THE UNIVERSITY of EDINBURGH
School of Biological Sciences

GENOME WIDE ASSOCIATION STUDIES OF MILK PRODUCTION TRAITS AND CLINICAL MASTITIS IN UK DAIRY GOATS

Mehmet KIZILASLAN

In partial fulfilment of the requirement for the Degree of
Master of Science in Animal Breeding and Genetics
at the University of Edinburgh
2015/2016

Supervisors: Dr. Joanne Conington and Dr. Sebastian Mucha

ACKNOWLEDGEMENTS

I would like to thank my supervisors Jo and Sebastian,for providing me with this opportunity of working in this project and for supporting me through all the process.

I would also like to thank the people of SRUC for such a friendly environment I have experienced throughout the course.Finally, I would like to thank my family, friends and especially Patri, whose support has always been with me.

FULL NAME: Mehmet KIZILASLAN

Matriculation Number: s1465247

Supervisor's Name: Dr. Joanne Conington and Dr. Sebastian Mucha

Name of programme: MSc Animal Breeding and Genetics (QGGA)

Submission Date: 18.08.2016 No of Pages:58 Word count:14,126

I have read and understood the University of Edinburgh guidelines on plagiarism and declare that this written dissertation is all my own work, except where I indicate otherwise by proper use of quotes and references. I confirm I have:

- Clearly referenced/listed all sources as appropriate ☐
- Referenced and put in inverted commas all quoted text of more than three words (from books, web, etc) ☐
- Given the sources of all pictures, data etc. that are not my own ☐
- Not made any use of the essay(s) of any other student(s) either past or present ☐
- Not sought or used the help of any external professional agencies for the work ☐
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources) ☐

I understand that any false claim for this work will be penalised in accordance with the University regulations.

Signature:.....

Date:.....

CONTENTS

Abstract

Animal health, welfare and milk production are strongly related aspects of dairy industry of almost all countries in the world. Therefore, factors affecting animal welfare and milk production can both be a big threat or a great benefit for animals, farmers, sellers, consumers and finally the economy of a country. Due to their fitness capacities and other particular driving forces, such as climate change, goats are considered to be a potential food source for feeding the world in the future. The goat genome was published in 2013 and the availability of Illumina caprine SNP-chip in 2014 has led to several new studies on goats including the data used for this study.

The objective of this study was to discover genomic variants that are hypothetically associated with milk production traits and also clinical mastitis, which is a major disease of lactating mammals. With this purpose, Genome-wide association studies (GWAS) using aprroximately ~7500 animals for milk production traits and ~3900 animals for clinical masititis were conducted with a mixed-breed dairy goat population. Various number of genome- and chromosome-wide SNPs were found. These were ALOX12, ASGR2, CSN1S1 (including CSN2, CSN1S2), PDE4B, DNAH2, KIF1C and EFTUD1 along with several others. Suggested pathways and causative genes are reported and compared with results from other species.

To our knowledge this was the first GWAS study on clinical mastitis of goats as well as being the most powerful study for goat milk production traits in the UK, regarding the sample size.

1. INTRODUCTION

1.1. Global position of goat farming

Animal health, welfare and milk production are strongly related aspects of dairy industry of almost all countries in the world. Therefore, factors affecting animal welfare and milk production can both be a big threat or a great benefit for animals, farmers, sellers, consumers and finally the economy of a country.

The projected increase in the world's population and competitive marketplace for dairy products puts pressure on the sector to obtain more milk and dairy products with less labour and expense. The estimated world population for 2005 was 6.5 billion and the expectation for 2050 is for 9.1 billion people (UN, 2004). World milk production has increased by 50 percent in the past three decades, from 482 billion kilograms in 1982 to 14,649 billion litres in 2014) (UK House of Commons Library, 2016). The growing demand for milk and meat may also increase the price of feed such as grains and maize. Apart from financial concerns, increasing global awareness towards climate changeand environmentalist concerns (Steinfield et al., 2006) are expected to promote the sector towards less consumption of clean water and other resources, as well as encouraging decreased occupation of land for livestock production. On top of it, social pressure to improve animal welfare and health may also encourage farmers to maximise production per animal with providing an acceptable level of animal welfare, health and relief for "the five freedoms", without incurring additional increasing expense of feed, building and veterinary costs (FAO 2013). These pressures are the drivers for more efficient and healthier livestock systems.

