ISSN 1652-6880
ISBN (print version) 978-91-576-8580-3
ISBN (electronic version) 978-91-576-8581-0

Mapping and fine-mapping of genetic factors affecting bovine milk composition Sandrine I. Duchemin

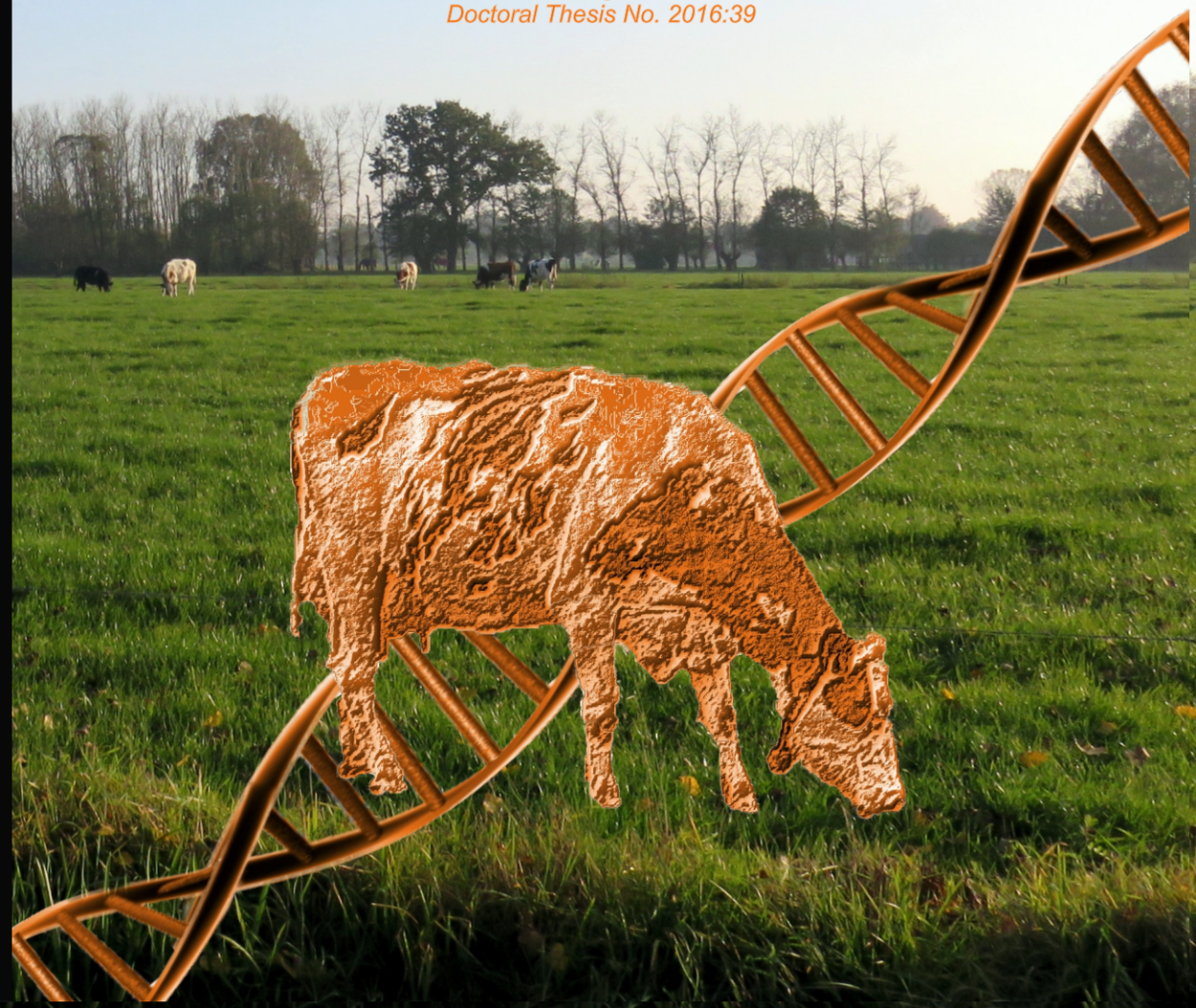
WAGENINGEN UR
For quality of life

SLU

Mapping and fine-mapping of genetic factors affecting bovine milk composition

Sandrine Isolde Duchemin

Acta Universitatis Agriculturae Sueciae
Doctoral Thesis No. 2016:39



Mapping and fine-mapping of genetic factors affecting bovine milk composition

Thesis committee

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

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

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

Dr. ir. W.F. Fikse
Senior researcher, Department of Animal Breeding and Genetics
Swedish University of Agricultural Sciences

Other members (assessment committee)
Prof. Dr. E.J.M. Feskens, Wageningen University
Prof. Dr. A.C.M. van Hooijdonk, Wageningen University
Prof. Dr. L. Andersson-Eklund, Swedish University of Agricultural Sciences, Sweden
Dr. D. Boichard, National Institute for Agricultural Research (INRA), France

The research presented in this doctoral thesis was conducted under the joint auspices of the Swedish University of Agricultural Sciences and the Graduate School Wageningen Institute of Animal Sciences of Wageningen University and is part of the Erasmus Mundus Joint Doctorate program “EGS-ABG".

Mapping and fine-mapping of genetic factors affecting bovine milk composition

Sandrine Isolde Duchemin

ACTA UNIVERSITATIS AGRICULTURAE SUECIAE
DOCTORAL THESIS № 2016:39

Thesis

submitted in fulfillment of the requirements for the degree of doctor from
Swedish University of Agricultural Sciences
by the authority of the Board of the Faculty of Veterinary Medicine and Animal Science and from
Wageningen University
by the authority of the Rector Magnificus, Prof. Dr. A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board of Wageningen University and the Board of the Faculty of Veterinary Medicine and Animal Science at the Swedish University of Agricultural Sciences
to be defended in public
on Monday May 30, 2016
at 4.00 p.m. in the Aula of Wageningen University

ISSN 1652-6880
ISBN (print version) 978-91-576-8580-3
ISBN (electronic version) 978-91-576-8581-0
ISBN 978-94-6257-730-5
DOI:10.18174/370103

Duchemin, S.I.
Mapping and fine-mapping of genetic factors affecting bovine milk composition.

Joint PhD thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden and Wageningen University, the Netherlands (2016)
With references, with summary in English

Abstract

Duchemin, S.I. (2016). Mapping and fine-mapping of genetic factors affecting bovine milk composition. Joint PhD thesis, between Swedish University of Agricultural Sciences, Sweden and Wageningen University, the Netherlands

Bovine milk is an important source of nutrients in Western diets. Unraveling the genetic background of bovine milk composition by finding genes associated with milk-fat composition and non-coagulation of milk were the main goals of this thesis. In ***Chapter 1***, a brief description of phenotypes and genotypes used throughout the thesis is given. In ***Chapter 2***, I calculated the genetic parameters for winter and summer milk-fat composition from ~2,000 Holstein-Friesian cows, and concluded that most of the fatty acids (**FA**) can be treated as genetically the same trait. The main differences between milk-fat composition between winter and summer milk samples are most likely due to differences in diets. In ***Chapter 3***, I performed genome-wide association studies (**GWAS**) with imputed 777,000 single nucleotide polymorphism (**SNP**) genotypes. I targeted a quantitative trait locus (**QTL**) region on Bos taurus autosome (**BTA**) 17 previously identified with 50,000 SNP genotypes, and identified a region covering 5 mega-base pairs on BTA17 that explained a large proportion of the genetic variation in de novo synthesized milk FA. In ***Chapter 4***, the availability of whole-genome sequences of keys ancestors of our population of cows allowed to fine-mapped BTA17 with imputed sequences. The resolution of the 5 mega base-pairs region substantially improved, which allowed the identification of the LA ribonucleoprotein domain family, member 1B (***LARP1B***) gene as the most likely candidate gene associated with de novo synthesized milk FA on BTA17. The LARP1B gene has not been associated with milk-fat composition before. In ***Chapter 5***, I explored the genetic background of non-coagulation of bovine milk. I performed a GWAS with 777,000 SNP genotypes in 382 Swedish Red cows, and identified a region covering 7 mega base-pairs on BTA18 strongly associated with non-coagulation of milk. This region was further characterized by means of fine-mapping with imputed sequences. In addition, haplotypes were built, genetically differentiated by means of a phylogenetic tree, and tested in phenotype-genotype association studies. As a result, I identified the vacuolar protein sorting 35 homolog, mRNA (***VPS35***) gene, as candidate. The VPS35 gene has not been associated to milk composition before. In ***Chapter 6***, the general discussion is presented. I start discussing the challenges with respect to high-density genotypes for gene discovery, and I continue discussing future possibilities to expand gene discovery studies, with which I propose some alternatives to identify causal variants underlying complex traits in cattle

List of Publications

This thesis is based on the work contained in the following papers:

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

II **Duchemin, S. I.**, Visker, M.H.P.W., Van Arendonk, J.A.M., and Bovenhuis, H. 2014. A quantitative trait locus on *Bos taurus autosome* 17 explains a large proportion of the genetic variation in de novo synthesized milk fatty acids. *J Dairy Sci* 97: 7276-7285.

III **Duchemin, S. I.**, Bovenhuis, H., Megens, H-J., Van Arendonk, J. A. M., and M. H. P. W. Visker. Fine-mapping of BTA17 using imputed sequences for associations with de novo synthesized fatty acids in bovine milk. (manuscript)

IV **Duchemin, S. I.**, Glantz, M., de Koning, D-J., Paulsson, M., and W.F. Fikse. 2016. Identification of QTL on chromosome 18 associated with non-coagulating milk in Swedish Red cows. *Front Genet* 7:57. doi: 10.3389/fgene.2016.00057.

1

General Introduction

1.1 Milk

Milk has fascinated mankind since the beginning of the ages. A clear example of this fascination is the Milky Way galaxy, which contains our Planet Earth. The Milky Way galaxy has its roots in the Greek-Roman Mythology. The word galaxy originates from galas, which is a synonym for milk in Greek language. According to the Mythology, the Milky Way galaxy was "drops of milk" spelt by goddess Hera, when breastfeeding Hercules, the bastard son of Zeus (Larousse encyclopedia, 2015). "The origin of the Milky Way" has been immortalized by the renaissance artist Jacopo Tintoretto circa 1575-1580 (National Gallery, London, UK; Figure 1.1), and the "Birth of the Milky Way" by the Flemish artist Peter Paul Rubens in 1637 (Museo del Prado, Madrid, Spain). In many civilizations, the Milky Way galaxy has been used as a metaphor for a splash of milk in the dark skies of our Universe. Essentially, this metaphor is a way of expressing the importance of milk for mankind. It is so important that from the very beginning of life, an infant receives milk as the primary source of nutrients.

