Turun yliopisto
University of Turku

SIGNALS OF SELECTION AND THE GENETIC BASIS OF MILK PRODUCTION IN CATTLE

Terhi Iso-Touru

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. A II osa - tom. 326 | Biologica - Geographica - Geologica | Turku 2016

University of Turku

Faculty of Mathematics and Natural Sciences

Department of Biology

From the Natural Resources Institute Finland (Luke)

Supervised by

Professor Johanna Vilkki
Green Technology, Genetic Research
Natural Resources Institute Finland (Luke)
Jokioinen, Finland

Reviewed by

Professor Outi Savolainen
Department of Biology
University of Oulu
Oulu, Finland

Dr James Kijas
Food and Health Sciences
CSIRO, Animal
Queensland, Australia

Opponent

Professor Miguel Pérez-Enciso
Centre for Research in Agricultural Genomics (CRAG)
Campus Universitat Autonoma Barcelona
Bellaterra, Spain

Cover picture: Lotta Koivuniemi

ISBN 978-951-29-6609-7 (PRINT)
ISBN 978-951-29-6610-3 (PDF)
ISSN 0082-6979 (PRINT)
ISSN 2343-3183 (ONLINE)
Painosalama Oy - Turku, Finland 2016

To my family and
in memory of my father

ABSTRACT

Terhi Iso-Touru

Signals of selection and the genetic basis of milk production in cattle

Domesticated species such as cattle have evolved under natural and artificial selection, leading to cattle breeds (both *Bos taurus* and *Bos indicus*) that display a broad phenotypic spectrum. As a result of intensive artificial selection together with artificial insemination, highly productive global cattle breeds have been developed that are replacing local, native cattle breeds. This study focuses on analyses of how selection has modified the genomes of different cattle populations having diverse breeding histories, especially with regard to milk production and adaptation. For that, both gene and genome level studies were conducted.

The molecular architecture of two quantitative trait loci (QTL) was investigated in different species and breeds using two candidate genes for milk yield, *GHR* and *PRLR*. The intracellular parts of the two genes were sequenced from over 10 cattle breeds and from different Artiodactyla species. The study revealed divergent selection pressures on *GHR* and *PRLR* genes among Artiodactyl species. The *GHR* gene was more divergent within genus Bos than between different species among the Bovinae linage. Nonsynonymous mutations have accumulated in the *PRLR* gene in pigs, possibly implying that *PRLR* has been either target of directional or artificial selection in pigs.

SNP markers covering the whole genome at medium density were used to search for effects of artificial selection in different types of cattle breeds and to compare the genetic relatedness of differently selected breeds. This revealed evidences that *GHR* gene has been a target of selection in certain cattle breeds. In addition, several other genomic regions were found to be targets of selection. Most of them were not shared between the breeds but a region on chromosome 16 was found to be under selection in six breeds. Clear genetic separation between the *turano-mongolicus* type breed and other *Bos taurus* breeds was found by both whole genome SNP data and the *GHR* gene sequence. The within breed diversity was relatively similar for all breeds even if the histories of the studied breeds varied substantially. The estimates of effective population sizes calculated from whole genome SNP data varied from extremely low (24) to moderately high (150).

In the last stage of the study, whole genome sequences were used for genome-wide association study (GWAS) to find genomic regions affecting milk, protein and fat yield in Nordic dairy cattle. The association study confirmed the existence of milk QTL on cattle chromosome 20, at the *GHR* gene, whereas no support for the QTL at *PRLR* gene was gained. Several thousand additional candidate SNPs with effect on milk production were located from eight cattle chromosomes. However, establishing the true causative variant remains challenging even when the densest possible marker map is used because of linkage disequilibrium.

Taken together, this thesis provides genetic information from various Northern Eurasian cattle breeds that can be used for example for conservation decisions and gives a map of selection signatures for them. These selected genome regions may contain variation that would provide valuable traits for changing climate conditions. The knowledge of the genetic background of milk production and the effect of artificial selection is essential when breeding organizations are making decisions how to maintain and improve their genetic material.

TIIVISTELMÄ

Terhi Iso-Touru

Valinnan jalanjäljet ja maidontuotannon geneettinen tausta naudalla

Ihmisen tekemä valinta yhdessä luonnonvalinnan kanssa on johtanut nautarotuihin, joiden ulkoasut vaihtelevat suuresti. Intensiivinen jalostusvalinta yhdessä keinosiemennyksen käyttöönoton kanssa on johtanut korkeatuottoisten rotujen maailmanlaajuiseen menestymiseen paikallisten nautarotujen kustannuksella. Tässä väitöskirjatyössä tutkittiin, kuinka valinta on vaikuttanut eri tavalla jalostettujen nautarotujen perimään, erityisesti keskittyen maidontuotantoon ja adaptaatioon.

Ensimmäisessä vaiheessa tutkittiin kahden kvantitatiivisiin ominaisuuksiin vaikuttavan lokuksen (QTL) molekyylirakennetta. Kahden maidontuotantoon vaikuttavan kandidaattigeenin, kasvuhormonireseptorin (*GHR*) ja prolaktiinireseptorin (*PRLR*) solunsisäisen osan koodaava alue sekvensoitiin eri tavalla jalostetuilta nautaroduilta sekä verrattiin saatuja sekvenssejä eri sorkkaeläinlajien vastaaviin sekvensseihin. Valintapaineen todettiin olleen erilainen näissä kahdessa geenissä, *GHR* geeni oli muuntelultaan rikkaampaa Bos suvun sisällä kuin lajien välillä Bovinae linjassa. Työssä selvitettiin, että sioilla *PRLR* geeniin on kerääntynyt useita aminohappomuutokseen johtavia mutaatioita, toisin kuin naudoilla. Sioilla *PRLR* geeni onkin voinut olla joko suoran tai ihmisen suorittaman jalostusvalinnan kohde.

Valinnan jalanjälkiä haettiin koko perimän kattavasta SNP-merkkiaineistosta käyttäen eri tavalla jalostettuja nautarotuja. Samaa aineistoa käytettiin nautarotujen geneettisen rakenteen selvittämiseen. Tässä työssä pystyttiin osoittamaan, että *GHR* geeni on ollut valinnan kohteena tietyissä nautaroduissa. Lisäksi löydettiin useita muita perimän alueita, joihin on kohdistunut valintaa. Useimmat alueista eivät ole samoja eri rotujen kesken, mutta naudan kromosomissa 16 on mielenkiintoinen alue, jonka osoitettiin olleen valinnan kohteena kuudessa eri rodussa. Työssä todettiin *turano-mongolicus* alatyyppiin kuuluvan nautarodun eroavan selvästi geneettisesti muista *Bos taurus* -tyyppisistä naudoista. Rotujen sisäisen geneettisen monimuotoisuuden todettiin olevan suhteellisen samanlaista, vaikka tutkittujen rotujen jalostushistoriat poikkeavatkin merkittävästi toisistaan. Teholliset populaatiokoot vaihtelivat nautarotujen välillä äärimmäisen alhaisesta (24) kohtalaisen korkeaan (150).

Viimeisessä vaiheessa koko perimän sekvenssien perusteella määritettyjä variaatioita käytettiin assosiaatiokartoituksessa, jonka avulla paikannettiin maito-, rasva- ja proteiinimääriin vaikuttavia perimän alueita pohjoismaisessa punaisessa lypsyrodussa. Tutkimus vahvisti *GHR* geenin olevan erittäin vahva kandidaatti havaitulle QTL vaikutukselle kromosomissa 20 kun taas *PRLR* geenistä vastaavaa ei todettu. Lisäksi löydettiin useita tuhansia maitotuotokseen assosioituneita SNP-merkkejä yhteensä kahdeksasta eri kromosomista. Vaikka käytössä oli teoreettisesti kaikki mahdolliset variaatiot naudan perimästä, kausatiivisen variaation tunnistaminen on haastavaa johtuen variaatioiden välisestä kytkentäepätasapainosta.

Yhteenvetona voidaan sanoa, että tutkimus tuo lisätietoa pohjoisen Euraasian nautarotujen geneettisestä taustasta mahdollisten suojelupäätösten tueksi ja antaa pohjan valinnan jalanjälkien tarkempaan tutkimukseen näillä roduilla. Eläinten geneettisten resurssien kartoittaminen on erityisen tärkeää muuttuvissa ilmasto-olosuhteissa. Maidontuotannon geneettisten taustojen selvittäminen on tärkeää jalostusohjelmissa ja tietoa voidaan käyttää jalostussuunnitelmien tukena.

TABLE OF CONTENTS

ABBREVIATIONS AND SYMBOLS

ABCG2	ATP binding cassette subfamily G member 2
AGTRAP	Angiotensin II receptor-associated protein
AI	Artificial insemination
AMOVA	Analysis of molecular variance
ARID3B	AT-rich interaction domain 3B
ARMC3	Armadillo repeat containing 3
BCE	Before common era
bp	base pair
BTA	Bovine chromosome
CLK3	CDC like kinase 3
CLL	Composite log likelihood
CLR	Composite likelihood ratio
COX5A	Cytochrome c oxidase subunit Va
CRISPR/Cas 9	Clustered Regularly Interspaced Short Palindromic Repeats/Carbonic anhydrase IX
CSK	C-src tyrosine kinase
CYP11A1	Cytochrome P450 family 11 subfamily A member 1
DGAT1	Diacylglycerol O-acyltransferase 1
DNA	Deoxyribonucleic acid
EBV	Estimated breeding value
EHH	Extended haplotype homozygosity
FABP4	Fatty acid binding protein 4, adipocyte
FCAMR	Fc receptor, IgA, IgM, high affinity
FIS	Fixation index or inbreeding coefficient
FY	Fat yield
GBS	Genotyping-by-sequencing
GEBV	Genomic breeding value
GHR	Growth hormone receptor
GO	Gene ontology
GS	Genomic selection
GWAS	Genome-wide association study
GZMB	Granzyme B
HKA	Hudson–Kreitman–Aguade
HMM	Hidden Markov model
Hs	Haplotype diversity
IGF2	Insulin-like growth factor II Insulin-like growth factor II Preptin
iHS	Integrated haplotype score

IL10	Interleukin-10 precursor
IL19	Interleukin 19
IL20	Interleukin 20
IL24	Interleukin 24
kb	kilobase
KIF1B	Kinesin family member 1B
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
LD	Linkage disequilibrium
LINE	Long interspersed element
LMM	Linear mixed models
M	Morgan
MAF	Minor allele frequency
MAS	Marker assisted selection
Mb	Mega base
MDS018	Phosphopantothenoylcysteine decarboxylase
MHC/BoLA	Major histocompatibility complex
MLPH	Melanophilin
MRC2	C-type mannose receptor 2 precursor
MY	Milk yield
NAV	Nordic cattle genetic evaluation
Ne	Effective population size
NGS	Next generation sequencing
NMNAT1	Nicotinamide nucleotide adenylyltransferase 1
NPPA	Natriuretic peptide A
PCA	Principal component analysis
PCR	Polymerase chain reaction
PIC	Polymorphic information content
PIGR	Polymeric immunoglobulin receptor
π	Nucleotide diversity
PIK3CD	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta
PML	Promyelocytic leukemia
POLL	Polymerase (DNA directed), lambda
PRLR	Prolactin receptor
PY	Protein yield
QTL	Quantitative trait locus
QTN	Quantitative trait nucleotide
r2	Linkage disequilibrium measurement
RERE	Arginine-glutamic acid dipeptide (RE) repeats
RFLP	Restriction fragment length polymorphism

ROH	Runs of homozygosity
SCAMP5	Secretory carrier membrane protein 5
SEMA7A	Semaphorin 7A, GPI membrane anchor
SLC25A33	Solute carrier family 25 (pyrimidine nucleotide carrier), member 33
SLC45A1	Solute carrier family 45 member 1
SNP	Single nucleotide polymorphism
SPSB1	SplA/ryanodine receptor domain and SOCS box containing 1
TALEN	Transcription activator-like effector nucleases
θ	Watterson's theta estimator
ULK3	Unc-51 like kinase 3
XP-CLR	Cross Population composite likelihood ratio
XP-EHH	Cross Population Extended Haplotype Homozygosity

LIST OF ORIGINAL PUBLICATIONS

This thesis is summary and discussion of the following articles, which are referred to in the text by their Roman numerals.

I Varvio, S.-L.*, **Iso-Touru, T.***., Kantanen, J., Viitala, S., Tapio, I., Mäki-Tanila, A., Zerabruk, M., Vilkki, J. 2008. Molecular anatomy of the cytoplasmic domain of bovine growth hormone receptor, a quantitative trait locus. *Proceedings of the Royal Society B: Biological Sciences* 275, 1642: 1525-1534. **equal contribution*

II **Iso-Touru, T.**, Kantanen, J., Li, MH., Gizejewski, Z., Vilkki, J. 2009. Divergent evolution in the cytoplasmic domains of PRLR and GHR genes in Artiodactyla. *BMC Evolutionary Biology* 9: 172-182.

III **Iso-Touru**, **T.**, Tapio, M., Vilkki, J., Kiseleva, T., Ammosov, I., Ivanova, Z., Popov, R., Ozerov, M., Kantanen, J. 2016. Genetic diversity and genomic signatures of selection among cattle breeds from Siberia, Eastern and Northern Europe. *Animal Genetics*, ePub. 2016 Sep 15. doi: 10.1111/age.12473.

IV **Iso-Touru, T.**, Sahana, G., Guldbrandtsen, B., Lund, MS., Vilkki, J. 2016. Genome-wide association analysis of milk yield traits in Nordic red cattle using imputed whole genome sequence data. *BMC Genetics,* 17:55

The original publications are reproduced with the permission of the copyright owners.

1. INTRODUCTION

1.1 Domestication history of cattle

The cow, a descendant of the wild ox (*Bos taurus primigenius)*, was domesticated at least in two different domestication centres around 10,000 years ago to provide nourishment and draught power (Bradley *et al.* 1996). It has now spread along with the dispersal of farming and animal husbandry to diverse environmental conditions (Felius 1995) and is subjected to artificial selection to improve milk and meat production and other economically important traits.

Cattle belong to the order *Artiodactyla*, suborder *Ruminantia* and is further divided into two subspecies, *Bos taurus* and *Bos indicus*. Morphologically *B. taurus* is referred as an European type of cattle whereas the characteristic feature of *B. indicus* is a fatty hump on the shoulders. The domestication centre for *B. taurus* is around the Fertile crescent in the Near East, for *B. indicus* it is the Indus valley (e.g. in Orozco-terWengel *et al.* 2015). The degree of polymorphism in taurine cattle is similar to humans whilst the diversity within indicine breeds is significantly higher (Bovine HapMap Consortium *et al.* 2009). These findings implicate the Indian continent as the major domestication centre and a source of predomestication diversity (Bovine HapMap Consortium *et al.* 2009). Archaeological and ancient mitochondrial DNA analyses of Neolithic to Iron Age Iranian domestic cattle samples combined with modern data analyses estimate that only around 80 female aurochs were initially domesticated in the Near East (Bollongino *et al.* 2012). Such an estimate is based on coalescent simulations and may therefore be biased. However, Bollongino *et al.* (2012) tested the estimation with various parameters and found that estimates varied within a relatively narrow range (from 66 to 128). If this estimate is accurate, then less than two females per generation were domesticated (Bollongino *et al.* 2012) highlighting that the domestication of cattle has not been a linear process. Unambiguous morphological evidence of domestication (e.g. rapid reduction in overall body size, changes in body conformation or horn size) found in archaeological samples date back upto 2,000 years later than evidence indicating the management of wild herds (Zeder 2008, Conolly *et al.* 2011). Bollongino *et al.* (2012) hypothesized that either the management of wild cattle was too challenging for a mobile human population or that the management of large, aggressive and territorial wild aurochs was too complex to be used more widely before breeding for more docile characteristics.

1.1.1 Breed formation and breeding

Natural and artificial selection, new mutations and backcrossing of domesticated animals with their wild ancestral species, in conjunction with isolation and genetic drift, have created numerous taurine cattle (*B. taurus*) breeds that display broad phenotypic and genetic variation (Gautier *et al.* 2010). Systematic breeding started in the 19th century at the same time when the first breed organizations and herd books were established (Weigel 2015a). Initially the goal of breeding was to harmonize the appearance of animals but soon after production traits were also considered. Modern breeding programs routinely use genetic information to select the best possible candidates for future usage. Intensive artificial selection has resulted in highly productive global cattle breeds that have replaced local breeds, leading to a situation where cattle have the highest number of breeds at risk among mammalian livestock (FAO 2013). Artificial selection has also created a situation where breeds are mainly used either for milk or meat production. The success of breeding is indisputable. For example in Finland (including all dairy breeds) the average milk production per cow per year has increased from 6,786 liters (2000) to 8,201 liters (2014), while fat and protein contents have remained fairly constant (Natural Resources Institute Finland). However, the increase in cattle productivity achieved through breeding and improved management has been accompanied by adverse effects on animal robustness raising ethical concerns over animal welfare (Rauw & Gomez-Raya 2015, Strucken *et al.* 2015).

1.2 Genetic variation

1.2.1 Markers

Genetic markers are, for example, utilized to distinguish individuals and populations from each other, for parentage testing and for mapping genomic regions influencing phenotype. Among the first genetic markers used in animal genetics were restriction fragment length polymorphisms (RFLP, Botstein *et al.* 1980). The advent of the polymerase chain reaction (PCR) in 1983 (Griffiths *et al.* 1999) enabled amplification of DNA fragments allowing sequence variations in different types of DNA to be used as genetic markers.

Laborious methods to detect RFLPs were replaced by microsatellites in the 1990s (Weber & May 1989). A microsatellite is a repetitive DNA region comprised of short nucleotide repeats with the number of repeats varying between alleles (Campbell *et al.* 1999). Microsatellites have been widely used for both population genetic and linkage mapping studies in animal genetics (e.g. Georges 2007, Groeneveld *et al.* 2010) and are still used, for example, in parentage testing and in population genetics. The release of the first draft of the human genome sequence (Venter *et al.* 2001) and the development of next generation sequencing (NGS) methodologies in the beginning of the 21st century (first NGS machine was commercially available in 2004) has substantially accelerated progress in genetic research. The first version of the cattle genome sequence was released in 2009 (Bovine Genome Sequencing and Analysis Consortium *et al.* 2009) and after that single nucleotide polymorphisms (SNPs) have begun to replace microsatellites in cattle research. The first commercial version of a SNP array covering the whole genome (more than 50,000 SNPs) came on the market in 2008 (Matukumalli *et al.* 2009) and a higher density version of the array (over 700,000 SNPs) followed in 2010. Now approximately 2 million dairy cattle have been genotyped with the genome-wide SNP array (Meuwissen *et al.* 2016). Whole-genome SNP arrays allow the study of genetic population histories and detection of chromosomal regions under selection more accurately than was previously possible (Lv *et al.* 2014).

The majority of the SNPs consist of two alleles. The disadvantage of SNPs compared with microsatellites is the limited number of alleles; polymorphic information content (PIC) for microsatellites is high compared with SNPs (McClure *et al.* 2012). Microsatellites are typically mostly neutral i.e. not causing a difference in phenotype but SNPs are potentially causative.

1.2.2 Structural variants

Deletions, insertions, segmental duplications, copy number variants, inversions and translocations are structural variants. Structural variants can influence phenotype (Bickhart & Liu 2014). For example, the distinctive coat color in Belgian Blue cattle (Li *et al.* 2016) and a sperm defect in Swedish Red cattle (Pausch *et al.* 2016) are phenotypes caused by deletion leading to a premature translation termination in the genes *MLPH* and *ARMC3*, respectively. In cattle, color sidedness is a result of two serial translocation events of the *KIT* gene (Durkin *et al.* 2012). A large deletion on cattle chromosome 12 is known to lower fertility but also associated with higher milk production (Kadri *et al.* 2014), and therefore a potential target for example for balancing selection.

1.3 Sequencing

Frederick Sanger invented the method for determining the nucleotide sequence of DNA in 1977 (Sanger *et al.* 1977). The detection of the nucleotide sequence is based on amplification with specific primers and the use of chain-terminating dideoxynucleotides. Currently, Sanger sequencing is performed with fluorescence labeled dideoxynucleotides and the sequence is determined by capillary electrophoresis using automated sequencers. However, Sanger sequencing is expensive per nucleotide sequenced and has a lower throughput compared with other methods, but has the advantage of a low error rate (Hoff 2009).

Next generation sequencing (NGS) methods provide high-throughput sequencing with a low cost per nucleotide within a relatively short time (Grada & Weinbrecht 2013). Several different methodologies are available, with the first based on pyrosequencing (Heather & Chain 2016), but many other techniques have been developed subsequently. Most of the widely used NGS methodologies (Heather & Chain 2016), are however, only capable of relatively short read lengths (150 – 500bp), require significant upfront investment in sequencing machines, and for some methods unable to sequence homopolymeric regions reliably. Nevertheless, opportunities to use NGS in genetic research are almost endless. One application of new NGS technologies is genotyping-by-sequencing (GBS) (Elshire *et al.* 2011) that represents a cost-effective genotyping method. It has been speculated that if the cost of sequencing continues to fall, GBS will be the most effective way to genotype individuals in the future (Gorjanc *et al.* 2015) . Projects such as the 1000 Bull Genomes Project (www.1000bullgenomes.com) use the strategy of sequencing key ancestors belonging to different breeds and then impute genotypes from sequenced animals for all other animals genotyped with SNP chips (Daetwyler *et al.* 2014). Such an approach generates large amounts of data that can be used for genome-wide association studies (GWAS), genomic prediction and in conservation genetics. Despite the high throughput of NGS methods, Sanger sequencing is still widely used and difficult to replace completely with NGS methods. For example, Sanger sequencing is used for small-scale projects, validation of NGS results and for producing longer reads (up to ~1000bp).