The world goat population has increased by more than 100 % in the last three decades reaching 996 million (FAOSTAT, 2014) and they supply around 2 per cent of the world total annual milk. Most goat breeds are under close surveillance of food security programmes such as "Feed the Future" initiative, since they are seen as a potential food source for the future, due to their adaptive capabilities and abovementioned driving forces (United States Department of Agriculture, USA). Most of the domestic population and milk production is in developing countries

(Chetroiu *et al.* 2013) and 15 percent of world total milk production is provided by Europe, despite having only 3 per cent of the world's goat population and being mainly used for cheese production. European countries with the greatest share of production are Spain, France and Greece. In the UK, goat population is approximately around 100,000 heads (DEFRA 2014), of which roughly 33,000 are dairy goats. In France, industry-wide breeding programmes are carried out (Montaldo and Manfredi, 2002) and recently genomic selection has become a tool of use in the breeding programmes of dairy goats (Carillier et al., 2013). In the UK, genomic selection was implemented for the first time in (Mucha et al., 2015).

Increasing global importance of both goat breeding and enhanced productivity whilst maintaining animal fitness, which includes aspects of animal health, promotes researchers to implement new technologies, such as Genome-Wide Association Studies (GWAS), on goat breeds. Ideally, the aim is to delve into inheritance of complex traits such as production and disease resistance, which are believed to show characteristics of polygenic inheritance. Utilising such technology on traits that are typically difficult or expensive to measure, such as mastitis; and traits holding a great share in the national markets, e.g. milk production traits, gives insight into the faster progress in genetic selection of traits of interest. Despite the fact that the genomic selection per se holds a great potential for breeding strategies, identification of quantitative trait loci could possibly help to deal with some limitations imposed by cost-effectiveness such as the need for genotyping large number of animals and repeating genotyping for after several generations (Taylor, 2012).

1.2. Mastitis

Masititis is one of the most prevalent dairy animal disorders, together with lameness and fertility problems, across almost all countries in the world. It is the inflammatory immune response of udder tissue to particular pathogens. It can be painful and if left untreated, may lead to death. It is an important disease and a major cause for reduction in milk yield, quality and ultimately final product quality (e.g., cheese, cream, butter etc.) (Leitner et al*.,* 2004; 2007). Overall mastitis prevalence in goats has been reported to be 36 %, 33 %, 18% respectively in the UK, France and Spain (Manser, 1986; Contreras et al., 1995; Poutrel and Lerondelle, 1983). Annual

incidence of clinical mastitis only, is usually lower than 5 % in small ruminants (Bergonier et al., 2003)

Clinical cases of mastitis are characterised as pain and swelling in the udder, tiredness and fatigue as a result of immune reaction, changes in the milk composition and the appearance. Subclinical mastitis is not readily detected by eye, so Somatic Cell Count is normally used as the phenotype in animal breeding programmes, as a proxy trait for mastitis in the dairy sector. Although results from international studies suggest wide variation on pathogens, the main causative bacteria are Gram-negative enterobacteria as well as Gram-positive streptococci and staphylococci. *Streptococcus agalactiae*, *Streptococcus dysgalactiae, Streptococcus uberis*, *Staphylococcus aureus*, and *Escherichia coli* are the most commonly observed microorganisms that are associated with the presence of condition (Zadoks and Fitzpatrick, 2009). These agents are also highly contagious, therefore detecting diseased animals promptly is of vital importance. Marker-assisted or genome-wide selection (GWS) offers a potential to detect animals that are more prone to disease, or that are more resistant to the condition. Both enable desired animals to be selected for breeding at a much earlier age and without waiting for the onset of disease. Identifying regions in the genome that are correlated with the existence or severity of the condition, gives us a possible gateway to step into genome based early selection for the trait. Current genomic technology and its potential to accelerate genetic improvement allows us to improve molecular markers for screening the allelic phase of a particular region in the genome to measure covariance between traits and their genetic predisposing factors. However, to be able to utilize such genomic information the initial is to have a clearly defined and precisely measured phenotypic observation of the trait.