Figure 1.1 – "The origin of the Milky Way" by Jacopo Tintoretto circa 1575-1580 (exposed in the National Gallery, London, UK)

The fascination exerted by Universe on mankind is understandable. By contemplating stars, mankind loses notion of time allowing deeper lessons to be learnt. When G. Galilei (*in:* Galilei and Van Helden, 1989) first observed the Milky Way galaxy through his telescope in 1610, he discovered that it was formed by many smaller groups of stars. Following the steps of G. Galilei (*in:* Galilei and Van Helden, 1989) a deeper look into the splash of milk in the dark skies might give us insights

into the composition of milk. The splash might represent the fluid part of milk. The small groups of stars composing this splash might represent the main components in milk, such as proteins and fatty acids. The interstellar dust accompanying these stars might represent the minerals in milk. In just a few instants, the composition of milk is described as an (scientific) idea that has been transmitted throughout centuries by a simple metaphor.

Metaphors with our Universe do not stop at the Milky Way galaxy. Mankind named constellations after species of animals (e.g., Taurus, Aries, and Pisces), just like cave men have represented wild animals in their cave drawings. From stone-age to modern times, domestication of animals has been one of the drivers for men's transition from hunters to farmers. During this process, the role of cattle was undeniable. By domesticating cows, mankind preserved through time important resources, such as the genetic variation of bovine species. The preservation of this genetic variation has important consequences for the current technological development of mankind. It is so important that from the beginning of every life, genetic variation will determine the future of all species.

By using metaphors, such as Milky Way galaxy and names of constellations, mankind transmitted more than just a simple image from cave to modern men. As intrinsic parts of the Milky Way galaxy, cave and modern men would be united forever as one student. For mankind, these metaphors have engraved in our collective memories a deep respect for our Planet Earth and its scarce resources. Resources beyond genetic variation have been translated. In our modern times, this deep respect is taught by uniting human needs (milk as a nutrient) and animal resources (genes affecting bovine milk composition) through Animal Breeding and Genetics.

The scope of my thesis was to investigate the genetic background of bovine milk composition. More specifically, my thesis focuses on the composition of milk-fat, and on non-coagulation of milk.

1.2 Milk-fat composition

Bovine milk fat is an important source of energy for mankind. The main bioactive lipids in bovine milk are fatty acids (**FA**). According to Jensen (2002), bovine milk-fat is composed of more than 400 individual FA, most occurring in amounts less than 1%. The individual FA in bovine milk-fat are organized in chain of carbons that vary in length from 4 to 22 carbons. According to their chain-lengths, these individual FA

are grouped as short-chain (C4:0 – C12:0), medium-chain (C14:0 – C16:0) and long-chain (C18:0 – C22:0) FA. In addition, individual FA can be either saturated or unsaturated. FA are saturated when a carbon is connected by a single bond to an adjacent carbon in the chain, and FA are unsaturated when a carbon is connected to an adjacent carbon in the chain by double or triple bonds. Differences in FA regarding their saturation are shown in Figure 1.2.

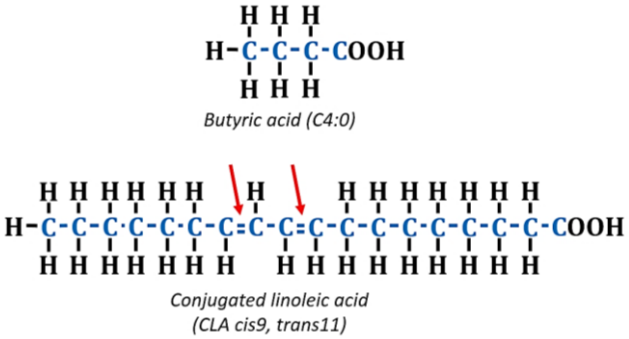


Figure 1.2 – Representation of fatty acids (FA). Butyric acid representing saturated FA, and conjugated linoleic acid representing unsaturated FA. Arrows in red point out the double bonds between adjacent carbons.

The biosynthesis of milk-fat occurs in the mammary gland of a cow. Individual FA in the mammary gland arise from circulating blood lipids and *de novo* synthesis. Circulating blood lipids originate from the feed of the cow or from the cow's body fat. Through the *de novo* synthesis, FA are elongated from precursors by adding C2:0. These precursors can be either acetate (C2:0), propionate (C3:0) or butyrate (C4:0). C2:0 and C3:0 originate from lipids in circulating blood, while C4:0 may either originate from blood lipids or the *de novo* synthesis itself (e.g., Craninx et al., 2008). Depending on the precursor, FA synthesized de novo may terminate at either C16:0 or C17:0. It is assumed that *de novo* synthesis produces the short-chain FA, C14:0 and 50% of C16:0 in milk, whereas the remaining 50% of C16:0 and the long-chain FA come from the lipids in circulating blood.

FA in bovine milk are relevant for human health. According to Calder et al. (2015), FA are essential for the well-being of humans, and they have important biological activities regarding the cell and tissue metabolism, as well as responsiveness to hormones and other signals in human cells. Stoop et al. (2008) indicated that FA in bovine milk are heritable, with heritability estimates between 0.22 and 0.71. These heritability estimates suggest that milk-fat composition can be improved by

breeding. In addition, Tzompa-Sosa et al. (2014) showed that increases in long-chain saturated FA can influence the thermal properties of milk-fat, which can lead to important changes in the quality of milk-fat derived products. Moreover, breeding could be used to reduce the concentration of certain FA in bovine milk-fat. For instance, low concentrations of C16:0 in bovine milk-fat would best meet infant requirements regarding the consumption of milk-fat derived products (e.g., Tzompa-Sosa et al., 2014). Therefore, increasing the biological knowledge regarding bovine milk-fat composition can be of great interest to the dairy industry.

1.3 Non-coagulation of milk

In addition to FA, bovine milk is an important source of proteins for mankind. The main proteins in bovine milk are the caseins, which account for almost 80% of the proteins in milk. There are four caseins in bovine milk: α_{s1}-,α_{s2}-, β-, and κ-casein. Most of these caseins are organized in micelles. These micelles are not soluble in water and can precipitate in the presence of rennet. This property is used in cheese production to induce coagulation of milk. In 2013, almost 30% of the total production of bovine milk in Sweden was destined to cheese production (LRF Dairy Sweden, 2015).

Besides the caseins, whey proteins account for the remaining 20% of the proteins in milk, of which β-lactoglobubin and α-lactalbumin are the most important ones. The whey proteins are considered by-products of cheese production. In contrast to caseins, whey proteins are soluble in water, and can only be denatured by heat. When heated, whey proteins can produce products such as ricotta and whey butter.

It is economically relevant for the cheese industry to reduce time and losses while producing cheese. In this sense, if caseins in bovine milk do not coagulate after rennet addition, the entire chain of cheese production is delayed, generating losses for this industry. Consequently, non-coagulation of milk can be considered as a new phenotype that accounts for the needs of the cheese industry. Non-coagulation (**NC**) of milk is prevalent among several dairy cattle breeds, such as Swedish Red, Finnish Ayrshire, Holstein-Friesian, and Italian Brown Swiss, to name a few (e.g., Frederiksen et al., 2011; Cecchinato et al., 2011, Gustavsson et al., 2014). The prevalence of NC milk varies among these breeds ranging from 4% in Italian Brown Swiss (Cecchinato et al., 2009) up to 13% in Finnish Ayrshires (Ikonen et al., 2004). A recent study reported the prevalence of NC milk at 18% in the Swedish Red cows (Gustavsson et al., 2014).

1.4 Genomic regions influencing bovine milk composition

Many genomic regions of the cattle genome have been associated with milk composition. While many of these genomic regions have not been studied in detail yet, some genes have been associated with milk-fat composition and non-coagulation of milk.

For bovine milk-fat composition, the main identified genes are: diacylglycerol O-acyltransferase 1 (*DGAT1*) located on *Bos taurus autosome* (**BTA**) 14, stearoyl-CoA desaturase 1 (*SCD1*) located on BTA26, acyl-CoA synthase short-chain family member 2 (*ACSS2*) located on BTA13, fatty acid synthase (*FASN*) located on BTA19, and 1-Acylglycerol-3-Phosphate O-Acyltransferase 6 (*AGPAT6*) located on BTA27. The association of the *DGAT1* and *SCD1* genes with milk-fat composition has been studied e.g., by Schennink et al. (2007, 2008). The association of the ACSS2, FASN and AGPAT6 genes with milk-fat composition has have been studied e.g., by Bouwman et al. (2011) and LittleJohn et al. (2014). The involvement of each of these genes occurs at different stages in the synthesis of milk-fat in the mammary gland of a cow: intracellular FA activation (*ACSS2*), fatty acid synthesis (*FASN*), unsaturation of FA (*SCD1*), and triacylglycerol synthesis (*AGPAT6*, *DGAT1*).

For bovine milk protein composition, the six major proteins in milk are encoded on the following chromosomes: α-lactalbumin on BTA5, the α_{s1}-,α_{s2}-, β-, and κ-caseins on BTA6, and β-lactoglobubin on BTA11. However, other chromosomal regions have been associated with milk protein composition (Schopen et al., 2011). These chromosomal regions encoding milk proteins seem to influence milk coagulation properties including non-coagulation of milk. Studies by Jensen et al. (2012) and by Gregersen et al. (2015) suggest that poor- and non-coagulation of milk are influenced by the milk protein variants of the k-casein gene. In contrast, study by Tyrisevä et al. (2008) and Gregersen et al. (2015) revealed that non-coagulation of milk can be influenced by other parts of the cattle genome too.

Promising genomic regions across the cattle genome in association with the desired trait can be identified with genetic markers. It is expected that associations with FA or non-coagulation of milk can be targeted to smaller chromosomal regions with sequences as compared to other panels of genetic markers, such as 50,000 (**50k**) and 777,000 (**777k**) single nucleotide polymorphism (**SNP**) markers. Sequences should contain all of the causal variants (Meuwissen and Goddard, 2010) that are believed

to be associated with the studied phenotype. The use of sequences for association studies has been enabled by the availability of an increasing number of sequenced animals (bulls and cows) from projects like the 1000Bull Genome Consortium (Daetwyler et al., 2014).

1.5 Aim and outline of this thesis

The present thesis aims at unraveling the genetic background of bovine milk composition by finding genes associated with milk-fat composition and non-coagulation of milk in targeted chromosomal regions. Throughout this thesis, there is a consistent increase in the number of genotypes analyzed, which have been useful to increase the resolution of some interesting genomic regions associated with bovine milk composition. In **Chapter 2**, we calculated the genetic correlations between the composition of bovine milk fat in winter and summer, and *DGAT1* and *SCD1* by season interactions. The conclusions of this work were further explored in Chapters 3 and 4. In **Chapter 3**, a quantitative trait locus on *Bos taurus autosome* (**BTA**) 17 explaining a large proportion of the genetic variation in de novo synthesized milk FA is mapped. In **Chapter 4**, we fine-mapped this QTL associated with de novo synthesized milk FA on BTA17 using imputed sequences. In **Chapter 5**, a similar fine-mapping methodology was used for the identification of a QTL on BTA18 associated with non-coagulation of milk in Swedish Red cows. In **Chapter 6**, challenges regarding the substantial increase in the number of genotypes used in this thesis, and the future possibilities to expand gene discovery are discussed.

