
C14:0	0.73	0.97 (0.02)	0.74
C16:0	0.71	0.92 (0.05)	0.74
C18:1	0.84	0.97 (0.04)	0.80
CLA	0.58	0.83 (0.10)	0.53

[1] As reported in Rutten et al. 2009.

The GC measured records were already checked for outliers in previous GWAS, while the MIR predicted records were not edited. Bivariate analysis shows that there was good concordance between GC measured C14:0 (GC-C14:0) and MIR predicted C14:0 (MIR-C14:0) phenotypes, while the MIR predictions for C16:0, C18:0 and CLA contained outliers with large prediction errors (Figure 6.3). One individual with GC measurements within the normal range of values appeared to have extreme values for MIR predicted C16:0 (MIR-C16:0; ~7 sd from mean) and C18:1 (MIR-C18:1; ~12 sd from mean). The MIR predicted records of this individual for C16:0 and C18:1 were set to missing in the GWAS. Single SNP GWAS were performed using the same univariate animal model in ASReml as described earlier in chapters 2, 3 and 4.

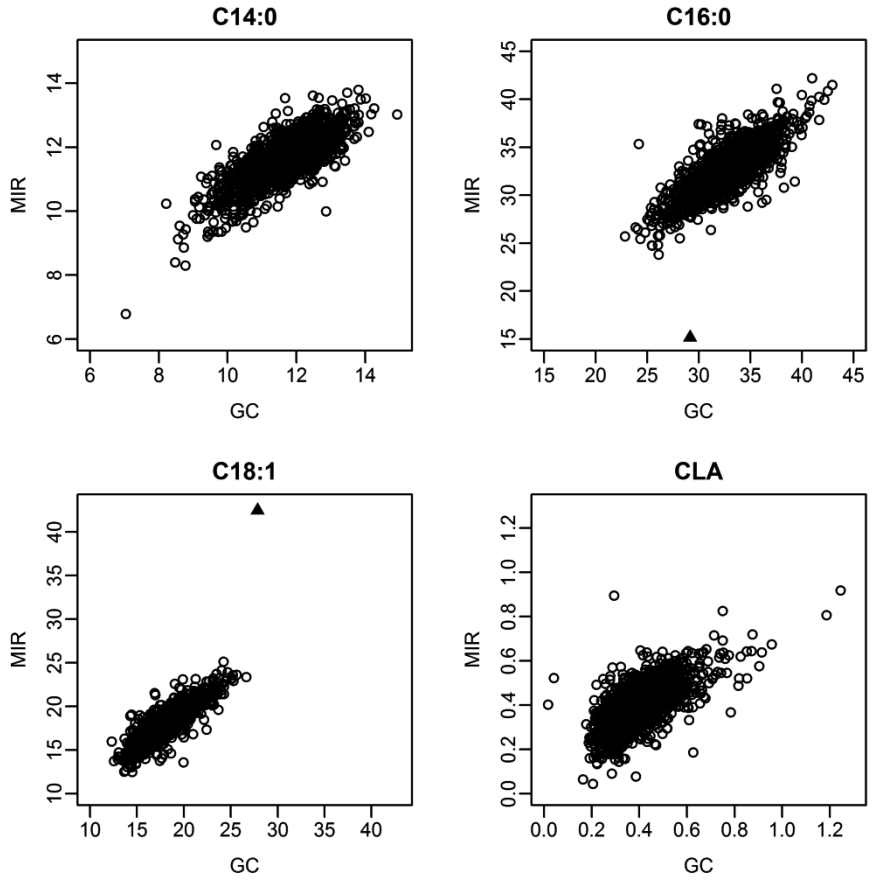


Figure 6.3 Fatty acid records measured with gas chromatography (GC) plotted against mid-infrared predicted records (MIR). Outliers due to high prediction errors are indicated as solid triangles.

Table 6.4 shows that GWAS for MIR-C14:0 and MIR-C16:0 resulted in smaller numbers of significant SNP compared to GC data, but almost all SNP declared significant using MIR predicted data were also significant using GC data. Single SNP GWAS for MIR-C18:1 resulted in a larger number of significant SNP compared to GC measured C18:1 (GC-C18:1), but almost all SNP declared significant using GC-C18:1 were also significant using the MIR predicted data. Single SNP GWAS for CLA using GC data (GC-CLA) resulted in only 8 SNP with a FDR of 5% and only 4 with a FDR of 0.1%. In contrast, GWAS for MIR predicted CLA (MIR-CLA) resulted in many more significant SNP, i.e. 100 at a FDR of 5% and 32 at a FDR of 0.1%, but included most SNP significant in the GWAS for GC-CLA.

Table 6.4 Number of SNP significant at false discovery rates (FDR) of 5% and 0.1% for genome-wide association based on gas chromatography measured (GC) fatty acids, mid-infrared predicted (MIR) fatty acids, and the overlap between both (both).

	FDR ≤ 0.05			FDR ≤ 0.001		
Trait	GC	MIR	both	GC	MIR	both
C14:0	372	157	136	167	46	46
C16:0	217	202	186	160	122	119
C18:1	207	232	183	118	136	114
CLA	8	100	7	4	32	4

The allele substitution effects of all (50k) SNP from GWAS with GC measured and MIR predicted C14:0, C16:0, C18:1, and CLA had a correlation of 0.74, 0.73, 0.78, and 0.53, respectively. These correlations are in good agreement with the prediction accuracies of the MIR spectra for GC measurement (Table 6.3). Similar relationship between the correlation of allele substitution effects of GWAS based on two measurements of the same population was shown by Barendse et al. (2011). As suggested by Barendse et al. (2011) a prediction accuracy of at least 0.95 would be required to obtain similar allele substitution effects for GWAS for both fatty acid phenotypes. Improvement of the prediction of fatty acids from MIR spectra might, thus, result in better concordance of GWAS results.

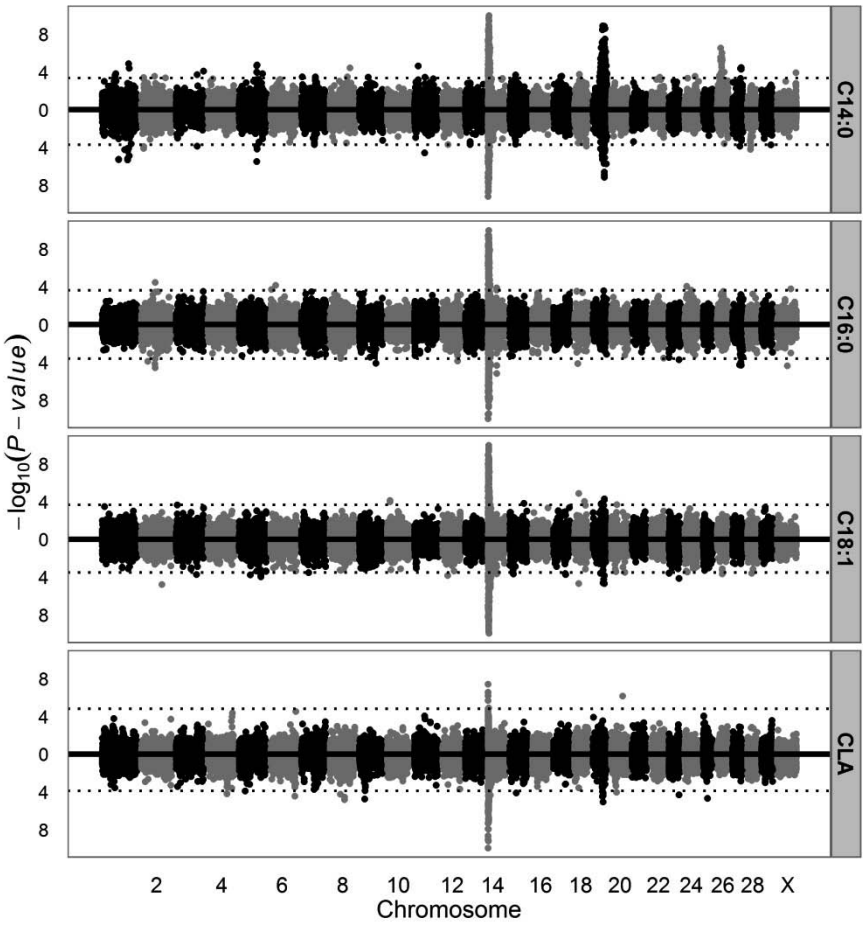


Figure 6.4 Manhattan plots for C14:0, C16:0, C18:1, and CLA comparing GWAS results based on gas chromatography (above x-axis) and mid-infrared (below x-axis). The dotted lines represent the false discovery rate threshold of 5%.

Figure 6.4 shows the results from GWAS for GC measured and MIR predicted milk fatty acids from the exact same milk samples. The major QTL (FDR of 0.1%) for GC-C14:0 on BTA14 and 19 were confirmed by MIR-C14:0; however, the major QTL on BTA26 for GC-C14:0 was not significant with MIR-C14:0. The QTL for GC-C14:0 on BTA1 (122 Mbp) and on BTA5 (82 Mbp) were confirmed by MIR-C14:0 with clear signals; in addition, some single SNP significant at a FDR of 5% were confirmed on BTA2 and 3. The QTL for GC-C14:0 on BTA27 (40 Mbp) was not detected with MIR-C14:0. The GWAS for MIR-C14:0 detected an additional QTL on BTA28 that was not significant with GC-C14:0.

The major QTL (FDR of 0.1%) for GC measured C16:0 (GC-C16:0) on BTA14 was confirmed by MIR-C16:0. In addition, the QTL for GC-C16:0 on BTA2 (64 Mbp) was confirmed by MIR-C16:0. The QTL for GC-C16:0 on BTA6 (4 Mbp) and BTA24 (8 Mbp) were not detected with MIR-C16:0. The GWAS based on MIR-C16:0 detected an additional QTL on BTA27 and a few SNP significant at a FDR of 5% on BTA2, 9, 12, 18, 23, and X.

The major QTL (FDR of 0.1%) for GC-C18:1 on BTA14 was confirmed by MIR-C18:1. In addition, the QTL for GC-C18:1 on BTA19 at 52 Mbp was confirmed by MIR-C18:1. The QTL for GC-C18:1 on BTA10 (126 Mbp) and BTA19 (37 Mbp) were not detected with MIR-C18:1. The GWAS for MIR-C18:1 detected a few SNP significant at a FDR of 5% on BTA1, 3, 5, 12, 15, 23, 26, and 28.

The major QTL (FDR of 0.1%) for GC-CLA on BTA14 was confirmed by MIR-CLA. Other than that QTL and a single significant SNP on BTA20 there were no significant regions for GC-CLA. The GWAS for MIR-CLA detected QTL on BTA4 (88 Mbp) and BTA6 (111 Mbp), GWAS for GC-CLA showed suggestive QTL for these regions just below the FDR threshold of 5% for GC-CLA. The GWAS for MIR-CLA detected additional QTL on BTA8 (67 Mbp), BTA19 (43-46 Mbp) and some single SNP significant at a FDR of 5% on BTA5, 9, 15, 20, 23, and 25.