1.4 Population genetics

1.4.1 Genetic factors altering populations

1.4.1.1 Mutations

Mutations are useful for selection and on the whole for evolutionary advantage. However, genetic mutation in nature is a rare event, such that very probable a mutation will be lost by chance even though it would be advantageous. Mutations can be classified by their nature as either a) structural (see section1.2.2) or b) to adjust the nucleotide sequence by replacing a nucleotide/s with another (e.g. SNPs, see section1.2.1) (Brown 1999). Most SNPs have no functional consequences, but if they influence protein structure or gene regulation, then an individual phenotype may change. A mutation is recessive when it only alters the phenotype when two copies of the mutated allele are present or dominant if an effect on phenotype is observed when only one copy of the mutated allele is present. Most mutations will disappear from a population rather quickly. However, if a mutation is beneficial then the frequency in a population can be increased by selection (natural or artificial) or by genetic drift. In breeding, mutations (either identified at the allelic level or by phenotype) having a positive effect on production or health traits will be transferred forward by choosing the animals carrying the favorable alleles as parents for the next generation.

1.4.1.2 Recombination

In addition to mutations, recombination events also alter the DNA sequence and lead to new allelic combinations. Recombination describes the process where DNA segments are exchanged between homologous chromosomes during meiosis (Brown 1999). A new mutation is linked to the adjacent loci/sequence until recombination breaks the connection. Recombination may enable genetic progress by creating new beneficial allele combinations but also hamper progress if a favorable allele combination (a haplotype) is broken down (Futuyma 2006). Recombination rate varies between different genome segments (Simianer *et al.* 1997), even though it is common to use an approximation of one million bases corresponding to 1 centiMorgan (i.e. 1Mb=0.01M). This approximation seems to be adequate for the estimation of effective population size (Flury *et al.* 2010).

1.4.1.3 Genetic drift

Random fluctuations in the frequencies of alleles or genotypes are referred to as genetic drift that may result in the fixation of two or more allele/genotype (Futuyma 2006). The strength of the genetic drift depends on the effective population size. Variants with a low frequency can be easily lost in a population with a low effective population size (Boichard *et al.* 2015). Cattle breeds are known to have low effective population sizes (e.g. Bovine HapMap Consortium *et al.* 2009). Up to 50% of the variants called from the whole genome sequence have a minor allele frequency of less than 5% in cattle (Daetwyler *et al.* 2014). Such variants may only be maintained through sustainable breeding practices, for example, by increasing the diversity of bulls used for artificial insemination (Boichard *et al.* 2015) or using genomic information to avoid inbreeding.

1.4.1.4 Gene flow

Transfer of alleles from one population to another can have a significant influence in natural populations whereas uncontrolled migration in domestic farm animals is rare. However, the value of global exports of live animals or bovine semen has more than doubled during the 21st century (FAO 2015) indicating accelerated gene flow between countries. Native breeds are typically protected from outside influences but the genetic content of commercial breeds (for example Finnish Ayrshire) can be influenced by genetic drift.

1.4.1.5 Population Bottlenecks

A severe and temporary reduction in the population size is termed a population bottleneck. In cattle, population bottlenecks occur due to domestication, breed formation, and more recently intensive use of artificial insemination (AI) and fewer numbers of terminal sires. Such events lead to a low effective population size (N_e) and decrease in genetic diversity (Daetwyler *et al.* 2014). N_e is used to describe the number of breeding individuals in an idealized population showing the same pattern of variation as the real population (Wright 1938).

1.4.1.6 Inbreeding

Inbreeding (i.e. mating between relatives) has three undesirable effects. It leads to inbreeding depression (loss of fitness, including an increase in the incidence of abnormalities caused by recessive deleterious alleles), a loss of genetic variance and random drift in the population mean (Brotherstone & Goddard 2005) that decreases the responsiveness to selection in breeding programs (Weigel 2001). Loss of genetic variance can lead to an excess of homozygous segments. Genetically such segments arise within individuals either because the parents have transmitted the same segment (originating from a common ancestor) to the offspring (identical-by-descent) or by chance both parents share an identical segment. Homozygous segments can be screened using statistical methods generally referred to runs of homozygosity (ROH). Recently formed ROH tend to be longer due to a lack of recombination or alternatively ROH can be long due to low local recombination rates (Kirin *et al.* 2010). A study with four cattle breeds concluded that almost one fifth of the cattle genome was located in ROH regions (Zhang *et al.* 2015). In cattle, both short and medium length ROH regions contain significantly more predicted deleterious mutations than long ROH regions (Zhang *et al.* 2015). Zhang *et al.* (2015) suggested that this is the result of long-term artificial selection that has enriched beneficial alleles in short and medium ROH regions, as well as hitch-hiking deleterious variants. Thus, inbreeding enables rare recessive diseases to be expressed at a population level and also magnifies the occurrence of mildly deleterious variants (Szpiech *et al.* 2013).

1.5 Selection

In cattle, as well as in other domesticated species, natural and artificial selection together with adaptations to various biogeographic regions and production conditions, have affected allele frequencies at loci associated with adaptation or variation in the selected traits. Methodologically differentiating the effects due to natural or artificial selection is challenging (Randhawa *et al.* 2016). Selection may lead to linkage disequilibrium (the nonrandom association of alleles at different loci, LD) and lower genetic variability in regions close to a favored allele (Nielsen *et al.* 2005, Slatkin 2008), resulting in detectable patterns that facilitate the localization of selective sweeps in the genome. Signals of ongoing selective sweeps indicate the presence of genetic variants likely to have an effect on phenotypes (Voight *et al.* 2006) but may also arise due to genetic drift or demographic processes, particularly in artificially selected species. Directional selection can either favour (positive selection) or discriminate against (negative or purifying selection) an allele causing the phenotype.

1.5.1 Positive selection

In the 1850's Darwin and Wallace came up with the concept of natural selection (Futuyma 2006). Darwin's theory of natural selection stated that "If variations useful to any organic being ever occur, assuredly individuals thus characterized will have the best chance of being preserved in the struggle for life; and from the strong principle of inheritance, these will tend to produce offspring similarly characterized. This principle of preservation, or the survival of the fittest, I have called natural selection" (Darwin 1859). If the phenotype increases the fitness of an individual, it becomes more frequent in a population over time. This phenomenon is called positive selection. The allele(s) behind the favorable phenotype will gradually become more frequent at the population level. The genomic signals of positive selection at the sequence level are characterized by decreased local variability (see 1.6.1), a deviated spectrum of allele frequencies (see 1.6.2) and specific linkage disequilibrium patterns (see 1.6.3).

1.5.2 Negative selection

Negative selection is referred to as background selection. Deleterious mutations are usually removed from the gene pool before they reach any detectable frequency within a population (Vitti *et al.* 2013). Genome regions, where no variations are tolerated, are under strong negative selection pressure and therefore usually highly conserved across species.

1.5.3 Balancing selection

When multiple alleles are maintained at an intermediate frequency in a population, such a phenomenon is called balancing selection. Balancing selection may happen due to heterozygote advantage (i.e. the heterozygote individual has higher fitness compared with either of the homozygotes) or frequency dependent selection (i.e. an allele has higher fitness when it is rare and many alleles will be maintained in population) (Vitti *et al.* 2013, Fijarczyk & Babik 2015). Balancing selection is the most challenging form of selection to detect and current methods suffer from low power and a high frequency of false positives (Fijarczyk & Babik 2015). Unambiguous evidence of balancing selection is seldom reported (Fijarczyk & Babik 2015). In cattle, balancing selection has been proven for an immune related major histocompatibility complex (*MHC*/bovine leukocyte antigen (*BoLA*) (e.g. Spurgin & Richardson 2010, Takeshima *et al.* 2014). Other examples of genes with indications of balancing selection can be found, including the milk production candidate locus *GHR* (study I, Blott *et al.* 2003) and *MRC2* responsible for the crooked tail syndrome in the Belgian blue (Sartelet *et al.* 2012). A deletion having a high frequency in livestock is also thought to be maintained by balancing selection (Kadri *et al.* 2014). Charlesworth (2015) proposed in the study of *Drosophila* that variability in the fitness of *Drosophila* populations is not maintained solely by a balance between the mutational input of deleterious variants and their elimination by selection, but rather some form of balancing selection. These findings suggests that balancing selection is not as uncommon as often predicted, but rather it cannot be reliably detected using current methods.

1.6 Methods to detect selection

Most methods for detecting selection have been developed to detect positive selection (Vitti *et al.* 2013, Utsunomiya *et al.* 2015) because it causes evident footprints on the genome. Detection of balancing selection is more challenging due to the rather subtle effects on the genome and negative selection is typically observed for conserved regions. In the era of genomic data it is now possible to infer adaptive processes in the absence of phenotypic data. Therefore selection signature methods are often described as "genome to phenotype" approaches that involve the statistical evaluation of population genomic data regardless of phenotype in order to identify likely targets of past selection (Qanbari & Simianer 2014). Signatures of selection can be found either from intergenic regions, coding regions or both depending on the test statistics. Some of the methods for identifying selection signals are briefly discussed in the following sections.

1.6.1 Local genetic diversity reduction

One way to observe a decrease in local genetic diversity for the detection of positive selection is to screen regions having reduced minor allele frequency (MAF). For example, this can be done using minor allele frequencies or by using runs of homozygosity (ROH) statistics. However, limited resolution and ascertainment bias of SNP arrays are the major drawbacks of both of methods, but these can be overcome by the analysis of sequence level variants. The ROH are organized into the genome in hot- and cold spots that can produce signals in selective sweeps (Utsunomiya *et al.* 2015, Metzger *et al.* 2015). However, for certain population, cattle in particular, ROH may arise from inbreeding due to artificial insemination (Zhang *et al.* 2015) such that the dissociation between true selective sweep and demographic effect represents a major challenge.

1.6.2 Changes in the allele frequency spectrum

Local genetic diversity depression can be detected from calculation of nucleotide diversity (π, Nei 1987) based on the average pairwise sequence differences as outlined in publications I and II. A widely used statistic is Tajima's D that compares two theta estimators, θ_T (calculated from number of pairwise differences) and θ_W (calculated from number of segregating sites) (Tajima 1989). Under neutrality, the Tajima's D value is assumed to be zero. Under positive selection there is an excess of rare polymorphisms such that the Tajima's D value becomes negative. However, negative D values can also occur due to population expansion. If there is balancing selection, intermediate frequency genetic variants are maintained and the Tajima's D value is positive. A statistic comparable to the Tajima's D are the Fu & Li's D* and F* (Fu & Li 1993), but there remains some uncertainty as to whether they are as statistically powerful as the Tajima's D statistic (Simonsen *et al.* 1995).

A method called the composite likelihood ratio (CLR) test (Kim & Stephan 2002) uses coalescent simulations to derive a distribution of the test statistic under the null hypothesis of no selection (Qanbari & Simianer 2014). The advantage of the CLR method is that it is possible to detect alleles already fixed. An extension of the CLR method is the composite log likelihood (CLL) test of differences in allelic frequencies between populations which has been used in studies of selection signals (Stella *et al.* 2010). Composite methods such as XP-CLR (Chen *et al.* 2010) typically combine individual scores from all markers within a specific region. The idea behind this approach is to reduce the number of false positives given that a contiguous region of positive markers is more likely to represent true selection than an individual signal from a single marker (Vitti *et al.* 2013).

1.6.3 Long-range haplotypes i.e. linkage disequilibrium patterns

Based on LD, Sabeti *et al.* (2002) proposed the concept of extended haplotype homozygosity (EHH). EHH detects recent positive selection by identifying long haplotypes that carry a so called "core allele" at a high frequency within the population. Haplotype homozygosity decays with increasing distance from the core allele. The frequency of these long haplotypes may rapidly increase due to selection as long as recombination has not been able to break them down, and therefore leads to strong and long-range LD (Voight *et al.* 2006, Utsunomiya *et al.* 2015, Sabeti *et al.* 2002). Based on the concept of extended haplotype homozygosity, Voight *et al.* (2006) proposed the integrated haplotype score (iHS). The iHS is calculated as the log-ratio between the integrated EHH for the haplotypes containing the ancestral (iHH_A) and the derived core allele (iHH_D) within one population: $iHS = \ln (iHH_A/iHH_D)$ (Voight *et al.* 2006, Utsunomiya *et al.* 2015). Sabeti *et al.* (2007) introduced an extended method of EHH and iHS, the Cross Population Extended Haplotype Homozygosity (XP-EHH). The XP-EHH metric compares long haplotypes among populations and detects selected alleles that have reached a high frequency or have been fixed in one but not all studied populations (Sabeti *et al.* 2007). The test controls the local genomic variation in recombination rates by comparing haplotype lengths across populations and normalizes genome-wide differences in haplotype length among populations. In XP-EHH iHH is calculated for the entire population instead of being partitioned between ancestral and derived alleles: $XP\text{-}EHH = \ln (iHH_{pop1}/iHH_{pop2})$ (Utsunomiya *et al.* 2015, Sabeti *et al.* 2007). All methods based on extended haplotype homozygosity are intended to identify recent selection events, but are unsuitable for detecting selection that occurred before speciation.

1.6.4 Methods based on population differentiation

The traditional method to discover genomic signals of selection is the F_{ST} statistic (Weir & Cockerham 1984) calculated from the variance of allele frequencies of genomic markers between populations. A disadvantage of the F_{ST} approach is the lack of known theoretical distribution under neutrality. However, an empirical null distribution can be computed by permutation based on random sampling of individuals or the random sorting of population labels (Utsunomiya *et al.* 2015). Extensions of the F_{ST} statistics, such as FLK (Lewontin & Krakauer 1973) and haploFLK test (Fariello *et al.* 2013) account for the effective population size and hierarchical population structure, and in contrast to F_{ST} statistics, have known distributional properties under neutrality (Utsunomiya *et al.* 2015).

The McDonald–Kreitman test (McDonald & Kreitman 1991) can be used for studying selection at the gene level. This test finds deviations from predictions assuming that if both synonymous and non-synonymous mutations are neutral, then the ratio of synonymous to nonsynonymous polymorphisms within a species will be similar to the ratio of synonymous to non-synonymous divergence between species. The Hudson–Kreitman–Aguadé (HKA, Hudson *et al.* 1987) test compares levels of diversity between loci. Under neutrality, the levels of polymorphism within a species and divergence between species should be proportional to the neutral mutation rate. The advantage of the HKA-test compared with others (e.g. McDonald-Kreitman test) is that it can be used for any genetic region, not only for those coding proteins, although the rate of neutral evolution is easier to infer from protein-coding regions (Vitti *et al.* 2013).

1.7 Quantitative traits and quantitative trait loci (QTL)

Most of the traits that have an economic value in animal breeding are quantitative. Quantitative traits are phenotypes that can be measured on a quantitative scale. They are controlled by large number of genes/other functional elements dispersed in the genome, each of them individually having a rather small effect to the phenotype with the interaction of the environment (Weigel 2015b, Remington 2015, www.nature.com/subjects/quantitative-trait). Quantitative traits can be distributed continuously (e.g. milk yield), as classes (e.g. number of eggs) or be binary (e.g. for females pregnant or open) (Rosa 2015). Different types of measures can be used as phenotypes in association studies in order to locate quantitative trait loci (QTL).

A QTL is a locus with allelic variants that affect a quantitative or complex trait (Remington 2015). In addition to resolving the genetic architecture underlying the trait, one goal behind QTL hunting is to find quantitative trait nucleotides (QTNs), genetic variants explaining phenotypic variations, and to understand how the phenotype is regulated. Only a few QTNs have been proven unequivocally to be causative in functional studies of production animals (Ron & Weller 2007). Among the best known examples in dairy cattle are the polymorphisms K232A in the *DGAT1* gene affecting milk yield and composition (Grisart *et al.* 2004a) and F279Y in the transmembrane domain of the *GHR* gene (Blott *et al.* 2003, Viitala *et al.* 2006) that influences milk yield.

1.7.1 Milk production i.e. lactation

Lactation is a mammalian specific character that provides nutrition and immune protection to the offspring (Strucken *et al.* 2015). Indirect evidence of human consumption ruminant milk dates back to 7th millennium BCE (Evershed *et al.* 2008). Direct evidence based on the presence of β-lactoglobulin in dental calcus specimens confirms milk as a food from the 3rd millennium BCE onwards in Europe and northern Southwest Asia (Warinner *et al.* 2014). Worldwide annual milk consumption per capita continues to increase, with the global demand for animal based foods expected to double by 2050 (FAO 2009), driven by both population growth and increased consumer preferences for meat and milk.

Breeding of cattle for milk production has been a success story when judged solely on the increase in milk production volume. The interest in milk production traits and the availability of large numbers of records lead to milk traits being among the first targets for QTL mapping (Georges *et al.* 1995). Thousands of milk related QTL and associated variants have been mapped to the cattle genome (Figure 1). The framework explaining endocrine regulation and physiology of milk production is well established but the genetic regulation underlying the dynamic processes of lactation remain poorly understood (Strucken *et al.* 2015).

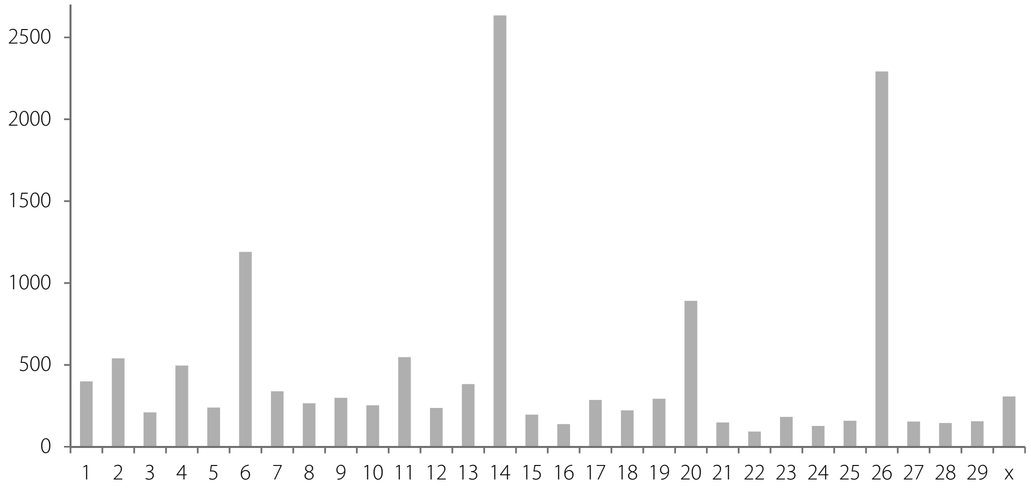

Figure 1. The number of milk related QTL and association variants per cattle chromosome. Data obtained from the Cattle QTL database (accessed on 5/2016) that contains overlapping results.

1.7.2 QTL mapping

QTL mapping is based on the association that exists between genetic markers and quantitative phenotypes. Several different statistical models have been developed to detect QTL with a combination of genotypes and phenotypes including approaches based on maximum likelihood and linear regression (Georges 2007). Breeding population structures generated by AI offer several alternatives for efficient linkage mapping such as paternal half-sister groups (daughter design) or paternal half-brothers with progeny test data (grand-daughter design, Weller *et al.* 1990, Georges 2007).

QTL mapping has changed with the development of new genotyping methods. Sparse sets of microsatellites have been replaced with genome-wide SNP panels, such that by necessity the methodology used has moved from linkage mapping towards genome-wide association studies.

1.7.3 Genome-wide Association Studies (GWAS)

Genome-wide association studies (GWAS) test associations between marker genotypes and a given trait. Implementing GWAS requires a population where the trait of interest is segregating and genotypes covering the whole genome at a sufficient density to be able to detect LD between any potential QTL and markers. A major challenge when using commercial chips is incomplete linkage disequilibrium (LD) between causal mutations and SNP markers (Kemper & Goddard 2012). Given the effort dedicated to genome sequencing projects (e.g. 1000 Bull Genomes Project, Daetwyler *et al.* 2014) and development of imputation methods, this is likely to change. The use of whole sequence variants will potentially allow all causative variants to be included and increase the possibility of discovering true variants responsible for phenotypic differences.

An important aspect of GWAS is accounting for possible population stratification in order to avoid spurious associations caused by hidden relatedness of analysed samples (Kang *et al.* 2010, Eu-Ahsunthornwattana *et al.* 2014, Weir *et al.* 2006). This can be achieved by including a pedigree- or marker-based relationship matrix in the statistical model. Marker-based relationship matrixes can be calculated by distance based methods (e.g. principal component analysis) or model-based methods (e.g. Markov Chain Monte Carlo methods, Pritchard *et al.* 2000) in order to calculate the relatedness of samples used to generate the genotype data. More recently, the usage of linear mixed models (LMMs, aka mixed linear models, MLMs) have become popular for modelling population structure and relatedness (Eu-Ahsunthornwattana *et al.* 2014).

Software package EMMAX (Efficient Mixed-Model Association eXpedited) uses an expedited mixed linear model to correct for sample structure (Kang *et al.* 2010). The method implemented in EMMAX calculates an approximation of the standard test statistics in linear mixed models at the expense of possible inaccurate P-values in the presence of a strong sample structure or a large marker effect (Zhou & Stephens 2012). Such an approach decreases calculation time and computing capacity (Eu-Ahsunthornwattana *et al.* 2014, Kang *et al.* 2010). Other software packages using the same strategy are also available (i.e. GenABEL, TASSEL, MERLIN, Eu-Ahsunthornwattana *et al.* 2014). Several nonlinear methods (e.g. BayesB, BayesR, BayesA, the LASSO) have been developed that allow estimation of all SNP effects simultaneously, while others also allow SNPs to have different effects (Bayes RC) to the trait in question (Meuwissen *et al.* 2016) that is biologically more relevant.