1.6 References

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

Calder, P. C. 2015. Functional roles of fatty acids and their effects on human health. J Parenter Enteral Nutr, 39.1: 18S-32S.

Cecchinato, A., De Marchi, M., Gallo, L., Bittante, G., and Carnier, P. 2009. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. J Dairy Sci 92, 5304–5313.

Cecchinato, A., Penasa, M., De Marchi, M., Gallo, L., Bittante, G., and Carnier, P. 2011. Genetic parameters of coagulation properties, milk yield, quality, and acidity: estimated using coagulating milk and noncoagulating information in Brown Swiss and Holstein cows. J Dairy Sci 94, 4205-4213.

Craninx, M., A. Steen, H. Van Laar, T. Van Nespen, J. Martin-Tereso, B. De Baets, and V. Fievez. 2008. Effect of lactation stage on the odd- and branched-chain milk fatty acids of dairy cattle under grazing and indoor conditions. J. Dairy Sci. 91:2662–2677.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brondum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerre, D., Bouchez, O., Rossignol, M-N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P.J., Coote, D. Chamberlain, A.J., Anderson, C., VanTassell, C.P., Hulsegge, I., Goddard, M.E., Guldbrandtsen, B., Lund, M.S., Veerkamp, R.F., Boichard, D.A., Fries, R., and Hayes, B. J. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46, 858–865.

Frederiksen, P. D., Andersen, K. K., Hammershøj, M., Poulsen, H. D., Sørensen, J., Bakman, M., Qvist, K.B., and Larsen, L.B. 2011. Composition and effect of blending of noncoagulating, poorly coagulating, and well-coagulating bovine milk from individual Danish Holstein cows. J Dairy Sci 94, 4787–4799.

Galilei, G., and Van Helden, A. 1989.*Sidereus Nuncius*, or the sidereal messenger. Chicago: University of Chicago Press.

Gustavsson, F., Glantz, M., Poulsen, N. A., Wadsö, L., Stålhammar, H., Andrén, A., Lindmark-Månsson, H., Larsen, L.B., Paulsson, M., and Fikse, W. F. 2014. Genetic parameters for rennet- and acid-induced coagulation properties in milk from Swedish Red dairy cows. J Dairy Sci 97, 5219–5229.

Gregersen, V. R., Gustavsson, F., Glantz, M., Christensen, O. F., Stålhammar, H., Andrén, A., Lindmark-Månsson, H., Poulsen, N. A., Larsen, L.B., Paulsson, M., and Bendixen, C. 2015. Bovine chromosomal regions affecting rheological traits in rennet-induced skim milk gels. J Dairy Sci 98, 1261-1272.

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

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

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

Larousse Encyclopedia. 2015. http://www.larousse.fr/encyclopedie, accessed on Nov 3rd, 2015.

Littlejohn, M.D., Tiplady, K., Lopdell, T., Law, T. A., Scott, A., Harland, C., Sherlock, R., Henty, K., Obolonkin, V.,Lehnert, K., MacGibbon, A., Spelman, R. J., Davis, S. R., and

Snell, R. G. 2014. Expression variants of the lipogenic AGPAT6 gene affect diverse milk composition phenotypes in *Bos taurus*. PLoS ONE 9: e85757.

LRF Dairy Sweden. 2015. http://www.lrf.se/globalassets/dokument/om-lrf/branscher/lrf-mjolk/statistik/milk_key_figures_sweden.pdf , accessed on Nov 3rd, 2015.

Meuwissen, T., and Goddard, M. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185, 623–631.

Rubens, P. P. 1637. Birth of the Milky Way. Museo del Prado, Madrid, Spain.

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

Schennink, A., J. M. L. Heck, H. Bovenhuis, M. H. P. W. Visker, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2008. Milk fatty acid unsaturation: genetic parameters and effects of Stearoyl-CoA Desaturase (*SCD1*) and Acyl CoA: Diacylglycerol Acyltransferase 1 (*DGAT1*). J. Dairy Sci. 91:2135-2143.

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

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

Tintoretto, J. (circa 1575-1580). Origins of the Milky Way. National Gallery, London, UK.

Tyrisevä, A. M., Elo, K., Kuusipuro, A., Vilva, V., Jänönen, I., Karjalainen, H., Ikonen, T., Ojala, M. 2008. Chromosomal regions underlying noncoagulation of milk in Finnish Ayrshire cows. Genetics 180, 1211–1220

Tzompa-Sosa, D. A., van Aken, G. A., van Hooijdonk, A. C. M., and van Valenberg, H. J. F. 2014. Influence of C16: 0 and long-chain saturated fatty acids on normal variation of bovine milk fat triacylglycerol structure. J Dairy Sci 97:4542-4551.

2

General discussion

2.1 Introduction

In this thesis, the genetic backgrounds of milk-fat composition and of non-coagulation of milk have been explored. Firstly, for bovine milk-fat composition, we investigated how genetic differences between winter and summer milk contributed to the observed phenotypic differences (Chapter 2). We showed that winter and summer milk-fat composition are largely genetically the same trait. Phenotypic differences between winter and summer milk-fat composition were mainly caused by dietary differences rather than by genetic differences. Furthermore, for most fatty acids (**FA**), no significant *DGAT1* and *SCD1* by season interactions were found. In case significant interactions were present, we showed that these interactions were likely caused by the scaling of the genotype effects. Secondly, for bovine milk-fat composition and for non-coagulation (**NC**) of milk, we explored their genetic variation by means of genome-wide association studies (**GWAS**). Through GWAS (in Chapters 3 and 5), we characterized promising chromosomal regions associated with the phenotypes. Subsequently, in Chapters 3, 4 and 5, these promising regions were fine-mapped with imputed 777k SNP genotypes and imputed sequence data. The fine-mappings refined the location of quantitative trait loci (**QTL**), and contributed to the identification of candidate genes for these QTLs.

In this general discussion, I discuss different perspectives regarding gene discovery in cattle. I had the opportunity to use a substantial number of genetic markers for gene discovery, and encountered some challenges. Therefore, firstly, I discuss the challenges with respect to high-density genotypes for gene discovery. Secondly, I discuss future possibilities to expand gene discovery studies, and I propose some alternatives to identify causal variants underlying complex traits in cattle.

2.2 Challenges with high-density genotypes for gene discovery

The two main challenges for gene discovery were the imputation to high-density genotypes and the annotation of the cattle genome. In general, the attainment of high-density genotypes (and herein, I include sequences as high-density genotypes) requires several expensive steps, such as genotyping DNA samples in laboratories, using bioinformatic tools plus programmers to handle the huge data sets, and storing data. In recent years in cattle, imputation has been used to reduce costs and to accelerate the attainment of these high-density genotypes for large groups of

animals. A recognized imputation strategy consists in genotyping influential ancestors in a population, and imputing the rest of the population to a higher density of genetic markers (e.g., Druet et al., 2014). After using imputation in Chapters 3, 4 and 5, the density of genetic markers increased while the distance between genetic markers decreased. Regarding the distance between genetic markers, it was reduced from 10 mega base-pairs (**bp**) with 50k SNP to ± 4 mega bp with (imputed) 777k SNP genotypes (Chapter 3), and to a few kilo bp with (imputed) sequences (Chapters 4 and 5). GWAS and fine-mapping using these imputed genotypes resulted in a substantial increase in the number of significant associations (in the thousands) with the phenotypes (Chapters 4 and 5). As a consequence, it became more difficult to identify among the thousands of significant associations which one is the causal mutation.

After finding thousands of significant associations with the phenotypes, the next step consisted in identifying candidate genes underlying these phenotypes. For this purpose, the annotation of the cattle genome is an important tool to pin-point candidate genes. The annotation of genomes including cattle is a dynamic process, hence, constantly changing over time. Currently, important developments regarding the assembly and the annotation of genomes including cattle are on their way. These developments, more specifically the FAANG Consortium (Andersson et al., 2015), will contribute to identify candidate genes and regulatory elements more efficiently than at present.

I will discuss in more detail the two challenges for gene discovery: imputation to high-density genotypes and the annotation of the cattle genome.

2.2.1 Imputation of high-density genotypes

A key feature in using GWAS with imputed high-density genotypes is the accurate imputation of genotypes. According to Marchini and Howie (2010), genotype imputation is a statistical method of predicting (i.e., imputing) genotypes in a sample based on a reference population (**RefPop**). The sample is a representation of a population, typically genotyped for a lower density of genetic markers (e.g., 50k SNP genotypes), and this sample has not been assayed for a higher density of genetic markers (e.g., 777k SNP genotypes). The RefPop consists of individuals that are related to the sampled population and that have been genotyped for a higher density of genetic markers (e.g., 777k SNP genotypes). Based on the RefPop, the sampled population is imputed to a higher density of genetic markers (see figure 2.1). The accuracies of the resulting imputed genotypes range from 0 (poorly imputed) to 1

(correctly imputed). In most cases, genotypes are imputed at accuracies lower than 1. Imputation accuracy is influenced by factors, such as the size of the RefPop, the genetic distance between the sampled population and the RefPop, the minor allele frequency (**MAF**), and the linkage disequilibrium (**LD**) between genetic markers (e.g., Zhang and Druet, 2010; Van Raden et al., 2013; Pausch et al., 2013; and Uemoto et al., 2015).

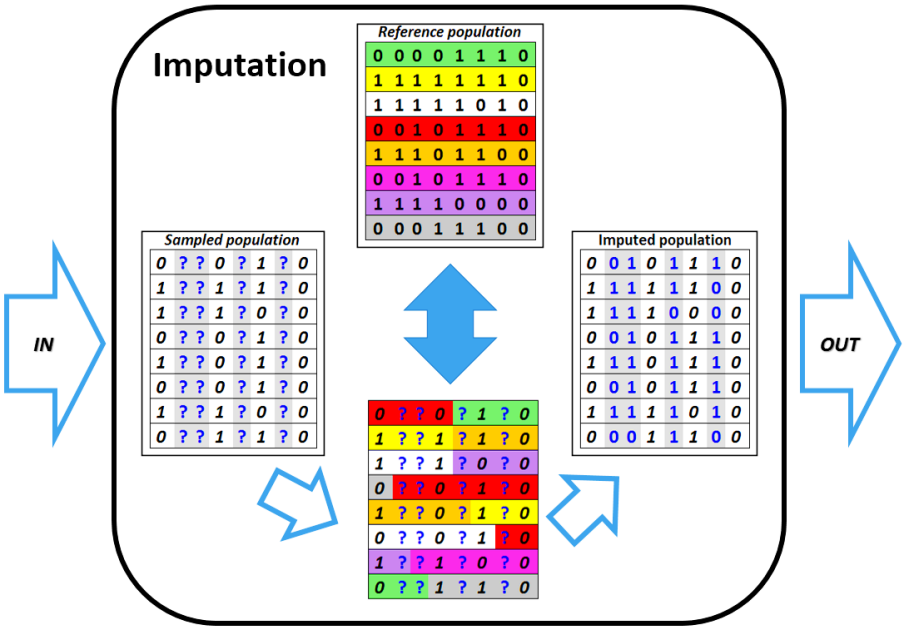


Figure 2.1 – Schematic representation of how imputation works. The sampled population is genotyped at a lower density of genetic markers. The reference population (**RefPop**) contains individuals related with the sampled population that are genotyped at a higher density of genetic markers. Based on the RefPop, the sampled population is imputed to a higher marker density.