In some cases the MIR predicted fatty acids resulted in QTL not observed with the GC measured fatty acids, e.g. the QTL on BTA19 for CLA. In other cases the QTL detected with GC measured fatty acids were not confirmed by MIR predicted fatty acids, e.g. the QTL on BTA26 for C14:0. The latter might be the result of reduced power of GWAS based on MIR data compared to GWAS based on GC data, because the predictions accuracies are lower than one. In the next section, I will discuss if reduced power of MIR predicted C14:0 can explain the failure to detect the major QTL for C14:0 on BTA26.

Power of GWAS

For QTL with major effect size it was expected they would be detected by GWAS with both GC and MIR data and, thus, would be robust against the prediction error of MIR predicted milk fatty acids. This expectation did not hold for C14:0, as three major QTL were detected based on GC data, while the GWAS for MIR predicted C14:0 detected only two of those major QTL (Figure 6.4). The major QTL on BTA26 was not detected using MIR predicted C14:0, but there was a suggestive QTL at approximately the same location. This suggests that the lack of confirmation of this QTL for C14:0 on BTA26 could be a power issue.

In practice many additional MIR predicted records are available which would resolve a power issue. However, in the current study no additional records were

available. Therefore, the GC dataset was reduced to mimic a situation where there are two or even four times as many MIR predicted records available as compared to GC records. At random 50% and 25% of the GC data was selected to perform a GWAS for C14:0 on BTA26. This was replicated 5 times for both 50% and 25% of the GC data. The average significance levels over the 5 replicates were used to compare with GWAS results of the full MIR predicted dataset.

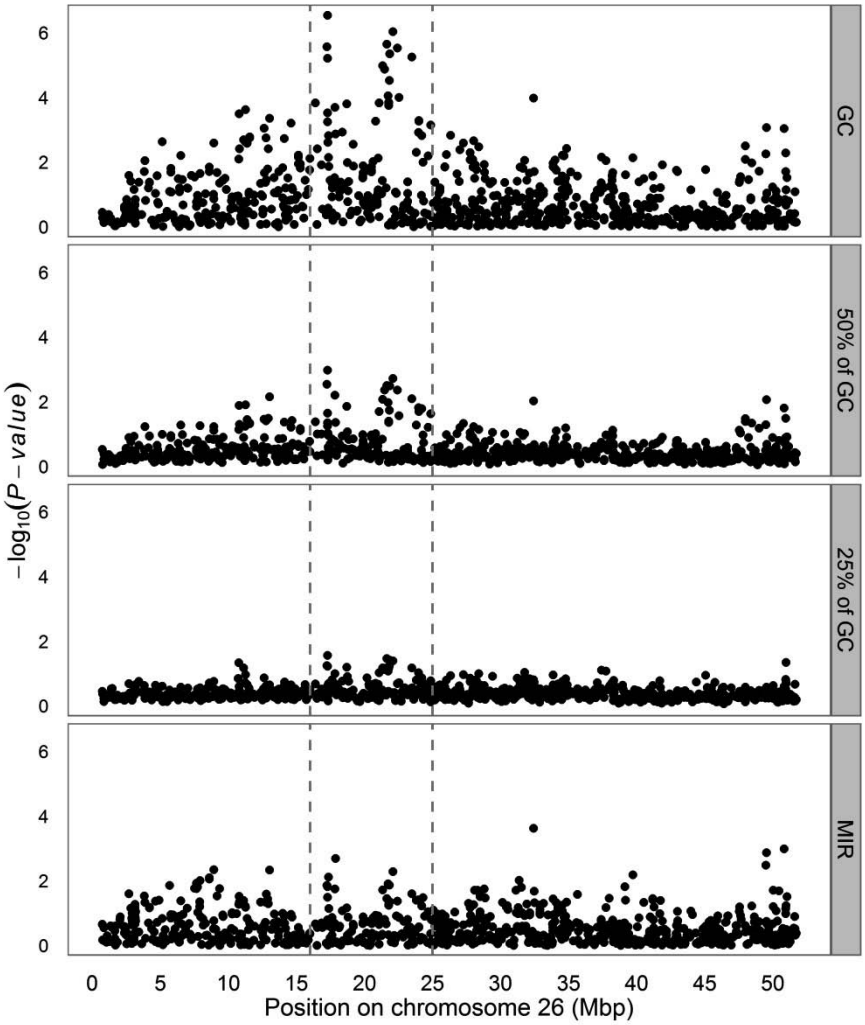


Figure 6.5 Manhattan plots for BTA26 comparing GWAS results for (100% of) the gas chromatography (GC) data, 50% of the GC data (average of 5 replicates) , 25% of the GC data (average of 5 replicates) and all the mid-infrared (MIR) data for C14:0. Region of QTL is indicated by the grey dashed lines.

The average significance over the 5 replicates with 50% of the GC data for C14:0 showed similar significance levels compared to the GWAS using the full MIR predicted dataset (Figure 6.5). However, the most significant SNP of the GWAS for the full GC dataset is not the same SNP. This SNP had only a $-\log_{10}(P\text{-value})$ of 1.81 in the GWAS for the full MIR predicted dataset, which is more similar to the significance level of the 5 replicates with 25% of the GC data (average $-\log_{10}(P\text{-value})$ = 1.56). These results suggest that using MIR predicted C14:0 severely reduces the power to detect this specific QTL. The loss in power is reflected in the prediction accuracy of 0.73 for C14:0, but shows to be more specific for this region instead of proportional over the whole genome. GWAS based on MIR predicted C14:0 did show a suggestive QTL on BTA26, indicating that MIR prediction captures at least part of the genetic variance of this major QTL. This indicates that for the GWAS design applied here two or even four times as many MIR predicted records are needed to obtain similar power as obtained with GC data for this particular QTL. Obtaining that many records would be feasible with routine measurement of the whole dairy cattle population. However, a more promising design can be applied too, by using daughter yield deviations or de-regressed proofs as phenotypes for sires in GWAS. Such phenotypes have a higher accuracy than own performance records used here, because they are based on observations on multiple daughters.

Such a design may average out extreme individual prediction errors over multiple daughters, but is probably not able to prevent the additional QTL observed here. In the next section, I will discuss a possible reason for the detection of additional QTL.

Additional QTL detect with MIR predicted fatty acids

The prediction accuracy of the milk fatty acids studied ranged from 0.58 to 0.84. Therefore, it was expected that some QTL would not be detected using MIR predicted data, but additional QTL were not expected. The additional QTL detected cannot be true QTL for the milk fatty acids studied, because the prediction equation is developed using GC data as 'true' phenotype for the same samples. For example, the QTL on BTA19 for MIR predicted CLA was very convincing, but not detected with GC measured CLA. Conjugated linoleic acid had relatively low prediction accuracy, relatively low genetic correlation between GC measured and MIR predicted records, and many individuals with large prediction errors (Table 6.3 and Figure 6.3). Regression of the MIR predicted records on GC measured records showed that the residual part (i.e. the prediction error) was associated with the significant SNP and not the part of the MIR prediction that actually explains the GC

measurement. So, MIR predicted fatty acids are probably associated with something else than the fatty acids it is supposed to represent.
Mid-infrared spectra represent resonance of all kinds of molecules, not only of fatty acids. Only certain wavelengths, selected e.g. by partial least squares, are used to predict a certain milk fatty acid. The predictions based on MIR spectra of milk samples could contain also information of other fatty acids and even of other milk components. Although speculative, this could imply that additional QTL detected using MIR data are QTL for other fatty acids or other milk components than the one studied. It is difficult to pinpoint where the associations exactly come from, but this complicates the search for candidate genes of the actual trait.

Phenotype definition for GWAS

Results presented here show that GC measurements and MIR predicted milk fatty acids are not the same trait but are correlated traits. Often different phenotypes are assumed to reflect the same trait. However, they are correlated traits. In human genetics, inconsistencies between phenotypes has been identified as one of the reasons why confirmation of initial results was unsuccessful (Chanock et al., 2007; König, 2011). Especially the misclassification of unaffected individuals as cases leads to extreme power reduction, while the misclassification of affected individuals as controls has little effect on the power (Edwards et al., 2005). For quantitative traits measurement errors have less extreme effects on power. In general, it is recognized that phenotypes are liable to measurement errors, but those errors are assumed to be random and, therefore, assumed to have no impact on GWAS results. Barendse (2011) showed that 2 independent measurements of back-fat thickness on the same carcasses, with the same equipment, within 24 hours, but by a different team, had a phenotypic correlation of 0.72 and gave different GWAS results. In their study, only 10% of the significant SNP were confirmed and a major QTL was shifted by 1 Mbp, which led to different candidate genes. An unambiguous phenotype description with a high repeatability is, therefore, very important for confirmation of QTL (Barendse, 2011; Edwards et al., 2005; Samuels et al., 2009).
In general, it has been recognized that GWAS need to be well-designed and for confirmation of QTL it is recognized that phenotypes need to be defined in the exact same way, because different definitions of phenotypes lead to different GWAS results, as is also the case with GC measured and MIR predicted milk fatty acids. It should be noted that this is only a requirement for confirmation and does not specify which phenotype should be studied in the original study. For example, GWAS results using MIR predicted fatty acid can be confirmed in a second GWAS

using the same Fourier-transform spectroscopy scanner and prediction equation, but might not lead to the same QTL using GC measurements, which are more likely to represent true associations due to more accurate phenotypes. The fact that different phenotype descriptions lead to different GWAS results is, however, troublesome and suggests that phenotype descriptions of initial GWAS should be chosen carefully when the aim is to search for causal variants that explain underlying biological mechanisms.

In dairy cattle we can see a trend towards predicted phenotypes, like the MIR predicted fatty acids, but also other milk components, cheese properties, methane emission and energy balance of cows can be predicted from MIR spectra (De Marchi et al., 2009; Dehareng et al., 2012; McParland et al., 2011; Rutten et al., 2009, 2011a, b; Soyeurt et al., 2006, 2007, 2009). This is an exciting development because the actual phenotypes are too difficult, expensive or invasive to record, and MIR spectra provide an opportunity for routine recording and breeding for predicted traits which are correlated to the actual phenotype of interest. Given that MIR profiles are already routinely recorded, MIR predicted milk fatty acids can be measured in large populations at low cost. The reduced power of using predicted phenotypes to detect QTL by GWAS can thus be overcome by using many more records for GWAS. Moreover, the routine MIR measurements on milk samples of cows can be used to produce de-regressed proofs or daughter yield deviations for genotyped sires. However, MIR predicted phenotypes add complexity to the genotype-phenotype relationship and, thus, enlarge the genotype-phenotype gap. Therefore, MIR predicted phenotypes are less appropriate to identify candidate genes and to infer the biological background of traits.

6.5 Conclusions

The genotype-phenotype gap can be narrowed by decomposing complex phenotypes into intermediate phenotypes and by inferring causal relationships among variables, whereas predicted phenotypes extend the genotype-phenotype gap further by adding more complexity.

References

References

Abe, T., J. Saburi, H. Hasebe, T. Nakagawa, T. Kawamura, K. Saito, T. Nade, S. Misumi, T. Okumura, K. Kuchida, T. Hayashi, S. Nakane, T. Mitsuhasi, K. Nirasawa, Y. Sugimoto, and E. Kobayashi. 2008. Bovine quantitative trait loci analysis for growth, carcass, and meat quality traits in an F2 population from a cross between Japanese Black and Limousin. Journal of Animal Science 86(11):2821-2832.