Since early detection of mastitis and monitoring presence are of vital importance for the sustainability of flock productivity, various methods have been improved and integrated to breeding systems. Currently, somatic cell count (SCC) is the most widely used method in dairy cattle breeding programmes. The log transformed SCC shows a genetic correlation of nearly 0.70 with clinical mastitis (Rupp and Boichard, 2003) and heritability of 0.15 in dairy cattle (Mrode and Swanson 1996). SCC has been used in a wide number of quantitative trait loci (QTL) mapping studies

examining genetic foundations of resistance to mastitis. However, there are studies signifying that the correlation between milk yield and SCC is not reliable for using it alone as an indicator of reduction in milk yield due to mastitis in dairy goats (Koop et al., 2010). Furthermore, goat milk normally has higher average somatic cell score even in the case of healthy mammary gland tissue and it is strongly affected by other factors such as parity and lactation stage (Paape and Capuco 1997). California Mastitis test (CMT) and electrical conductivity measurements are another techniques used for early detection (Contreras et al*.,* 1996; Kitchen, 1981). Electrical conductivity measurement systems can be automated and used at each milking occasion (Maatje et al., 1983). Automation and on-line integration of this system allows early detection of mastitis and therefore leads to lower costs associated with mastitis by early detection of affected animals. Studies show that 77 % sensitivity and 100% specificity can be obtained for detection of mastitis (Nielen et al., 1995). Therefore currently direct observations of clinical mastitis and conductivity results can be conceived as reliable indicators of mastitis existence in dairy goats.

1.3. Genetic parameter estimates

Several studies have estimated genetic parameters for milk production traits (such as milk yield, protein yield, fat yield and lactose yield) and for mastitis traits (such as clinical mastitis (binary), SCC and EC) in a variety of goat populations. A study has been carried out by (Mucha et al*.*, 2014) to estimate genetic parameters of milk yield in a mixed-breed (Alpine, Saanen, Toggenburg) dairy goat population with the pedigree of 28,184 individuals. The highest heritability (0.45) estimated at 200 and 250 days in milk (DIM) in the first lactation, whereas it is between 0.34-0.25 in the subsequent three lactations. It is observed that after 300 days in milk heritability decreases to 0.23 and 0.10 at 400 days in milk for the first and following lactations, respectively. Heritability of milk yield (MY), protein yield (PY) and fat yield (FY) were estimated using animal models via pedigree of 33,725 US dairy goats, which were obtained as 0.37, 0.37, 0.38 respectively (Castañeda-Bustoset al., 2014). Estimation of genetic correlations between aforementioned milk production traits were done via 783 Alpine and Saanen goats, resulting in below table :

Table 1. Genetic correlations between milk production traits (Brito et al., 2011)

Traits	Milk yield	Protein yield	Fat yield	Lactose yield
Milk yield	0.19	0.96	0.86	0.98
Protein yield	--	0.12	0.93	0.94
Fat yield	--	--	0.10	0.88
Lactose yield	--	--	--	0.15

Milk yield (Milk yield at 270d)

Numbers on the diagonal represent heritability, genetic correlations above the diagonal.

Estimated genetic correlations among milk production traits are high, which may suggest similar genetic background for the traits.

Estimates of heritability in dairy cattle for clinical mastitis as a binary observed trait are quite low, being 0.02-0.04 with linear models and 0.07-0.10 with threshold models.The heritability of SCC has been estimated as being 0.20 and 0.24 for 67,882 Alpine and 49,709 Saanen goats respectively (Rupp *et al.*, 2011).

In another study conducted on sheep data, genetic correlation between mastitis status and milk yield, was estimated to be 0.59, demonstrates also the antagonistic relationship between udder health and milk yield (Tolone et al., 2013).