1.8 Positive selection signatures in cattle

The key for understanding phenotypic diversity is to identify the genetic architecture behind the phenotype. This can be done with QTL studies (e.g. GWAS) or using methods dedicated to find signatures of selection. Earlier studies on selection signatures were made with microsatellite markers (Li *et al.* 2010) or limited number of SNPs (I and II). The availability of whole genome-wide genotyping arrays and whole genome sequences has extended research capability in the study of genetics. For example, the 1000 Genomes project in humans (www.1000genomes.org/) has produced a public database (http://hsb.upf.edu/) listing signatures of selection based on different methodologies.

In cattle, evolutionarily important genomic regions are those that are associated with domestication and adaptation. For example, variations in coat colour in cattle are a domestication-related feature. Qanbari *et al.* (2014) found that regions associated with coat colour significantly overlapped regions found to be selected using the iHS and CLR methodologies. Ramey *et al.* (2013) used minor allele frequencies to define genomic regions exposed to selective sweeps (at least five SNPs spanning at least 200kb having no SNPs with MAF>0.01) and identified the *POLL* locus known to control horn development (Georges *et al.* 1993). Kemper *et al.* (2014) argued that genomic signatures from the selection of simple traits such as coat colour or horn development have left detectable patterns in the genome, but for more complex traits, selection pressure at individual loci may be too weak to be detected.

However, several studies have provided some evidence of selection signatures close to known QTL, such that by combining results from genotype based selection signature studies and phenotype-based GWAS studies it would be possible to validate both approaches. Barendse *et al.* (2009) used F_{ST} statistics and concluded that combining analyses of genome wide selection signatures and GWAS helps to define the trait under selection or the population group in which the QTL is likely to be segregating. Flori *et al.* (2009) also used F_{ST} to detect signatures of selection for genes (e.g. *GHR*) known to affect milk production. Zhao *et al.* (2015) combined F_{ST} and iHS methods for the analysis of data from seven different cattle breeds. They reported signatures of selection from several known candidate genes affecting production and reproduction but also identified novel regions. Other studies (e.g. Bahbahani *et al.* 2015, Sorbolini *et al.* 2015) have combined F_{ST} with the other selection methods and reported signatures of selection near known

QTL regions. Other common methodologies used in cattle research include EHH (Pan *et al.* 2013, Qanbari *et al.* 2009, Bomba *et al.* 2015) and XP-EHH (Rothammer *et al.* 2013, Noyes *et al.* 2011, III). Gutierrez-Gil *et al.* (2015) compiled data from 21 selection studies in the European *B. taurus*. Randhawa *et al.* (2016) reviewed 64 studies on selective sweeps in cattle and constructed a meta-assembly of 16,158 selection signatures from 56 genome-wide scans. Randhawa *et al.* (2016) provided a consensus profile of 263 genomic regions under selection, with some found across multiple populations that included known major genes or QTL.

1.9 Progeny testing and marker assisted selection

Historically progeny testing of bulls was used routinely to obtain breeding values for elite sires to be used in AI. Progeny testing has been particularly important in dairy cattle because the main economic phenotype, lactation, is sex-limited and only measurable from females. This has required a high number of offspring and a constant measurement of phenotype. Testing takes 6 to 7 years and the costs for one progeny tested sire can be rather high at approximately 30,000$ to 35,000$ (Georges 2014, Funk 2006). Approved progeny tested sires have been used extensively across the world and have had a major influence on the genetics of the global cattle population (Weigel 2015a).

Genetic information was first included in breeding programs via marker assisted selection (MAS) in 1995 for the German Holstein (Szyda *et al.* 2005) that was adopted in other countries, for example year 2000 in France (Boichard *et al.* 2002). In MAS, breeders used markers linked to QTL in addition to traditional phenotypic evaluation with a focus on a few individual genes with large effects (for example *DGAT1* variant *K232A* (Grisart *et al.* 2002, Weigel 2015b). The inherent complication of MAS is that the majority of quantitative traits are not expressed by a single gene or several genes with large effects, but rather multiple genes each having a small effect on the trait.

1.9.1 Genomic Selection (GS)

The concept of genomic selection (GS) was first introduced by Meuwissen *et al.* (2001). In GS, the effect on the quantitative trait of small chromosome segments is estimated by the haplotypes of marker alleles that they carry (Meuwissen *et al.* 2001, Meuwissen *et al.* 2016). The effects of chromosomal segments are estimated in a progeny tested, genotyped reference population, preferably comprised of thousands of animals. The estimated effects are further combined to a genomic breeding value (GEBV) that can be used as the basis for selection of young (genotyped) animals before phenotypic measurements or progeny testing. Potentially GS captures all QTL contributing to the phenotype because genomic breeding values are calculated as the sum of the effects of genome-wide markers (Hayes *et al.* 2009a). The ongoing 1000 Bull Genomes project will have a major impact on the future development of GS (Georges 2014). It has been proposed that eventually millions of animals will have sequence level data to be used in GS (Hickey 2013). When GS is combined with reproduction technologies such as embryo testing (Machaty *et al.* 2012), genetic gain can be accelerated due a shortening of the generation interval, but at the same time, use of GS may increase the rate of inbreeding (Meuwissen *et al.* 2016). Nevertheless, GS has been described as the most remarkable advance in cattle breeding since the advent of artificial insemination (Georges 2014, Weigel 2015a). One challenge to the implementation of GS is the persistence of the genetic gain. If the LD is incomplete, fixing the marker will not fix the QTL. In such cases, fixation of the SNP does not allow all QTL variance to be captured by GS (Hayes *et al.* 2009a).

2. AIMS OF THE STUDY

The aim of the study was to analyse how (artificial) selection has modified the genomes of different cattle populations having diverse breeding histories by

1. Characterizing population structures and genetic diversity within commercial and native cattle breeds (I, II, III)

2. Studying the genetic architecture of two milk QTL (I, II)

3. Finding signals of selection caused by domestication and breeding (I, II, III, IV)

4. Locating genomic regions having an effect on milk yield in commercial dairy cattle and comparing this information with information from selection signatures (III, IV)

3. MATERIALS AND METHODS

3.1 Samples and DNA extraction

Cattle samples from a total of 25 breeds from different biogeographic regions in Eurasia with various breeding histories and production environments were used (Table 1, Figure 2). More details of the breeds used are described in papers I, II, III and IV. Additional species belonging to the order Artiodactyla (sheep (I, II); pig domestic (I, II) and pig wild (II); yak (I, II); American bison (I, II), European bison (II) and reindeer (I)) were used as reference species.

Genomic DNA was extracted from semen or blood samples in the main using phenol-chloroform extraction according to Miller *et al.* (1988). DNA from hair follicles (wild boars) was extracted from the lysing of hair roots.

Table 1. Cattle breeds used in studies I, II, III and IV.

BREED	PURPOSE	ORIGN	GENES SEQUENCED	STUDY
Belorussian Red, *B. taurus*	Dairy	Byelorussia	*GHR, PRLR*	II
Jersey, *B. taurus*	Dairy	Denmark	*GHR*	I
Jutland cattle, *B. taurus*	Dairy	Denmark	*GHR*	I
Danish Red, *B. taurus*	Dairy	Denmark	*GHR*	I, IV
Barka*, zebu*	Dairy-beef	Ethiopia	*GHR, PRLR*	I, II
Raya*, sanga*	Dairy-beef	Ethiopia	*GHR, PRLR*	I, II
Fogera*, zebu-sanga*	Dairy-beef	Ethiopia	*GHR, PRLR*	I, II
Finnish Ayrshire*, B. taurus*	Dairy	Finland	*GHR, PRLR*	I, II, III, IV
Western Finncattle, *B. taurus*	Dairy	Finland	*GHR, PRLR*	I, II, III
Northern Finncattle, *B. taurus*	Dairy	Finland	*GHR, PRLR*	I, II, III
Eastern Finncattle, *B. taurus*	Dairy	Finland	*GHR, PRLR*	I, II, III
Finnish Holstein-Friesian*, B. taurus*	Dairy	Finland	*GHR, PRLR*	I, II
Aberdeen Angus, *B. taurus*	Beef	Great Britain	*GHR*	I
Charolais*, B. taurus*	Beef	Great Britain	*GHR*	I
Hereford*, B. taurus*	Beef	Great Britain	*GHR*	I
Yarovslavskaya*, B. taurus*	Dairy	Russia		III
Kalmykian cattle, *B. taurus*	Beef	Russia		III
Kholmogor*, B. taurus*	Dairy	Russia	*GHR, PRLR*	II
Bestuzhev*, B. taurus*	Dairy	Russia	*GHR, PRLR*	II
Yakutian cattle, *B. taurus*	Dairy-beef	Russia	*GHR, PRLR*	II, III
Busha, *B. taurus*	Dairy-beef-draft	Serbia	*GHR, PRLR*	II, III
Podolian cattle, *B. taurus*	Draft	Serbia	*GHR, PRLR*	II, III
Swedish Red, *B. taurus*	Dairy	Sweden		IV
Ukrainian grey, *B. taurus*	Beef- draft	Ukraine	*GHR, PRLR*	II, III

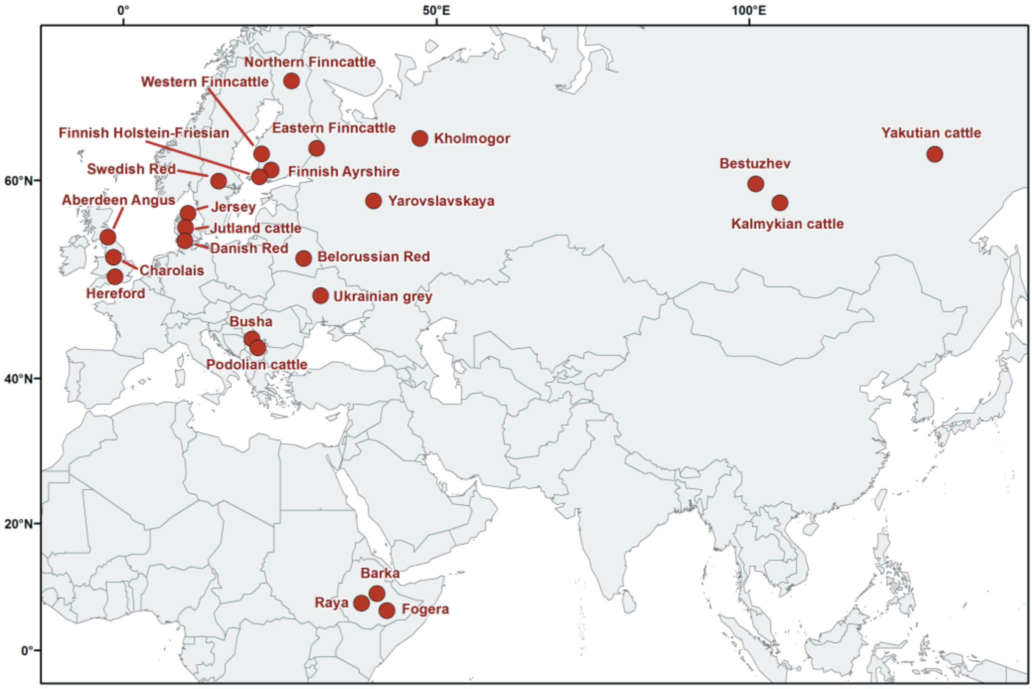

Figure 2. Geographic locations of the studied cattle breeds (I-IV). (Figure by Timo Pitkänen).

3.2 Genotyping

3.2.1 SNP genotyping by Sanger sequencing

Markers for the *GHR* (I) and *PRLR* (II) genes were genotyped by Sanger sequencing (Sanger *et al.* 1977). For both genes, the sequence of the last exon (about 1000bp) coding for the intracellular signalling part of the receptor was amplified with PCR in two fragments and directly sequenced in forward and reverse directions. In addition, the candidate causative SNP (F279Y, Blott *et al.* 2003) in the *GHR* gene was genotyped by sequencing approximately 500bp around the variant (I).

3.2.2 SNP genotyping by chips

In papers III and IV, SNPs were genotyped using either Illumina BovineSNP50 BeadChip version 1 or 2 (III, IV) or Illumina BovineHD chip (IV). Quality controls applied for different data sets varied depending on the starting material and data usage.

3.2.3 Whole genome sequences

Whole genome sequences and variant callings were produced within the 1000 Bull Genomes project and in Aarhus University (Höglund *et al.* 2014).

3.3 Genetic Analyses

3.3.1 Data quality

Visual inspection of sequence quality was used to investigate the quality of data in papers I and II. For whole genome SNP panels quality controls were applied both at an individual and marker level. Specific details are given in papers III and IV but in principle, call rates for individuals and loci (III, IV), minor allele frequencies (III, IV), deviations from the Hardy-Weinberg equilibrium (IV) and imputation accuracies (IV) were calculated and used for data filtering. Deviation from Hardy-Weinberg equilibrium was not taken into account for the dataset in paper III because the data was derived from multiple populations.

3.3.2 Haplotyping

3.3.2.1 Statistical haplotyping

GHR and *PRLR* gene haplotypes (I, II) were statistically inferred from SNP genotype data using the Bayesian haplotype reconstruction method implemented in the program PHASE (Stephens *et al.* 2001). Whole genome level SNP data (III, IV) was phased to estimate the regions under selection (III) and to impute SNP data from the BovineHD chip to whole genome variants (IV). Both datasets used in studies III and IV were phased with the Beagle software (Browning & Browning 2007) that uses a hidden Markov model (HMM) to infer the most-likely haplotype pairs.

For the purposes of this summary, the frequency spectrum of the *GHR* exon 10 haplotypes was investigated with a larger dataset than used in the published papers I and II. Variants were extracted from the 1000 Bull Genomes project database and phased with data from studies I and II using the PHASE v2.1.1 software (Stephens *et al.* 2001).

3.3.2.2 Haplotyping by cloning

For *GHR* and *PRLR* genes, amplified sequence fragments were cloned for samples with a low statistical haplotype prediction (I, II). Cloning was performed using the fragments amplified by PCR. These fragments were ligated to a vector and transformed into *Escherichia coli* cells. Cells that included the insert were selected after cultivation and inserted fragments were directly sequenced with universal primers.

3.3.3 Imputation

Imputation to whole genome variants (IV) was conducted at Aarhus University, Denmark using a two-step approach. First genotypes from the BovineSNP50 BeadChip were imputed to high-density genotypes from the Illumina BovineHD chip. These imputed

HD genotypes were further imputed to the whole genome sequence level using a multi-breed reference panel consisting of 1,228 animals. The number of genotypes per animal after two imputation steps was over 22,000,000 covering the 29 autosomal chromosomes of cattle.

3.3.4 Phenotypes

The phenotypes from three milk production traits (milk yield, fat yield and protein yield) were used as deregressed breeding values (IV) since the use of estimated breeding values (EBVs) may lead to higher numbers of false positives (type I error) (Ekine *et al.* 2014). Phenotypes were obtained by routine genetic evaluation (Nordic cattle genetic evaluation, NAV, www.nordicebv.info/production).

3.3.5 Analysis of population structures

3.3.5.1 Genetic diversity

Genetic diversity was estimated with several statistical methods. Nucleotide diversity (π) (Nei 1987) based on the average pairwise sequence differences was calculated from the haplotypes obtained from *GHR* and *PRLR* exon 10 data (II). Similarly, Watterson's theta estimator (θ) (Watterson 1975) was calculated from haplotypes for both genes (II). A sliding window plot for the estimates of nucleotide diversity (Nei 1987) of the *GHR* and *PRLR* exon 10 haplotypes obtained from the studies (I, II) and background sequences from the GenBank was created to reveal areas of low genetic diversity within vertebrate species. Haplotype diversity (H_S) was estimated for *PRLR* (II), *GHR* (II) and for 4-SNP haplotypes calculated from SNPs in the BovineSNP50 BeadChip (III).

3.3.5.2 Genetic distance of the populations

Phylogenetic reconstruction of *GHR* haplotypes was performed with the median joining network allowing reticulations using the NETWORK v. 4.2.0.1 program from Fluxus Technology Ltd. (I).To generate a cladogram from the *PRLR* gene DNA sequences a statistical parsimony method that finds the tree that requires the fewest evolutionary changes was used. This methodology is implemented in the TCS1.21 program (II)

To infer the most probable number of genetic clusters (K) using the data from paper III, a model-based Bayesian clustering method implemented in the Structure program was used (Pritchard *et al.* 2000). Evaluation of the K-values was undertaken by plotting the LnPD, Evanno DeltaK and an AIC type measure. Principal component analysis (PCA) using smartpca in EIGENSOFT 5.0.1 (Patterson *et al.* 2006) was also performed on the same set of SNPs used in the Structure calculations.

3.3.5.3 Linkage disequilibrium (LD) and effective population size (N_e)

Linkage disequilibrium (LD) between the causative allele F278Y in *GHR* exon 8 and exon 10 haplotypes as well as individual SNPs within the exon 10 was calculated to identify possible interactions of the intracellular domain with variants in the transmembrane domain (II).

Genome-wide LD was estimated by calculating r^2 ($r^2 = D^2/[p_1*p_2*q_1*q_2]$, where p_1 and q_1 are the frequencies of allele I for the respective markers, (de Koning 2015)) to visualize the relationship of r^2 with genetic distance among breeds in paper III.

Effective population size (III) was estimated based on the relationship between linkage disequilibrium (*r^2*), effective population size (N_e), and recombination rate (1Mb = 0.01 Morgan).

3.3.6 Selection

In order to explore whether *GHR* or *PRLR* genes have been targets of selection, the Tajima's D (I, II), Fu & Li's D* (I, II), Fu & Li's F* (II), McDonald-Kreitman (I) and Hudson-Kreitman-Aguadé (I) –tests were calculated from the haplotype data.

Patterns of selection signatures from the genome-wide SNP data were searched with the XP-EHH among eight cattle breeds originating from Northern and Eastern Europe and Siberia. Two cattle breeds (Podolian cattle and Busha) were excluded from the XP-EHH analysis due to the low number of samples left after excluding closely related duos or a strong within-breed structure, respectively (III). Two different reference populations were chosen for two independent XP-EHH runs, either the Finnish Ayrshire (the most intensively selected dairy breed in our studies) or Yakutian cattle (the most divergent local breed).

3.4 Genome-wide association analysis

EMMAX (Kang *et al.* 2010) was chosen for the GWAS analysis of the milk production traits. Significance of the associations was tested with the Bonferroni correction (IV).

3.5 Consequences of the variants

Missense variants causing amino acid substitutions were analysed with SIFT (Ng & Henikoff 2003) when they were i) predicted to be statistically significant for milk, protein or fat yield (IV), ii) located within genomic regions possibly under selection (III), or iii) were in the region of a milk QTL (II). SIFT uses multiple sequence alignments to predict the impact of amino acid variants on protein structure. In paper II, an alternative method, PolyPhen (Ramensky *et al.* 2002), was used for same purpose. All statistically significant variants from paper IV were annotated with the variant effect predictor tool (McLaren *et al.* 2010). The Biomart tool (Kinsella *et al.* 2011) embedded in www.ensembl.org was used to find genes within genomic regions indicating selection signatures (III).

To further predict the possible effect of variants, gene ontology (GO) term enrichment analyses were conducted (III, IV). GO terms associated with genes found in genomic regions showing selection were tested for enrichment, with specific emphasis on the enrichment of production or adaptation related traits (III). Genes within the QTL peak areas with statistically significant associations for each trait (IV) were analysed with Qiagen's Ingenuity® pathway analysis (IPA®, Qiagen Redwood City, www.ingenuity.com) to generate gene networks on the basis of their connectivity.

The SNPs significantly associated with milk production traits (IV) were compared to the results obtained from a study of dairy cow fertility (Höglund *et al.* 2015) to explore possible links between milk production and reproductive efficiency.

4. RESULTS AND DISCUSSION

4.1 Genetic diversity and population structures

Knowledge of genetic diversity, genetic distinctiveness and genetic population structure provides critical information for the conservation and management of animal genetic resources. Understanding the genetic basis of phenotypic diversity is one of the fundamental goals for conservation genetics and also for animal breeding. If phenotypes and genotypes show no variation between individuals, selection for important breeding traits will not be successful. Therefore, it is important to maintain genetic variation within the global cattle population to provide genetic resources to meet future societal challenges including food security, increased competition for land and greater variation in climatic conditions.

4.1.1 Genetic diversity between breeds is fairly constant

The genetic diversity of the studied cattle breeds was measured by average minor allele frequency, fixation index and gene diversity (at SNP and haplotype level). The within breed diversity was relatively similar for all breeds even if the histories of studied breeds varied substantially. The results are in agreement with a previous study: the average heterozygosity of 0.297 in study III was similar to 0.267 reported by Gautier *et al.* (2010), where 47 cattle breeds were analysed using the same SNP chip. Gautier *et al.* (2010) noted that European cattle breeds exhibited higher heterozygosity than breeds originating from Africa. The ascertainment bias in the construction of the SNP chip is a likely cause for this phenomenon. As is shown in study III, the SNP diversity is overestimated for the Finnish Ayrshire and underestimated for Yakutian cattle. This is most likely explained by the closer genetic relatedness of the Finnish Ayrshire with breeds used to develop the array. The heterozygosity estimates calculated from a single gene haplotype (II) show elevated levels of heterozygosity for breeds with an African origin compared with European breeds. In general, taurine cattle have been found to have lower genetic diversity compared with indicine cattle (Bovine HapMap Consortium *et al.* 2009) partly due to breed formation, artificial selection and geographic distance from the domestication centre (Bovine HapMap Consortium *et al.* 2009, Loftus *et al.* 1999). Edea *et al.* (2015) used the SNP panel consisting of SNPs derived mainly from *B. indicus* and found high within-breed variation among Ethiopian cattle. Such findings confirm that the availability of unbiased variant data is essential for the estimation of diversity. In this regard, projects such as the 1000 Bull Genomes are invaluable to support future developments in cattle breeding and in the study of population and conservation genetics.