Abe, T., J. Saburi, H. Hasebe, T. Nakagawa, S. Misumi, T. Nade, H. Nakajima, N. Shoji, M. Kobayashi, and E. Kobayashi. 2009. Novel Mutations of the FASN Gene and Their Effect on Fatty Acid Composition in Japanese Black Beef. Biochemical Genetics 47(5):397-411.

Agarwal, A. K. and A. Garg. 2003. Congenital generalized lipodystrophy: significance of triglyceride biosynthetic pathways. Trends in Endocrinology & Metabolism 14(5):214-221.

Alexander, L. J., M. D. MacNeil, T. W. Geary, W. M. Snelling, D. C. Rule, and J. A. Scanga. 2007. Quantitative trait loci with additive effects on palatability and fatty acid composition of meat in a Wagyu–Limousin F2 population. Animal Genetics 38(5):506-513.

Almeida, M., P. Oliveira, T. Pereira, J. Krieger, and A. Pereira. 2011. An empirical evaluation of imputation accuracy for association statistics reveals increased type-I error rates in genome-wide associations. BMC Genetics 12(1):10.

Ashes, J. R., S. K. Gulati, and T. W. Scott. 1997. Potential to Alter the Content and Composition of Milk Fat Through Nutrition. J Dairy Sci 80(9):2204-2212.

Ashwell, M. S., D. W. Heyen, T. S. Sonstegard, C. P. Van Tassell, Y. Da, P. M. VanRaden, M. Ron, J. I. Weller, and H. A. Lewin. 2004. Detection of Quantitative Trait Loci Affecting Milk Production, Health, and Reproductive Traits in Holstein Cattle. J Dairy Sci 87(2):468-475.

Barendse, W. 2011. The effect of measurement error of phenotypes on genome wide association studies. BMC Genomics 12(1):232.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263-265.

Bastin, C., N. Gengler, and H. Soyeurt. 2011. Phenotypic and genetic variability of production traits and milk fatty acid contents across days in milk for Walloon Holstein first-parity cows. J Dairy Sci 94(8):4152-4163.

Bastin, C., N. Gengler, H. Soyeurt, S. McParland, E. Wall, and M. Calus. 2012. Genome-wide association study for milk fatty acid composition using cow versus bull data. in Proc. Proceeding of the 63rd Annual Meeting of the European Federation of Animal Science, Bratislava, Slovakia.

Beaulieu, A. D. and D. L. Palmquist. 1995. Differential Effects of High Fat Diets on Fatty Acid Composition in Milk of Jersey and Holstein Cows. J Dairy Sci 78(6):1336-1344.

Beavis, W. D. 1998. QTL analyses: power precision and accuracy. Pages 145-162 in Molecular dissection of complex traits. P. AH, ed. CRC Press, New York.

Bionaz, M. and J. Loor. 2008. Gene networks driving bovine milk fat synthesis during the lactation cycle. BMC Genomics 9(1):366.

Blott, S., J.-J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, D. Johnson, L. Karim, P. Simon, R. Snell, R. Spelman, J. Wong, J. Vilkki, M. Georges, F. Farnir, and W. Coppieters. 2003. Molecular Dissection of a Quantitative Trait Locus: A Phenylalanine-to-Tyrosine Substitution in the Transmembrane Domain of the Bovine Growth Hormone Receptor Is Associated With a Major Effect on Milk Yield and Composition. Genetics 163(1):253-266.

Bouwman, A., H. Bovenhuis, M. Visker, and J. Van Arendonk. 2011. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genetics 12(1):43.

Bouwman, A., M. H. Visker, J. A. van Arendonk, and H. Bovenhuis. 2012. Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. BMC Genetics 13(1):93.

Browning, B. L. and S. R. Browning. 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics 84(2):210-223.

Chaibub Neto, E., M. P. Keller, A. D. Attie, and B. S. Yandell. 2010. Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. Ann. Appl. Stat. 4(1):320-339.

Chanock, S. J., T. Manolio, M. Boehnke, E. Boerwinkle, D. J. Hunter, G. Thomas, J. N. Hirschhorn, A. Goncalo, D. Altshuler, J. E. Bailey-Wilson, L. D. Brooks, L. R. Cardon, M. J. Daly, Donnelly, J. F. J. Fraumeni, N. B. Freimer, D. S. Gerhard, C. Gunter, A. E. Guttmacher, M. S. Guyer, E. L. Harris, J. Hoh, R. Hoover, C. A. Kong, K. R. Merikangas, C. C. Morton, L. J. Palmer, E. G. Phimister, J. P. Rice, J. Roberts, C. N. Rotimi, M. A. Tucker, K. J. Vogan, S. Wacholder, E. M. Wijsman, D. M. Winn, and F. S. Collins. 2007. Replicating genotype–phenotype associations. Nature 447(7145):655-660.

Civelek, M. and A. J. Lusis. 2014. Systems genetics approaches to understand complex traits. Nat Rev Genet 15(1):34-48.

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, A. E.-v. d. Wind, J.-H. Lee, J. K. Drackley, M. R. Band, A. G. Hernandez, M. Shani, H. A. Lewin, J. I. Weller, and M. Ron. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Research 15(7):936-944.

Conte, G., M. Mele, S. Chessa, B. Castiglioni, A. Serra, G. Pagnacco, and P. Secchiari. 2010. Diacylglycerol acyltransferase 1, stearoyl-CoA desaturase 1, and sterol regulatory element binding protein 1 gene polymorphisms and milk fatty acid composition in Italian Brown cattle. J Dairy Sci 93(2):753-763.

Coppieters, W., J. Riquet, J.-J. Arranz, P. Berzi, N. Cambisano, B. Grisart, L. Karim, F. Marcq, L. Moreau, C. Nezer, P. Simon, P. Vanmanshoven, D. Wagenaar, and M. Georges. 1998. A QTL with major effect on milk yield and composition maps to bovine Chromosome 14. 9(7):540-544.

Daetwyler, H. D., F. S. Schenkel, M. Sargolzaei, and J. A. B. Robinson. 2008. A Genome Scan to Detect Quantitative Trait Loci for Economically Important Traits in Holstein Cattle Using Two Methods and a Dense Single Nucleotide Polymorphism Map. J Dairy Sci 91(8):3225-3236.

Dehareng, F., C. Delfosse, E. Froidmont, H. Soyeurt, C. Martin, N. Gengler, A. Vanlierde, and P. Dardenne. 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. animal 6(10):1694-1701.

de la Fuente, A., N. Bing, I. Hoeschele, and P. Mendes. 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20(18):3565-3574.

de los Campos, G., D. Gianola, and B. Heringstad. 2006. A Structural Equation Model for Describing Relationships Between Somatic Cell Score and Milk Yield in First-Lactation Dairy Cows. J Dairy Sci 89(11):4445-4455.

de Marchi, M., C. C. Fagan, C. P. O'Donnell, A. Cecchinato, R. Dal Zotto, M. Cassandro, M. Penasa, and G. Bittante. 2009. Prediction of coagulation properties, titratable acidity, and pH of bovine milk using mid-infrared spectroscopy. J Dairy Sci 92(1):423-432.

de Maturana, E. L., X.-L. Wu, D. Gianola, K. A. Weigel, and G. J. M. Rosa. 2009. Exploring Biological Relationships Between Calving Traits in Primiparous Cattle with a Bayesian Recursive Model. Genetics 181(1):277-287.

DePeters, E. J., J. F. Medrano, and B. A. Reed. 1995. Fatty acid composition of milk fat from three breeds of dairy cattle. Canadian Journal of Animal Science 75(2):267-269.

Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume, D. Derbala, D. Zelenika, D. Lechner, C. Charon, D. Boichard, I. G. Gut, A. Eggen, and M. Gautier. 2008. Fine Mapping of Quantitative Trait Loci Affecting Female Fertility in Dairy Cattle on BTA03 Using a Dense Single-Nucleotide Polymorphism Map. Genetics 178(4):2227-2235.

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

Durinck, S. and W. Huber. 2012. biomaRt: Interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene). version 2.6.0. ed. R package

Edwards, B., C. Haynes, M. Levenstien, S. Finch, and D. Gordon. 2005. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genetics 6(1):18.

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(7):4114-4129.

Farnir, F., B. Grisart, W. Coppieters, J. Riquet, P. Berzi, N. Cambisano, L. Karim, M. Mni, S. Moisio, P. Simon, D. Wagenaar, J. Vilkki, and M. Georges. 2002. Simultaneous Mining of Linkage and Linkage Disequilibrium to Fine Map Quantitative Trait Loci in Outbred Half-Sib Pedigrees: Revisiting the Location of a Quantitative Trait Locus With Major Effect on Milk Production on Bovine Chromosome 14. Genetics 161(1):275-287.

Fievez, V., B. Vlaeminck, M. S. Dhanoa, and R. J. Dewhurst. 2003. Use of principal component analysis to investigate the origin of heptadecenoic and conjugated linoleic acids in milk. J Dairy Sci 86(12):4047-4053.

Gautier, M., R. R. Barcelona, S. Fritz, C. Grohs, T. Druet, D. Boichard, A. Eggen, and T. H. E. Meuwissen. 2006. Fine Mapping and Physical Characterization of Two Linked Quantitative Trait Loci Affecting Milk Fat Yield in Dairy Cattle on BTA26. Genetics 172(1):425-436.

German, J. B. and C. J. Dillard. 2006. Composition, Structure and Absorption of Milk Lipids: A Source of Energy, Fat-Soluble Nutrients and Bioactive Molecules. Critical Reviews in Food Science and Nutrition 46(1):57-92.

Gianola, D. and D. Sorensen. 2004. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. Genetics 167(3):1407-1424.

Gilmour, A. R., Gogel, B.J., Cullis, B.R., and Thompson, R. . 2006. ASReml User Guide Release 2.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Goddard, M. E. and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10(6):381-391.

González, J. R., L. Armengol, X. Solé, E. Guinó, J. M. Mercader, X. Estivill, and V. Moreno. 2007. SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23(5):654-655.

Govignon-Gion, A., S. Fritz, H. Larroque, M. Brochard, C. Chantry, F. Lahalle, and D. Boichard. 2012. Detection of QTL affecting milk fatty acid composition in three French dairy cattle breeds. in Proc. 63rd Annual Meeting of the European Federation of Animal Science, Bratislava, Slovakia.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Research 12(2):222-231.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J.-J. Kim, A. Kvasz, M. Mni, P. Simon, J.-M. Frère, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proceedings of the National Academy of Sciences of the United States of America 101(8):2398-2403.

Guan, Y. and M. Stephens. 2008. Practical Issues in Imputation-Based Association Mapping. PLoS Genet 4(12):e1000279.

Haavelmo, T. 1943. The Statistical Implications of a System of Simultaneous Equations. Econometrica 11(1):1-12.

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

Heck, J. M., H. J. van Valenberg, H. Bovenhuis, J. Dijkstra, and T. C. van Hooijdonk. 2012. Characterization of milk fatty acids based on genetic and herd parameters. Journal of Dairy Research 79(01):39-46.