1.4. Genomics in goats

Currently the processing of sequence data with observations from large number of animals is computationally demanding. Using genetic markers to potentially uncover significant molecular regions of interest for subsequent sequencing to detect the causative mutations is a useful strategy to target specific genomic regions.

The relative importance of those markers is determined by the amount of variance explained by them and cost-effectiveness. Currently, Single Nucleotide Polymorphisms (SNPs) are the most convenient molecular markers for tracking and interpreting sequence information. The recent advance of high throughput genomic technologies and computational statistics has allowed rapid and multiple detection of

those SNP's and therefore provided a means of genotyping those molecular markers in a genome–wide scale. This has facilitated tracking and evaluation of quantitative traits and complex diseases segregating within populations via genome-wide association studies (GWAS). Moreover, it allows genomic selection to be a tool of use for breeding aims in the livestock industry.

Exploitation of genomic data in goats has progressed slowly compared to dairy cattle due to the absence of proper molecular, statistical tools and lack of investment. Recently, the International Goat Genome Consortium has led to the development of a 50K SNP-chip via using 97 animals of six meat and dairy goat breeds (Alpine, Boer, Creole, Katjang, Saanen and Savanna) and it has been validated also for breeds Angora and Skopelos (Tosser-Klopp et al., 2014). Manufactured byIllumina, 50K Caprine SNP-chip, with evenly distributed SNP's across whole genome, facilitates genome-wide association studies in goats, leading to the insight of traits such as disease resistance, milk yield and production, fertility and eventually it lightens the path of genetic selection on goats, especially of interest, dairy goats.

1.5. Genome wide association studies

Genome-wide association is currently the most widely used statistical method of relating phenotypic observations to the correlated regions in the genome. Linkage disequilibrium phenomenon and genome-widedistributed markers are employed for detecting the association between possible genetic variants and phenotypic observation of a particular trait.

The main assumption behind the LD based mapping is that the marker or markers that appear to be statistically significant for this 'co-variation' analysis are closely linked to the QTL affecting the trait, with a possibility of being directly causal variant. The resolution achievable via making use of LD and a high density genome-wide dispersed SNP-marker panel is < 0.1 cM (~ 100 kb) (Rannala and Reeve 2001), which reduces the candidate region of interest for further analysis and implementation (such as positional identification of effective gene regions via molecular techniques, gene editing etc.) comparing to the linkage analysis approach. As well as giving higher resolution, it is also more powerful than the

traditional linkage mapping approach in regards of detecting small QTL effects (Risch and Merikangas 1996). The main factors affecting the statistical power ofassociation mapping are sample size, effect size of the putative QTL (i.e., increasedpower with bigger effect size), the difference between allele frequencies of QTL and marker allele in LD (i.e., smaller the difference higher the statistical power) and overall genome-wide LD, since low LD needs denser marker spacing for detection. It is of vital importance to note that a detected "association" does not necessarily indicate causality. Apart from detected marker or markers being causally related to the trait, significance may be due to strong LD between marker locus and the causal variant. On the other hand, the results could be still spurious due to confounding factors such as population stratification and cryptic relatedness. Furthermore, confounding can be sourced by genotyping error rates varying among phenotypic observations, since a statistical bias resembling to population structure can be observed (Clayton et al., 2005). Both population stratification and cryptic relatedness, as a whole, can be perceived as the two ends of the same spectrum: the unobserved pedigree that unveils true relationships among all subjects of the study (Astle and Balding, 2009). As a result of implementing association analysis with populations showing any of these characteristics, the main assumption of independence of test statistics for each marker locus is compromised. Therefore, an inflation of test statistics occurs and high false positive rate becomes inevitable. Inflation due to population structure and relatedness is quantifiable and it isreferred as genomic inflation factor (λ-lambda) (Devlin and Roeder, 1999). In the existence of inflation, the test statistics are distributed as λ. $\boldsymbol{X}^2$with one degree of freedom.