4.1.2 Yakutian cattle differs from other *Bos taurus* cattle breeds

The population structure of cattle breeds was studied with the whole genome SNP panel (III) and also partially at the gene sequence level (I and II). Table 1 indicates which breeds were used in different parts of the study. In general, results revealed a clear separation between Yakutian cattle (*turano-mongolicus* type breed) and other *B. taurus* breeds as indicated by both the study of whole genome data and analysis of the *GHR* gene.

The data used in paper III indicated six distinct breed groups (PCA and Structure analyses, Figure 3). According to Felius (1995) and Li & Kantanen (2010) the dairy breeds used in III could be divided into three different subgroups; namely i) the North-European polled and Celtic breeds (Eastern Finncattle, Western Finncattle and Northern Finncattle), ii) Longhorned dairy breeds of Scandinavia and Scotland (Finnish Ayrshire), and iii) the West and North European Black Pied and Red Pied Lowland Dairy breeds and breeds originating from Central and Eastern Europe (Yarovslavskaya). In the studies of Felius (1995) and Li & Kantanen 2010, other breeds were divided as follows: Podolian and Ukrainian Grey to Podolian breeds of Italy and Eastern Europe, Busha to Illyrian Shorthorn breeds of the Balkans and Greece, and Kalmykian cattle and Yakutian cattle to Turano-Mongolian breeds of Central and Northeast Asia, the yak and yak-cattle hybrids. The SNP data from the BovineSNP50 BeadChip is consistent with these findings, with the exception that Kalmykian cattle did not group with Yakutian cattle. This may be due to sample structure as the effective population size of Kalmykian cattle samples analysed in paper III is rather low. However, other breeds with similarly low effective population sizes did group as expected. Kalmykian cattle have not been studied using other markers (such as microsatellites) but the result from study III indicates that re-evaluation of the Kalmykian breed phylogenetic position would be worthwhile.

The native breeds used in the present study are all present as small populations and are partly endangered. Some of the breeds show close genetic relatedness (for example Northern, Western and Eastern Finncattle, Figure 3) and would possibly benefit of controlled crossbreeding. In addition, as was suggested in paper III and other studies (Kantanen *et al.* 2000a, Hiemstra *et al.* 2010), pedigree recording and *in vitro/in vivo* conservation programs should be implemented together with careful monitoring of the number of parents used for future generations. The value of individual breeds is not only one of economic benefit but also part of a cultural heritage which should be taken into account when programmes such as crossbreeding are considered. There is an urgent need to adjudge whether crossbreeding of native breeds would be more advantageous than planning a breeding programme utilizing genetic information based on the genotyping of all individuals to ensure that populations remain as diverse as possible as has been recommended by others (Meszaros *et al.* 2015).

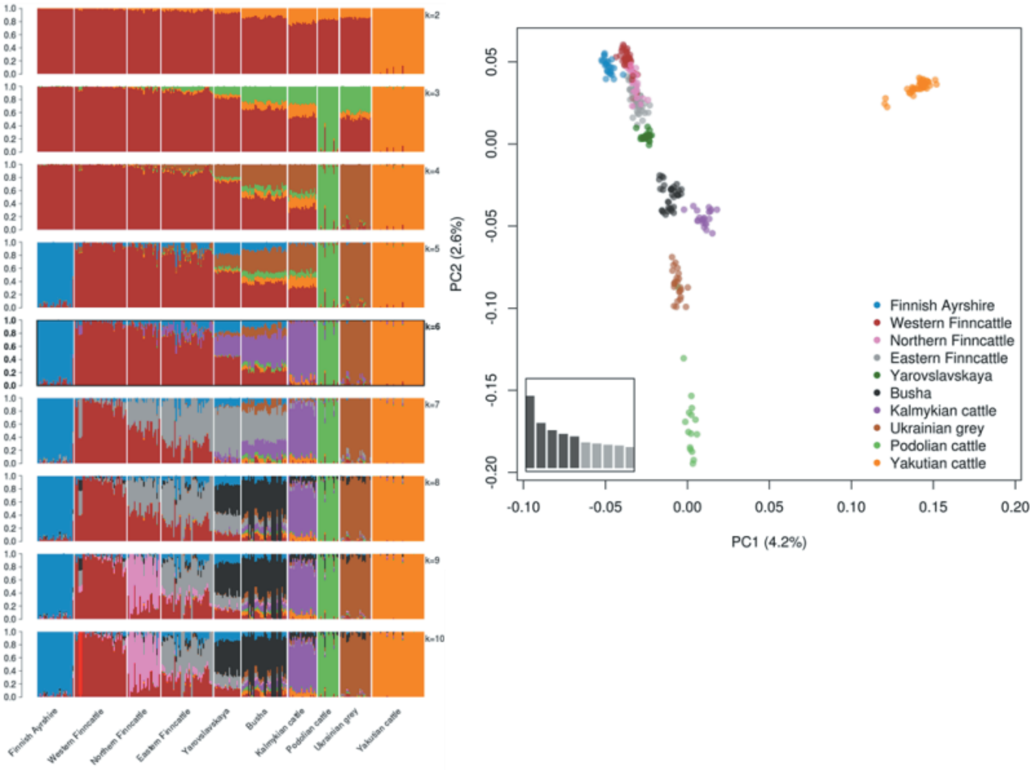

Figure 3. Genetic differences among ten cattle breeds as revealed by clustering and principal component analyses (III).

The fixation index or inbreeding coefficient (F_{IS}) was positive for three breeds (Western Finncattle, Eastern Finncattle and Busha) indicating the presence of inbreeding. This might be explained by their small sample size, although highly related duos were removed from the dataset used to estimate F_{IS}, to generate more reliable results. The increased inbreeding coefficient is one of the major problems facing small native cattle breeds (Mastrangelo *et al.* 2016) but also a concern within commercial breeds (Meuwissen *et al.* 2016, Zhang *et al.* 2015). Some estimates suggest that the annual inbreeding rate has increased in Holstein cattle because of GS (reviewed by Meuwissen *et al.* 2016). It is therefore critically important to monitor the rate of inbreeding in all breeds to maintain a genetically diverse cattle population. The estimation of F_{IS} in paper III was made using 4-SNP haplotypes but it would be worthwhile to estimate F_{IS} from ROHs. Zhang *et al.* (2015) concluded that the BovineSNP50 BeadChip can be used to detect ROHs in order to estimate inbreeding coefficient, but the values generated are influenced by marker density. This is a problem that becomes particularly important in the study of native breeds. For example in paper III, the BovineSNP50 BeadChip was found to be biased leading to a distortion in diversity estimates.

4.1.3 Genome-wide linkage disequilibrium (LD) pattern reflects breed history

Domestication, breed formation and selection have influenced the level of LD in cattle making it extend longer than for humans (Kemper & Goddard 2012). Potential bottlenecks during breed formation should leave detectable LD patterns when estimated at the genome-wide level (Bovine HapMap Consortium *et al.* 2009). The population history, breeding system and geographical subdivisions are reflected in the genome-wide LD, whereas LD in individual genomic regions reflect the history of natural selection, gene conversion, mutation and other forces that cause gene-frequency evolution (Slatkin 2008).

Genome-wide LD (measured as r^2) in cattle reaches a plateau at around 200kb (Figure 4, III) consistent with earlier observations (Gautier *et al.* 2010, Bovine HapMap Consortium *et al.* 2009). Genome-wide LD diminished rapidly when breeds were pooled implicating independent haplotype structures in each population (Figure 4). The low r^2 values in short and long distances indicate heterogenic ancestry of a population/breed as a result of genetic admixture or existence of subpopulations within the population (Li *et al.* 2007). Such an effect was detected in Eastern Finncattle and Busha in study III. Busha may have experienced admixture (Ramljak *et al.* 2011 and J. Kantanen, personal communication) which would explain the observed pattern. Eastern Finncattle have been reformed in the 1980s from several isolated founder herds (Kantanen *et al.* 2000b) that can be seen in the LD pattern since the genome-wide LD is low in both short and long distances.

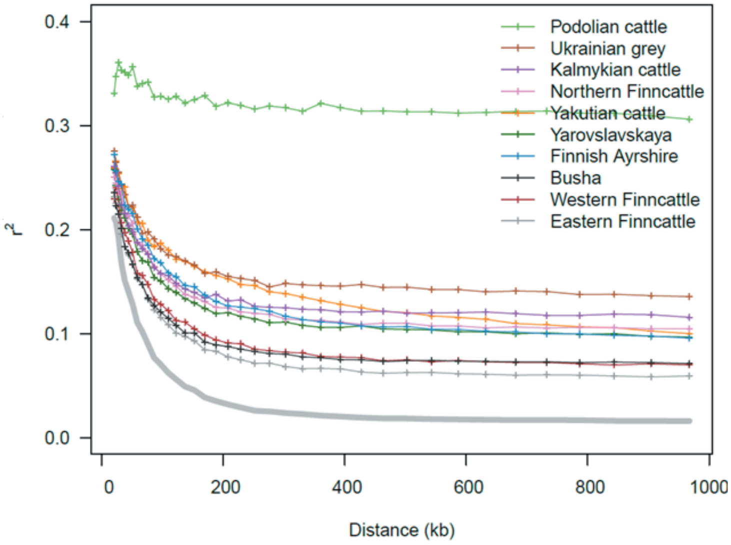

Figure 4. Change in linkage disequilibrium (r^2) between marker pairs with increasing distance for all studied breeds. The solid grey line indicates the change in r^2 when breeds are pooled.

A notable exception is the Podolian cattle (Figure 4) that has higher r^2 values both in short and long distances compared with all other studied breeds (III). This is partly due to the low number of samples, as after excluding closely related duos, only 5 samples remained. However, the breed has extremely few individuals (see below), and is now classified as critically endangered. Genetic diversity of Podolian cattle is estimated to be low when measured with microsatellite markers (Ramljak *et al.* 2011). Collectively the findings from the present study (III) provide further evidence of a genetically critical status of the Podolian cattle breed.

4.1.4 Effective population sizes (N_e) vary from low to moderate

The estimates of effective population sizes calculated from SNP data (III) varied from extremely low (24, Podolian cattle) to moderately high (150 for Yarovslavskaya). Pedigree based estimations of N_e are only available for a limited number of the studied breeds due to a lack of herdbook information. The Western Finncattle has a pedigree based N_e estimation with a harmonic mean of 171 (Toro *et al.* 2011). In comparison, the estimate based on genetic information from 39 samples (III) was 108.

The demographic way to estimate N_e is to use information on the number of breeding females and males. The Podolian cattle population includes 286 breeding females and seven breeding males (in year 2014) according to the DAD-IS database (accessed 4/2016, DAD-IS). By using the classical population genetic theory (Wright 1931), then N_e is approximately:

$$\frac{4N_m N_f}{N_m N_f}$$

where N_m is the number of breeding males and N_f is the number of breeding females. Thus the estimation of N_e based on demographic information would be 27 for Podolian cattle. However, this estimation ignores annual fluctuations in the number of breeding animals and is therefore not a particularly useful measure for cattle populations.

In general, the N_e estimates reported in paper III are likely biased due to the (low) sampling size and sparse marker density. The method chosen to estimate N_e was based on the relationship between linkage disequilibrium (r^2), N_e and recombination rate (Barbato *et al. 2015*), factors all dependent on the quality of marker panel used. Nevertheless, the information generated in study III is the only available estimate for many of the breeds studied. The availability of whole genome sequence data for some of the breeds (Finnish Ayrshire, Western Finncattle, Yakutian cattle) will, however, change this situation in the near future. New methods have recently been developed to simultaneously infer robust N_e estimates from several complete genomes (Boitard *et al.* 2016).

4.2 Molecular anatomy/evolution of two QTL (*GHR* and *PRLR*)

Many of the livestock species used for the production of human foods belong to the order Artiodactyla. The first Artiodactyla species to be domesticated was the sheep followed by goats, pigs and cattle (compiled by Larson *et al.* 2014). Meat, skin and horns were first materials to be used, but gradually humans learnt to use animals without killing them (milking, wool etc.). The selection pressure has varied for different Artiodactyla species. Traditionally cattle have been used as a source of draft power, but in most developed countries specialized cattle breeds are predominantly used for milk or meat production. Sheep are used in the main for meat and wool production, and to a lesser extent for milk production in certain countries, whereas pigs are used exclusively for meat production. Half-tamed and wild Artiodactyla species (such as bison, yaks and reindeers, wild boar) are farmed for meat and fur production and have not been intensively selected.

Studies I and II addressed the research question: has the divergent selection pressure in different cattle breeds and/or Artiodactyla species left detectable signals in form of sequence variation to two evolutionary related and closely located genes known to have roles in growth, reproduction and lactation (e.g. Blott *et al.* 2003, Viitala *et al.* 2006).

The genes studied were the growth hormone receptor gene (*GHR*, 20: 31,890,736-32,199,996) and the prolactin receptor gene (*PRLR*, 20: 39,073,246-39,137,480). Both the *GHR* and *PRLR* gene belong to the family of cytokine receptors and have three domains (extracellular, transmembrane and cytoplasmic). The molecular anatomy of these genes was studied based on the analysis of the cytoplasmic domains, since intracellular signal transmission of *GHR* and *PRLR* genes is directed via this domain and the JAK-Stat signalling pathway (Frank 2001, Bole-Feysot *et al.* 1998, Forsyth & Wallis 2002).

4.2.1 Conserved regions harbour most of the missense variants in the cattle *GHR* gene

Several SNPs were found from exon 10 in the *GHR* and *PRLR* from all species studied, although *Bison* and pigs were monomorphic for *GHR* (Tables 2 and 3). Sheep exhibited a similar number of nonsynonymous variants in both the *GHR* and *PRLR* gene. Nucleotide diversity was calculated for the studied regions of *GHR* and *PRLR* using reference sequences from the different species. In *GHR*, most of the cattle missense mutations are located in regions having low nucleotide divergence among reference species suggesting a functional importance of variant, whereas the same is only true for one of the cattle missense mutations in the *PRLR* gene. Individual polymorphisms within the *GHR* and *PRLR* intracellular parts were investigated for deviation from neutrality. The test statistics for different neutrality indexes were not significant. However, these tests are based on several assumptions, including random mating and a large and constant population size that are obviously violated in livestock populations.

Table 2. Heterozygous variants in the cytoplasmic domain of the *GHR* gene across different species.

GHR														
	Varvio et al.		AA		European	African			Wild		B.		Rein-
rsID	2008	Position	change	SIFT	cattle	cattle	Sheep	Pig	boar	B. bison	bonasus	Yak	deer
rs210410103	Nt1095	31891618			x	x							
rs516176074	Nt1134	31891579				x							
rs446292000	Nt1317	31891397	S439N	Tolerated		x							
rs477542099	Nt1428	31891285				x							
NA	Nt1458	31891315				x							
rs109240320	Nt1482	31891231			x	x							
rs721387762	Nt1557	31891158	P519S	Tolerated		x							
rs380310659	Nt1569	31891146	N523D	Tolerated	x		x						
NA	Nt1575	31891138				x							
rs110265189	Nt1584	31891130	N528T	Deleterious	x	x							
rs209676814	Nt1608	31891107	A536T	Tolerated	x								
rs209323588	Nt1623	31891092	A541S	Tolerated	x								
rs109136815	Nt1635	31891078			x	x							
rs109300983	Nt1665	31891050	S555G	Tolerated	x	x							
NA	Nt1293		P431S				x						
NA	Nt1740		H580N				x						

Table 3. Heterozygous variants in the cytoplasmic domain of *PRLR* gene across different species

PRLR												
	Iso-Touru et		AA		European	African					B.	
rsid	al. 2009	Position	change	SIFT	cattle	cattle	Sheep	Pig	Wild boar	B. bison	bonasus	Yak
rs209364409	Nt1088	39136179			x	x						
rs480522564	Nt1104	39136195	P340T	Tolerated	x							
rs442785003	Nt1218	39136309	E378K	Tolerated	x	x						
rs524800635		39136492	V439M	Tolerated		x						
rs110971500	Nt1427	39136518			x	x						
rs458818443		39136667	L497R	Tolerated		x						
rs527077702	Nt1622	39136713			x	x						
NA	Nt1682	39136773			x							
rs524620576	Nt1693	39136784	A536V	Tolerated	x							
rs382007362	Nt1754	39136845			x	x						
rs522802924	Nt1769	39136860				x						
rs440053154	Nt1775	39136866			x	x						
rs524756765	Nt1817	39136908			x							
NA												
NA			E384K							x	x	
NA			D588E							x		
NA			Q397K								x	
NA			M446V								x	
NA	Nt1730										x	
NA	Nt1007						x					
NA	Nt1160						x					
NA	Nt1217						x					
NA	Nt1400						x					
NA			E387K				x					
NA			A476T				x					
NA			S480R				x					
NA	Nt1620							x	x			
NA			L406P					x				
NA			D428A					x				
NA			A461G					x				
NA			K480R					x				
NA			M510L					x				
NA			G534S					x				
NA			G597S					x	x			
NA			A601V					x				

4.2.2 *GHR* and *PRLR* haplotypes are telling different stories in cattle

Variants of the *GHR* and *PRLR* gene were used further to statistically infer haplotypes in order to investigate haplotype frequencies and to construct phylogenetic trees from the genes for all species separately. The main *GHR* haplotype for the dairy breeds that have a *B. taurus* background is BOS3 (Appendix 1), with the exception of Yakutian cattle. To analyse the haplotype frequency in a bigger sample group, data from the 1000 Bull Genomes project was also used increasing the sample size to 1,800. The BOS3 haplotype remained the most frequent with a total frequency of 0.53 (unpublished, Appendix 1). Eight breeds did not have the BOS3 haplotype. Three of them were of African origin with a *B. indicus* background (reported in I) and the rest were Yakutian cattle, Romagnola (1000 Bull), Salers (1000 Bull), Belted Galloway(1000 Bull) and one crossbreed (Gelbvieh x Limousine, 1000 Bull). Romagnola belongs to the Podolian group of grey cattle and Salers is thought to be the one of the oldest and most genetically pure of all European breeds (www.ansi.okstate.edu/breeds/cattle/salers). However, the number of samples per Romagnola, Salers, Belted Galloway and crossbreeds in the 1000 Bull dataset is rather low (n = 1-2), and therefore these breeds are not well represented. BOS3 was found to differ from bison and yak haplotypes by only one synonymous substitution (I, Figure 5a). It is tempting to speculate that this could be an ancient haplotype from the *B. taurus* lineage. The hypothesis put forward in study I with respect to the BOS1 haplotype originating from *B. indicus* gains more support from the analysis of the 1000 Bull data. Only the samples of African origin (included in studies I and II) were found to have the BOS1 haplotype after the addition of a substantial amount of data from samples of *B. taurus* origin.

The haplotype structure of *PRLR* was also found to differ from *GHR*. Two major haplotypes are shared between European and African cattle, such that major unique haplotypes are absent (Appendix 2). In contrast to the *GHR* gene where differences within European cattle breeds and between European/African breeds were larger than between different species (cattle, yak, American and European Bison) (II), the phylogenetic network constructed from the *PRLR* gene haplotype sequences corresponds well with the known history of Bovinaes (Figure 5b).

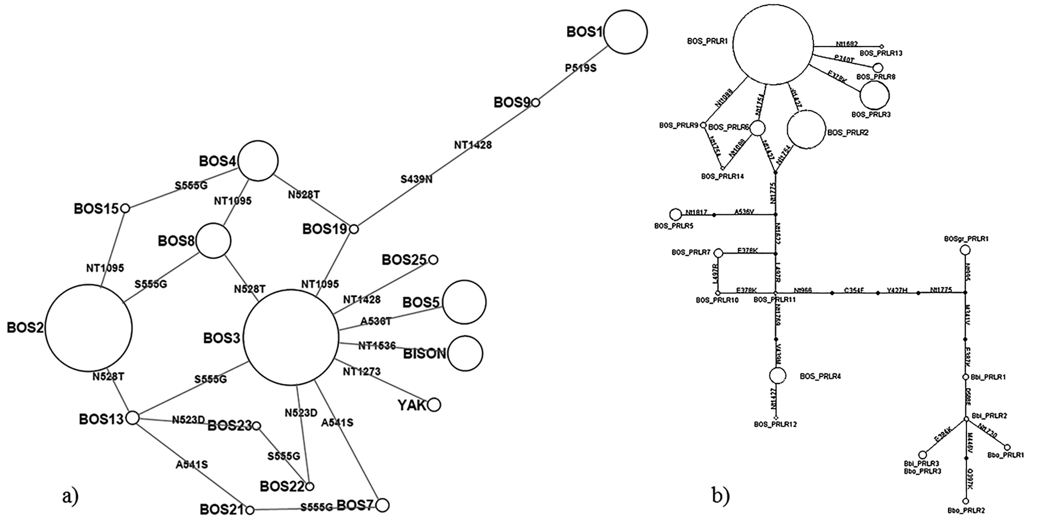

Figure 5. Haplotype networks constructed from a) *GHR* haplotypes and b) *PRLR* haplotypes. The *PRLR* gene network provided a better fit to the known phylogenetic structure of Bovinae lineage, with taurus and indicus breeds being more closely related than other Bovinaes (*Bos grunniens* i.e. yaks and *Bison bison/bonasus* i.e. bisons). BOSgr/YAK refers to the yak, Bbi/BISON to the American bison and Bbo to the European bison haplotypes.

4.2.3 *GHR*; a possible target of selection during domestication and breeding?