Size of the reference population and the genetic distance between the sampled and the reference population. The 1000 Bulls Genome Consortium (Daetwyler et al., 2014) is a world-wide collaborative initiative that aims at sequencing animals from the cattle population, and at creating a multi-breed RefPop. Using this multi-breed RefPop, a substantial increase in the density of genetic markers is currently available for imputation giving the opportunity to impute genotypes to whole-genome sequences (**WGS**). The WGS are available for more than 15 breeds, and each breed is represented by a number of key sequenced influential ancestors. Recently the 1,000 Bull Genome Consortium increased the

number of sequenced animals, and has included sequences of influential cows in this multi-breed RefPop. By accounting for influential cows and bulls, more relationships between the sampled population and the RefPop are considered. Consequently, the accuracy of imputed genotypes should increase.In this multi-breed RefPop, the Hostein-Friesian (**HF**) breed is well represented with 450 HF sequenced ancestors (the latest *Run5*). In contrast, the Swedish Red (**SR)** breed is represented with 16 SR sequenced ancestors and the Finnish Ayrshire (**FAY**) breed is represented with 17 FAY sequenced ancestors.

In Chapter 4, we aimed at imputing the imputed 777k SNP genotypes of HF cows to WGS level. Therefore, only HF sequences (N=450) from the multi-breed RefPop were used to impute genotypes to WGS level, and due to the size of the RefPop, at high accuracies (> 0.9). In contrast, in Chapter 5, a rather limited number (N=33) of sequenced SR and FAY were available for the imputation to WGS level. The 33 sequenced SR and FAY bulls have a large impact in the SR cow population. To make the best possible use of the multi-breed RefPop, our approach in Chapter 5 consisted of imputing a variant three times, each time with a different RefPop (33 SR and FAY sequences, 284 dairy-breeds sequences, and 429 beef- and dairy-breeds sequences). Subsequently, we were able to impute the genotypes of SR cow population to WGS. Based on the findings of Chapter 5, the accuracies of imputed genotypes in smaller breeds (e.g., SR) will only improve if the addition of sequenced animals in the multi-breed RefPop is tailored toward smaller breeds.

Minor Allele Frequency. According to Daetwyler et al. (2014) imputation errors for low MAF (< 0.05) genetic markers are high when imputing a cow population to WGS level. If an allele segregates at low MAF, then there is a relatively small number of sequenced ancestors in the RefPop carrying this low MAF variant. Hence, the imputation of this low MAF variant in the sampled population will be more difficult, and there is a high probability that this variant will be poorly imputed. Therefore, the interpretation of GWAS findings needs more caution when significant associations concern imputed low MAF variants. GWAS detects QTL with genetic markers at a certain power. This detection occurs under the assumption that a genetic marker is correlated with the QTL. MAF at the QTL is an important determinant of power because the heritability of a QTL is directly proportional to the frequencies of the alleles at the QTL locus (Sham and Purcell, 2014). In this context, the power of detecting a QTL segregating at low MAF is low. In addition, the power of detecting this QTL becomes even lower when using imputed low MAF variants, especially if their imputation accuracy is low. If a variant has low MAF, low imputation accuracy

and is strongly correlated with the QTL, this implies that QTL effect size needs to be sufficiently high to be detected by GWAS. In Chapter 4, the 8 strongest associations with milk-fat composition segregate at a MAF=0.44. For the findings of Chapter 4, the imputation accuracy of low MAF variants was not an important issue. In Chapter 5, the 3 strongest associations with NC milk were segregating at a MAF=0.03 and explained more than 10% of the phenotypic variance. This strong signal, which was first detected in SR cows genotyped for 777k SNP genotypes, can be explained by a large QTL effect of more 1 phenotypic standard deviation. This illustrates that rare variants should not by default be considered sequencing errors and therefore excluded from GWAS.

The inclusion of pedigree information can improve the accuracy of imputation of low MAF genetic markers. This approach focuses on imputing identical-by-descent genetic markers that segregate from parents to offspring instead of using information on LD between genetic markers. However, this approach is computationally time-consuming. Some examples of softwares with implemented algorithm that account for simple pedigree information (i.e., duos and trios) are Beagle, fastPHASE, and Fimpute. Recently, a method that imputes SNP combining LD and identical-by-descent information has been proposed (iBLUP, Yang et al., 2014). In general, accounting for pedigree information is expected to impute low MAF genetic markers more accurately than without pedigree information.

Linkage disequilibrium. The non-random association between two loci is defined as LD. Two sampling processes cause LD to arise in a population according to Hill and Weir (1980). First, the sampling of gametes from parents to offspring, and this process depends on the effective population size. Second, the number of individuals sampled from a finite population. In the case of cattle, crossbreeding, mutation, drift, and small population size are events that create LD. Imputation uses LD present in the RefPop to impute the genotypes of the sampled population. One of the problems is that LD can exist between an (imputed) marker and QTL in one family but not in other families (Goddard and Hayes, 2012). For Chapters 3 and 4, the sires of the sampled population of HF cows were included in the RefPop, and in Chapter 5, this was also the case with the 33 sequenced ancestors of SR and FAY. However, in Chapter 5, we also used two other imputation scenarios that included different breeds, for which the sequenced SR and FAY have no common ancestors. In this case, LD in SR and FAY breeds can be different than LD in other breeds. LD across-breeds is expected to be smaller than LD within a breed because more recombination events separate individuals from different breeds (De Roos et al., 2008). Therefore,

imputation accuracy is probably influenced by the differences in LD within- and across- breeds, which might result in lower imputation accuracies for genotypes in small breeds compared with large breeds.

In both cases, for milk-fat composition and for NC milk, imputation to high-density genotypes was challenging. The factors affecting imputation and their consequences on the interpretation of GWAS and fine-mapping results cannot be solved with the data at hand. Only through validation studies it will be possible to confirm the findings reported in this thesis. Validation studies would further help to ascertain if the strongest associations identified in Chapters 3, 4 and 5, and thus the most likely candidate genes, can be confirmed. If a validation study would be based on multiple breeds and these associations persist across breeds, the genetic markers are likely to be very close to the QTL, because of the limited extent of LD across-breeds (e.g., De Roos et al., 2008; Goddard and Hayes, 2012). However, we cannot exclude the possibility that the QTL might not segregate in other breeds (Goddard and Hayes, 2012). Nonetheless, by attempting to validate our associations, it would lead us closer to the identification of the causal variants for the QTL identified in Chapters 3, 4 and 5.

2.2.2 Annotation of the cattle genome

The second major challenge encountered in Chapters 3, 4 and 5 was the limited, hence, incomplete annotation of the cattle genome. The cattle genome contains the genetic information organized in chromosomes, which include the genes for the protein coding regions, and the DNA sequences for the non-protein coding regions. The genome annotation attaches to these genes and DNA sequences the biological information of an organism (Stein, 2001). In Chapter 3, the QTL region located between 29 and 34 mega bp on BTA17 contained 29 genes. A total of 18 out of the 29 genes had not been annotated yet. Among these 18 genes, the non-annotated *LOC515517* was the gene closest to our strongest association on BTA17, and was pointed out as a suggestive candidate gene in Chapter 3. *LOC515517* was assigned this symbol because the investigation of all orthologs for this gene was incomplete. Orthologs are genes in different species that evolved from a common ancestral gene by speciation. The full determination of orthologs assist in the annotation of a gene. Two years later, this QTL region was re-analyzed with imputed sequences (Chapter 4). In these two years, the non-annotated *LOC515517* has been annotated as the *LARP1B* gene in the cattle genome. In Chapter 4, the *LARP1B* gene became our primary candidate gene because 6 out of the 8 strongest associations were located in this gene. In two years, a clear improvement has been made on the annotation of

genes and their biological functions, at least for BTA17. The lesson taken from Chapters 3 and 4 is that the limited annotation of the cattle genome should not be a reason to discard suggestive candidate genes.

The annotation of the genome of domesticated animal species is a slow and complex process. In the last decade, the annotation of the genome of domesticated animal species has been extrapolated from the annotation of the human genome, through actions such as the encyclopedia of DNA elements (**ENCODE**). ENCODE is a global initiative to identify functional variants in high-quality sequences of humans. It is the aim of ENCODE to improve the annotation of structural and regulatory variants as well as non-coding genes in humans. The ENCODE initiative has been very successful in humans, and was expanded to other species like mouse (Shen et al., 2012; Yue et al., 2014). However, the idea of extrapolating gene-expression and its regulation network from human to mouse was not successful because of substantial divergence between these two species (Yue et al., 2014). This genetic diversity between species contributes to the complexity and the slow annotation of the domesticated animal species genomes.

The genetic diversity of domesticated animal species is the focus of the recently started functional annotation of the animal genomes (**FAANG**) consortium. The FAANG consortium aims at identifying all functional elements in the genome of domesticated animal species (Andersson et al., 2015), and involves a collaboration between several research groups worldwide. In a first stage, many different tissues across domesticated animal species will be sampled, such as skeletal muscle, adipose and liver tissues, and in addition, samples of reproductive, immune and nervous systems will be collected. These sampled tissues and systems are necessary to perform functional studies. These studies enable the prediction of the function encoded in sequences. Andersson et al. (2015) argue that filling the genotype-to-phenotype gap requires functional genome annotation of species with substantial phenotype information. The FAANG initiative aims at improving the annotation of the genome of domesticated animal species by creating standardized protocols for sampling, storing, and analyzing the information among the participating research groups (Clarke et al., 2015). The samples will be analyzed by some of the following protocols: transcribed loci (using RNA sequencing), chromatin accessibility and architecture (the link between gene-expression and nuclear organization of cells), and histone modification marks (to identify regulatory elements; Andersson et al., 2015). In a second stage, other tissues will be sampled, such as rumen tissues from ruminant species, mammary tissue from mammals, among others (Andersson et al.,

2015). As pointed out by Zhou et al. (2015), the genomes of chicken, cow and pig have been assembled, but limited information is available on the enhancers, promoters, and other elements of the genome of these species. The identification of these elements and their biological roles will improve the annotation of these three genomes. I expect that it will take some time (> 5 years) to gather and analyze all this information, in order to produce a comprehensive and better annotated genome for each of the domesticated animal species, including cattle. Therefore, the identification of candidate genes will be more efficient in the near future.