To overcome this problem several methods have been proposed. Genomic Control (Devlin and Roeder,1999), structured association (Pritchardet al*.,* 2000), EIGENSTRAT method (Price et al*.*, 2006) and linear mixed models (LMM) (Yu et al., *2006)* are the well-known methods. Although all these methods have proven quite useful, all have their limitations. Genomic control adjusts test statistics of all loci with the same inflation factor (λ), regardless of the degree of differentiation between allele frequencies of sub-populations among different markers. Since some loci may be differentiated strongly, this method fails to be robust enough for certain conditions. Structured association is not computationally efficient for allowing co-membership of individuals to different sub-populations whereas EIGENSTRAT

method (principal components analysis based) seems to have solution for both these methods. However, the increased complexity of relationships sets limit to the power of this method. Among the proposed approaches, mixed linear models account for the genetic background effect via a relationship matrix of individuals and therefore deals with structural characteristics of sample data better in the existence of complex population structure and familial relatedness (Henderson, 1984). However, a combination of these methods rather than all per se is possible and in the case of strong cryptic relatedness and population stratification might be the best option to deal with inflation.

Two problems arise related to the usage of mixed models in GWA analysis: excessive computational complexity due to maximum likelihood estimations for mixed model with SNPs fitted fixed effects (Aulchenko et al., 2007a) and incorrect or missing pedigree information. Classical pedigree-based mixed model approach proves less efficient estimating relative kinship between individuals when pedigrees have missing or inaccurate information. Furthermore, even in the case of complete pedigree, estimated kinship coefficients are inherently only the 'expectations' of real relationships.The percentage of relatives' alleles inherited by individuals is not traceable with information from only pedigree. However, since for genome-wide association studies genotypes at thousands of loci would already be obtained, it is therefore possible to estimate the kinship between individuals from marker data. Twice the kinship coefficients obtained from marker data represent the 'realised' coefficients of relationship between individuals since the kinship estimates capture Mendelian sampling terms.

The two step method proposed by (Chen and Abecasis, 2007), which can be named as Family Based Score Test Approximation (FASTA), with the genomic kinship matrix calculated from marker information appears to be managing both problems simultaneously. Furthermore, access to the method and ease of use is relatively simple since it is implemented by GenABEL software package on R environment. Another advantage of the referred method is that it allows a "genotype probability score" to be calculated from flanking markers for the loci with missing genotypes among study subjects. This feature has increasingly great importance since it is known that statistical power increases with available sample size and subjects with missing genotype data is ignored by computational progress due to matrix

estimation problems. Briefly, the two step approach used by Chen and Abecasis (implemented in GenABEL) is a family-based association test that unlike transmission disequilibrium tests and its extensions (Spielman and Ewens, 1998; Abecasis et al., 2000), instead of focusing on allele transmission from heterozygous parents; its main focus is estimation of ancestry of study subjects from genotype data.

It was demonstrated that different LMM approaches such as those implemented in the packages GenABEL (including FASTA), EMMAX, FAST-LMM,GEMMA and MMM provide a proper and robust approach for family-based genome-wide analysis of both quantitative and binary observations (Eu-ahsunthornwattana et al., 2014). Furthermore, they show promising results for controlling the overall genomic inflation at a convenient level and generally offer higher power than traditional family-based association analysis approaches such as those carried out in FBAT. Similar results are also provided by alternative approaches implemented in other specific software packages such as MASTOR, Mendel, MQLS, ROADTRIPS. It is clear that results propose a concordance between different LMM methods. Therefore choice of method, in a sense, depends on ease of use, access to the software, dealing way of chosen software with large data sets and some other inherent advantages. GenABEL has proved to be efficient for dealing with very large data sets with missing information both on genotypes and phenotypes and its relative ease of use renders it as a handy genomic tool of use (Aulchenko and Karssen, 2015).