The genomic region harbouring *GHR* has been identified to be under (positive) selection in several studies (compiled by Randhawa *et al.* 2016). Typically lowered variability is a signature of positive selection (e.g. Gutierrez-Gil *et al.* 2015). Decreases in variability can be due to intensive artificial selection combined with a low effective population size. However, there are reports indicating that nucleotide variability may not necessarily be affected by domestication. For example Ojeda *et al.* (2008) did not detect any apparent reduction in nucleotide variability after domestication in the porcine *IGF2* gene (increases lean muscle content) region. Similarly, our exon10-based study (I, II) did not show signatures of reduced variability at *GHR* in cattle, but rather a high level of polymorphism. The reasons for the existing polymorphism in the *GHR* gene of cattle could be rather diverse. The hypothesis presented in I speculated that one explanation could be that the *GHR* intracellular domain has evolved under relaxed functional constraints because of artificial selection, and thereby able to capture amino acid altering mutations. This suggests that owing to the persistence of polymorphisms, cattle have been responsive to artificial selection for growth and lactation traits. Balancing selection may well explain the persistent polymorphism in the *GHR* gene. Signals that could indicate balancing selection according to Fijarczyk & Babik (2015) are:

a) Shared polymorphisms between species as a result of long gene genealogies.

b) Increased diversity around the target of selection.

c) Excess of nonsynonymous polymorphisms segregating at intermediate frequencies.

d) Distribution of allele frequencies that is more even than expected under neutrality.

e) Differentiation between populations departing from the genome-wise average.

f) Increased LD around the target of recent selection.

However, the criteria listed above are only partly fulfilled. One of the non-synonymous SNP is shared between Bos and Ovis species. This finding in isolation is not enough to provide a clear indication of balancing selection. Proofs to parts b – e were not unambiguous and there was no increased LD around the target region. The lines of evidence supporting balancing selection were inconclusive, suggesting an alternative explanation. One possible explanation presented in study I was related to the ruminant specific tyrosine residue that is surrounded by the polymorphic amino acid sites in cattle. Tyrosine residues are phosphorylated by JAK2 transphosphorylation and are considered critical to intracellular signalling (Frank 2001). The ruminant specific tyrosine site could be an additional target of phosphorylation facilitating additional protein interactions. Polymorphisms around the target site may affect three-dimensional protein structures and serve as a more amenable substrate for protein interactions.

One objective of this research was to establish whether differentially bred and selected breeds exhibit a different pattern of sequence level variation in such a gene. Iso-Touru (2004) did not find statistically significant difference in AMOVA analysis between dairy and beef breeds when using haplotype sequences from the *GHR* gene. These observations were supported by the findings of study I. However, the largest differences in sliding window average difference in allele frequencies between beef and dairy cattle have been found on BTA20 in the same region of the *GHR* gene (Hayes *et al.* 2009b), although Kemper *et al.* (2014) did not observe large differences in allele frequencies in the *GHR* gene between beef and dairy breeds. Results thus far are contradictory and leave open the fundamental question of whether artificial selection for different breeding goals (milk vs. meat) within one species is sufficiently strong to leave detectable signals to a major QTL locus.

4.2.4 Divergent selection pressure within Artiodactyl species on *GHR* and *PRLR* genes

As was reported in papers I and II, Artiodactyl (especially pigs and cattle) genes have responded differently to different selection pressures based on the analysis of the *GHR* and *PRLR* gene. Unlike other species, sheep appear to harbour an equal amount of nonsynonymous variants in both genes. More nonsynonymous variants were detected in the *PRLR* gene among pigs than cattle (Table 3), contrary to the findings from the *GHR* gene (Table 2). Of particular interest is that samples from the pig, wild boar and *Bison* species were monomorphic for the *GHR* gene, whereas these species, especially the domestic pig, had multiple non-synonymous variants in the *PRLR* gene (Tables 2 and 3). This might implicate opposite selection pressures towards these genes in pigs (purifying selection/selective sweep for *GHR* and positive selection for *PRLR*). A high degree of polymorphism within a gene in pigs is not uncommon. For example, the *FABP4* gene associated with fat metabolism in mammals, is expressed with an unusually high polymorphism in pigs, but only intronic and synonymous variants were found (Ojeda *et al.* 2006). Measured from *PRLR* haplotypes, wild boars were less divergent (H_d 0.21 vs domestic pig had H_d 0.63), that differs from the findings in the *FABP4* gene (Ojeda *et al.* 2006). One hypothesis explaining the observed variation pattern is that the accumulation of mutations in the *PRLR* gene of pigs is due to human influence. The breeding process has most likely favoured pigs with a high number of piglets, whereas in cattle increased hazard for mortality is associated with carrying multiple calves (Shahid *et al.* 2015). Variants in the porcine *PRLR* gene have been found to be associated with total number of piglets born, number of piglets born alive and age of puberty in a Landrace-Duroc-Yorkshire composite population (Rempel *et al.* 2010) and with back fat thickness in Italian Large White sows (Fontanesi *et al.* 2012). However, no association of litter size with *PRLR* haplotypes have been found in the Finnish Yorkshire population (Sironen *et al.* 2012), although a LINE insertion downstream of the *PRLR* gene has been shown to down-regulate the *PRLR* gene in the ovary, oviduct and uterus of LINE homozygous and carrier sows (Sironen *et al.* 2014).

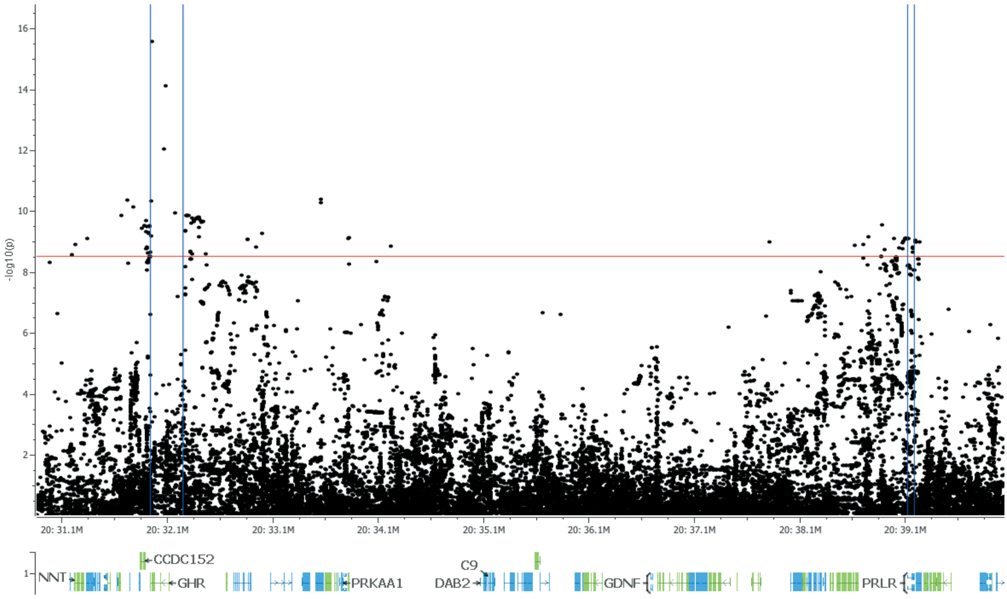

Figure 6. -$\log_{10}(p)$ –values for milk yield plotted against the genomic region of *GHR* and *PRLR* genes (BTA20:30 – 40Mb). The horizontal red line indicates the genome-wide significance level (-$\log_{10}(p) = 8.50$) corresponding to an error rate of 0.05 after correction for multiple testing using a Bonferroni correction and vertical blue lines indicate the positions of the *GHR* (left) and *PRLR* (right) genes.

4.2.5 The *GHR* or *PRLR* exon 10 variants are not significantly associated with milk yield in the current breeding population

Results from study IV were screened to investigate if the SNPs discovered in studies I and II were associated with milk production traits. Since fat or protein yield were not associated with the variants on BTA20 (IV), only measurements of milk yield were used. Some of the variants in studies I or II were not found in the data generated in study IV (e.g. *indicus* specific SNPs), but 19 SNPs that were present in both datasets were used for further exploration. None of them were significantly associated with milk yield, but the putative causative variant (F279Y, Blott *et al.* 2003) was the top SNP for the milk QTL on BTA20 (Refer to Figure 6 and Section 4.4). It would be useful to repeat the association analysis at the haplotype level. The findings from study I indicated that the F-allele (associated with higher protein and fat yield, Blott *et al.* 2003) was not in LD with any of the *GHR* haplotypes, whereas the Y-allele (associated with more milk yield, Blott *et al.* 2003) was linked with divergent haplotypes suggesting either recurrent mutation or intragenic recombination.

4.3 Selection signatures at the genome level

Selection signatures were in the first instance screened at candidate locus/gene level (as performed for the *GHR* and *PRLR* gene in studies I and II). The availability of genome-wide data enables screening for signatures without presumptions. This leads to a rather complicated situation however, since genomes are not perfectly annotated and it is difficult to establish the cause of selection signatures. Different databases can be used. For example, regions showing signatures of selection can be screened for the presence of QTL to confirm whether signatures are overlapping genomic regions with QTL. However, such an approach is limited since most QTL have a small effect on the phenotype (Kemper *et al.* 2014) and selection acts on several loci, such that the changes in the allele frequency changes may not be particularly rapid or drastic, leading to signatures of selection being typically rather weak.

Signals produced by noncommercial traits, such as adaptation to local climate, are more difficult to demonstrate because of a lack of available data. Librado *et al.* (2015) reported enrichment of genes involved in hair development, body size and metabolic and hormone signalling in Yakutian horses that are maintained in the same harsh conditions (winter -50° C) as Yakutian cattle. Librado *et al.* (2015) stated that those genes and pathways are an essential part of the adaptive genetic toolkit in the Yakutian horse. It is possible that the same genes could be good candidates for Yakutian cattle, but thus far, there is no evidence to support a clear functional role.

4.3.1 Selection signatures around genomic regions affecting milk production

Yakutian cattle were used as the reference in one XP-EHH run with the expectation that more milk related signals would be detected. Only two dairy breeds (Finnish Ayrshire and Eastern Finncattle) provided clear evidence of selection around *GHR* based on the screening of the top 5% candidate regions (see Table S1 in III). Several studies have identified traces of selection near or at the *GHR* gene (Randhawa *et al.* 2016, Stella *et al.* 2010, Kemper *et al.* 2014, Qanbari *et al.* 2009, Pintus *et al.* 2014), that is contrary to the gene level study (I) where no deviation from neutrality was observed. The *ABCG2* gene (6:37,913,110-38,030,583) and the casein gene cluster (6:87,141,556 – 87,392,750) on BTA6 are two established QTL for milk production. The dairy breed Yarovslavskaya shows selection signatures within and nearby the *ABCG2* gene, and there is some evidence of selection signatures nearby the casein gene cluster in the Finnish Ayrshire (refer to Tables S1 and S2 in III). Another possible candidate chromosome for milk production is BTA26 (Figure 1). Many of the QTL located on the BTA26 chromosome are associated with fat yield and cluster around 40Mb, whilst Western Finncattle, Northern Finncattle, Eastern Finncattle and Kalmykian cattle show selection signals nearby in the present analysis (Figure 7b).

The *DGAT1* gene region, functionally proven to affect milk production (Grisart *et al.* 2004b), did not show signs of selection in study III. This may be explained by the sparse nature of the SNP chip used. In other studies, the *DGAT1* mutation has showed signals of selection, but only in beef breeds (Kemper *et al.* 2014, Zhao *et al.* 2015). Different explanations for these findings have been proposed that include the direction of selection has changed towards the ancestral allele (increases milk fat instead of milk volume), signals are not detected by test statistics because the ancestral allele is likely to carry a variety of haplotypes (Kemper *et al.* 2014), the mutant allele is not segregating in the population or that the mutant allele has unfavourable pleiotropic effects that prevent the frequency from increasing (Zhao *et al.* 2015).

4.3.2 Immunity and adaptation: linked through selection signatures?

A rather different question was addressed by the XP-EHH analysis using the Finnish Ayrshire as the reference population. It was anticipated that under these circumstances more signals related to adaptation would be found in all other breeds.

Notably, a region containing immuno-related genes (*IL10, IL19, IL20, PIGR, FCAMR, IL24)* on BTA16 (between positions 4,116,037 – 4,616,037) indicated signatures of selection in many native cattle breeds (Figure 7a). The *IL24* gene on BTA16 also showed signatures of selection in the Bovine Genome Sequencing and Analysis Consortium *et al.* (2009). On BTA21 the region with significant selection signals in the Ukrainian Grey and Yakutian cattle (between positions 33,802,673 – 35,302,673, refer to Table S2 in III) includes several genes playing a role in immune system processes (*CSK, GZMB, PML* and *SEMA7A*) and reproduction (*SCAMP5, CSK, CYP11A1, COX5A, CLK3, MDS018, ARID3B* and *ULK3*).

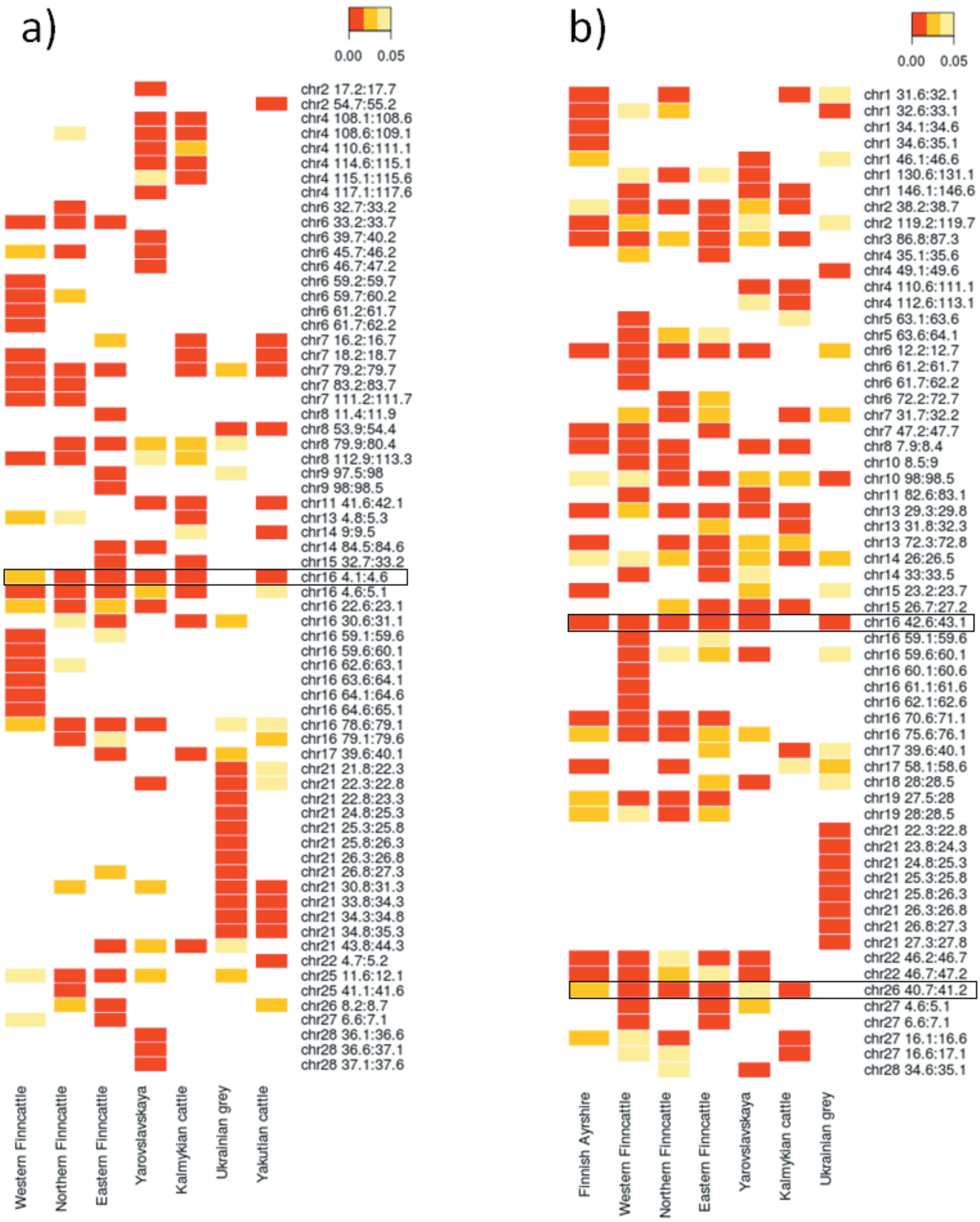

Figure 7. Heat map of the top 10 segments of the each breed when a) Finnish Ayrshire or b) Yakutian cattle are used as the reference population. Only experimental *P*-values ≤0.05 are shown (study III). Highlighted regions indicate selection signatures around milk production QTL (BTA26:40.7-41.7; Section 4.3.1), immune-related genes (BTA16:4.1-4.6; Section 4.3.2) or a regional hotspot for selection signatures (BTA16:42.6-43.1; Section 4.3.3).

Enrichment analysis revealed statistically significant enrichment of genes related to viral processes in the regions with selection signatures in Yakutian cattle. Viral processes include infection of a host cell, replication of the viral genome, assembly of progeny virus particles and in some cases, viral genetic material integration into the host genome. Immunity can be under selection for example due to microbial fermentation in the rumen (higher microbial pressure) or due to herd structure (denser population, exposure to more diseases) (Bovine Genome Sequencing and Analysis Consortium *et al.* 2009). Local veterinarians have reported that Yakutian cattle have lower or no incidence of tuberculosis, leucosis or brucellosis (Kantanen *et al.* 2009).This can be either due an adaptation to climate or indicative of higher resistance to infectious diseases, which could potentially explain the selection signals gained from the immuno-related regions or the lack of infections in extreme climatic conditions.

4.3.3 Selection signature on BTA16 is found across breeds and studies

The study of Gutierrez-Gil *et al.* (2015) compiled results from 21 selection signature searches performed for different *B. taurus* breeds. The overlaps revealed an intriguing region on BTA16 (around 40Mb to 44Mb). It was reported to have been under selection in 19 different cattle breeds with varying purpose of use. In a more recent study (Randhawa *et al.* 2016) involving a meta-assembly of selection signatures in cattle based on the results from 64 different studies of the global cattle population, the same gene rich region on BTA16 was identified to be a regional hotspot for selection signatures. Six out of eight breeds in study III had signatures of selection in that particular genome segment (Figure 7b), that included the Finnish Ayrshire, Eastern Finncattle, Western Finncattle, Northern Finncattle Yarovslavskaya and the Ukrainian Grey. All but one breed (Ukrainian Grey) are used in the most part for milk production.

As the region harbours several genes, a number of suggestions have been put forward to explain the causes of selection signatures. Gene *NPPA* was highlighted in study III because it is associated with female pregnancy via GO annotation. The genes *AGTRAP* (mammary gland function), *KIF1B* (under strong selection in Holstein dairy cattle), *NMNAT1* and *RERE* (candidates of positive selection for embryonic growth and reproductive development as reviewed by Randhawa *et al.* 2016) are good candidates to be the target of observed selection. Immunorelated genes *SLC25A33, SLC45A1, PIK3CD* and *SPSB1* are also located in that region and may be associated with the observed selection signatures (reviewed by Randhawa *et al.* 2016).

The true source of selection signatures remains to be elucidated among the possible group of candidate genes. This region would be an interesting candidate to study with the data from the 1000 Bull Genomes Project. A denser marker map with all possible variants could help to narrow down the genomic region and the candidate gene list to make more precise predictions of functionality. The region is found to be selected in European, African and Zebu breeds (Randhawa *et al.* 2016), but this does not infer the same causative gene or common variant.

Table 6 QTL regions for each trait. The top SNP for each QTL are shown including position, $\log_{10}(p)$-values, minor allele frequency (MAF), gene information, annotation of the top SNP, allele substitution effect (b.value) and standard error of the b.value (SE).