2.3 From GWAS to causal variants

The typical outcomes of GWAS are large chromosomal regions, and many polymorphisms that are statistically associated with phenotypes. In Chapter 3, GWAS with imputed 777k SNP genotypes identified a QTL region covering 5 mega bp that contained 29 genes. Subsequent fine-mapping with imputed sequences (Chapter 4) refined the QTL region and reduced the number of candidate genes from 29 to 14. Although this characterization of chromosomal regions associated with our phenotypes (Chapters 3, 4 and 5) was successful, what remains unclear from GWAS and subsequent fine-mapping is whether a polymorphism is the actual causal variant. For complex traits, such as bovine milk composition, it would be interesting to identify causal variants. It would increase biological knowledge, and specifically, help to understand how these causal variants influence our phenotypes. Consequently, it would be possible to predict potential pleiotropic effects on non-(routinely) recorded traits with consequences on the selection of the next-generation of cows. According to Falconer and Mackay (1996), quantitative genetic theory will become more realistic when the numbers and the properties of genes are known because it would improve the methods to studying complex traits. If this is the case, we need to find causal variants to confirm that the identified genes influence the phenotypes. Therefore, in this section, I propose several possibilities to identify causal variants. In more detail, I explore the possibilities of using targeted gene-expression studies, gene-editing, and gene knockouts in livestock to identify causal variants.

2.3.1 Exploring alternatives to identify causal variants

As indicated by Das et al. (2011), the causality of a polymorphism is difficult to be determined by GWAS and fine-mapping. In practice, when GWAS and fine-mapping identify significant associations with the phenotype, the associated variants can be located within protein-coding regions. When this happens, the gene is declared a

candidate gene and the polymorphism might be a causal variant. If the variant is causal, it is possible to predict changes to the encoded-protein, thus predicting functional changes to the phenotype (e.g., Freedman et al., 2011). Consequences on the phenotype can be straightforward for monogenic diseases in humans, such as the Duchenne muscular dystrophy. This disease is caused by large deletions of one or more exon(s) in the dystrophin gene causing severe muscular dystrophy in about 60% of male infants (Hoffman et al., 1987). However, consequences on complex traits are more difficult to interpret than for monogenic diseases. In Chapters 4 and 5, many associations with milk FA composition and with NC milk were identified within and outside protein-coding regions. In Chapters 4 and 5, the *LARP1B* and the *VPS35* genes were nominated as positional candidate genes, after these genes were found expressed in bovine mammary tissue (Bionaz et al., 2012), and during different stages of lactation in humans (Lemay et al., 2013). Figure 2.2 (A and B) illustrates the strongest associations with milk FA composition in the LARP1B gene and with NC milk in the *VPS35* gene. Although we limited the number of candidate genes to only 2, the interpretation of possible functional changes of these 2 genes on milk FA composition and on NC milk are unclear.

Furthermore, two other complications arise. First, the strongest identified associations with milk FA composition and with NC milk are in strong LD (figure 2.2-A and B). Hence, we cannot disentangle which of these associations would promote changes to the phenotypes. Second, some of these correlated associations are intron variants in these candidate genes (figures 2.2 A and B). Particularly in livestock species, there might be a bias in declaring candidate genes toward well–annotated genes (Taşan et al., 2015) because non-coding protein regions still need to be characterized (Andersson et al., 2015). Consequently, associations identified in non-protein coding regions are often ignored. To understand the possible changes to the phenotypes, I hypothesize that the causal variants are among one of the significant associations with the *LARP1B* and *VPS35* genes. If this is the case, this hypothesis can serve as research question for further studies, such as targeted gene-expression studies.

2.3.2 Targeted gene-expression studies

Gene-expression is the process by which functional gene products are formed. Gene products have been studied in many species including mice, rats and humans, and in different cell types (e.g., de Koning et al., 2007; Civelek and Lusis, 2014). Gene

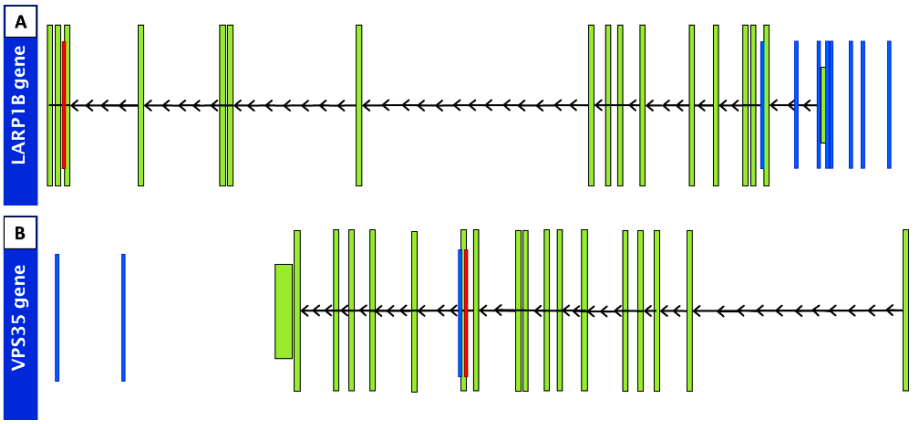


Figure 2.2 – Schematic view of the *LARP1B* and the *VPS35* genes. The green boxes represent the exons connected by a black line and small arrows showing the protein coding direction of the genes. Blue boxes represent the location of the strongest associations, and the red boxes represent the splice region variants. (**A**) The *LARP1B* gene, and its eight strongest associations with multiple fatty acids on Bos Taurus Autosome (BTA) 17 [at –log10 (P-value) = 7.66, and linkage disequilibrium between the eight markers = 1]. (**B**) The *VPS35* gene, and its three strongest associations with non-coagulation of milk on BTA18 [at –log10 (P-value) = 14.12, and linkage disequilibrium between the three markers = 1].

products can be transcripts of genes (mRNA) but equally protein abundance and metabolite levels. The most often analyzed gene products are mRNA rather than protein abundance or metabolite levels (e.g., Albert and Kruglyak, 2015). Typically, the mRNA expression is constantly changing over time (e.g., Jiang et al., 2013). After establishing that most genes are quantitatively expressed, Jansen and Nap (2001) proposed the "genetical genomics" approach. G*enetical genomic*s combines the (quantitative) gene-expression and the genetic variation from related individuals in segregating populations (as a representation of genetic markers).

In *genetical genomics* (or its equivalent genome-wide association of gene-expression studies – **eQTL**), the mRNA abundance is treated as the quantitative phenotype, and the genomic regions influencing gene-expression result in the detection of eQTL (e.g., Jansen and Nap, 2001; Jansen, 2003). According to Jansen and Nap (2001), the eQTLs can act in two ways: a) in *cis* by influencing the expression of the closest gene nearby (also known as *locale* QTL); or b) in *trans* by influencing the expression of genes in other parts of the genome (also known as *distant* eQTL). In animal breeding, Kadarmideen et al. (2006) indicated that eQTLs contribute to the refinement of the identified traditional QTL, candidate gene and SNP discovery. Furthermore, de Koning et al. (2007) combined eQTL and fine-mapping to reduce the confidence

interval of functional trait loci in poultry. As a consequence, the chromosomal region under investigation and the number of candidate genes were reduced. This targeted eQTL approach allows the identification of cis-acting eQTL rather than trans-eQTL. Targeted eQTL are especially important when there is no obvious biological reason supporting a significant association with the phenotypes. The reason being that eQTL can provide further insights into the function, regulation and pathways of genes underlying a complex trait (e.g., Jansen, 2003; de Koning et al., 2007; Lowe and Reddy, 2015). For instance, the *LARP1B* and the *VPS35* genes have not been associated to bovine milk composition before the present thesis. Further insights into the function, regulation and pathways would clarify the functional role of the *LARP1B* and the *VPS35* genes in relation to their respective phenotypes.

According to Hassan and Saeij (2014), if a genetic variant influences the mRNA abundance of a nearby gene, which in turn modulates a complex trait, this cis-eQTL can co-localize with the QTL identified by traditional GWAS. When a common chromosomal region identified by cis-eQTL co-localizes with the QTL from traditional GWAS at the same genetic variant, it provide strong evidence that the underlying candidate gene is correctly identified (Schadt et al., 2005). In addition, this co-localization (if observed) would suggest that the causal variant is associated with the gene-expression and with the phenotype simultaneously (Schadt et al., 2005). Based on these findings, targeted eQTL focused on the expression of the *LARP1B* and the *VPS35* genes would help confirm that the candidate genes were correctly assigned, and help determine the most likely causal variants for these phenotypes.

Nonetheless, targeted eQTL on the expression of *LARP1B* and *VPS35* genes can point out variants in regulatory elements. In humans, some studies have suggested that multiple correlated associations can influence the activity of multiple enhancers (regulatory elements). When the activity of these regulatory elements is coordinated, their effects can alter gene-expression (e.g., Corrandin et al., 2014; Lowe and Reddy, 2015). Albert and Kruglyak (2015) indicated that many polymorphisms identified in human GWAS are over-represented in regulatory regions. In addition, Parikshak et al. (2015) indicated that these regulatory elements are located in non-protein coding regions of the genome. In our case, multiple significant associations with the *LARP1B* and the *VPS35* genes are in strong LD and are located in non-protein coding regions (figure 2.2 A and B). I would investigate if the co-localization of the cis-eQTL with the QTL from a traditional GWAS would occur at one of the variants located in the non-protein coding regions of *LARP1B* and *VPS35* genes. If this would happen, the position of the regulatory element showing the cis-

eQTL effect could be accurately determined based on the sequence data. One limitation, however, is that the regulatory elements of the cattle genome are not annotated yet. In summary, it is possible that the significant associations in strong LD for the *LARP1B* and the *VPS35* genes are regulatory elements.

A step further from targeted eQTL would be to investigate the proteins encoded by the genes directly. This approach would be interesting because of a highly regulated mechanism known as alternative splicing (Hassan and Saeij, 2014). Through this process, introns and exons in genes are re-arranged creating the opportunity for mRNA to synthesize different protein variants (isoforms) that may have different cellular functions (Wang et al., 2008). This process occurs at a specific site known as splice junction (or splice variant). Interestingly, the *LARP1B* and the *VPS35* genes contain splice-region variants (figure 2.2 A and B). Using RNA-sequencing technology, it is possible to distinguish between the transcript abundance from alternative splicing and regular transcript abundance (Trapnell et al., 2010). According to Wickramasinghe et al. (2014), RNA-sequencing technology is the method of choice for studying RNA transcripts, and this technology shows great ability in studying allele-specific expression and non-coding RNA. In a further study, it might be worth investigating the different isoforms resulting from the splice-variants found in the *LARP1B* and the *VPS35* genes with RNA-sequencing.