Henderson, C. R. and R. L. Quaas. 1976. Multiple trait evaluation using relatives' records. Journal of Animal Science 43(6):1188-1197.

Heyen, D. W., J. I. Weller, M. Ron, M. Band, J. E. Beever, E. Feldmesser, Y. Da, G. R. Wiggans, P. M. VanRaden, and H. A. Lewin. 1999. A genome scan for QTL influencing milk production and health traits in dairy cattle. Physiological Genomics 1(3):165-175.

Hirschhorn, J. N. and M. J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6(2):95-108.

Houle, D., D. R. Govindaraju, and S. Omholt. 2010. Phenomics: the next challenge. Nat Rev Genet 11(12):855-866.

Hu, Z.-L., C. A. Park, E. R. Fritz, and J. M. Reecy. 2010. QTLdb: a comprehensive database tool building bridges between genotypes and phenotypes. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production. Leipzig, Germany.

Igl, B. W., I. R. König, and A. Ziegler. 2009. What Do We Mean by 'Replication' and 'Validation' in Genome-Wide Association Studies? Hum Hered 67(1):66-68.

Ishii, A., K. Yamaji, Y. Uemoto, N. Sasago, E. Kobayashi, N. Kobayashi, T. Matsuhashi, S. Maruyama, H. Matsumoto, S. Sasazaki, and H. Mannen. 2013. Genome-wide association study for fatty acid composition in Japanese Black cattle. Animal Science Journal:n/a-n/a.

Jamrozik, J., J. Bohmanova, and L. R. Schaeffer. 2010. Relationships between milk yield and somatic cell score in Canadian Holsteins from simultaneous and recursive random regression models. J Dairy Sci 93(3):1216-1233.

Janss, L. 2010. Bayz manual. Janss Biostatistics, Leiden, The Netherlands.

Jenkins, T. C. 1993. Lipid Metabolism in the Rumen. J Dairy Sci 76(12):3851-3863.

Jensen, R. G. 2002. The Composition of Bovine Milk Lipids: January 1995 to December 2000. J Dairy Sci 85(2):295-350.

Karijord, Ø., N. Standal, and O. Syrstad. 1982. Sources of variation in composition of milk fat. Zeitschrift für Tierzüchtung und Züchtungsbiologie 99(1-4):81-93.

Karim, L., H. Takeda, L. Lin, T. Druet, J. A. C. Arias, D. Baurain, N. Cambisano, S. R. Davis, F. Farnir, B. Grisart, B. L. Harris, M. D. Keehan, M. D. Littlejohn, R. J. Spelman, M. Georges, and W. Coppieters. 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nature Genetics 43(5):405-413.

Kelsey, J. A., B. A. Corl, R. J. Collier, and D. E. Bauman. 2003. The Effect of Breed, Parity, and Stage of Lactation on Conjugated Linoleic Acid (CLA) in Milk Fat from Dairy Cows. J Dairy Sci 86(8):2588-2597.

Kennedy, B. W., M. Quinton, and J. A. van Arendonk. 1992. Estimation of effects of single genes on quantitative traits. Journal of Animal Science 70(7):2000-2012.

Kgwatalala, P. M., E. M. Ibeagha-Awemu, J. F. Hayes, and X. Zhao. 2009. Stearoyl-CoA desaturase 1 3'UTR SNPs and their influence on milk fatty acid composition of Canadian Holstein cows. Journal of Animal Breeding and Genetics 126(5):394-403.

Khatib, H., I. Zaitoun, J. Wiebelhaus-Finger, Y. M. Chang, and G. J. M. Rosa. 2007. The Association of Bovine PPARGC1A and OPN Genes with Milk Composition in Two Independent Holstein Cattle Populations. J Dairy Sci 90(6):2966-2970.

Khatkar, M., P. Thomson, I. Tammen, and H. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genetics Selection Evolution 36(2):1-28.

Kim, E.-S., X. Shi, O. Cobanoglu, K. Weigel, P. J. Berger, and B. W. Kirkpatrick. 2009. Refined mapping of twinning-rate quantitative trait loci on bovine chromosome 5 and analysis of insulin-like growth factor-1 as a positional candidate gene. Journal of Animal Science 87(3):835-843.

König, I. R. 2011. Validation in Genetic Association Studies. Briefings in Bioinformatics 12(3):253-258.

König, S., X. L. Wu, D. Gianola, B. Heringstad, and H. Simianer. 2008. Exploration of Relationships Between Claw Disorders and Milk Yield in Holstein Cows via Recursive Linear and Threshold Models. J Dairy Sci 91(1):395-406.

Lawless, F., C. Stanton, P. L'Escop, R. Devery, P. Dillon, and J. J. Murphy. 1999. Influence of breed on bovine milk cis-9, trans-11-conjugated linoleic acid content. Livestock Production Science 62(1):43-49.

Li, C., N. Aldai, M. Vinsky, M. E. R. Dugan, and T. A. McAllister. 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):93-97.

Li, R., S-W Tsaih, K. Shockley, I. M. Stylianou, J. Wergedal, B. Paigen, and G. A. Churchill. 2006. Structural model analysis of multiple quantitative traits. PLoS Genet 2(7):e114.

Lillehammer, M., B. J. Hayes, T. H. E. Meuwissen, and M. E. Goddard. 2009. Gene by environment interactions for production traits in Australian dairy cattle. J Dairy Sci 92(8):4008-4017.

Liu, B., A. de la Fuente, and I. Hoeschele. 2008a. Gene network inference via structural equation modeling in genetical genomics experiments. Genetics 178(3):1763-1776.

Liu, Y.-J., C. J. Papasian, J.-F. Liu, J. Hamilton, and H.-W. Deng. 2008b. Is Replication the Gold Standard for Validating Genome-Wide Association Findings? PLoS ONE 3(12):e4037.

Liu, Y., X. Qin, X.-Z. Song, H. Jiang, Y. Shen, K. J. Durbin, S. Lien, M. Kent, M. Sodeland, Y. Ren, L. Zhang, E. Sodergren, P. Havlak, K. Worley, G. Weinstock, and R. Gibbs. 2009. Bos taurus genome assembly. BMC Genomics 10(1):180.

MacGibbon, A. K. H. and M. W. Taylor. 2006. Composition and Structure of Bovine Milk Lipids. Pages 1-42 in Advanced Dairy Chemistry Volume 2 Lipids. P. F. Fox and P. L. H. McSweeney, ed. Springer US.

Mackay, T. F. C. 2001. Quantitative trait loci in Drosophila. Nature Reviews Genetics 2:11-20.

Mai, M. D., G. Sahana, F. B. Christiansen, and B. Guldbrandtsen. 2010. A genome-wide association study for milk production traits in Danish Jersey cattle using a 50K single nucleotide polymorphism chip. Journal of Animal Science 88(11):3522-3528.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-753.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39(7):906-913.

Marchini, J. and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11(7):499-511.

Massart-Leën, A. M. and D. L. Massart. 1981. The use of clustering techniques in the elucidation or confirmation of metabolic pathways. Application to the branched-chain fatty acids present in the milk fat of lactating goats. . Biochem. J. 196(2):611-618.

Maurice-Van Eijndhoven, M. H. T., S. J. Hiemstra, and M. P. L. Calus. 2011. Short communication: Milk fat composition of 4 cattle breeds in the Netherlands. J Dairy Sci 94(2):1021-1025.

McParland, S., G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J Dairy Sci 94(7):3651-3661.

Medrano, J., G. Rincon, and I.-T. A. 2010. Comparative analysis of bovine milk and mammary gland transcriptome using RNA-Seq. in Proc. 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany.

Mele, M., G. Conte, B. Castiglioni, S. Chessa, N. P. P. Macciotta, A. Serra, A. Buccioni, G. Pagnacco, and P. Secchiari. 2007. Stearoyl-Coenzyme A Desaturase Gene Polymorphism and Milk Fatty Acid Composition in Italian Holsteins. J Dairy Sci 90(9):4458-4465.

Mele, M., R. Dal Zotto, M. Cassandro, G. Conte, A. Serra, A. Buccioni, G. Bittante, and P. Secchiari. 2009. Genetic parameters for conjugated linoleic acid, selected milk fatty acids, and milk fatty acid unsaturation of Italian Holstein-Friesian cows. J Dairy Sci 92(1):392-400.

Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard. 2002. Fine Mapping of a Quantitative Trait Locus for Twinning Rate Using Combined Linkage and Linkage Disequilibrium Mapping. Genetics 161(1):373-379.

Meyer, K. and M. Kirkpatrick. 2008. Perils of Parsimony: Properties of Reduced-Rank Estimates of Genetic Covariance Matrices. Genetics 180(2):1153-1166.

Moioli, B., G. Contarini, A. Avalli, G. Catillo, L. Orrù, G. De Matteis, G. Masoero, and F. Napolitano. 2007. Short Communication: Effect of Stearoyl-Coenzyme A Desaturase Polymorphism on Fatty Acid Composition of Milk. J Dairy Sci 90(7):3553-3558.

Morris, C., N. Cullen, B. Glass, D. Hyndman, T. Manley, S. Hickey, J. McEwan, W. Pitchford, C. Bottema, and M. Lee. 2007. Fatty acid synthase effects on bovine adipose fat and milk fat. Mammalian Genome 18(1):64-74.

Morris, C. A., C. D. K. Bottema, N. G. Cullen, S. M. Hickey, A. K. Esmailizadeh, B. D. Siebert, and W. S. Pitchford. 2010. Quantitative trait loci for organ weights and adipose fat composition in Jersey and Limousin back-cross cattle finished on pasture or feedlot. Animal Genetics 41(6):589-596.

Nafikov, R. A., J. P. Schoonmaker, K. T. Korn, K. Noack, D. J. Garrick, K. J. Koehler, J. Minick-Bormann, J. M. Reecy, D. E. Spurlock, and D. C. Beitz. 2013. Sterol regulatory element binding transcription factor 1 (SREBF1) polymorphism and milk fatty acid composition. J Dairy Sci 96(4):2605-2616.

Neville, M. C. and M. F. Picciano. 1997. Regulation of milk lipid secretion and composition. Annual Review of Nutrition 17(1):159-184.

Oh, D., Y. Lee, B. La, J. Yeo, E. Chung, Y. Kim, and C. Lee. 2012. Fatty acid composition of beef is associated with exonic nucleotide variants of the gene encoding FASN. Molecular Biology Reports 39(4):4083-4090.

Olsen, H., H. Nilsen, B. Hayes, P. Berg, M. Svendsen, S. Lien, and T. Meuwissen. 2007. Genetic support for a quantitative trait nucleotide in the ABCG2 gene affecting milk composition of dairy cattle. BMC Genetics 8(1):32.

Ordovás, L., R. Roy, S. Pampín, P. Zaragoza, R. Osta, J. C. Rodríguez-Rey, and C. Rodellar. 2008. The g.763G>C SNP of the bovine FASN gene affects its promoter activity via Sp-mediated regulation: implications for the bovine lactating mammary gland. Physiological Genomics 34(2):144-148.

Palmquist, D. L., A. Denise Beaulieu, and D. M. Barbano. 1993. Feed and Animal Factors Influencing Milk Fat Composition. J Dairy Sci 76(6):1753-1771.