In GWA studies, information type of observations (**y**) carries as much importance as genotype information since any biased source of observation could easily lead to spurious associations. As well as phenotypic records; pseudo-phenotypes such as estimated breeding values (EBVs) and daughter yield deviation (VanRaden and Wiggans, 1991) are also commonly used in QTL mapping studies. The need for de-regressing estimated breeding values that are going to be used in QTL discovery and fine-mapping studies was clearly delivered by (Garrick et al., 2009). Since EBVs ($\hat{\boldsymbol{g}}$) are not true genetic merit of animals (**g**) but only estimates, training with EBVs can be considered as training with true breeding values with some prediction error. Inclusion of this error term to the BLUP modelsgives rise to two issues, which are sourced from properties of BLUP estimators (Henderson, 1975). Firstly, a reduction in

total variance takes place due to prediction errors being negatively correlated with **g**, which leads to under-evaluation of superior animals but over-evaluation of low merit animals (Garrick et al., 2009). Reduction in variance in turn affects GWAS studies via increased false positive rates. The second problem is that BLUP, as a shrinkage estimator, shrinks the observations towards the mean regarding the extent of available information (from own performance and/or relatives), which results in varying shrinkage with the reliability (r_i^2)of each EBV($\hat{\boldsymbol{g}}_i$). Therefore, inconsistency arises in the situation where EBVs have different reliabilities as it is in most of studies. As a solution to both of these problemsinflating the EBVs were suggested, which in its simplest form is done via $\hat{\boldsymbol{g}}_i/r_i^2$. A de-regressed EBV covers all available information on an individual and its relatives which is inherently corrected for fitted systematic environmental effects (e.g. herd, season etc.). Its usage in genomic predictions has been proved more reliable as a source of response variable in genomic prediction in comparison with traditional EBV (Ostersen et al., 2011). There are alternative methods of weighting, which proved to be successful for example, in the case of using offspring information (Fikse and Banos, 1997).

1.6. Properties of the Goat Genome

The genome of a goat is spread over 29 autosomes and 1 sex chromosome. De novo sequencing of a domestic goat genome (Capra hircus) provided a reference sequence for further studies of goat populations (Donget al.,2013). A female Yunnan black goat genome sequence has been obtained as ~ 2.66 Gb via utilizing combination of two methods: short-read sequencing and optical mapping with a high-throughput genome sequencer. The same group also annotated 22,175 protein-coding genes from RNA-seq data of 10 different tissues.

The most comprehensive study on the extent of linkage disequilibrium in goat populations have recently been published by (Brito et al., 2015). 9 particular goat breeds from Canada and Australia were genotyped using Illumina 50K Caprine SNP-chip. Average linkage disequilibrium (r^2) at the given distances for different goat breeds are presented below in the table 2:

Table 2. Average LD between adjacent SNPs

Breed	LD (r^2)	Distance (Mb)
Alpine	0.144	0.0530
Saanen	0.153	0.0534
LaMancha	0.194	0.0621
Nubian	0.276	0.0733
Toggenburg	0.248	0.0731
Boer (Canada)	0.287	0.0674
Cashmere	0.182	0.0620
Rangeland	0.110	0.0545
Boer (Australia)	0.289	0.0648

According to the study, the breeds LaMancha, Nubian, Toggenburg and Boer resulted in r^2 values that are useful (r^2> 0.2) for using them with genomic prediction puposes via mixed models.

Another study on the UK dairy goats, showed a mean linkage disequilibrium (r^2) of 0.18 at 50 kb SNP distance (the average distance between markers of 50 K Caprine SNP-chip), 0.14 at 100 kb, 0.09 at 1000 kb and finally 0.07 at 2000 kb respectively (Mucha et al., 2015).

1.7. Mapped Quantitative Trait Loci and GWAS

Goat genomics is a relatively new area of interest and therefore there is not many QTL mapping study conducted on goats comparing to cattle and sheep breeds (Amills, 2014).

To date, there is no GWAS or linkage mapping study on clinical mastitis of goats. However, A number of QTL have been found using SCC as a correlated trait. QTL detection study with 2254 Alpine and Saanen goats has been conducted by (Rupp

et al., 2014). According to the results, 2 genomic regions are found to be significant via linkage mapping methods and 16 regions found to be significant via linkage disequilibrium (GWAS) methods. Among the regions, the ones on chromosome 16, 19, 21 were of strong interest since they were observed in both breeds and had quite low p-values. Region on chromosome 19 was in a homologous region for both goat and sheep. Regarding clinical mastitis there are large amount of studies conducted on cattle genome, however results are condensed on chromosome 5, 9, 14 and 20 (Animal genome QTL database).