The contribution of RNA-sequencing is not limited to studying gene-expression. RNA-sequencing can also be used for SNP and gene discovery, as well as gene ontology and pathway analysis. The RNA-sequencing approach is different than genetical genomics. Using RNA-sequencing and gene-expression of bovine milk retrieved from somatic cells, the different isoforms of interesting genes are tested for associations directly with the phenotypes. When a significant association is identified, if this association is identified within the isoforms, then SNP and candidate genes can be identified. Several studies have used this approach to identify candidate genes associated with bovine milk composition (e.g., Cánovas et al.,2010; Wickramasinghe et al., 2012; and Cánovas et al., 2013). It is important to acknowledge the substantial contribution of the RNA sequencing technology for studying bovine milk composition.

2.3.3 Gene-editing and gene knockouts in livestock

A complementary approach to gene-expression studies is targeting genes in mouse models. Targeting a gene in mouse models means to disrupt a specific gene in the genome of a mouse, thus creating a knockout mouse for that specific gene. In the

last 50 years, gene targeting by means of homologous recombination combined with the refinement of protocols (e.g., microinjection of purified DNA, electroporation, and positive selection enrichments) and the subsequent transmission to mouse germlines have led to knockout more than 7,000 genes in transgenic mouse models (Capecchi, 2005). The "principles for introducing specific gene modifications in mice by the use of embryonic stem cells" have made Dr. Capecchi, Dr. Evans and Dr. Oliver winners of the Nobel Prizes in Physiology or Medicine in 2007. This refinement of methods and protocols has substantially accelerated the biological knowledge of genes, and has led to the development of gene-editing.

Gene-editing. Although gene targeting has required the introgression of exogenous DNA into the genome of a mouse, gene-editing with site-specific nucleases is an alternative to target specific genes without the introgression of exogenous DNA (e.g., Capecchi, 2005; Carlson et al., 2014). According to Cappechi (2005), the use of these site-specific nucleases allow to target a series of alleles in the same gene, thus manipulating any chosen allele in mouse models. There are at least three known site-specific nucleases: the zinc-finger nucleases (Kim et al., 1996), the transcription activator-like effector nucleases (Boch et al., 2009; Moscou and Bogdanove, 2009), and the clustered regularly interspaced short palindromic repeats associated endonuclease cas9 (**CRISPR/Cas9**; Cong et al., 2013; Mali et al., 2013). My focus will be on the most recent, the CRISPR/Cas9 system.

The CRISPR/Cas9 system is part of the protection mechanism against viruses that has been identified from the immune system of bacteria. The CRISPR/Cas9 was first described by Cong et al. (2013) and by Mali et al. (2013), as a RNA-guided site-specific DNA cleavage technique. According to Cong et al. (2013), the Cas9 nuclease can direct short RNAs to induce precise cleavage at DNA loci, facilitating the knockout of targeted genes. Initially, the CRISPR/Cas9 technique was intended to understand genes, their regulation and their biological functions because of its easiness of programmability and of usage (Cong et al., 2013). Gene-editing has the potential of targeting a single gene as well as multiple genes simultaneously. Gene-editing can be used to obtain cell-specific knockdown (one copy of the gene inactivated) or knockout (both copies of a gene inactivated) as well as gene specific mutations using rodent models (Shalem et al., 2015). For this reason, it has become an important ally to study genes underlying complex traits, such as bovine milk composition. For bovine milk composition, gene-editing has the potential to accelerate knowledge discovery (about genes, their biological function, and their influence at the

phenotypic level). On this regard, gene-editing is substantially contributing to improve the annotation of domesticated animal species genomes, including cattle.

Gene knockouts in livestock. With gene-editing, some gene knockouts in livestock have been successfully produced. With the zinc-finger nuclease, the knockout of the *PPARγ* gene in pigs (Yang et al., 2011) and of the β-*LG* gene in cattle (Yu et al., 2011) was possible. However, Carlson et al. (2014) indicated that proprietary algorithms were responsible for impeding the use of this zinc-finger nuclease. With the transcription activator-like effector nucleases, Proudfoot et al. (2015) reports the gene-editing of the myostatin (**MSTS**) gene in sheep and in cattle with successful results. In the future, using gene-editing with the CRISPR/Cas9 technique, knockout cows are likely to be produced. The resulting (functional) changes will be interpretable at the phenotypic level. It would be useful to understand the extent of changes from one or multiple genes on bovine milk composition, but also on the important physiologic changes faced by cows at parturition. For phenotypes such as bovine milk, I foresee in the coming future gene knockout cows being widely produced, kept and challenged in a commercial environment. I can also foresee the knockdown of one or multiple alleles in the *LARP1B* and the *VPS35* genes, as well as the knockout of these genes in gene-edited cows.

While gene-editing with the CRISPR/Cas9 technique will become widely used in the future, functional changes in bovine milk composition can already be studied using a lactating bovine mammary epithelial cell (**bMEC**) model. Zhao et al. (2010) and Jedrzejczak and Szatkowska (2014) indicated that bMEC models are suitable to study bovine milk synthesis. Instead of using bMEC sampled from tissues through biopsy, Boutinaud et al. (2002) isolated mRNA directly from somatic cells, which are naturally released in milk during lactation. Using RNA sequencing, Medrano et al. (2010) and Cánovas et al. (2014), both concluded the viability of using milk somatic cells and milk fat globules to study mammary gland expression. For bovine milk composition, functional changes to be phenotypes can already be assessed by studying the gene-expression of *LARP1B* and the *VPS35* genes directly from milk samples. In addition, it is also a possibility to target one or multiple alleles in a single gene (e.g., the *LARP1B* and the *VPS35* genes) using bMEC models.

In summary, there are many opportunities to transform the significant associations identified from traditional GWAS and fine-mapping in research questions for further studies. All the approaches discussed in this section would, a priori, help to identify causal variants underlying complex traits such as bovine milk composition, and a

posteriori, help to understand the function of genes and their biological role in bovine milk.

2.4 References

Albert, F. W., and Kruglyak, L. 2015. The role of regulatory variation in complex traits and disease. Nature Rev Genet 16: 197-212.

L. Andersson, Archibald, A. L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., et al. 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol 16:57

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

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. Science 326: 1509–12. doi:10.1126/science.1178811.

Boutinaud, M., and Jammes, H. 2002. Potential uses of milk epithelial cells: a review. Reprod Nutr Dev 42:133-147.

Cánovas, A., Rincon, G., Islas-Trejo, A., Wickramasinghe, S., and Medrano, J. F. 2010. SNP discovery in the bovine milk transcriptome using RNA-Seq technology. Mamm genome 21: 592-598.

Cánovas, A., Rincón, G., Islas-Trejo, A., Jimenez-Flores, R., Laubscher, A., and Medrano, J. F. 2013. RNA sequencing to study gene expression and single nucleotide polymorphism variation associated with citrate content in cow milk. J Dairy Sci 96: 2637-2648.

Cánovas, A., Rincón, G., Bevilacqua, C., Islas-Trejo, A., Brenaut, P., Hovey, R. C., et al. 2014. Comparison of five different RNA sources to examine the lactating bovine mammary gland transcriptome using RNA-Sequencing. Sci Rep 4. doi:10.1038/srep05297

Capecchi, M. R. 2005. Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. Nat Rev Genet 6: 507-512.

Carlson, D. F., Tan, W., Hackett, P. B., and Fahrenkrug, S. C. 2014. Editing livestock genomes with site-specific nucleases. Reprod Fertil Dev 26: 74-82.

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

Clarke, L, Archibald, A. L.,Flicek, P., Burt, D. ,Hume, D., Vernimmen, D., et al. 2015. The functional annotation of animal genomes, data standards, annotation and sharing. In Plant and Animal Genome XXIII Conference. Plant and Animal Genome.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. Science 339:819-823.

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., et al. 2014. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res 24: 1–13.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R. F., et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46, 858–865. doi:10.1038/ng.3034.

Das, S. K., and Sharma, N. K. 2014. Expression quantitative trait analyses to identify causal genetic variants for type 2 diabetes susceptibility. World J Diabetes 5:97–114.

de Koning, D. J., Cabrera, C. P., and Haley, C. S. 2007. Genetical genomics: combining gene expression with marker genotypes in poultry. Poultry Sci 86:1501-1509.

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

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

Falconer, D. S., and Mackay, T. F. C. 1996. Introduction to Quantitative Genetics. Correlated characters: genotype-environment interaction. Pages 321-325. Fourth edition, ed. Longman Greens, Harlow, Essex, UK.

Freedman, M. L., Monteiro, A. N., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C., et al. 2011. Principles for the post-GWAS functional characterization of cancer risk loci. Nat genet 43:513-518.

Goddard, M. E., and Hayes, B. J. 2012. Bovine Genomics. Linkage disequilibrium in cattle. Pages 192-210. Ed. John Wiley and Sons, Inc. West Sussex, UK.

Hassan, M. A., and Saeij, J. P. 2014. Incorporating alternative splicing and mRNA editing into the genetic analysis of complex traits. BioEssays, 36:1032-1040.

Weir, B. S., and Hill, W. G. 1980. Effect of mating structure on variation in linkage disequilibrium. Genetics 95:477-488.

Hoffman, E. P., Brown, R. H., and Kunkel, L. M. 1987. Dystrophin: the protein product of the Duchenne muscular dystrophy locus. Cell, 51:919-928.

Jansen, R. C., and Nap, J. P. 2001. Genetical genomics: the added value from segregation. Trends Genet 17:388-391.

Jansen, R. C. 2003. Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 4: 145-151.

Jedrzejczak, M., and Szatkowska, I. 2014. Bovine mammary epithelial cell cultures for the study of mammary gland functions. In Vitro Cell Dev Biol Anim 50: 389-398.

Jiang, J., Cui, W., Vongsangnak, W., Hu, G., and Shen, B. 2013. Post genome-wide association studies functional characterization of prostate cancer risk loci. BMC genomics 14:S9.

Kadarmideen, H. N., von Rohr, P., and L. L. Janss. 2006. From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. Mamm Genome 17:548-564.

Kim, Y. G., Cha, J., and Chandrasegaran, S. 1996. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proc. Natl Acad. Sci. USA 93:1156–1160. doi:10.1073/PNAS.93.3.1156

Lemay, D. G., Ballard, O. A., Hughes, M. A., Morrow, A. L., Horseman, N. D., and Nommsen-Rivers, L. A. 2013. RNA sequencing of the human milk fat layer transcriptome reveals distinct gene expression profiles at three stages of lactation. PloS one 8:e67531.

Lowe, W. L., and Reddy, T. E. 2015. Genomic approaches for understanding the genetics of complex disease. Genom Res 25:1432-1441.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., et al. 2013. RNA-guided human genome engineering via Cas9. Science 339:823-826.