Palmquist, D. L. 2006. Milk Fat: Origin of Fatty Acids and Influence of Nutritional Factors Thereon. Pages 43-92 in Advanced Dairy Chemistry Volume 2 Lipids. P. F. Fox and P. L. H. McSweeney, ed. Springer US.

Pausch, H., B. Aigner, R. Emmerling, C. Edel, K.-U. Gotz, and R. Fries. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genetics Selection Evolution 45(1):3.

Pearl, J. 2009. Causality: Models, Reasoning and Inference. Cambridge University Press.

Pereira, S. L., A. E. Leonard, and P. Mukerji. 2003. Recent advances in the study of fatty acid desaturases from animals and lower eukaryotes. Prostaglandins, Leukotrienes and Essential Fatty Acids 68(2):97-106.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News 6(1):7-11.

Pryce, J. E., S. Bolormaa, A. J. Chamberlain, P. J. Bowman, K. Savin, M. E. Goddard, and B. J. Hayes. 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J Dairy Sci 93(7):3331-3345.

Riquet, J., W. Coppieters, N. Cambisano, J.-J. Arranz, P. Berzi, S. K. Davis, B. Grisart, F. Farnir, L. Karim, M. Mni, P. Simon, J. F. Taylor, P. Vanmanshoven, D. Wagenaar, J. E. Womack, and M. Georges. 1999. Fine-mapping of quantitative trait loci by identity by descent in outbred populations: Application to milk production in dairy cattle. Proceedings of the National Academy of Sciences 96(16):9252-9257.

Ron, M. and J. I. Weller. 2007. From QTL to QTN identification in livestock – winning by points rather than knock-out: a review. Animal Genetics 38(5):429-439.

Roy, R., L. Ordovas, P. Zaragoza, A. Romero, C. Moreno, J. Altarriba, and C. Rodellar. 2006. Association of polymorphisms in the bovine FASN gene with milk-fat content. Animal Genetics 37(3):215-218.

Rutten, M. J. M., H. Bovenhuis, K. A. Hettinga, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J Dairy Sci 92(12):6202-620.

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

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011b. Prediction of β-lactoglobulin genotypes based on milk Fourier transform infrared spectra. J Dairy Sci 94(8):4183-4188.

Samuels, D. C., D. J. Burn, and P. F. Chinnery. 2009. Detecting new neurodegenerative disease genes: does phenotype accuracy limit the horizon? Trends in Genetics 25(11):486-488.

Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. 37(7):710-717.

Schennink, A., W. M. Stoop, M. H. P. W. Visker, J. M. L. Heck, H. Bovenhuis, J. J. Van Der Poel, H. J. F. Van Valenberg, and J. A. M. Van Arendonk. 2007. DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. Animal Genetics 38(5):467-473.

Schennink, A., J. M. L. Heck, H. Bovenhuis, M. H. P. W. Visker, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2008. Milk Fatty Acid Unsaturation: Genetic Parameters and Effects of Stearoyl-CoA Desaturase (SCD1) and Acyl CoA: Diacylglycerol Acyltransferase 1 (DGAT1). J Dairy Sci 91(5):2135-2143.

Schennink, A., H. Bovenhuis, K. M. Léon-Kloosterziel, J. A. M. Van Arendonk, and M. H. P. W. Visker. 2009a. Effect of polymorphisms in the FASN, OLR1, PPARGC1A, PRL and STAT5A genes on bovine milk-fat composition. Animal Genetics 40(6):909-916.

Schennink, A., W. M. Stoop, M. H. P. W. Visker, J. J. van der Poel, H. Bovenhuis, and J. A. M. van Arendonk. 2009b. Short communication: Genome-wide scan for bovine milk-fat composition. II. Quantitative trait loci for long-chain fatty acids. J Dairy Sci 92(9):4676-4682.

Schrooten, C., R. Dassonneville, R. Brondum, J. Chen, Z. Liu, and T. Druet. 2012. Error rate for imputation from BovineSNP50 to BovineHD. in Proc. 63rd Annual Meeting of the European Federation of Animal Science (EAAP), Bratislava, Slovakia.

Schumacker, R. and G. Marcoulides. 1998. Interaction and nonlinear effects in structural equation modeling. Lawrence Erlbaum Associates, Inc., Publishers, Mahwah, New Jersey, USA.

Shipley, B. 2002. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge, UK.

Šidák, Z. 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association 62(318):626-633.

Smith, S. 1994. The animal fatty acid synthase: one gene, one polypeptide, seven enzymes. The FASEB Journal 8(15):1248-1259.

Smith, S., A. Witkowski, and A. K. Joshi. 2003. Structural and functional organization of the animal fatty acid synthase. Progress in Lipid Research 42(4):289-317.

Sørensen, L., L. Janss, P. Madsen, T. Mark, and M. Lund. 2012. Estimation of (co)variances for genomic regions of flexible sizes: application to complex infectious udder diseases in dairy cattle. Genetics Selection Evolution 44(1):1-15.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating Fatty Acid Content in Cow Milk Using Mid-Infrared Spectrometry. J Dairy Sci 89(9):3690-3695.

Soyeurt, H., A. Gillon, S. Vanderick, P. Mayeres, C. Bertozzi, and N. Gengler. 2007. Estimation of Heritability and Genetic Correlations for the Major Fatty Acids in Bovine Milk. J Dairy Sci 90(9):4435-4442.

Soyeurt, H., D. Bruwier, J. M. Romnee, N. Gengler, C. Bertozzi, D. Veselko, and P. Dardenne. 2009. Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. J Dairy Sci 92(6):2444-2454.

Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini. 2009. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. PLoS Genet 5(5):e1000477.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64(4):583-639.

Spirtes, P. 1993. Directed cyclic graphs, conditional independence, and non-recursive linear structural equation models. Department of Philosophy Technical Report CMU-Phil-35.

Spirtes, P., C. Glymour, and R. Scheines. 2000. Causation, Prediction, and Search. 2 ed. Massachusetts Institute of Technology, Cambridge, MA.

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(1):385-394.

Stoop, W. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2009a. Effect of lactation stage and energy status on milk fat composition of Holstein-Friesian cows. J Dairy Sci 92(4):1469-1478.

Stoop, W. M., A. Schennink, M. H. P. W. Visker, E. Mullaart, J. A. M. van Arendonk, and H. Bovenhuis. 2009b. Genome-wide scan for bovine milk-fat composition. I. Quantitative trait loci for short- and medium-chain fatty acids. J Dairy Sci 92(9):4664-4675.

Storey, J. D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100(16):9440-9445.

Stull, J. W. and W. H. Brown. 1964. Fatty Acid Composition of Milk. II. Some Differences in Common Dairy Breeds1. J Dairy Sci 47(12):1412.

Takeuchi, K. and K. Reue. 2009. Biochemistry, physiology, and genetics of GPAT, AGPAT, and lipin enzymes in triglyceride synthesis. American Journal of Physiology - Endocrinology and Metabolism 296(6):E1195-E1209.

Taniguchi, M., T. Utsugi, K. Oyama, H. Mannen, M. Kobayashi, Y. Tanabe, A. Ogino, and S. Tsuji. 2004. Genotype of stearoyl-CoA desaturase is associated with fatty acid composition in Japanese Black cattle. Mammalian Genome 15(2):142-148.

Uemoto, Y., T. Abe, N. Tameoka, H. Hasebe, K. Inoue, H. Nakajima, N. Shoji, M. Kobayashi, and E. Kobayashi. 2011. Whole-genome association study for fatty acid composition of oleic acid in Japanese Black cattle. Animal Genetics 42(2):141-148.

Valente, B. D., G. J. M. Rosa, G. de los Campos, D. Gianola, and M. A. Silva. 2010. Searching for recursive causal structures in multivariate quantitative genetics mixed models. Genetics 185(2):633-644.

Valente, B. D., G. J. M. Rosa, M. A. Silva, R. Teixeira, and R. Torres. 2011. Searching for phenotypic causal networks involving complex traits: an application to European quail. Genetics Selection Evolution 43(1):1-12.

Valente, B. D., G. J. M. Rosa, D. Gianola, X.-L. Wu, and K. A. Weigel. 2013. Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? Genetics.

van den Oord, E. J. C. G. 2008. Controlling false discoveries in genetic studies. Am. J. Med. Genet. 147B(5):637-644.

van der Sluis, S., M. Verhage, D. Posthuma, and C. V. Dolan. 2010. Phenotypic Complexity, Measurement Bias, and Poor Phenotypic Resolution Contribute to the Missing Heritability Problem in Genetic Association Studies. PLoS ONE 5(11):e13929.

Vergnes, L., A. P. Beigneux, R. Davis, S. M. Watkins, S. G. Young, and K. Reue. 2006. Agpat6 deficiency causes subdermal lipodystrophy and resistance to obesity. Journal of Lipid Research 47(4):745-754.

Verma, T. and J. Pearl. 1990. Equivalence and synthesis of causal models. Pages 255-268 in Proc. Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, Cambridge, USA.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57(2):307-333.

Wang, X., C. Wurmser, H. Pausch, S. Jung, F. Reinhardt, J. Tetens, G. Thaller, and R. Fries. 2012. Identification and Dissection of Four Major QTL Affecting Milk Fat Content in the German Holstein-Friesian Population. PLoS ONE 7(7):e40711.

Weikard, R., C. Kühn, T. Goldammer, G. Freyer, and M. Schwerin. 2005. The bovine PPARGC1A gene: molecular characterization and association of an SNP with variation of milk fat synthesis. Physiological Genomics 21(1):1-13.

Weller, J. I. and M. Ron. 2011. Invited review: Quantitative trait nucleotide determination in the era of genomic selection. J Dairy Sci 94(3):1082-1090.

Wibowo, T., C. Gaskins, R. Newberry, G. Thorgaard, J. Michal, and Z. Jiang. 2008. Genome Assembly Anchored QTL Map of Bovine Chromosome 14. Int J Biol Sci 4(6):406-414.

Wilmink, J. B. M. 1987. Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. Livestock Production Science 16(4):335-348.

Wright, S. 1921. Correlation and Causation. Journal of Agricultural Research 20:557-585.

Wu, X. L., B. Heringstad, Y. M. Chang, G. de los Campos, and D. Gianola. 2007. Inferring Relationships Between Somatic Cell Score and Milk Yield Using Simultaneous and Recursive Models. J Dairy Sci 90(7):3508-3521.

Yao, J., S. E. Aggrey, D. Zadworny, J. F. Hayes, and U. Kühnlein. 1996. Sequence Variations in the Bovine Growth Hormone Gene Characterized by Single-Strand Conformation Polymorphism (SSCP) Analysis and Their Association with Milk Production Traits in Holsteins. Genetics 144(4):1809-1816.

Zhang, S., T. J. Knight, J. M. Reecy, and D. C. Beitz. 2008. DNA polymorphisms in bovine fatty acid synthase are associated with beef fatty acid composition1. Animal Genetics 39(1):62-70.