The identification of genomic regions associated with milk yield has been carried out with 4563 goats and using multi-locus mixed models (Mucha et al., 2016). Only one SNP on chromosome 19 exceeded genome-wise significance level, whereas SNP's on chromosomes 4, 8, 14, 28, 29 reached chromosome-wise significance level. Candidate gene studies for milk production traits in goats have been focused on goat chromosome 6, especially regions covering caseine coding genes, since apparently milk yield and content is strongly affected by their activities (Amills, 2014).

Results from QTL mapping studies are quite prominent in terms of increasing genetic gain in a targeted population. Accuracy of selection for a particular trait increases with marker assisted selection, whose markers are determined according to detected quantitative trait loci.Furthermore marker assisted disease monitoring provides earliest possible information about flock health.Therefore, possible QTL regions obtained from mapping studies have a wide range of application areas including animal genetic improvement, disease resistance and genetic monitoring of flocks.

1.8. Objective

With the objective of uncovering possible QTL regions, in this study GWAS on certain milk production traits (i.e., milk yield, milk fat yield, milk lactose yield and milk protein yield) and clinical mastitis have been carried out. De-regressed EBVs were used as pseudo-phenotypes and linear mixed model based association analyses were implemented with genotypes obtained from Illumina caprine SNP-chip. Finally,

putative QTL regions were further investigated with NCBI and Ensembl genome browser.

2. MATERIAL and METHODS

2.1. Characterisation of phenotypes and genetic information

2.1.1. Pedigree details, traits and phenotype data

Data were provided by a commercial company that keeps the largest share of the UK dairy goats as well as having an electronic identification system which enhances data recording. The synthetic breed of the data population was formed in 1985 by crossing 3 different goat breeds: Alpine, Saanen, Toggenburg. There was not any particular breeding strategy utilized. In every generation, the best-performing animals were chosen for creating the next generation. The pedigree file contained 42,193 animals of which 478 bucks and 16728 does were recorded to be sires and dams respectively.

Phenotype data had binary observations on clinical mastitis (CM), information on potential clinical mastitis covariates (such as year and season of observation, age of doe at kidding, lactation stage of observation and lactation stage group of observation). Moreover, de-regressed estimated breeding values (de-regressed EBVs) -inherently corrected for systematic environmental effects- for particular milk production traits (milk yield (MY), milk fat yield (MFY), milk protein yield (MPY) and finally milk lactose yield (MLY)) were present. Traits were measured as percentages from milk yield and then multiplied by the total yield to express each traits as yield.

For milk yield, within those genotyped animals 7993 goats (7623 does and 370 bucks), which had one de-regressed EBV per animal, were used in the analyses, whereas for MFY, MPY and MLY number of animals with available observation was 7367 (7003 does and 338 bucks). The number of animals with available phenotypic records was 3980 goats for CM.All observations for CM were recorded for the first lactation of animals. There were records for 140 case and 3840 controls in the analysis.

For association analyses, R statistical analysis environment with several libraries (GenABEL, LDHeatmap, car, cgmisc packages) were utilized (R Development Core Team, 2006).

2.1.2. Calculation of de-regressed breeding values

For the estimation of breeding values of all milk production traits the following model was used:

$$\mathbf{y = Xb + Za + Wp + e},$$

where **y** is the vector of observations (test day measurements in kg); **b** is the vector of fixed effects which are year–season, herd test day, lactation curves - modelled utilizing Legendre polynomials of fourth order (Kirkpatrick et al., 1990)- and age at kidding; **a** is the vector of random animal effects and **p** is the vector of permanent environment effects. Both are 1 x 3 vector of random regression coefficients resulting from Legendre polynomials of second order. Finally **e** is the vector of random residual errors with **X** being matrix relating fixed effects to the observations while **Z** and **W** being matrices relating random animal and permanent environment effects to the observations. All random effects were assumed to be drawn from a normal distribution with zero mean and the following (co)variance structure:

$$Var\begin{bmatrix}\mathbf{a}\\ \boldsymbol{p}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{AxG} & 0 & 0\\ 0 & \boldsymbol{IxP} & 0\\ 0 & 0 & \boldsymbol{I\sigma^2 e}\end{bmatrix}$$