Marchini, J., and Howie, B. 2010. Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499-511.

Medrano, J.F., Rincon, G.,and Islas-Trejo, A. 2010. Comparative analysis of bovine milk and mammary gland transcriptome using RNA-Seq. In: 9th World congress on genetics applied to livestock production, Leipzig, Germany, August 1-6, 2010, Paper no 0852.

Moscou, M. J., and Bogdanove, A. J. 2009. A simple cipher governs DNA recognition by TAL effectors. Science 326:1501-1501.

Parikshak, N. N., Gandal, M. J., and Geschwind, D. H. 2015. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nat Rev Genet 16:441-458.

Pausch, H., Aigner, B., Emmerling, R., Edel, C., Götz, K. U., and Fries, R. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet Sel Evol 45:10-1186.

Proudfoot, C., Carlson, D. F., Huddart, R., Long, C. R., Pryor, J. H., and King, T. J., et al. 2015. Genome edited sheep and cattle. Transgenic Res 24:147-153.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., et al. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710-717.

Shalem, O., Sanjana, N. E., and Zhang, F. 2015. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet 16:299-311.

Sham, P. C., and Purcell, S. M. 2014. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet, 15:335-346.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., et al. 2012. A map of the cis-regulatory sequences in the mouse genome. Nature 488:116-120.

Stein, L. 2001. Genome annotation: from sequence to biology. Nat Rev Genet 2:493-503.

Taşan, M., Musso, G., Hao, T., Vidal, M., MacRae, C. A., and Roth, F. P. 2015. Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. Nat Methods, 12:154-159.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. 2010. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat Biotechnol 28: 511–515.

Uemoto, Y., Sasaki, S., Sugimoto, Y., and Watanabe, T. 2015. Accuracy of high-density genotype imputation in Japanese Black cattle. Anim Genet 46: 388-394.

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., et al. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci 96:668-678.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456:470-476.

Wickramasinghe, S., Rincon, G., Islas-Trejo, A., and J. F. Medrano. 2012. Transcriptional profiling of bovine milk using RNA sequencing. BMC genomics 13:1

Wickramasinghe, S., Cánovas, A., Rincón, G., and Medrano, J. F. 2014. RNA-sequencing: a tool to explore new frontiers in animal genetics. Livest Sci, 166: 206-216.

Yang, D., Yang, H., Li, W., Zhao, B., Ouyang, Z., Liu, Z., et al. 2011. Generation of PPARγ mono-allelic knockout pigs via zinc-finger nucleases and nuclear transfer cloning. Cell Res 21:979.

Yang, Y., Wang, Q., Chen, Q., Liao, R., Zhang, X., Yang, H., et al. 2014. A new genotype imputation method with tolerance to high missing rate and rare variants. PloS one 9:e101025.

Yu, S., Luo, J., Song, Z., Ding, F., Dai, Y., and Li, N. 2011. Highly efficient modification of beta-lactoglobulin (BLG) gene via zinc-finger nucleases in cattle. Cell Res. 21:1638–1640. doi:10.1038/CR.2011.153

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515:355-364.

Zhang, Z., and Druet, T. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J Dairy Sci 93:5487-5494.

Zhao, K., Liu, H. Y., Zhou, M. M., and Liu, J. X. 2010. Establishment and characterization of a lactating bovine mammary epithelial cell model for the study of milk synthesis. Cell Biol Int 34: 717-721.

Zhou, H., Ross, P.J., Korf, I., Delany, M. E., Cheng, H., Medrano, J. F., et al. 2015. Annotation of functional regulatory elements in livestock species. In ADSA-ASAS 2015 Midwest Meeting. Asas.

Summary

Summary

The present thesis aims at unraveling the genetic background of bovine milk composition by finding genes associated with milk-fat composition and non-coagulation of milk. The fine-mapping was realized by increasing the number of genotypes analyzed in the targeted chromosomal regions. This allowed to increase the resolution for these genomic regions and pin-point candidate genes associated with bovine milk composition.

In **Chapter 2**, we analyzed milk fat composition in winter and summer and estimated in both seasons' genetic parameters, the effects of acyl-CoA: diacylglycerol acyltransferase1 (***DGAT1***) K232A and stearoyl-CoA desaturase1 (***SCD1***) A293V polymorphisms. Furthermore, we estimated genetic correlations between winter and summer milk fatty acids and tested for genotype by season interactions of *DGAT1* K232A and *SCD1* A293V polymorphisms. Phenotypes consisted of gas chromatography measurements (%w/%w) of seventeen individual fatty acids (C4:0 to C18:0, C10:1 to C18:1cis-9, C18:1trans-11, C18:2cis-9,trans-11 (**CLA**), C18:2cis-9,12 and C18:3cis-9,12,15), groups of fatty acids (saturated FA (**SFA**), unsaturated FA (**UFA**) and the ratio SFA to UFA), and six unsaturation indices (C10 index – CLAindex). These phenotypes were available for 2,001 cows in winter and in summer milk samples. We showed that the genetic correlations between winter and summer milk FA were very high, and these indicated that milk-fat composition in winter and in summer can largely be considered as genetically the same trait. We showed that effects of *DGAT1* K232A and *SCD1* A293V polymorphism were very similar in winter and in summer milk for most FA. At last, we tested for genotype by season interactions, and demonstrated significant *DGAT1* K232A by season interaction for some FA. A *SCD1* A293V by season interaction was only found for C18:1trans-11. These genotype by season interactions were due to scaling of genotype effects.

In **Chapter 3** and in **Chapter 4**, we used a subset of the fatty acids analyzed in **Chapter 2**. This subset consisted of six individual FA (C4:0 - C14:0) were available for winter and for summer milk samples.

In **Chapter 3**, a quantitative trait locus (**QTL**) on *Bos taurus autosome* (**BTA**) 17 explaining a large proportion of the genetic variation in de novo synthesized milk FA was fine-mapped. This QTL region has been identified previously using 50k SNP genotypes. We fine-mapped this QTL region with imputed 777k single nucleotide polymorphism (**SNP**) genotypes to identify candidate genes associated with milk FA composition. Single-SNP analyses showed that several SNP in a region located

between 29.0 and 34.0 mega base-pairs were in strong association with C6:0, C8:0, and C10:0. This region was further characterized based on haplotypes, and these analyses suggested the presence of one causal variant. Although many genes are present in this QTL region on BTA17, the strongest association was found close to the progesterone receptor membrane component 2 (***PGRMC2***) gene. This gene has not been associated previously to milk FA composition.

In **Chapter 4**, the chromosomal region associated with de novo synthesized milk FA on BTA17 was further re-fined using imputed whole-genome sequences (**WGS**). WGS were available for 450 Holstein-Friesian (**HF**) animals (the 1000 bull genome consortium (*Run5*) and 45 HF sequenced animals from the Dutch Milk Genomics Initiative. Based on these 495 HF sequences, all cows were imputed from (imputed) 777k SNP genotypes to sequence level. Single-marker analyses identified many significant associations (in the thousands) with c6:0, c8:0, c10:0, c12:0 and c14:0. Most significant associations were detected in a region covering 5 mega base-pairs and in this region a total of 14 genes could be identified. Six out of the 8 SNP that showed the strongest associations were located in the LA ribonucleoprotein domain family, member 1B (***LARP1B***) gene. This candidate gene has not been associated with milk-fat composition before.

In **Chapter 5**, firstly, we performed a GWAS using 777k SNP genotypes to identify the most promising genomic regions associated with non-coagulation (**NC**) of milk in Swedish Red cows. Secondly, we fine-mapped the most promising genomic region using imputed sequences. Individual morning milk samples were available for the 382 Swedish Red cows that were also genotyped using a 777k SNP array. Using 429 sequences from the 1000 bull genome consortium (*Run 3*), all cows were imputed from 777k to sequence level. Single-marker analyses identified 14 associations with NC milk in a 7 mega base-pairs region on BTA18. For this region, our strongest association explained almost 34% of the genetic variation in NC milk. Haplotypes were built, genetically differentiated by means of a phylogenetic tree, and tested in phenotype-genotype association studies. A candidate gene is the vacuolar protein sorting 35 homolog, mRNA (***VPS35)*** gene, for which one of our strongest association is an intron SNP in this gene. The *VPS35* gene belongs to the mammary gene sets of pre-parturient and of lactating cows, and has not been associated to milk composition yet.

In **Chapter 6**, the general discussion is presented. Firstly, I discuss the imputation to high-density genotypes and the annotation of the cattle genome. I discuss what

imputation is, the factors which affect imputation accuracy, and the consequences of using imputed genotypes for GWAS and fine-mapping studies. Regarding the annotation of the cattle genome, I discuss the major difficulties in finding candidate genes with the current annotation, and discuss future initiatives that will contribute for a better annotation of genomes in the future.

Secondly, the future possibilities to expand gene discovery are discussed. In this section, the discussion starts with the importance of identifying causal variants underlying complex traits. The discussion continues by exploring possibilities, such as targeted gene-expression studies, eQTL, gene editing and knockout cows, to identify the causal variants underlying complex traits

Acknowledgements

Acknowledgements

To God: Thank you for this third opportunity.

To my friends and colleagues: *Acknowledgements are always a very difficult task to write. And throughout this PhD, lots of people have contributed directly and indirectly to this achievement. I would like to say thank you to each and every one of you who contributed, but in a different way.*

This is the year 2009 and I am decided to make some changes. Yet, I have no idea what is to come. Guided by my will, this idea grows stronger and stronger inside my heart. After a few clicks and a directed search on the internet, I find EM-ABG. The advertisement seem too good to be true. Never mind: I subscribe. The road ahead is unknown, and one of the most important journeys of my life is about to start.

Exactly three days after I subscribed, I receive an e-mail from the captain of ABGC, Johan Van Arendonk, asking me if I would like to apply for a scholarship that would cover my living expenses while on board. I will never forget that I really thought it was a phishing attempt. After successfully getting the scholarship, I travel to this far distant new world called the Netherlands. In my luggage, some pieces of clothes and a heart full of hope and eager for adventure. After 26 hours of travel, I finally arrive to this beautiful place called Wageningen Bay.

What an exciting first view! Beyond the main deck of ABGC, I can see Forum Building as the harbor that connects all the other ships. The joy and the excitement are suddenly cut by the voice of the captain: "You have the opportunity and the privilege to be part of this diverse and multicultural team. Enjoy the training, the trip, and have fun!". After a few introductions, EM-ABG are sent to the hold of ABGC ship, where during two years, me and my colleagues will struggle with codes, cleaning data and learning all aspects of the genetic architecture of traits in Animal Breeding and Genetics. As final exam, I am challenged to sail across these beautiful and calm waters of Wageningen Bay. The final result is priceless! After two unforgettable years, the training is completed.