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

Zou, X., J. Huang, Q. Jin, Z. Guo, Y. Liu, L. Cheong, X. Xu, and X. Wang. 2013. Lipid Composition Analysis of Milk Fats from Different Mammalian Species: Potential for Use as Human Milk Fat Substitutes. Journal of Agricultural and Food Chemistry 61(29):7070-7080.

Summary

Summary

This thesis describes the research on the genetic background of bovine milk fat composition that aimed to detect, confirm and fine-map quantitative trait loci (QTL) for milk fatty acids in Dutch Holstein Friesian cattle. Identification of genomic regions, and preferably individual genes, responsible for genetic variation in bovine milk fat composition enhances the understanding of biological pathways involved in fatty acid synthesis and is expected to increase opportunities for changing bovine milk fat composition by means of selective breeding.

Chapter 2 describes a genome-wide association study (GWAS) using 50,000 single nucleotide polymorphisms (SNP) for even-chain saturated fatty acids (C4:0-C18:0), even-chain *cis9* monounsaturated fatty acids (C10:1-C18:1), and the polyunsaturated C18:2*cis9,trans11* (Conjugated linoleic acid; CLA). Fatty acids were measured by gas chromatography on approximately 2,000 dairy cows in winter milk samples. A total of 54 regions on 29 chromosomes were significantly associated with one or more fatty acids. Bos taurus autosomes (BTA) 14, 19, and 26 showed highly significant associations with seven to ten fatty acids each, explaining a relatively large percentage of the total additive genetic variation. Many additional regions were significantly associated with fatty acids. Some of the associated regions harbor genes that are known to be involved in fat synthesis or have previously been identified as underlying QTL for fat yield or content.

The GWAS for fatty acids from winter milk was followed up by a GWAS for fatty acids from summer milk samples from the same cows and is described in chapter 3. The GWAS for summer milk samples resulted in 51 regions on 24 chromosomes that were significantly associated with one or more milk fatty acids. Results from the GWAS for summer milk fatty acids were in agreement with most associations that were previously detected in the GWAS for fatty acids from winter milk samples. For SNP that were found significant in both GWAS high correlations were found between the levels of significance ($-\log_{10}$(*P*-values)) in winter and summer as well as between SNP effects in both seasons. This implies that the effects of the SNP were similar on winter and summer milk fatty acids. Associations that were in agreement between both GWAS are more likely to be true compared to regions detected in only one GWAS and are, therefore, worthwhile to pursue in fine-mapping studies.

Chapter 4 describes how the location of the QTL detected in chapters 2 and 3 on BTA19 was refined using a denser set of markers. Opportunities for fine mapping were provided by imputation from 50,000 genotyped SNP to a high density SNP panel with 777,000 SNP. The QTL region was narrowed down to a linkage

disequilibrium block formed by 22 SNP covering 85,007 bp, from 51,303,322 bp to 51,388,329 bp on BTA19. This linkage disequilibrium block contained 2 genes: coiled-coil domain containing 57 (*CCDC57*) and fatty acid synthase (*FASN*). There is not much known about the function of *CCDC57* and this gene has not been associated with bovine milk fat previously, but is expressed in the mammary gland. The gene *FASN* has been associated with bovine milk fat and fat in adipose tissue in other studies. This gene is a likely candidate for the QTL on BTA19 because of its involvement in de novo fat synthesis.

Chapter 5 describes the use of causal inference to provide more insight into the relationships between individual fatty acids. The aim of this study was to find the causal network best supported by the milk fatty acid data and to use this in a structural equation model (SEM). An inductive causation (IC) algorithm can be used to search for causal structures. The IC algorithm adapted to mixed models settings was applied to study 14 correlated bovine milk fatty acids, resulting in an undirected network. The undirected pathway from C4:0 to C12:0 resembled the de novo synthesis pathway of short and medium-chain saturated fatty acids. By using prior knowledge, directions were assigned to that part of the network. The resulting causal structure was used as condition for fitting a SEM. Structural equation models can be used to quantify causal relations between traits and allows prediction of outcomes of interventions applied to such a network. Structural coefficients ranged from 0.85 to 1.05. The deviance information criterion indicated that the SEM was more plausible than a standard multi-trait model for these milk fatty acids. This changed the focus from marginal associations between traits to direct relationships, thus, towards relationships that may result in changes when external interventions are applied.

Chapter 6 is the general discussion. The first part addresses the insights that were gained from GWAS for milk fatty acids. Moving from linkage analysis toward GWAS confirmed and refined the size of QTL regions and resulted in new QTL regions. Genome-wide association studies provide a picture of the genetic architecture of a trait by scanning the whole genome. This is preferred over candidate gene studies that may lead to wrong conclusions because of associations due to linkage disequilibrium.

The second part of the general discussion concentrates on intermediate phenotypes. Performing GWAS based on individual fatty acids resulted in additional QTL as compared to GWAS based on fat percentage or yield. This shows that refinement of complex phenotypes into underlying components results in better

links between genes and phenotypes, and is a way to fill part of the so called genotype-phenotype gap.

The next section of the general discussion deals with causal relationships between the QTL on BTA19 and the de novo fatty acids. For C4:0, C6:0, C8:0, C10:0 and C14:0 there is evidence for a direct effect of the QTL on BTA19. The QTL indirectly affects C12:0 through its effect on C10:0. Furthermore, analyses show that the QTL has no effect on C16:0.

The last part of the general discussion is about prediction of milk fat composition with mid-infrared (MIR) spectra as alternative for gas chromatography (GC) measurements. Mid-infrared predicted fatty acids are cheap as compared to the expensive and time consuming GC measurements. The potential of GWAS based on MIR predicted fatty acids was explored by comparing GWAS results for GC measured and MIR predicted milk fatty acids on the same milk samples. Part of the QTL detected based on GC were not detected using MIR predicted fatty acids due to reduced power. The loss in power is reflected in the accuracy of predicting milk fatty acids based on MIR. The GWAS for MIR predicted phenotypes resulted also in QTL that were not detected based on GC measurements. These QTL probably do not reflect true QTL for the milk fatty acids studied but may relate to other milk components. Therefore, MIR predicted phenotypes add complexity to the genotype-phenotype relationship, and renders MIR predicted phenotypes less appropriate to identify candidate genes and to infer the biological background of traits.

Samenvatting

Samenvatting

Dit proefschrift beschrijft mijn onderzoek naar de genetische achtergrond van de samenstelling van het vet in koeienmelk. Het doel van het onderzoek was het opsporen van gebieden in het DNA die verantwoordelijk zijn voor verschillen tussen koeien in de samenstelling van het melkvet. Deze DNA gebieden zijn daarna bevestigd en de locaties zijn verfijnd om genen betrokken bij de synthese van melkvet te identificeren. Deze identificatie van DNA gebieden en onderliggende genen die verantwoordelijk zijn voor verschillen in vetsamenstelling van koeienmelk is belangrijk voor het inzicht in de biologische mechanismen van de synthese van melkvet. Bovendien vergroot het de mogelijkheden voor het veranderen van de melkvetsamenstelling door middel van fokkerij.

Voor dit onderzoek zijn van ongeveer 2.000 zwartbonte Holstein-Friesian koeien gegevens verzameld. De samenstelling van het vet is gemeten met behulp van gas chromatografie in zowel winter- als zomermelk. We hebben de 14 vetzuren met de hoogste concentraties in melkvet bestudeerd. Dit zijn de verzadigde vetzuren C4:0, C6:0, C8:0, C10:0, C12:0, C14:0, C16:0, en C18:0, de enkelvoudig onverzadigde vetzuren C10:1, C12:1, C14:1, C16:1 en C18:1, en het meervoudig onverzadigde C18:2*cis9,trans11* (Geconjugeerd linolzuur; CLA). Daarnaast zijn van de koeien ook DNA profielen bestaande uit 50.000 DNA merkers bepaald.

Hoofdstuk 2 beschrijft een genoom-wijde associatie studie (GWAS) voor de samenstelling van het vet in wintermelk. Hierbij is per DNA merker gekeken of deze significant geassocieerd is met de individuele melkvetzuren. In totaal waren 54 DNA gebieden op 29 chromosomen significant geassocieerd met een of meerdere vetzuren. De gebieden op de chromosomen 14, 19 en 26 toonden zeer significante associaties met zeven tot tien vetzuren elk, en verklaarden een relatief groot deel van de genetische verschillen tussen de koeien. Daarnaast waren vele andere DNA gebieden significant geassocieerd met de vetzuren. Sommige van deze DNA gebieden bevatten genen waarvan bekend is dat ze betrokken zijn bij de synthese van vet, van andere gebieden is bekend dat ze verantwoordelijk zijn voor verschillen in de hoeveelheid vet in melk, en weer andere gebieden zijn nieuwe gebieden waar nog weinig over bekend is.

De GWAS voor de samenstelling van het vet in wintermelk is opgevolgd door een GWAS voor de samenstelling van het vet in zomermelk van dezelfde koeien, en staat beschreven in hoofdstuk 3. De GWAS voor de zomermelk resulteerde in 51 DNA gebieden op 24 chromosomen die significant geassocieerd waren met een of meerdere vetzuren. De resultaten van de GWAS voor vetsamenstelling in zomermelk waren grotendeels in overeenstemming met de eerder gevonden

associaties voor vetsamenstelling in wintermelk. De DNA gebieden die significant waren voor zowel winter- als zomermelk bleken een vergelijkbaar effect te hebben op de vetsamenstelling in de beide seizoenen. Deze DNA gebieden die overeenstemmen tussen beide GWAS zijn de moeite waard om verder te bestuderen.

Op chromosoom 19 vonden we een relatief groot DNA gebied dat zeer significant geassocieerd was met de samenstelling van het vet in wintermelk en in zomermelk. Van dit gebied was nog niet met zekerheid bekend welk gen betrokken is bij de synthese van melkvet. Het doel van de studie beschreven in hoofdstuk 4 was om dit DNA gebied beter te karakteriseren met behulp van een groter aantal DNA merkers. Het aantal DNA merkers op chromosoom 19 werd daarom uitgebreid van 1.454 naar 18.893. Hierdoor werd de resolutie van het gebied groter en konden we beter inzoomen op de vele genen die in dit gebied liggen. Zodoende bleek dat 2 genen een mogelijke rol spelen: coiled-coil domain containing 57 (*CCDC57*) en fatty acid synthase (*FASN*). Er is niet veel bekend over *CCDC57*; dit gen is nooit eerder geassocieerd met melkvet, maar is wel actief in het uier van de koe. Het gen *FASN* is eerder geassocieerd met zowel melkvetzuren als vetzuren in het vetweefsel van koeien. Bovendien is *FASN* betrokken bij de synthese van korte en middellange verzadigde vetzuren, en daarom een goede kandidaat voor het DNA gebied op chromosoom 19.