Chromosome	Start (bp)	End (bp)	Length of the QTL region (bp)	Significant SNPs in region	No. of genes with significant SNPs within the QTL region	Top SNP	Position of the top SNP (bp)	$\log_{10}(p)$	MAF	Gene	Annotation of the top SNP	b.value	SE
FAT YIELD													
5	92,372,732	94,425,668	2,052,936	330	3	rs209818856	93,945,694	27.49	0.38	*MGST1*	intron variant	-2.807	0.253
14	1,448,510	2271832	823,322	509	49	rs136783505	1,807,140	42.01	0.07	*DGAT1/HSF1*	downstream variant/intron variant	-6.709	0.484
23	28,567,796	28,591,530	23,734	5	1	rs381390819	28,567,796	9.36	0.45	*TRIM26*	intron variant	1.148	0.184
25	8,222,347	11,507,986	3,285,639	883	16	rs379546164	9,870,005	14.40	0.24	*CLEC16A*	intron variant	-1.769	0.224
25	36,226,978	36,227,132	154	2		rs109480808	36,226,978	8.72	0.16		intergenic variant	1.635	0.272
26	22,144,777	24,793,744	2,648,967	595	32	rs438420348	24,379,571	14.28	0.16	*NEURL1*	intron variant	-2.273	0.290
26	44,802,991	44,802,991	0	1		rs135624939	44,802,991	9.72	0.28		intergenic variant	-1.474	0.231
MILK YIELD													
5	112,343,204	11,2450,860	107,656	2	1	rs383553819	112,343,204	8.70	0.36	*MKL1*	intron variant	1.437	0.239
14	1,448,510	2,271,832	823,322	455	48	rs133033480	1,743,939	33.01	0.09	*CPSF1/ADCK5*	downstream variant/splice region variant, intron variant	6.266	0.513
16	1,322,611	1,322,611	0	1	1	rs108979795	1,322,611	8.58	0.28	*LAX1*	upstream variant	-1.451	0.243
19	61,447,138	61,449,096	1,958	5		rs210324693	61,449,096	8.93	0.32		intergenic variant	1.490	0.244
20	30,531,217	32,952,019	2,420,802	74	5	rs385640152	31,909,478	15.56	0.11	*GHR*	missense variant	-3.877	0.472
20	37,766,226	39,183,141	1,416,915	34	3	NA	38,828,254	9.54	0.16		intergenic variant	-2.250	0.356
25	2,669,704	2,669,704	0	1		rs209691835	2,669,704	9.21	0.13		intergenic variant	-2.855	0.460
25	3,494,706	3,516,671	21,965	13	4	rs110749311	3,498,960	9.32	0.41	*PAM16/GLIS2*	downstream variant	1.225	0.196
PROTEIN YIELD													
5	112,450,860	112,450,860	0	1		rs109041054	112,450,860	8.88	0.48		intergenic variant	-1.473	0.242
14	1,802,667	1,802,667	0	1	2	NA	1,802,667	8.52	0.06	*DGAT1/HSF1*	intron variant/downstream variant	3.354	0.564
25	1,094,996	1,257,612	162,616	12	3	rs136085792	1,103,856	10.83	0.22	*UNKL*	intron variant	1.694	0.250
25	3,306,363	3,516,671	210,308	40	8	rs110749311	3,498,960	11.70	0.41	*PAM16/GLIS2*	downstream variant	1.427	0.202

4.4 Association analysis for milk production traits

Three milk production traits, milk yield (MY), fat yield (FY) and protein yield (PY) were analysed with the imputed whole genome variants derived from Nordic Red Cattle (IV). Generally, fewer QTL were detected for protein yield than for milk and fat yield. This might implicate protein synthesis being controlled by more genes with a smaller effect compared with the two traits that may be regulated by fewer genes with much larger effects (Lemay *et al.* 2009). Table 6 lists QTL regions identified for each trait. Those QTL with a large associated region were examined for existence of several QTL within the region by fixing the top SNPs. Peaks remaining after fixation were considered as potential additional QTL. Milk QTL were compared with the QTL from study of Höglund *et al.* (2015) to establish possible overlaps with genomic regions associated to fertility.

4.4.1 Confirmed QTL on BTA14 and BTA20

Collectively, seven, eight and four QTL were found for fat, milk and protein yield, respectively (Figure 8, Table 6). The most apparent QTL is located on BTA14. The QTL is shared between all studied traits even the top SNP is not common to all (Table 6). A known QTL affecting milk yield and composition, thought to be caused by a functional variant K232A (14:1,802,266) in the *DGAT1* gene (Grisart *et al.* 2002, Grisart *et al.* 2004b) is located within this region. The K232A mutation was not the variant with the lowest P-value in our data or in studies of Fleckvieh and Holstein bulls (Daetwyler *et al.* 2014). However, when the effect of K232A variant was fixed, no additional significant SNP effects remained (Figure 9). Whilst variant K232A is a rather convincing QTN, it would be necessary to evaluate the haplotype structure in and around the *DGAT1* gene using for example, data available from the 1000 Bull Genomes Project. It would also be possible to investigate the evolutionary history of the *DGAT1* gene by adding sequence information from other Artiodactyla species, since both *DGAT1* and *DGAT2* are ubiquitous in most eukaryotic organisms, and therefore assumed to be very ancient enzymes (Turchetto-Zolet *et al.* 2011).

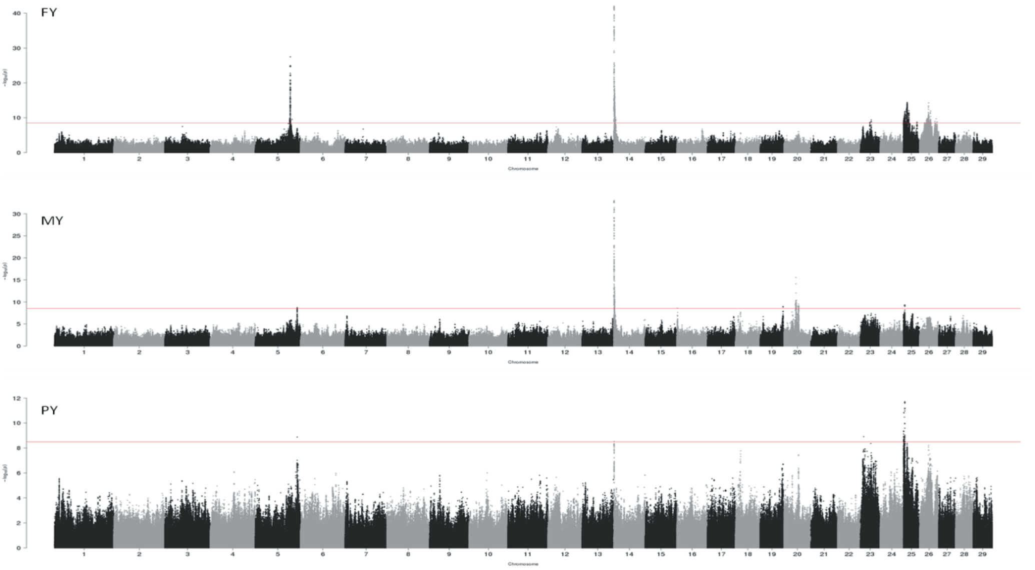

Figure 8. Genome-wide Manhattan plots for fat yield (FY), milk yield (MY) and protein yield (PY). The red line indicates the genome-wide level of significance.

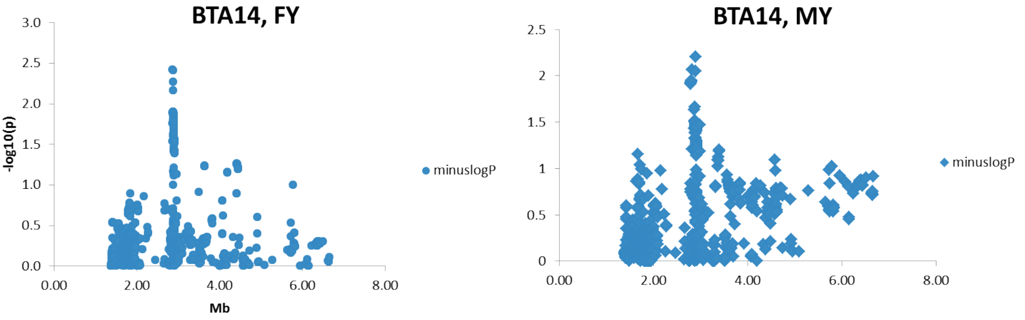

Figure 9. –log(p) –values for fat yield (FY) and milk yield (MY) plotted against the genomic location around the *DGAT1* gene on BTA14 when the effect of a causative variant (K232A) is fixed.

On BTA20, a statistically significant effect of the known and most likely causative variant (F279Y, Blott *et al.* 2003) in the *GHR* gene was seen for milk yield, although elevated –$\log_{10}(p)$-values were also detected for fat yield (Figure 8, Table 6). Even though this study did not provide strong evidence of an association to fat and protein yield, a recent study (Kadri *et al.* 2015) reported a strong association of F279Y with fat and protein yields as well as milk yield. The frequency of the Y-allele that increases milk yield at the expense of milk fat and protein percentage, was 0.101 in the Finnish Ayrshire and 0.053 in the Danish Red based on imputed data from study IV (unpublished). Corresponding values of 0.08 and 0.01 were obtained based on the 1000 Bull dataset. Samples from the Finnish Ayrshire used in study I were older than those used in study IV or in the 1000 Bull dataset and show a higher frequency for the Y-allele (0.13). However, older samples from the Danish Red in study I had a lower Y-allele frequency (0.03) compared with those analysed in study IV (Appendix 1). Currently, the Nordic Red cattle is evaluated under a joint breeding value evaluation system (NAV, www.nordicebv.info/Forside.htm) and the genetic material is shared between Finland, Sweden and Denmark. This may account for the fluctuation in Y-allele frequency. In general, the frequency of the Y-allele was 0.08 among European breeds used in study I, whereas African breeds were homozygous for the F-allele (I). The highest frequency (0.545) of the Y-allele was detected in the Hereford beef breed but when the frequency was calculated from the 1000 Bull data, the Hereford (n = 47) had Y-allele frequency of 0.160 (Appendix 1). This compares with the Y-allele frequency for the entire 1000 Bull dataset of 0.121.

The other QTL for milk yield on BTA20 was located in the intragenic region (IV). Viitala *et al.* (2006) suggested that the variant S18N may influence protein and fat yield. However, the results from study IV did not support the causality of S18N and it has been proposed that variant S18N is more likely linked to the causative mutation, rather than being causative *per se* (e.g. Pausch *et al.* 2015). Nonetheless, the *PLRLR* gene has an important role in milk production. As indicated by the gene network analysis, *GHR*, *PRLR* and *DGAT1* act via a common signalling network (Figure 10). The mechanism of how the identified intragenic candidate variant influences milk production is however, uncertain.

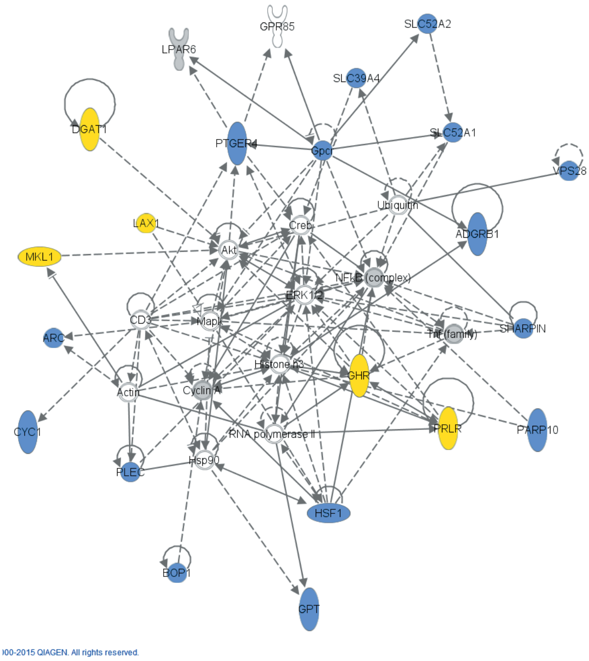

Figure 10. Gene networks generated by the IPA© platform for milk yield. Genes marked with blue exhibit variant associated with milk yield. A yellow colour represents genes with a candidate causative variant for milk yield. Genes marked in white or grey are added by IPA to connect the network. Dotted lines indicate indirect interactions and solid lines indicate a direct interaction between specific genes (Paper IV).

4.4.2 Statistical and methodological choices affect results

Previously a QTL for milk, fat and protein yield has been detected on BTA12 in the Finnish Ayrshire by Viitala *et al.* (2003). More recent studies have confirmed the occurrence of the same QTL in the Nordic red population (Kadri *et al.* 2014). This QTL not detected in the population analysed in study IV has been located in the immediate vicinity of a 660kb deletion that is embryonically lethal (Kadri *et al.* 2014). The genotypes used in the study of Kadri *et al.* (2014) were obtained by genotyping and not imputed. Such a deletion may affect imputation accuracy, confounding the interpretation of variants associated to the BTA12 milk QTL, which may well explain why this was not discovered in study IV. However, the two-step approach (from 50K -> HD -> sequence level) used in study IV to impute whole genome variants should improve the accuracy of imputation compared with direct imputation of sequence level data from the 50K SNP-panel (van Binsbergen *et al.* 2014).

The significance of the QTL findings were investigated by correcting P-values for multiple testing using a Bonferroni correction to an error rate of 0.05 ($-\log_{10}(p)$ 8.50). The Bonferroni correction is inherently conservative leading to many true associations being discarded, simply because the correction is performed for all SNPs in the panel, even if many are in LD. Study population also has an effect, such that the larger the sample group the more loci can be expected to reach the significance threshold (Wellcome Trust Case Control Consortium 2007). Meuwissen *et al.* (2016) has proposed that the focus should be on the estimation of the effects of all markers rather than calculating significances for individual SNP markers. It is evident that the method of analysis influences the results. The EMMAX method systematically underestimates the most significant P –values based on the implicit assumption that each SNP has only a minor effect on the desired trait.

The strength of association between a sampled SNP and a causative site depends on both the history of recombination events separating them and on where each mutation occurred in the coalescent tree (Remington 2015). An obvious drawback of association studies based on SNP chips is that only when every coding or regulatory variant affecting the phenotype is in complete LD with a given SNP is it possible to capture the entire phenotypic effect. In theory, when using whole genome variants, it should be possible to detect causative variants. Due to imputation limitations (for example deletions and insertions being poorly covered and low frequency markers being filtered out due to quality requirements) the complete set of variants is inevitably missed. When using whole genome variants, LD imposes formidable challenges. Phenotypes are typically available for commercial populations having elevated rates of LD (i.e. Sodeland *et al.* 2011) due to artificial insemination. Thus, LD complicates defining the true variant as several variants are linked to the causative mutation. As performed in studies I, II and IV, the predicted functional consequences of variants (for example Ng & Henikoff 2003) may help to identify true causative mutations. Alternatively, the likelihood of finding causative variants could be increased using Bayesian methods ((Bayes A, B, R) (Meuwissen *et al.* 2016). Recently MacLeod *et al.* (2016) proposed a new method Bayes RC that enables the inclusion of *a priori* biological information of variants in the model that simultaneously improves QTL discovery and genomic prediction accuracy. However, the reliability of *a priori* information is deeply connected with the accuracy of genome annotations that whilst good, may not be sufficiently detailed in non-model organisms such as cattle.

4.4.3 Milk production and fertility are linked through gene networks

One of the major drawbacks of selecting for higher milk production in dairy cattle has been a concomitant decline in cow fertility (Atashi *et al.* 2012, Butler 2013, Dochi *et al.* 2010), traits that are often assumed to be connected. The Nordic breeding program has included fertility in genomic evaluations for several years. Over this time the decline in dairy cow fertility in the Nordic countries has been arrested and in some cases even refracted (www.sweebv.info/ba52nycknav.aspx). Results in Holstein and Jersey breeds have indicated little or no overlap between genomic regions associated with milk yield and fertility (Minozzi *et al.* 2013, Aliloo *et al.* 2015). Data from (Höglund *et al.* 2015) and study IV was used to see if similar observations also held true in Nordic Red cattle.

No common SNPs were found associated with milk production traits and fertility consistent with the observations of Minozzi *et al.* (2013) and Aliloo *et al.* (2015). However, a few SNPs associated to fertility were located in close proximity to the QTL regions on BTA20 (30,531,217 – 33,773,311 and 38,572,674 – 39,183,141) detected for milk yield. When a more detailed analysis of gene networks was performed, a common gene network pathway for milk production traits and fertility was identified (Figure 11). It is therefore recommended that the use of functional gene information of networks and pathways should be explored to pinpoint interacting genes as possible candidates for phenotypic effects. Accurate genomic prediction of phenotypes is essential for animal breeding. However, even the most recent methodologies do not yet relate the phenotype to molecular pathways and gene networks involved in the regulation of homeostasis, development and function. Characterization of causal sequence variants and an improved understanding of the underlying biology has the potential to increase the efficacy of genomic selection compared with anonymous markers alone (Meuwissen *et al.* 2013), such that a greater understanding of the genetic mechanisms underlying milk production traits could simultaneously improve milk yield and health and fertility traits (Daetwyler *et al.* 2014).

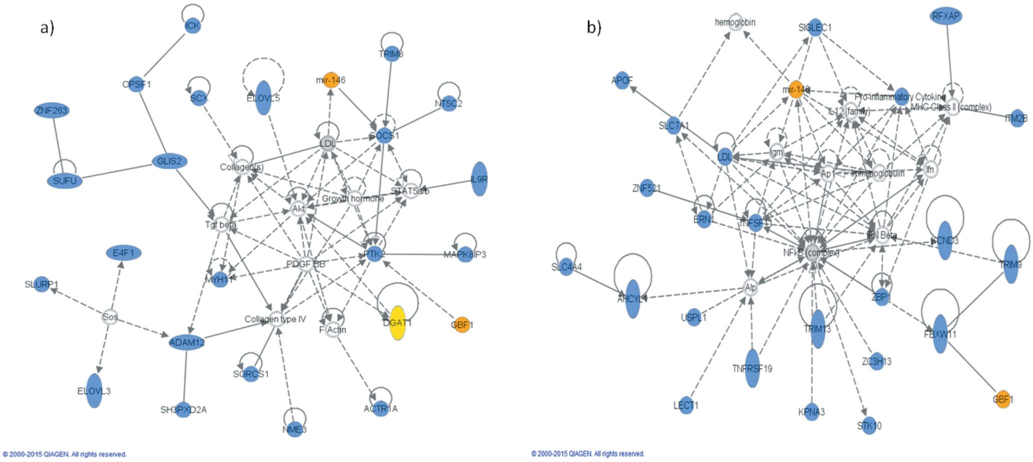

Figure 11 Gene networks generated by the IPA® platform for fat yield (a) and fertility index (b). Genes marked with blue exhibit variants with a statistically significant association with fat yield or the fertility index. The yellow colour represents genes that have candidate causative variant for fat yield; genes indicated in orange have SNPs significantly associated with fertility and fat yield. Genes with white or grey colour have been included by IPA to connect the network. Dotted lines indicate indirect interactions and the solid lines indicate direct interaction between the genes.

5. CONCLUSIONS AND FUTURE PROSPECTS

The 'Omics' evolution advanced substantially during the course of the research outlined in this thesis. Whole genome sequences became available and genome annotations for many livestock species were developed. This partly led to the atypical progress of studies documented in this thesis. Typically QTL mapping is done from a genome scan to more detailed analysis of QTL loci. Here a few selected candidate QTL loci were first investigated in detail followed by association analysis using all possible SNPs from whole genome sequences from the breed in question.

The rapid escalation of genomic data has revolutionised the transition of population genetics into population genomics. Sparse microsatellite panels are or will be replaced with whole genome SNP/sequence information increasing the accuracy of the results because the new information does not rely on only a few loci. Furthermore, as noted in study III and other studies, commercial SNP chips are not free of ascertainment bias. This leads to underestimations of genetic diversity in native populations (as for Yakutian cattle) and overestimations in commercial populations (such as the Finnish Ayrshire, III). This bias can be diminished using denser SNP chips, but ascertainment bias is only fully prevented using whole genome sequencing. Even though the price of whole genome sequencing has decreased dramatically it is still the factor limiting the availability of genome level sequence information. Hence imputation could be used as a robust and cost-effective way to expand available information not only for the purposes of breeding (IV) but also for conservation biology. Nevertheless, imputation is not accurate for rare variants that are important to maintain if the goal is to maintain the highest level of diversity possible. Furthermore, structural genomic variants (e.g. deletions, gains, copy number variations), introduce additional challenges which are not easily accommodated in the analyses even if they may influence the phenotype.

Investigation of the candidate genes *GHR* and *PRLR* among Artiodactyl species revealed divergent selection pressure towards these genes. An unexpected level of nonsynonymous variation was found to accumulate or persist in these genes from different Artiodactyla species. The *GHR* gene was more divergent within genus Bos than between different species among the Bovinae lineage and was shown to have been selected when selection signatures were searched for at the genomic level (III). It was striking that the *PRLR* gene has accumulated in pigs, particularly nonsynonymous mutations during the domestication process. Possible explanations for the observed diversity patterns include: selective sweeps before domestication (*GHR* in pigs) or before species divergence (*GHR* in Bison), directional/artificial selection (*PRLR* in pigs) or functional switching (*GHR* in cattle). However, the reason for the persistence of variation at *GHR* in cattle is not known.

Additional candidate loci for milk production were located collectively from eight cattle chromosomes (IV) using imputed whole genome variants. New candidate loci were identified together with those previously identified. However, establishing the true causative variant remains challenging even when the densest possible marker map is used because of linkage disequilibrium. The occurrence of LD limits the possibilities to pinpoint just one functional variant but does allow haplotypes most likely being partly responsible of the phenotypic effect to be explored. In general, all livestock species are suffering from an incomplete annotation of the genome leading to only tentative (*in silico*) predictions of the effects of observed variants. Non-coding regions for gene regulation and function as well the impact of the synonymous variants requires further studies to establish how these may influence the phenotype.

Identification and functional confirmation of the causative variants is hugely demanding. Only very recently (5/2016) the first empirical validation of a putative causative allele in livestock was published when Carlson *et al.* (2016) used genome editing (using TALENs methodology) to produce hornless cattle. It can be assumed that in the future genome editing techniques (such as CRISPR/Cas9 and TALENs) will be used more widely to provide proof of the causativity of variants *in vitro/vivo* when a particular phenotype is caused and/or strongly affected by a single mutation.

It remains unclear as to whether genome editing techniques will be used routinely in animal breeding programs. In addition to legal and ethical issues, many technical matters (such as off-target effects) need to be resolved before genome editing methods can be applied in practice. Furthermore, the impacts of other genetic or non-genetic factors such as gene-to-gene interactions (epistasis), epigenomics (incl. DNA methylation and histone modifications), nutrigenomics (immediate and direct effects of nutritional factors on gene expression, Soller 2015) and the microbiome (microbial population in digestive tract) all contribute to the observed phenotype, and yet the influence of these factors is far from clear. If these elements together with genomic information from both sexes could be implemented to the genomic selection concept, then breeding programs would benefit substantially.