I would like to kindly thank Johan, Dieuwertje, Patricia, Marleen, Aniek, Ada, Gerda, Piet, Eduardo, Christelle, Andrés, Guillaume and all the teachers for their support, guidance and friendship during EM-ABG. I would kindly thank the Koepon family for the amazing opportunity that they offered me.

This is the year 2011, and new challenges have been announced: there is a possibility of subscribing to EGS-ABG. The catchy advertisement comes with a difficult mission: sailing to the North in the open sea. Without hesitation, I subscribe. ☺ ."All on

board", shouts Captain Johan! EGS-ABG gathers together for the first time. The main deck is a huge promotion for most of us. Some came with more experience than others, and the group is very diverse. At first sight, this is going to be challenging. The main deck is indeed a huge responsibility. But we are not alone, at least we think so! All PhD receive specific jobs, but our destination remains unknown. Only the captain and his crew know the direction ABGC ship is heading for. The sails are lifted, and in no time, we leave the quiet and calm waters of Wageningen Bay!

Under the supervision of Colonel Henk and Major Marleen, I happily start my task. After a few months at sea, the excitement has been replaced by a tedious and continuous routine. Asreml, Excel, Linux and R are just part of the job, which is complemented with endless meetings with Colonel Henk and Major Marleen. To keep the spirit alive, some strategic stops are planned, like harbors Pub-Quiz, WE-day and ABGC day-outs. Ahead of us, the first storm in sight: the huge storm coined "Paper One". Paper One Storm soon brings lots of bumpy waves and strong winds. Winds from the North and South reviewers that seemed to battle endlessly with us on the main desk. I almost was thrown out of the main deck. Colonel Henk shouting endless orders, followed by obedient Major Marleen, and a beaten up PhD Sandrine. "Pull the sails down!" shouts Colonel Henk, "The reviewers are angry", he continues. "We need to hold ourselves, 'cause these winds are too strong!!!!". Milk Genomics meetings, presentations, minutes, discussions, posters, endless shift hours, few sets of brilliant ideas, a list of new suggestions, and frustration stepping in at high speed. These were unusual times for me, and all my expectations changed. Would I be able to continue? At these times, the excellent team of PhDs is like an island of comfort in these troubled waters. After discussing and sharing our deepest fears and frustrations, the morale of the PhDs substantially improves. Motivated as I have never been before, I think: "Let's go through this storm, let's do this!". Welcome meetings, presentations, minutes, discussions, posters, QDG, TLMs! Finally, Paper One Storm has passed; and I remember thinking: "OUF, I survived!".

I would like to kindly thank Johan, Henk and Marleen for their guidance and support throughout the PhD. Yes, I do not come with a manual, but neither do you. ☺. I would kindly thank CRV for their financial support for the last year of my PhD. I would kindly say thank you to Erik Mullaart for your constant interest in my work, Daylan, Elsa, Kasper, and Hein for the nice discussions within Milk Genomics. I would like to say thank you to Mahlet, Marzieh, Yogesh, Hooiling, Troncg, Susan, Ewa, Katrijn, Naomi, Gabriel, Marcos, Hamed, Mirte, Bert, Kimberly, Sabine, Tessa, Jovana, Sonia, Maria, Zih-Hua, Anoop, Maulik, Vinicius, Coralia, Amabel, Mathijs, Claudia, Kasper, Saskia, Mathieu, Floor, Qiuyu, Mandy, Wosseni, Robert, Haibo, Shuwen, Yvonne, Esther, Ilse,

Anouk, Aniek, Jérémie, Alex, Rosilde and Maya. I would also like to say thank you to Pim, Henry, Jan, Piter, Liesbeth, John, Richard, Martin, and all the other staff members for all the discussions at QDG and at lunch breaks.

In subsequent years, ABGC ship came across some other important storms. I can say Paper One Storm prepared me for the next storms that were still to come. However, nothing was as frightening as in 2013 when the sea started shaking so much that I was sea-sick. This has never happened before. After receiving a lot of help from my good friend Marshall Dieuwertje, I discover that I have to go back to Rio de Janeiro Bay and stay some time recovering while on land. Before I left, Captain Johan was very supportive "Sandrine", he said, "Take your time, health is more important than anything. When you are fully recovered you come back." How grateful I am to have this kind of support. I leave ABGC ship thinking: "I will be back before you know it".

A few months later, I return to ABGC ship. A part of me is excited. I miss being at ABGC, I miss the EGS-ABG gang, all the other PhDs, I miss the Marshalls, the nice friends and colleagues, and I miss the blue Sea of Knowledge that lies in front of ABGC ship. The other part of me is different. I have deeply changed after the sickness, and things do not look the same. It seems that time has continued for everyone, and it has stopped for me. Caught in my thoughts, I hear this voice behind me, "Oh dear, don't be sad, everything is going to be fine". I look back, and see Marshall Ada. She continues: "Your program has been upgraded. You just need time to get used to it. All will be fine at the end. You will see, relax, and no worries". I am so grateful to be hearing this. And Marshall Lisette adds a little more: "No worries, we, 1975 are the best! I am sure you will recover in no time. Hey girl, we are '75s! Uh-u!". My heart is feeling lighter again, and I think proudly to myself: "Yes, 'mam. I am a '75s. Go for it!".

Dieuwertje, I will never forget how much you helped me. Thank you! For all the support and help on this difficult phase, I acknowledge Dr. Cafure and his family, my family, Johan, Marleen, and Henk. I would like to say thank you for the amazing support and hard work that Ada and Lisette did. "Lieve Dames, dank jullie wel!"

This is the year 2014, and on this very sunny day, Captain Johan, Colonel Henk and Major Marleen altogether announce my final destination: "Sandrine", said captain Johan, "You are going to the North Pole. There, you will spend some time in a ship called SLU. The captain is a good friend of mine and you can learn lots of things from him and his crew. I argued back: "Captain, my Captain! These are dangerous waters. I am going to freeze to death!" "Naja", says Captain Johan, "You just need some good clothes, then it will be OK!". Colonel Henk watching me worried, says "Sandrine, keep

an eye on polar bears. Beware of sliding bears! They can swipe you out of the deck!". "Safe trip! "said Major Marleen. After waving goodbye to all colleagues and friends, and gathering nice tips from my fellow PhD Dianne, my puzzlement was replaced by the eagerness of discovering this new boat, place and crew.

It is on a summer sunny day when I finally reach ship SLU. This boat was somewhat surprising; the main deck was round. I was a little lost at first, especially because so many people around me were saying "Fiiiikka!". I could not stop thinking: "What a strange language!". "AH, AH", says this voice at the far end of the deck. "You made it! Welcome, welcome to the main deck of the SLU ship. By the way, I am Captain DJ and this is my crew: Major Freddy, Lieutenants Fernando and Lisa. You also know Nancy and André!". It was so nice to see these familiar faces. Very supportive PhDs Nancy and André helped me settling in very fast. In no time, the round deck became a very familiar place. But there was that dark side of the deck. I turn to André, and ask: "Hey bro, what is on that dark side of the deck?". "Sandrine, follow me", he said. In no time, we step into the dark side, and André says: "Meet the SLU Mafia!". "Hey, bro! Who is THAT? You are not supposed to bring strange people in." says this PhD to André. She turns to me and says: "My name is Agnese, and I am sort of the leader of the SLU mafia! And these are Merina, Chrissy, Bingjie, Ahmed, Thu, Shizhi, Xiaowei, and all the others! This is where all the PhD gather and organize many parties and all sorts of activities! You are most welcome to join! By the way, Fiiiiikkka." I thought "And here we go again". ☺

It was mid-October 2014 and strong winds were bringing very dark clouds that marked the beginning of the winter. The forecast was announcing light snow for the evening, and at the main deck, I noticed that the days were getting shorter quite rapidly. Captain DJ in his usual good shoes was sort of inspired: "Sandrine, the weather is not an issue, we are inside the ship. For some months, the main deck will remain closed, and we will be stuck in the North Pole until spring, next year." I say: "WHAT????? Spring is in April, we are gonna die!" Major Freddy and Lieutenant Fernando started their usual jokes "Ah, Ah, we are gonna die indoors, so we will go out to ski, ice-skate and all sorts of nice things! It will be fun! You will see!". The next morning the weatherman announces: "Yesterday, it only snowed one meter of snow." "Whow, this winter is gonna be promising", I thought.

I would like to say thank DJ and Freddy for all the support that you gave me while in Sweden and afterwards. Thank you Maria and Marie for all the nice comments. A special thank you to Lisa because you let me stay two months in your house, and I am really grateful for this. Fernando, Karl and Cano thanks for keeping me smiling. A

special thank you to the SLU Mafia. Thank you for all the amazing stuff we did together: "Guys, you rock!". Thanks Sofia, Kim, Emilie, Thomas, Eva and Valentina for the nice discussions.

Thank you Dianne for the tips before I went to Sweden, especially the one to go to Kiruna! It was fantastic! Nancy and André: Hey you two! I shall never forget you! After a PhD, we shared so many moments! I can only say: thanks for everything.

This is the year 2015 and spring makes its way in this rather dark room. This new ship ABGC 2.0 is located in the middle of this rather dark forest. A change that I notice, especially after spending sometime at the North Pole. This is the last chapter of this tremendous adventure called EGS-ABG to me. I have experienced so much, and many PhD have harvested their thesis already. The direction set for me now is towards the sun. I am heading full speed towards the final stage of every training: the Aula. This period of time is intense, and everything has to be ready before spring 2016. Courses have to be finalized, all the Storm Papers are mastered by now, and the final challenge makes its entrance in no time: Hurricane General Discussion. Winds much stronger than expected and waves just look like mountains of waters in front of ABGC 2.0 ship. Everything is so dark, and suddenly caught off guards, I fell in the sea. "Woman at Sea", shouts the Captain. I am safe and sound. I am quite lucky because new EGS-ABG and PhDs have started their training.

So nice to meet them with their high spirits and hearts full of determination. The nice and quiet main deck is suddenly taken by their voices, bringing a new sense of hope. They do not realize, but they came to the rescue right on … "DRING, DRING", I am immediately transposed at the computer behind my desk at Radix building. "DRING, DRING", insists the phone. "Bonjour Maman, Bonjour Papa!"…

Para minha Família: *Merci Maman et Papa! Merci pour tous ce que vous avez fait pour moi et de m'avoir enseignée ce que l'amour inconditionnel est. Je vous aime! Merci Yvan et Stéphane, pour les visites, voyages et vos soucis. Obrigada Maria-Claudia e Sophia pelo carinho. Obrigada à tia Carmen, tio Reimar , Alexandra, Simão, Felipe, Mariana, Fernando e à falecida tia Margitte por todo o carinho, interesse e apoio.*

Afinal, um PhD não é fruto de coincidências. É fruto de muito trabalho, e dedicação. Por isso, estendo os meus agradeçimentos a todos os meus professores da FAA-Valença/RJ, e em especial à minha amiga Aparecida e ao meu amigo Generoso.