Hoofdstuk 5 beschrijft de studie van de onderlinge relaties tussen de individuele vetzuren in melkvet. Melkvetzuren zijn sterk gecorreleerd, maar het is niet bekend of deze correlaties direct of indirect zijn. Directe relaties worden ook wel oorzakelijke of causale relaties genoemd. Bij causale relaties tussen vetzuren is een verandering in de concentratie van het ene vetzuur de oorzaak van een verandering in de concentratie van een ander vetzuur. Het doel van deze studie was om een causaal netwerk te vinden voor melkvetzuren. Het resultaat was een netwerk van C4:0 tot C12:0 dat overeenkomt met de synthese-route van korte en middellange verzadigde vetzuren. Op basis van de analyses kon geen richting worden gegeven aan dit netwerk. Echter, met behulp van bestaande kennis over de synthese van melkvet kon dit netwerk alsnog worden voorzien van een richting. Het gerichte netwerk is vervolgens gebruikt om de omvang van de causale relaties tussen deze vetzuren vast te stellen. Hierdoor kunnen uitkomsten van interventies toegepast op het netwerk worden voorspeld.

Hoofdstuk 6 is de algemene discussie. Het eerste deel richt zich op de inzichten verkregen door GWAS voor de samenstelling van melkvet. De GWAS heeft geleid tot het bevestigen en verfijnen van DNA gebieden betrokken bij melkvet-samenstelling en heeft bovendien geresulteerd in de identificatie van nieuwe

gebieden. Genoom-wijde associatie studies geven een goed beeld van de genetische architectuur van een kenmerk omdat het volledige genoom wordt onderzocht. Dit heeft de voorkeur boven het bestuderen van individuele genen die worden gekozen op basis van bestaande kennis over het kenmerk.

Het tweede deel van de algemene discussie richt zich op de verfijning van complexe kenmerken in onderliggende componenten. Genoom-wijde associatie studies op basis van individuele melkvetzuren resulteerden in meer DNA gebieden in vergelijking met GWAS op basis van melkvetpercentage of kilo's melkvet. Dit bevestigd dat verfijning van complexe kenmerken in onderliggende componenten meer inzicht geeft in de verbanden tussen genen en kenmerken.

De volgende sectie van de algemene discussie gaat over causale relaties tussen het DNA gebied op chromosoom 19 en de korte en middellange verzadigde vetzuren. Dit DNA gebied is geassocieerd met meerdere vetzuren die onderling causale relaties vertonen. Het DNA gebied kan dus een direct effect hebben op elk vetzuur ($y_1 \leftarrow DNA \rightarrow y_2$), of een indirect effect waarbij het DNA een effect heeft op een ander vetzuur dat vervolgens een effect heeft op het bestudeerde vetzuur ($DNA \rightarrow y_1 \rightarrow y_2$). Er is bewijs voor een direct effect van het DNA gebied op chromosoom 19 op C4:0, C6:0, C8:0, C10:0 en C14:0, en een indirect effect op C12:0 via C10:0. Bovendien is er geen effect van dit gebied op C16:0.

Het laatste deel van de algemene discussie gaat over het gebruik van mid-infra rood (MIR) spectra voor het voorspellen van de samenstelling van het melkvet als alternatief voor melkvetsamenstelling gemeten met gas chromatografie (GC). Het voorspellen van de concentratie van melkvetzuren op basis van MIR is veel goedkoper dan de dure en tijdrovende GC bepalingen. De mogelijkheden van GWAS op basis van MIR voorspelde samenstelling van het melkvet zijn onderzocht door GWAS resultaten van GC bepaalde en MIR voorspelde vetsamenstelling van dezelfde melkmonsters te vergelijken. Een deel van de DNA gebieden gevonden op basis van GC zijn niet gevonden met de MIR voorspelde melkvetsamenstelling. Dit is waarschijnlijk het gevolg van de onnauwkeurigheid van de MIR voorspelling van melkvetsamenstelling. De GWAS op basis van MIR voorspellingen resulteerde bovendien in DNA gebieden die niet gevonden zijn op basis van GC. Deze gebieden vertegenwoordigen waarschijnlijk geen DNA gebieden die echt geassocieerd zijn met melkvetsamenstelling, maar zijn mogelijk gerelateerd aan andere componenten in melk die ook worden opgepikt uit het MIR profiel. Deze analyses laten zien dat MIR voorspelde kenmerken complexiteit toevoegen aan de verbanden tussen genen en kenmerken. Dit maakt MIR voorspelde kenmerken minder geschikt voor het opsporen van genen betrokken bij het kenmerk en voor het vergroten van het inzicht in de biologische mechanismen achter het kenmerk.

Dankwoord

Dankwoord

Op de cover van mijn thesis staat Elly 24 (Olympic x Russel) van de familie Wientjes. Dit is de enige koe die ik in levende lijve bestudeerd heb voor de totstandkoming van deze thesis. De data van de 2.000 koeien die ik heb onderzocht heb ik simpelweg uit de Milk Genomics database gehaald. Ik wil dan ook mijn voorgangers Anke, Marianne, Jeroen, Ghyslaine en het team om hen heen bedanken voor het verzamelen en opschonen van de data. Gelukkig kreeg Milk Genomics een vervolg en ik ben blij dat ik onderdeel mocht uitmaken van dit enthousiaste inter-disciplinaire team. I would like to thank all people that have been a part of Milk Genomics for the interesting discussions about milk, their enthusiasm about new results, and the pleasant collaboration, especially Patrick, Jan, Marc, Róbert, Hein, Kasper, Lu, Etske, Daylan and Sandrine.
Johan, dank je wel voor de mogelijkheid om een promotieonderzoek te doen binnen Milk Genomics en binnen het Animal Breeding and Genomics Centre. Beste Henk, inhoudelijk had ik het niet zonder jou af gekund. Dank je wel voor je begeleiding, de discussies, je geduld en peptalks, maar ook voor de gezelligheid. Marleen, jouw begeleiding en excellente schrijfvaardigheid hebben mijn voortgang en publicaties sterk verbeterd. Dank je wel voor het delen van jouw inzichten over de wonderlijke wereld die wetenschap heet, maar vooral voor het verjagen van de vele beren die ik op mijn weg zag. Sorry dat ik letterlijk de beer in California heb weggejaagd voordat je ook maar de kans had om hem te zien.

I would like to thank Guilherme Rosa and Bruno Valente for the opportunity to work with them at the University of Wisconsin in Madison, USA. Thank you very much for all your support, discussions, causality lessons, paper revisions, and the fun we had in and out the office. I had a great time in Madison! I would like to thank all the people I met for making me feel welcome and for all (‘old skool’) social activities. Especially my roommates Vera and Marina: ‘muito obrigado’ for sharing so much with me.

Ook wil ik Luc Janss graag bedanken voor zijn hulp met de convergentie van mijn modellen, zonder dat was er geen paper over causale relaties tussen vetzuren geweest. Albart, bedankt voor de introductie in R, ik kan met trots zeggen dat ik inmiddels een ervaren gebruiker ben.

Many thanks to all colleagues at the Animal Breeding and Genomics Centre for the discussions and social activities we had. It has been a great pleasure to be part of such a diverse, active and social group.

I would like to thank my paranymphs for being on my side this day. Dear Panya, you have been on my side since we started our MSc here in Wageningen. Thank you for your endless support, the social activities we have enjoyed, and the greatly timed smiley's. You always believed in me!
Lieve Inge, dank je wel voor je vriendschap en nooit aflatende interesse in mijn proefschrift. Ik ben blij dat ik met mijn Nederlandse samenvatting eindelijk heb kunnen verhelder waar mijn onderzoek nou over gaat.

Bij deze wil ik ook graag Lucia en Roel bedanken voor de motivatie en de ruimte om mijn thesis af te ronden naast mijn baan bij Wageningen UR Livestock Research. Bedankt voor de lessen knopen doorhakken, ook al heb ik er nog een paar nodig.

Dank aan familie en vrienden voor alle steun en interesse. Pap en mam, welke weg ik ook koos jullie zijn er altijd voor mij geweest, bedankt daarvoor! Hilly en Martin, bedankt voor de nodige culturele en sportieve ontspanning, maar ook voor het luisterend oor. Freek, je bent de enige thuis met wie ik over QTL, SNP en GWAS kan praten. Ik ben dan ook erg blij dat onze onderzoekswerelden elkaar raken. Chris, bedankt voor de momenten waarop je me op een andere manier naar dingen laat kijken. Lieve Lieke en Veerle, dank jullie wel voor al jullie vrolijke uitspattingen.
'Op Heile' heb ik heel wat uurtjes gewerkt aan de afronding van dit proefschrift, ik wil dan ook de familie Calus en de Dekkertjes bedanken voor de steun en interesse in mijn onderzoek, maar ook voor de gezelligheid.
Tot slot, Mario, bedankt voor je steun en geduld. Jou verfrissende, maar kritische kijk tijdens de vele discussies hebben ook een belangrijke bijdrage geleverd. Bedankt voor je positieve kijk op alles (zelfs als ik daardoor negatief lijk)!

Aniek

Curriculum Vitae

About the author

Aniek Bouwman was born on 10 January 1983 in Heerlen and raised in Oirsbeek, the Netherlands. In 2001, she graduated from high school Sint-Janscollege, Hoensbroek. In 2002, Aniek started a Bachelor in Animal Husbandry & Animal Care at HAS Den Bosch. During her bachelor, she spent three months abroad at the department of Animal Sciences of the University of Adelaide (Roseworthy, Australia) where she tested the presence of a specific myostatin SNP in several beef breeds. She also gained practical experience in agricultural accountancy at A&A Groep (currently Accon) in Maastricht. For her bachelor thesis, she developed a tool for dairy farmers to assess the animal welfare on their own farm. In 2006, she received her bachelor diploma and started a master in Animal Sciences at Wageningen University. For her specialization in Animal Breeding and Genetics she performed two theses at the department of Animal Breeding and Genetics, Wageningen University. For her minor thesis within the Dutch Milk Genomics project she performed a genome-scan for several milk parameters. For her major thesis, in collaboration with TOPIGS Research Centre IPG, she studied the contribution of social interactions to heritable variation in average daily gain of piglets. In 2008, she finished her master and started a PhD within the Dutch Milk Genomics project at the Animal Breeding and Genomics Centre, Wageningen University. The results of this research about the genetic background of bovine milk fat composition are described in this thesis. During her PhD, she spent four months at the department of Animal Sciences of the University of Wisconsin (Madison, USA) to work on causality and structural equation models with Bruno Valente and Guilherme Rosa. Since November 2012, Aniek is working as a post-doc at Wageningen UR Livestock Research.

Over de auteur

Aniek Bouwman werd geboren op 10 Januari 1983 te Heerlen en groeide op in Oirsbeek. In 2001 behaalde zij haar VWO diploma aan het Sint-Janscollege te Hoensbroek, waarna ze in 2002 begon aan een bachelor Dier- en Veehouderij aan de HAS Den Bosch. Tijdens deze studie deed ze een buitenlandse onderzoeksstage bij het departement Animal Sciences aan de University of Adelaide (Roseworthy, Australië) waar ze heeft gekeken naar de aanwezigheid van een specifieke DNA variant in het myostatine gen bij verschillende rundvee rassen. Ook liep zij stage bij accountants-kantoor A&A Groep (overgenomen door Accon) te Maastricht. Haar bachelor thesis betrof het ontwikkelen van een instrument voor melkveehouders om het welzijn van de dieren op hun eigen bedrijf te toetsen, de Welzijnswijzer Melkvee. In 2006 heeft Aniek haar bachelor diploma behaald en is ze begonnen aan een master Dierwetenschappen aan Wageningen University. Voor haar specialisatie in 'Fokkerij en Genetica' heeft ze twee afstudeeropdrachten volbracht bij de leerstoelgroep Fokkerij en Genetica, Wageningen University. Voor haar minor thesis heeft Aniek binnen het Milk Genomics project een genoom scan uitgevoerd voor verschillende melkparameters. Haar major thesis ging over het effect van sociale interacties op erfelijke variatie in de groei van biggen, en was een samenwerking met TOPIGS Research Centre IPG. In 2008, heeft Aniek haar master afgerond en is ze begonnen als promovenda bij het Animal Breeding and Genomics Centre aan Wageningen University. De resultaten van het onderzoek naar de genetische achtergrond van melkvetsamenstelling van koeien zijn beschreven in dit proefschrift. Tijdens haar promotie verbleef Aniek vier maanden bij het departement Animal Sciences van de University of Wisconsin (Madison, USA) waar zij met Bruno Valente and Guilherme Rosa aan causale relaties heeft gewerkt. Sinds November 2012 is Aniek werkzaam als onderzoeker bij Wageningen UR Livestock Research.