6. ACKNOWLEDGEMENTS

This research was carried out at the MTT Agrifood Research Finland, Jokioinen and at the Natural Resources Institute Finland (Luke), Jokioinen. The work was financially supported by the Ministry of Agriculture and Forestry in Finland, Finnish graduate School in Population Genetics, Finnish Cultural foundation, Häme Regional fund, Emil Aaltonen foundation, Betty Väänänen foundation and Doctoral Programme in Biology, Geography and Geology (University of Turku).

First of all I want to express my sincerest and deepest gratitude to my supervisor Professor Johanna Vilkki. Thank you for your excellent guidance during this journey that in a way started already from the course "Molecular Genetics in Animal Breeding" you kept ages ago in the university. Without your support, patience and encouraging attitude I would have never been able to finish this thesis.

I am grateful to Professor Juha Kantanen that he has offered me possibilities to work in his projects. I feel privileged because I have had the opportunity to work with the unique cattle samples you have collected. Those samples and your enthusiasm have been essential to this work.

Professor Craig Primmer from the University of Turku is warmly thanked for being my research director and helping with my studies and last steps of this process. The reviewers of this thesis, Professor Outi Savolainen and Dr James Kijas, are gratefully acknowledged for their valuable comments and suggestions. I am grateful to Professor Kevin Shingfield (sadly passed away on September 11th, 2016) for the language revision of my thesis.

This work would not have been possible without my co-authors; your contribution is greatly acknowledged. I wish to thank Professor Asko Mäki-Tanila for his contribution to the first paper and for the excellent working facilities he provided in the former Animal production research unit in MTT. Dr Miika Tapio and Dr Goutam Sahana, I owe you both a lot. Miika, your contribution to the third paper was invaluable and Goutam, collaboration with the fourth paper was truly a privilege; you showed how smooth and easy it can be to work together even not being in the same country or the institute.

All the past and current members of Animal genomics team are greatly acknowledged. I want to thank all of you for your help and friendship in work related and not work related things. Anu Sironen, Ilma Tapio and Daniel Fischer, our coffee room discussions have been priceless during the last year! Anu, thank you for your friendship and all your help at the work. Ilma, your positive attitude and warm personality have saved many working days. Daniel, I am grateful to you for your help regarding the 1000 Bull Genomes data. Anneli Virta, Tiina Jaakkola, Jonna Tabell and Jouni Virta, you are the ones who taught me how to work in the lab, thank you! Professor Esa Mäntysaari (Biometrical genetics team in Luke), thank you for patiently answering my questions regarding the animal breeding.

My friends from childhood, school, university and adulthood, Outi, Janika, Mervi, Minna, TerhiP, TerhiS, Mirja, Saija, Elina, you are all part of the environmental factor of my phenotype, I wouldn't be me without you girls! Especially vorssatytöt; those countless hours used for talking, hanging around and card playing in high school library and moments we have spent together after school years are irreplaceable!

Anneli ja Antti, kiitos tuesta ja lastenhoitoavusta, jota olemme teiltä saaneet, se on korvaamatonta. Äiti, kiitos siitä tuesta, jonka olen saanut kotoa. Kiitos myös kaikesta siitä avusta jonka olet antanut arjen pyörittämiseen ja lasten hoitoon, ilman niitä ei tätä kirjaa olisi kirjoitettu. My sister Johanna, your unconditional support means lot to me. Jussi, thank you for your love and support, they mean more to me than I can say. My little superstars, Lotta and Ossi, thank you for showing me what fundamentally is important in life.

7. REFERENCES

Aliloo H., Pryce J.E., Gonzalez-Recio O., Cocks B.G. & Hayes B.J. (2015). Validation of markers with non-additive effects on milk yield and fertility in Holstein and Jersey cows. *BMC genetic*s 16, 89-015-0241-9.

Atashi H., Zamiri M.J., Sayyadnejad M.B. & Akhlaghi A. (2012). Trends in the reproductive performance of Holstein dairy cows in Iran. *Tropical animal health and productio*n 44, 2001-2006.

Bahbahani H., Clifford H., Wragg D., Mbole-Kariuki M.N., Van Tassell C., Sonstegard T., et al. (2015). Signatures of positive selection in East African Shorthorn Zebu: A genome-wide single nucleotide polymorphism analysis. *Scientific report*s 5, 11729.

Barbato ,M, Orozco-terWengel P., Tapio ,M. & Bruford MW. (2015) SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics* 6,109

Barendse W., Harrison B.E., Bunch R.J., Thomas M.B. & Turner L.B. (2009). Genome wide signatures of positive selection: the comparison of independent samples and the identification of regions associated to traits. *BMC genomic*s 10, 178.

Bickhart D.M. & Liu G.E. (2014). The challenges and importance of structural variation detection in livestock. *Frontiers in genetic*s 5, 37.

Blott S., Kim J.J., Moisio S., Schmidt-Kuntzel A., Cornet A., Berzi P., et al. (2003). Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetic*s 163, 253-266.

Boichard D., Fritz S., Rossignol M.N., Boscher M.Y., Malafosse A. & Colleau J.J. (2002) Implementation of marker-assisted selection in French dairy cattle. *Proc. 7th World Congr. Genet. Appl. Livest. Prod, Montpellier, France Communication no. 22-0*3.

Boichard D., Ducrocq V. & Fritz S. (2015). Sustainable dairy cattle selection in the genomic era. *Journal of animal breeding and genetics* 132, 135-143.

Boitard S., Rodriguez W., Jay F., Mona S. & Austerlitz F. (2016). Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLoS genetic*s 12, e1005877.

Bole-Feysot C., Goffin V., Edery M., Binart N. & Kelly P.A. (1998). Prolactin (PRL) and its receptor: actions, signal transduction pathways and phenotypes observed in PRL receptor knockout mice. *Endocrine review*s 19, 225-268.

Bollongino R., Burger J., Powell A., Mashkour M., Vigne J.D. & Thomas M.G. (2012). Modern Taurine Cattle Descended from Small Number of Near-Eastern Founders. *Molecular biology and evolutio*n.

Bomba L., Nicolazzi E.L., Milanesi M., Negrini R., Mancini G., Biscarini F., et al. (2015). Relative extended haplotype homozygosity signals across breeds reveal dairy and beef specific signatures of selection. *Genetics, selection, evolution* 47, 25-015-0113-9.

Botstein D., White R.L., Skolnick M. & Davis R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetic*s 32, 314-331.

Bovine Genome Sequencing and Analysis Consortium, Elsik C.G., Tellam R.L., Worley K.C., Gibbs R.A., Muzny D.M., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324, 522-528.

Bovine HapMap Consortium, Gibbs R.A., Taylor J.F., Van Tassell C.P., Barendse W., Eversole K.A., et al. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324, 528-532.

Bradley D.G., MacHugh D.E., Cunningham P. & Loftus R.T. (1996). Mitochondrial diversity and the origins of African and European cattle. *Proceedings of the National Academy of Sciences of the United States of Americ*a 93, 5131-5135.

Brotherstone S. & Goddard M. (2005). Artificial selection and maintenance of genetic variance in the global dairy cow population. *Philosophical transactions of the Royal Society of London.Series B, Biological science*s 360, 1479-1488.

Brown T.A. (1999) Genomes. BIOS Scientific Publisher Ltd.

Browning S.R. & Browning B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetic*s 81, 1084-1097.

Butler S.T. (2013). Genetic control of reproduction in dairy cows. *Reproduction, fertility, and developmen*t 26, 1-11.

Campbell N., Reece J. & Mitchell L. (1999) Biology, 5th Edition. Jim Green.

Carlson D.F., Lancto C.A., Zang B., Kim E.S., Walton M., Oldeschulte D., et al. (2016). Production of hornless dairy cattle from genome-edited cell lines. *Nature biotechnolog*y 34, 479-481.

Charlesworth B. (2015). Causes of natural variation in fitness: evidence from studies of Drosophila populations. *Proceedings of the National Academy of Sciences of the United States of Americ*a 112, 1662-1669.

Chen H., Patterson N. & Reich D. (2010). Population differentiation as a test for selective sweeps. *Genome researc*h 20, 393-402.

Conolly J., Sue Colledge S., Dobney K., Vigne J., Peterse J., Stopp B., Manning K. & Shennan S. (2011) Meta-analysis of zooarchaeological data from SW Asia and SE Europe provides insight into the origins and spread of animal husbandry. *Journal of Archaeological Scienc*e, 538.

DAD-IS. Domestic Animal Diversity Information System, http://dad.fao.org/. Accessed 4/2016

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. *Nature genetic*s 46, 858-865.

Darwin C. (1859) The Origin of Species by Means of Natural Selection, Or the Preservation of Favored Faces in the Struggle for Life. Modern Library, New York.

de Koning D. (2015) Genome-wide Association Studies in Pedigreed Populations. In: *Molecular and Quantitative Animal Genetics*. (ed. by H. Khatib), pp. 155. John Wiley Sons, Inc.

Dochi O., Kabeya S. & Koyama H. (2010). Factors affecting reproductive performance in high milk-producing Holstein cows. *The Journal of reproduction and developmen*t 56 Suppl, S61-5.

Durkin K., Coppieters W., Drogemuller C., Ahariz N., Cambisano N., Druet T., et al. (2012). Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Natur*e 482, 81-84.

Edea Z., Bhuiyan M.S., Dessie T., Rothschild M.F., Dadi H. & Kim K.S. (2015). Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. *Animal* 9, 218-226.

Ekine C.C., Rowe S.J., Bishop S.C. & de Koning D.J. (2014). Why breeding values estimated using familial data should not be used for genome-wide association studies. *G3* 4, 341-347.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS on*e 6, e19379.

Eu-Ahsunthornwattana J., Miller E.N., Fakiola M., Wellcome Trust Case Control Consortium 2, Jeronimo S.M., Blackwell J.M., et al. (2014). Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS genetic*s 10, e1004445.

Evershed R.P., Payne S., Sherratt A.G., Copley M.S., Coolidge J., Urem-Kotsu D., et al. (2008). Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Natur*e 455, 528-531.

FAO (2015) The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture. (ed. by B.D. Scherf & D. Pilling). FAO Commission on Genetic Resources for Food and Agriculture Assessments. Rome (available at http://www.fao.org/3/a-i4787e/index.html).

FAO (2013). In vivo conservation of animal genetic resources. FAO Animal Production and Health Guidelines. No. 14. Rome

FAO (2009). Proceedings of the Expert Meeting on How to Feed the World in 2050. 24-26 June 2009, FAO Headquarters, Rome

Fariello M.I., Boitard S., Naya H., SanCristobal M. & Servin B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetic*s 193, 929-941.

Felius M. (1995) Cattle Breeds - an Encyclopedia. Trafalgar Square Books.1st edition.

Fijarczyk A. & Babik W. (2015). Detecting balancing selection in genomes: limits and prospects. *Molecular ecology* 24, 3529-3545.

Flori L., Fritz S., Jaffrezic F., Boussaha M., Gut I., Heath S., et al. (2009). The genome response to artificial selection: a case study in dairy cattle. *PloS on*e 4, e6595.

Flury C., Tapio M., Sonstegard T., Drogemuller C., Leeb T., Simianer H., et al. (2010). Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *Journal of animal breeding and genetics* 127, 339-347.

Fontanesi L., Galimberti G., Calò D.G., Fronza R., Martelli P.L., Scotti E., et al. (2012). Identification and association analysis of several hundred single nucleotide polymorphisms within candidate genes for back fat thickness in Italian Large White pigs using a selective genotyping approach. *Journal of animal breeding and genetics* 90, 2450-2464

Forsyth I.A. & Wallis M. (2002). Growth hormone and prolactin--molecular and functional evolution. *Journal of mammary gland biology and neoplasi*a 7, 291-312.

Frank S.J. (2001). Growth hormone signalling and its regulation: preventing too much of a good thing. *Growth hormone & IGF research* 11, 201-212.

Fu Y.X. & Li W.H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133, 693-709.

Funk D.A. (2006). Major advances in globalization and consolidation of the artificial insemination industry. *Journal of dairy science* 89, 1362-1368.

Futuyma D. (2006) Evolution. Sinauer Associates, INC.

Gautier M., Laloe D. & Moazami-Goudarzi K. (2010). Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *PloS one* 5, e13038.

Georges M. (2014). Towards sequence-based genomic selection of cattle. *Nature genetics* 46, 807-809.

Georges M. (2007). Mapping, fine mapping, and molecular dissection of quantitative trait Loci in domestic animals. *Annual review of genomics and human genetics* 8, 131-162.

Georges M., Nielsen D., Mackinnon M., Mishra A., Okimoto R., Pasquino A.T., et al. (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* 139, 907-920.

Georges M., Drinkwater R., King T., Mishra A., Moore S.S., Nielsen D., et al. (1993). Microsatellite mapping of a gene affecting horn development in Bos taurus. *Nature genetics* 4, 206-210.

Gorjanc G., Cleveland M.A., Houston R.D. & Hickey J.M. (2015). Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genetics, selection, evolution* 47, 12-015-0102-z.

Grada A. & Weinbrecht K. (2013). Next-generation sequencing: methodology and application. *The Journal of investigative dermatology* 133, e11.

Griffiths A., Gelbart W., Miller J. & Lewontin R. (1999) Modern Genetic Analysis. W. H. Freeman and Company.

Grisart B., Farnir F., Karim L., Cambisano N., Kim J.J., Kvasz A., et al. (2004a). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America* 101, 2398-2403.

Grisart B., Farnir F., Karim L., Cambisano N., Kim J.J., Kvasz A., et al. (2004b). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America* 101, 2398-2403.

Grisart B., Coppieters W., Farnir F., Karim L., Ford C., Berzi P., et al. (2002). Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome research* 12, 222-231.

Groeneveld L.F., Lenstra J.A., Eding H., Toro M.A., Scherf B., Pilling D., et al. (2010). Genetic diversity in farm animals--a review. *Animal Genetics* 41 Suppl 1, 6-31.

Gutierrez-Gil B., Arranz J.J. & Wiener P. (2015). An interpretive review of selective sweep studies in Bos taurus cattle populations: identification of unique and shared selection signals across breeds. *Frontiers in genetics* 6, 167.

Hayes B.J., Bowman P.J., Chamberlain A.J. & Goddard M.E. (2009a). Invited review: Genomic selection in dairy cattle: progress and challenges. *Journal of dairy science* 92, 433-443.

Hayes B.J., Chamberlain A.J., Maceachern S., Savin K., McPartlan H., MacLeod I., et al. (2009b). A genome map of divergent artificial selection between Bos taurus dairy cattle and Bos taurus beef cattle. *Animal Genetics* 40, 176-184.

Heather J.M. & Chain B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* 107, 1-8.

Hickey J.M. (2013). Sequencing millions of animals for genomic selection 2.0. *Journal of animal breeding and genetics* 130, 331-332.

Hiemstra S., Mäki-Tanila A. & Gandini G. (2010) Recommendations for the management of local cattle breeds in Europe. In: *Local Cattle Breeds in Europe : Development of Policies and Strategies for Self-Sustaining Breeds.* (ed. by S. Himestra, Y. de Haas, A. Mäki-Tanila & G. Gandini), pp. 142-150.

Hoff K.J. (2009). The effect of sequencing errors on metagenomic gene prediction. *BMC genomics* 10, 520-2164-10-520.

Höglund J.K., Buitenhuis B., Guldbrandtsen B., Lund M.S. & Sahana G. (2015). Genome-wide association study for female fertility in Nordic Red cattle. *BMC genetics* 16, 110-015-0269-x.

Höglund J.K., Sahana G., Brondum R.F., Guldbrandtsen B., Buitenhuis B. & Lund M.S. (2014). Fine mapping QTL for female fertility on BTA04 and BTA13 in dairy cattle using HD SNP and sequence data. *BMC genomics* 15, 790-2164-15-790.

Hudson R.R., Kreitman M. & Aguade M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* 116, 153-159.

Iso-Touru T. (2004) Valinnan Vaikutus Kasvuhormonireseptorin Geenin Vaihteluun Nautaroduissa. Master's thesis, University of Turku.

Kadri N.K., Guldbrandtsen B., Lund M.S. & Sahana G. (2015). Genetic dissection of milk yield traits and mastitis resistance QTL on chromosome 20 in dairy cattle. *Journal of dairy science*.

Kadri N.K., Sahana G., Charlier C., Iso-Touru T., Guldbrandtsen B., Karim L., et al. (2014). A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. *PLoS genetics* 10, e1004049.

Kang H.M., Sul J.H., Zaitlen N.A., Kong S., Freimer N.B., Sabatti C., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* 42, 348-354.

Kantanen J., Ammosov I., Li M.H., Osva A. & Popov R. (2009) A cow of the permafrost. In: *Sakha Ynaga : Cattle of the Yakuts* (ed. by L. Granberg, K. Soini & J. Kantanen), pp. 19. *Finnish Academy of Science and Letters*.

Kantanen J., Olsaker I., Brusgaard K., Eythorsdottir E., Holm L.E., Lien S., et al. (2000a). Frequencies of genes for coat colour and horns in Nordic cattle breeds. *Genetics, selection, evolution* 32, 561-576.

Kantanen J., Olsaker I., Holm L.E., Lien S., Vilkki J., Brusgaard K., et al. (2000b). Genetic diversity and population structure of 20 North European cattle breeds. *The Journal of heredity* 91, 446-457.

Kemper K.E., Saxton S.J., Bolormaa S., Hayes B.J. & Goddard M.E. (2014). Selection for complex traits leaves little or no classic signatures of selection. *BMC genomics* 15, 246-2164-15-246.

Kemper K.E. & Goddard M.E. (2012). Understanding and predicting complex traits: knowledge from cattle. *Human molecular genetics* 21, R45-51.

Kim Y. & Stephan W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160, 765-777.

Kinsella R.J., Kahari A., Haider S., Zamora J., Proctor G., Spudich G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011, bar030.

Kirin M., McQuillan R., Franklin C.S., Campbell H., McKeigue P.M. & Wilson J.F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PloS one* 5, e13996.

Larson G., Piperno D.R., Allaby R.G., Purugganan M.D., Andersson L., Arroyo-Kalin M., et al. (2014). Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences of the United States of America* 111, 6139-6146.

Lemay D.G., Lynn D.J., Martin W.F., Neville M.C., Casey T.M., Rincon G., et al. (2009). The bovine lactation genome: insights into the evolution of mammalian milk. *Genome biology* 10, R43-2009-10-4-r43. Epub 2009 Apr 24.

Lewontin R.C. & Krakauer J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74, 175-195.

Li M.H., Iso-Touru T., Lauren H. & Kantanen J. (2010). A microsatellite-based analysis for the detection of selection on BTA1 and BTA20 in northern Eurasian cattle (Bos taurus) populations. *Genetics, selection, evolution* 42, 32.

Li M.H. & Kantanen J. (2010). Genetic structure of Eurasian cattle (Bos taurus) based on microsatellites: clarification for their breed classification. *Animal Genetics* 41, 150-158.

Li M.H., Tapio I., Vilkki J., Ivanova Z., Kiselyova T., Marzanov N., et al. (2007). The genetic structure of cattle populations (Bos taurus) in northern Eurasia and the neighbouring Near Eastern regions: implications for breeding strategies and conservation. *Molecular ecology* 16, 3839-3853.

Li W., Sartelet A., Tamma N., Coppieters W., Georges M. & Charlier C. (2016). Reverse genetic screen for loss-of-function mutations uncovers a frameshifting deletion in the melanophilin gene accountable for a distinctive coat color in Belgian Blue cattle. *Animal Genetics* 47, 110-113.

Librado P., Der Sarkissian C., Ermini L., Schubert M., Jonsson H., Albrechtsen A., et al. (2015). Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proceedings of the National Academy of Sciences of the United States of America* 112, E6889-97.

Loftus R.T., Ertugrul O., Harba A.H., El-Barody M.A., MacHugh D.E., Park S.D., et al. (1999). A microsatellite survey of cattle from a centre of origin: the Near East. *Molecular ecology* 8, 2015-2022.

Lv F.H., Agha S., Kantanen J., Colli L., Stucki S., Kijas J.W., et al. (2014). Adaptations to climate-mediated selective pressures in sheep. *Molecular biology and evolution* 31, 3324-3343.

Machaty Z., Peippo J. & Peter A. (2012). Production and manipulation of bovine embryos: techniques and terminology. *Theriogenology* 78, 937-950.

MacLeod I.M., Bowman P.J., Vander Jagt C.J., Haile-Mariam M., Kemper K.E., Chamberlain A.J., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC genomics* 17, 144-016-2443-6.

Mastrangelo S., Tolone M., Di Gerlando R., Fontanesi L., Sardina M.T. & Portolano B. (2016). Genomic inbreeding estimation in small populations: evaluation of runs of homozygosity in three local dairy cattle breeds. *Animal* 10, 746-754.

Matukumalli L.K., Lawley C.T., Schnabel R.D., Taylor J.F., Allan M.F., Heaton M.P., et al. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PloS one* 4, e5350.

McClure M., Sonstegard T., Wiggans G. & Van Tassell C.P. (2012). Imputation of microsatellite alleles from dense SNP genotypes for parental verification. *Frontiers in genetics* 3, 140.

McDonald J.H. & Kreitman M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351, 652-654.

McLaren W., Pritchard B., Rios D., Chen Y., Flicek P. & Cunningham F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26, 2069-2070.

Meszaros G., Boison S.A., Perez O'Brien A.M., Ferencakovic M., Curik I., Da Silva M.V., et al. (2015). Genomic analysis for managing small and endangered populations: a case study in Tyrol Grey cattle. *Frontiers in genetics* 6, 173.

Metzger J., Karwath M., Tonda R., Beltran S., Agueda L., Gut M., et al. (2015). Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC genomics* 16, 764-015-1977-3.

Meuwissen T., Hayes B. & Goddard M. (2016). Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers* 6.

Meuwissen T., Hayes B. & Goddard M. (2013). Accelerating improvement of livestock with genomic selection. *Annual review of animal biosciences* 1, 221-237.

Meuwissen T.H., Hayes B.J. & Goddard M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819-1829.

Miller S.A., Dykes D.D. & Polesky H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids research* 16, 1215.

Minozzi G., Nicolazzi E.L., Stella A., Biffani S., Negrini R., Lazzari B., et al. (2013). Genome wide analysis of fertility and production traits in Italian Holstein cattle. *PloS one* 8, e80219.

Natural Resources Institute Finland. Milk and Milk Products Statistics, http://stat.luke.fi/en/milk-and-milk-product-statistics (Accessed 1/2016)

Nei M. (1987) Molecular Evolutionary Genetics. Columbia University Press, New York.

Ng P.C. & Henikoff S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids researc*h 31, 3812-3814.

Nielsen R., Williamson S., Kim Y., Hubisz M.J., Clark A.G. & Bustamante C. (2005). Genomic scans for selective sweeps using SNP data. *Genome researc*h 15, 1566-1575.

Noyes H., Brass A., Obara I., Anderson S., Archibald A.L., Bradley D.G., et al. (2011). Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection. *Proceedings of the National Academy of Sciences of the United States of Americ*a 108, 9304-9309.

Ojeda A., Huang L.S., Ren J., Angiolillo A., Cho I.C., Soto H., et al. (2008). Selection in the making: a worldwide survey of haplotypic diversity around a causative mutation in porcine IGF2. *Genetic*s 178, 1639-1652.

Ojeda A., Rozas J., Folch J.M. & Perez-Enciso M. (2006). Unexpected high polymorphism at the FABP4 gene unveils a complex history for pig populations. *Genetic*s 174, 2119-2127.

Orozco-terWengel P., Barbato M., Nicolazzi E., Biscarini F., Milanesi M., Davies W., et al. (2015). Revisiting demographic processes in cattle with genome-wide population genetic analysis. *Frontiers in genetic*s 6, 191.

Pan D., Zhang S., Jiang J., Jiang L., Zhang Q. & Liu J. (2013). Genome-wide detection of selective signature in chinese holstein. *PloS on*e 8, e60440.

Pausch H., Venhoranta H., Wurmser C., Hakala K., Iso-Touru T., Sironen A., et al. (2016). A frameshift mutation in ARMC3 is associated with a tail stump sperm defect in Swedish Red (Bos taurus) cattle. *BMC genetic*s 17, 49-016-0356-7.

Pausch H., Wurmser C., Reinhardt F., Emmerling R. & Fries R. (2015). Short communication: Validation of 4 candidate causative trait variants in 2 cattle breeds using targeted sequence imputation. *Journal of dairy scienc*e, 4162-4167.

Pintus E., Sorbolini S., Albera A., Gaspa G., Dimauro C., Steri R., et al. (2014). Use of locally weighted scatterplot smoothing (LOWESS) regression to study selection signatures in Piedmontese and Italian Brown cattle breeds. *Animal Genetic*s 45, 1-11.

Pritchard J.K., Stephens M. & Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetic*s 155, 945-959.

Qanbari S., Pausch H., Jansen S., Somel M., Strom T.M., Fries R., et al. (2014). Classic selective sweeps revealed by massive sequencing in cattle. *PLoS genetic*s 10, e1004148.

Qanbari S. & Simianer H. (2014). Mapping signatures of positive selection in the genome of livestock. *Livestock Scienc*e 166, 133-143.

Qanbari S., Pimentel E.C., Tetens J., Thaller G., Lichtner P., Sharifi A.R., et al. (2009). A genome-wide scan for signatures of recent selection in Holstein cattle. *Animal Genetic*s. 41, 377-389.

Ramensky V., Bork P. & Sunyaev S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic acids researc*h 30, 3894-3900.

Ramey H.R., Decker J.E., McKay S.D., Rolf M.M., Schnabel R.D. & Taylor J.F. (2013). Detection of selective sweeps in cattle using genome-wide SNP data. *BMC genomic*s 14, 382-2164-14-382.

Ramljak J., Ivankovic A., Veit-Kensch C.E., Forster M. & Medugorac I. (2011). Analysis of genetic and cultural conservation value of three indigenous Croatian cattle breeds in a local and global context. *Journal of animal breeding and genetics* 128, 73-84.

Randhawa I.A., Khatkar M.S., Thomson P.C. & Raadsma H.W. (2016). A Meta-Assembly of Selection Signatures in Cattle. *PloS on*e 11, e0153013.

Rauw W.M. & Gomez-Raya L. (2015). Genotype by environment interaction and breeding for robustness in livestock. *Frontiers in genetic*s 6, 310.

Remington D.L. (2015). Alleles vs. mutations: Understanding the evolution of genetic architecture requires a molecular perspective on allelic origins. *Evolution* 69, 3025-3038.

Rempel L.A., Nonneman D.J., Wise T.H., Erkens T., Peelman L.J. & Rohrer G.A. (2010). Association analyses of candidate single nucleotide polymorphisms on reproductive traits in swine. *Journal of animal scienc*e 88, 1-15.

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

Rosa G.J.M. (2015) Basic Genetic Model for quantitative traits. In: *Molecular and Quatitative Animal Genetics*. (ed. by H. Khatib), pp. 33-42. John Wiley & Sons, Inc., Hoboken, New Jersey.

Rothammer S., Seichter D., Forster M. & Medugorac I. (2013). A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. *BMC genomic*s 14, 908-2164-14-908.

Sabeti P.C., Varilly P., Fry B., Lohmueller J., Hostetter E., Cotsapas C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Natur*e 449, 913-918.

Sabeti P.C., Reich D.E., Higgins J.M., Levine H.Z., Richter D.J., Schaffner S.F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Natur*e 419, 832-837.

Sanger F., Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes C.A., et al. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Natur*e 265, 687-695.

Sartelet A., Klingbeil P., Franklin C.K., Fasquelle C., Geron S., Isacke C.M., et al. (2012). Allelic heterogeneity of Crooked Tail Syndrome: result of balancing selection? *Animal Genetic*s 43, 604-607.

Shahid M.Q., Reneau J.K., Chester-Jones H., Chebel R.C. & Endres M.I. (2015). Cow- and herd-level risk factors for on-farm mortality in Midwest US dairy herds. *Journal of dairy scienc*e 98, 4401-4413.

Simianer H., Szyda J., Ramon G. & Lien S. (1997). Evidence for individual and between-family variability of the recombination rate in cattle. *Mammalian genome* 8, 830-835.

Simonsen K.L., Churchill G.A. & Aquadro C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetic*s 141, 413-429.

Sironen A., Fischer D., Laiho A., Gyenesei A. & Vilkki J. (2014). A recent L1 insertion within SPEF2 gene is associated with changes in PRLR expression in sow reproductive organs. *Animal Genetic*s 45, 500-507.

Sironen A., Uimari P., Iso-Touru T. & Vilkki J. (2012). L1 insertion within SPEF2 gene is associated with increased litter size in the Finnish Yorkshire population. *Journal of animal breeding and genetics* 129, 92-97.

Slatkin M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetic*s 9, 477-485.

Sodeland M., Kent M., Hayes B.J., Grove H. & Lien S. (2011). Recent and historical recombination in the admixed Norwegian Red cattle breed. *BMC genomic*s 12, 33-2164-12-33.

Soller M. (2015). If a bull were a cow, how much milk would he give? *Annual review of animal bioscience*s 3, 1-17.

Sorbolini S., Marras G., Gaspa G., Dimauro C., Cellesi M., Valentini A., et al. (2015). Detection of selection signatures in Piemontese and Marchigiana cattle, two breeds with similar production aptitudes but different selection histories. *Genetics, selection, evolution* 47, 52-015-0128-2.

Spurgin L.G. & Richardson D.S. (2010). How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings. Biological sciences / The Royal Societ*y 277, 979-988.

Stella A., Ajmone-Marsan P., Lazzari B. & Boettcher P. (2010). Identification of selection signatures in cattle breeds selected for dairy production. *Genetic*s 185, 1451-1461.

Stephens M. & Donnelly P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* 73, 1162-1169.

Stephens M., Smith N.J. & Donnelly P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* 68, 978-989.

Strucken E.M., Laurenson Y.C. & Brockmann G.A. (2015). Go with the flow-biology and genetics of the lactation cycle. *Frontiers in genetics* 6, 118.

Szpiech Z.A., Xu J., Pemberton T.J., Peng W., Zollner S., Rosenberg N.A., et al. (2013). Long runs of homozygosity are enriched for deleterious variation. *American Journal of Human Genetics* 93, 90-102.

Szyda J., Liu Z., Reinhardt F. & Reents R. (2005). Estimation of quantitative trait loci parameters for milk production traits in German Holstein dairy cattle population. *Journal of dairy science* 88, 356-367.

Tajima F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585-595.

Takeshima S.N., Miyasaka T., Polat M., Kikuya M., Matsumoto Y., Mingala C.N., et al. (2014). The great diversity of major histocompatibility complex class II genes in Philippine native cattle. *Meta gene* 2, 176-190.

Toro M.A., Meuwissen T.H., Fernandez J., Shaat I. & Maki-Tanila A. (2011). Assessing the genetic diversity in small farm animal populations. *Animal* 5, 1669-1683.

Turchetto-Zolet A.C., Maraschin F.S., de Morais G.L., Cagliari A., Andrade C.M., Margis-Pinheiro M., et al. (2011). Evolutionary view of acyl-CoA diacylglycerol acyltransferase (DGAT), a key enzyme in neutral lipid biosynthesis. *BMC evolutionary biology* 11, 263-2148-11-263.

Utsunomiya Y.T., Perez O'Brien A.M., Sonstegard T.S., Solkner J. & Garcia J.F. (2015). Genomic data as the "hitchhiker's guide" to cattle adaptation: tracking the milestones of past selection in the bovine genome. *Frontiers in genetics* 6, 36.

van Binsbergen R., Bink M.C., Calus M.P., van Eeuwijk F.A., Hayes B.J., Hulsegge I., et al. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genetics, selection, evolution* 46, 41-9686-46-41.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., et al. (2001). The sequence of the human genome. *Science* 291, 1304-1351.

Viitala S., Szyda J., Blott S., Schulman N., Lidauer M., Maki-Tanila A., et al. (2006). The role of the bovine growth hormone receptor and prolactin receptor genes in milk, fat and protein production in Finnish Ayrshire dairy cattle. *Genetics* 173, 2151-2164.

Viitala S.M., Schulman N.F., de Koning D.J., Elo K., Kinos R., Virta A., et al. (2003). Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle. *Journal of dairy science* 86, 1828-1836.

Vitti J.J., Grossman S.R. & Sabeti P.C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics* 47, 97-120.

Voight B.F., Kudaravalli S., Wen X. & Pritchard J.K. (2006). A map of recent positive selection in the human genome. *PLoS biology* 4, e72.

Warinner C., Hendy J., Speller C., Cappellini E., Fischer R., Trachsel C., et al. (2014). Direct evidence of milk consumption from ancient human dental calculus. *Scientific reports* 4, 7104.

Watterson G.A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology* 7, 256-276.

Weber J.L. & May P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* 44, 388-396.

Weigel K. (2015a) Genetic Improvement Program for Dairy Cattle. In: *Molecular and Quantitative Animal Genetics.* (ed. by H. Khatib), pp. 85-96. John Wiley Sons, Inc.

Weigel K. (2015b) Genomic selection, inbreeding and crossbreeding in dairy cattle. In: *Molecular and Quantitative Animal Genetics.* (ed. by H. Khatib). pp. 25-31. John Wiley Sons, Inc.

Weigel K.A. (2001). Controlling Inbreeding in Modern Breeding Programs. *Journal of Dairy Science*, E177-E184.

Weir B.S. & Cockerham C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358-1370.

Weir B.S., Anderson A.D. & Hepler A.B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature reviews. Genetics* 7, 771-780.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-678.

Weller J.I., Kashi Y. & Soller M. (1990). Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of dairy science* 73, 2525-2537.

Wright S. (1938). Size of a population and breeding structure in relation to evolution. *Science* 87, 430.

Wright S. (1931). Evolution in Mendelian Populations. *Genetics* 16, 97-159.

Zeder M.A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proceedings of the National Academy of Sciences of the United States of America* 105, 11597-11604.

Zhang Q., Guldbrandtsen B., Bosse M., Lund M.S. & Sahana G. (2015). Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC genomics* 16, 542-015-1715-x.

Zhao F., McParland S., Kearney F., Du L. & Berry D.P. (2015). Detection of selection signatures in dairy and beef cattle using high-density genomic information. *Genetics, selection, evolution* 47, 49-015-0127-3.

Zhou X. & Stephens M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* 44, 821-824.

8. APPENDIXES

Appendix 1 Frequencies of the *GHR* BOS haplotypes. Breeds included in either study I or II are marked with *. Other breed information originates from the 1000 Bull Genomes Project (unpublished data).

		GHR BOS haplotypes																										F279Y	
breed	**purpose**	**n**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**F**	**Y**
Limousin Holstein		1			0.50		0.50																					0.50	0.50
Salers	Beef, (milk, draft)	1		0.50				0.50																				0.50	0.50
Scotish Highland		2			0.75		0.25																					0.50	0.50
Stabilizer		2		0.25	0.25		0.50																					0.50	0.50
Jersey		66		0.51	0.42	0.02		0.01				0.01				0.02										0.01		0.67	0.33
Hereford Polled		4			0.63		0.25						0.13															0.75	0.25
Simmental x Holstein		5		0.20	0.80																							0.80	0.20
Swedish Red	Dairy	16		0.13	0.47	0.13	0.13	0.03	0.13																			0.81	0.19
Angus	Beef	141		0.17	0.56	0.01	0.17	0.01				0.01	0.01			0.00				0.00	0.00					0.01		0.83	0.17
Hinterwalder		3		0.50	0.17	0.17	0.17																					0.83	0.17
Simmental	Dairy	12		0.67	0.29	0.04																						0.83	0.17
Simmental x Angus Red		3		0.33	0.33	0.17															0.17							0.83	0.17
Holstein	Dairy	450		0.10	0.86	0.01	0.02	0.00	0.01			0.00			0.00	0.00												0.84	0.16
Hereford		37		0.15	0.66		0.18					0.01																0.84	0.16
Holstein x Charolais		18		0.22	0.50		0.06	0.03	0.03						0.03	0.03										0.03		0.86	0.14
Brown Swiss		97		0.27	0.35	0.25	0.11	0.01				0.01										0.01						0.89	0.11
Orig. Braunvieh xBrow nSwiss		18		0.42	0.25	0.25	0.08																					0.89	0.11
Norwegian Red	Dairy	24		0.06	0.60	0.13	0.08		0.02						0.02							0.04						0.90	0.10
Angus Red	Beef	16		0.31	0.25		0.25					0.19																0.91	0.09
Simmental		74		0.48	0.26	0.19	0.01	0.02				0.01					0.01									0.01		0.91	0.09
Normande		24		0.38	0.54	0.02	0.02	0.02				0.02																0.91	0.09
Beef Booster	Beef	29		0.24	0.40	0.07	0.24	0.02				0.02																0.91	0.09
Finnish Ayrshire	Dairy	25		0.06	0.62	0.30	0.02																					0.92	0.08

		GHR BOS haplotypes																										F279Y	
breed	purpose	n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	F	Y
Guelph Composite		33		0.36	0.33	0.05	0.17	0.02				0.05								0.02								0.92	0.08
Alberta Composite		30		0.22	0.38	0.02	0.28	0.02				0.02		0.02		0.02												0.94	0.06
Charolais	Beef	39		0.35	0.39	0.05	0.06					0.04				0.03										0.01		0.94	0.06
Marchigiana		8		0.56	0.19	0.19																						0.94	0.06
Orig. Braunvieh		8		0.19	0.50	0.25	0.06																					0.94	0.06
Fleckvieh	Dual	145		0.38	0.25	0.27	0.02	0.01				0.05			0.00	0.00	0.01					0.00				0.01		0.94	0.06
Simmental x Fleckvieh x Pezzatarossa		43		0.49	0.16	0.29						0.04				0.01												0.94	0.06
Belgian Blue	Beef	10		0.15	0.35	0.15	0.30					0.05																0.95	0.05
Guernsey		20		0.18	0.55	0.25																						0.98	0.03
Montbeliarde		28		0.54	0.30	0.11	0.02									0.02	0.02											0.98	0.02
Gelbvieh	Beef	41		0.55	0.18	0.17	0.05					0.01				0.02	0.01											0.99	0.01
Danish Red	Dairy	44		0.10	0.64	0.07	0.15		0.01					0.01			0.01											0.99	0.01
Angler		5		0.20	0.60	0.10	0.10																					1.00	0.00
Belted Galloway	Beef	1					1.00																					1.00	0.00
Eringer	beef	2		0.25	0.50																					0.25		1.00	0.00
Galloway	"Very ancient", beef	1			1.00																							1.00	0.00
Gelbvieh x Limousine	Beef	1				0.50						0.50																1.00	0.00
Jersey x Limousine		2		0.50	0.25	0.25																						1.00	0.00
Limousin		31		0.31	0.42	0.02	0.10					0.03	0.02		0.02											0.08		1.00	0.00
Piedmontese		5		0.70	0.20																					0.10		1.00	0.00
Piedmontese x Normande		1			1.00																							1.00	0.00
Romagnola	Draft/beef	2		1.00																								1.00	0.00
Simmental x Angus		1		0.50	0.50																							1.00	0.00
Tyrolean Grey		2		0.75	0.25																							1.00	0.00
Vorderwalder		3		0.33	0.17	0.33			0.17																			1.00	0.00
Angus*	Beef	13		0.19	0.65		0.15																					0.86	0.14
Barka*	Dairy-beef	16	0.50	0.25				0.16		0.06																	0.03	1.00	0.00
Belorussian Red*	Dairy	14		0.29	0.43			0.07					0.07						0.11							0.04		0.8¥	0.2¥
Bestuzhev*	Dairy	11		0.41	0.41		0.18																					NA	

		GHR BOS haplotypes																										F279Y	
breed	**purpose**	**n**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**F**	**Y**
Busha*	Dairy-beef-draft	13		0.35	0.38	0.19											0.08											1¥	0¥
Charolais*	Beef	15		0.33	0.40	0.13							0.07											0.03	0.03			0.96	0.04
Danish Jersey*	Dairy	18		0.81	0.14										0.06													0.97	0.03
Eastern Finncattle*	Dairy	14		0.29	0.54	0.04						0.07			0.04						0.04							0.97	0.03
Finnish Ayrshire*	Dairy	19		0.16	0.61	0.11			0.05			0.05				0.03												0.88	0.13
Fogera*	Dairy-beef	17	0.32	0.47				0.09		0.06		0.06																1.00	0.00
Hereford*	Beef	11		0.55	0.32		0.14																					0.45	0.55
Holstein Friesian*	Dairy	19			0.66		0.16	0.03				0.05	0.03	0.05						0.03								0.93	0.08
Jutland cattle*	Dairy	18		0.03	0.58		0.25							0.06				0.06		0.03								0.90	0.10
Kholmogor*	Dairy	19		0.26	0.42	0.24			0.05			0.03																0.98¥	0.02¥
Northern Finncattle*	Dairy	19		0.37	0.42	0.05			0.05								0.05				0.03		0.03					1.00	0.00
Pechora*	Dairy	17		0.26	0.32	0.24	0.06		0.06			0.03					0.03											NA	
Podolian cattle*	Draft	18		0.89	0.03	0.06						0.03																1¥	0¥
Raya*	Dairy-beef	14	0.32	0.32				0.14		0.07	0.04	0.07															0.04	1.00	0.00
Red Danish*		13			0.62		0.35																			0.04		1.00	0.00
Ukrainian Grey*	Beef-draft	20		0.23	0.75	0.03																						1¥	0¥
Western Finncattle*	Dairy	20		0.18	0.35	0.35										0.08						0.05						0.97	0.03
Yakutian cattle*	Dairy-beef-draft	16		0.84		0.06	0.09																					1¥	0¥
Yarovslavskaya*	Dairy	18		0.28	0.44	0.17	0.03							0.03	0.03		0.03											0.87¥	0.13¥

¥ data not included in studies I or II (unpublished)

Appendix 2. Frequencies of *PRLR* haplotypes within 15 cattle breeds.

Breed	n	origin	BOS_PRLR1	BOS_PRLR2	BOS_PRLR3	BOS_PRLR4	BOS_PRLR5	BOS_PRLR6	BOS_PRLR7	BOS_PRLR8	BOS_PRLR9	BOS_PRLR10	BOS_PRLR11	BOS_PRLR12	BOS_PRLR13	BOS_PRLR14
Finnish Ayrshire	17	Taurus	0.88	0.06	0.00			0.06								
Holstein-Friesian	20	Taurus	0.73	0.20	0.03			0.03								0.03
Wetern Finncattle	20	Taurus	0.75	0.18	0.00					0.08						
Northern Finncattle	19	Taurus	0.74	0.26												
Eastern Finncattle	19	Taurus	0.79	0.13				0.05							0.03	
Kholmogor	20	Taurus	0.78		0.23											
Busha	14	Taurus	0.82	0.11	0.07											
Podolian cattle	17	Taurus	0.97		0.03											
Yakutian cattle	20	Taurus	0.80	0.20												
Ukrainian grey	20	Taurus	0.38	0.25	0.15		0.05	0.03	0.15							
Bestuzhev	14	Taurus	0.54	0.18			0.29									
Belarussian Red	16	Taurus	0.63	0.28	0.06					0.03						
Barka	6	Indicus	0.58	0.17		0.25										
Raya	10	Indicus	0.35	0.20		0.25					0.15			0.05		
Fogera	6	Indicus	0.08	0.25		0.42					0.00	0.17	0.08			