Peer reviewed publications

Bouwman A.C., J.M. Hickey, M.P.L. Calus, R.F. Veerkamp, 2014. Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. Genetics Selection Evolution 46 (1), 6.

Bouwman A.C., B.D. Valente, L.L.G. Janss, H. Bovenhuis, G.J.M. Rosa, 2014. Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. Genetics Selection Evolution 46 (1), 2.

Bouwman A.C., M.H.P.W. Visker, J.A.M. van Arendonk, H. Bovenhuis, 2014. Fine mapping of a quantitative trait locus for bovine milk fat composition on *Bos taurus* autosome 19. Journal of Dairy Science 97 (2), 1139-1149.

Duchemin S., H. Bovenhuis, W.M. Stoop, A.C. Bouwman, J.A.M. Van Arendonk, M.H.P.W. Visker, 2013. Genetic correlation between composition of bovine milk fat in winter and summer, and *DGAT1* and *SCD1* by season interactions. Journal of Dairy Science 96 (1), 592-604.

Rutten M.J.M, A.C. Bouwman, R.C. Sprong, J.A.M. van Arendonk, M.H.P.W. Visker, 2013. Genetic variation in vitamin B-12 content of bovine milk and its association with SNP along the bovine genome. PloS One 8 (4), e62382.

Bouwman A.C., M.H.P.W. Visker, J.A.M. van Arendonk, H. Bovenhuis, 2012. Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. BMC Genetics 13 (1), 93.

Bouwman A.C., H. Bovenhuis, M.H.P.W. Visker, J.A.M. van Arendonk, 2011. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genetics 12 (1), 43.

Bouwman A.C., L.L.G. Janss, H.C.M. Heuven, 2011. A Bayesian approach to detect QTL affecting a simulated binary and quantitative trait. BMC Proceedings 5 (Suppl 3), S4.

Bouwman A.C., R. Bergsma, N. Duijvesteijn, P. Bijma, 2010. Maternal and social genetic effects on average daily gain of piglets from birth until weaning. Journal of Animal Science 88 (9), 2883-2892.

Bouwman A.C., G.C.B. Schopen, H. Bovenhuis., M.H.P.W. Visker, J.A.M. van Arendonk, 2010. Genome-wide scan to detect quantitative trait loci for milk urea nitrogen in Dutch Holstein-Friesian cows. Journal of Dairy Science 93 (7), 3310-3319.

Conference proceedings

Bouwman A.C., Hickey, J.M., Calus, M.P.L., Veerkamp, R.F. Imputation of non-genotyped individuals based on genotyped relatives: a real case scenario. The 64th Annual Meeting of the European Association of Animal Production, Nantes, France, 26 - 30 August 2013.

Bouwman A.C., M.H.P.W. Visker, J.A.M. van Arendonk, H. Bovenhuis. Genome regions involved in the fatty acid composition of summer and winter milk. The 9th International Symposium Milk Genomics and Human Health, Wageningen, the Netherlands, October 2012.

Duchemin S.I., H. Bovenhuis, W.M. Stoop, A.C. Bouwman, J.A.M. van Arendonk, M.H.P.W. Visker. Genetic relation between composition of bovine milk fat in winter and summer. The 9th International Symposium Milk Genomics and Human Health, Wageningen, the Netherlands, October 2012.

Visker M.H.P.W., M.J.M. Rutten, R.C. Spronk, A.C. Bouwman, J.A.M. van Arendonk. Opportunities for improving vitamin B12 content in bovine milk. The 9th International Symposium Milk Genomics and Human Health, Wageningen, the Netherlands, October 2012.

Bouwman A.C., B.D. Valente, H. Bovenhuis, G.J.M. Rosa. Structural equation models to study causal relationships between bovine milk fatty acids. The 63th Annual Meeting of the European Association of Animal Production, Bratislava, Slovakia, 27 - 31 August 2012.

Duchemin S.I., H. Bovenhuis, W.M. Stoop, A.C. Bouwman, J.A.M. van Arendonk, M.H.P.W. Visker. Genetic correlation between composition of bovine milk fat in winter and summer, and DGAT1 and SCD1 by season interactions. The 63th Annual Meeting of the European Association of Animal Production, Bratislava, Slovakia, 27 - 31 August 2012.

Rosa G.J.M., B.D. Valente, F. Peñagaricano, A.C. Bouwman. Inferring causal phenotype networks using structural equation models and genomic information. The 6th International Conference on Genomics, Shenzhen, China, 12 - 15 November 2011.

Bouwman A.C., M.H.P.W. Visker, J.A.M. van Arendonk, H. Bovenhuis. Genome-wide association of fatty acids from summer milk of Dutch dairy cattle. The 62th Annual Meeting of the European Association of Animal Production, Stavanger, Norway, 26 August - 2 September 2011.

Bouwman A.C., H. Bovenhuis. Genome-wide association of myristic and palmitic acid in milk of Dutch dairy cattle. The 7th International Symposium Milk Genomics and Human Health, Davis, USA, October 2010.

Bouwman A.C., H. Bovenhuis, M.H.P.W. Visker, J.A.M. van Arendonk. Genome-wide association of the ratio of saturated to unsaturated milk fatty acids in Dutch dairy cattle. Proceedings of the 9th World Congress on Genetic Applied to Livestock Production, Leipzig, Germany, 1 - 6 August 2010.

Bouwman A.C., R. Bergsma, N. Duijvesteijn, P. Bijma. The contribution of social effects to heritable variation in average daily gain of piglets from birth till weaning. The 60th Annual Meeting of the European Association of Animal Production, Barcelona, Spain, August 2009.

Sellick G.S., H. McGrice, A.C. Bouwman, B. Kruk, C.D.K. Bottema, Polymorphisms within the cattle myostatin gene. Proceedings of the 30th International Conference on Animal Genetics (ISAG), Porto Seguro, Brazil, 20 - 25 August 2006.

Training and supervision plan

The Basic Package (3 ECTS)	**Year**
WIAS Introduction Course	2008
Ethics and Philosophy of Animal Science	2009

Scientific Exposure (20 ECTS)

International conferences (8 ECTS)

60th annual meeting EAAP, Barcelona, Spain	2009
9th WCGALP, Leipzig, Germany	2010
7th International symposium milk genomics and human health, Davis, USA	2010
62th annual meeting EAAP, Stavanger, Norway	2011
63th annual meeting EAAP, Bratislava, Slovakia	2012
9th International symposium Milk Genomics and Human Health, Wageningen	2012

Seminars and workshops (5 ECTS)

Fokkerij en Genetica connectie dagen, Vught, the Netherlands (2x)	2008/2010
WIAS Science Day, Wageningen, the Netherlands (4x)	2009-2012
Friends or Fiends? Consequences of social interactions for artificial breeding programs and evolution in natural populations, Wageningen	2009
QTL MAS workshop, Wageningen, the Netherlands	2009
Genetics of milk quality, Wageningen, the Netherlands	2009
Developments in genome-wide evaluation and genomic selection, Wageningen, the Netherlands	2009
QTL MAS workshop, Poznan, Poland	2010
Workshop milk, lactation biology & high throughput RNA sequencing, Davis, USA	2010
Genomics and animal breeding, Wageningen, the Netherlands	2011

Presentations (7 ECTS)

Heritable social effects in piglets, 60th EAAP Barcelona (oral)	2009
Genome wide association for fat composition in Dutch dairy cattle, 9th WCGALP, Leipzig, Germany (oral)	2010
Genome-wide association of myristic and palmitic acid in milk of Dutch dairy cattle, 7th IMGC, Davis, USA (poster)	2010
Genome-wide association of milk fatty acids of Dutch dairy cattle, WIAS Science Day, Wageningen, the Netherlands (oral)	2011
Genomic regions associated with bovine milk fatty acids, 62th EAAP, Stavanger, Norway (oral)	2011
Structural equation models to study causal relationships between bovine milk fatty acids, 63th EAAP, Bratislava, Slovakia (oral)	2012
Genomic regions associated with both summer and winter bovine milk fatty acids, 9th IMGC, Wageningen, the Netherlands (oral)	2012

In-Depth Studies (9 ECTS)

Disciplinary and interdisciplinary courses (7 ECTS)

Nutrient density of milk, Wageningen, the Netherlands	2009
Population genetic data analysis, Summer Institute in Statistical Genetics, Liège, Belgium	2009
QTL mapping, Summer Institute in Statistical Genetics, Liège, Belgium	2009
Association mapping, Summer Institute in Statistical Genetics, Liège, Belgium	2009
Quantitative Genetics of Selection Response, Wageningen, the Netherlands	2010
Genomic selection in Livestock, Wageningen, the Netherlands	2011
Identity by Descent Approaches to Genomic Analyses of Genetic Traits, Wageningen, the Netherlands	2012

PhD students' discussion groups (2 ECTS)

Quantitative genetics discussion group (QDG), Wageningen, the Netherlands	2008-2012

Professional Skills Support Courses (3 ECTS)

Course Supervising MSc thesis work	2010
Techniques for Writing and Presenting a Scientific paper	2010
Career assessment	2012
Voice matters	2012
Effective behavior in your professional surroundings	2012

Research Skills Training (6 ECTS)

Introduction to R for statistical analysis, Wageningen, the Netherlands	2008
Preparing own PhD research proposal	2009
External training period at University of Wisconsin-Madison, USA	2011

Didactic Skills Training (6 ECTS)

Supervising practicals and excursions (4 ECTS)

Genetic Improvement of Livestock, ABG-31306, Wageningen University (2x)	2009-2010

Supervising theses (2 ECTS)

MSc Thesis Animal Breeding and Genetics, Wageningen University	2010

Tutorship (0.3 ECTS)

Research Master Cluster, YAS-60312, Wageningen University	2009

Education and Training Total	**47 ECTS**

Colophon

The research described in this thesis is part of the Dutch Milk Genomics Initiative and the project 'Melk op Maat', funded by Wageningen University (the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), the cooperative cattle improvement organization CRV (Arnhem, the Netherlands), the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands).

Artwork on cover by Aniek Bouwman

Printed by GVO drukkers en vormgevers B.V. / Ponsen & Looijen, Ede, the Netherlands