A is the additive relationship matrix, **P** and **G** are 3 x 3 (co)variance matrices for random permanent environment and animal effect. $\boldsymbol{\sigma^2 e}$is the residual variance and **I**is the identity matrix. (Co)variance components were estimated via REML algorithm implemented by ASReml (Gilmour et al., 2009).

De-regression was done via MIX99 software package (Lidauer et al., 2011) with full animal pedigree. Effective offspring contributions (EOC) (Fikse and Banos, 1997) calculated below, were used as weighting factors:

$$EOC_i = \frac{rel_i . kdau}{1 - rel_i}$$

$$kdau = \frac{4-h^2}{h^2},$$

where h^2 = heritability of the trait analysed in the population and rel_i is the reliability of each single EBVs. Obtained de-regressed EBVs were single values per animal which are corrected for systematic environmental effects and inflated via a proper inflation factor due to the reasons mentioned in "Introduction". It was done in a way that removes the bias sourced by the variation in offspring number of animals.

2.1.3. Genotyping and quality control procedure

All animals used in the study were genotyped via Illumina Caprine 50K BeadChip (Illumina Inc., San Diego, CA: Tosser-Klopp et al., 2014) at GeneSeek, Inc., a Neogen Company (Edinburgh, UK). Samples for DNA extraction were collected nasally and vaginally via swabs manufactured by DNA Genotek Inc. (Ottawa, CANADA). SNPs on the chip have an average spacing of 40 kb in between, which corresponds to approximately 25 marker SNPs per mega base pairs (Mb).

Initial genotypes in the analysis had been already checked for certain quality control and assurance thresholds and were 44,907 SNPs with X- related markers being removed according to the thresholds applied by (Mucha et al., 2016). Further quality control (QC) checks were implemented using two rounds with GenABEL R package (Aulchenko et al., 2007). In the first round, the number of SNPs was initially 44,907 base pairs segregating throughout 29 autosomes. QC checks were performed using 'check.marker' function of GenABEL as the following: In the first round, minor allele frequency (MAF) cut-off was set to 0.05 to eliminate rare variants. Call rate threshold for SNPs was 95 % and for overall genotype call rate of animals was 90 % to keep the reliability of genotypes in the analyses relatively higher. Between the two rounds, SNPs were out of Hardy-Weinberg Equilibrium possibly as a result of stratification was investigated using Principal Component Analysis (PCA). Since no clear grouping of animals was observed in the PCA, they were excluded via a threshold corrected with Bonferroni method for multiple testing, which was set as 0.05/M where M is the number of markers that'survived' the first round of the data cleaning process. This was done including this threshold on top of the other

thresholds and applying another QC check, so that all thresholds could be applied at the same time.

2.1.4. Multidimensional scaling analysis

Since unknown breed composition might cause spurious associations and misleading genetic parameter estimations, multidimensional scaling analysis was implemented. Multidimensional scaling, which is a form of principal coordinates analysis (Gower, 1966) and strongly related to the PCA, was performed via 'cmdscale' function in 'stats' package of R environment. Possible population sub-structure was visualised via reducing dimensionality of the data. This was obtained with a 'distance' matrix derived from genomic kinship matrix. Furthermore, with the same function, eigenvectors were obtained and plotted to quantify the variance explained by each vector.

2.1.5. Linkage Disequilibrium Analysis and Visualisation

Pairwise linkage disequilibrium (LD) was measured as squared correlation of alleles at two loci (r^2) with 'r2fast' function of GenABEL (Hao et al., 2006). Formula used was described by (Hill and Robertson, 1968):

$$r^2 = \frac{[f(AB) - f(A)f(B)]^2}{f(A)f(a)f(B)f(b)},$$

where $f(A), f(a), f(B), f(b)$ are frequencies of alleles and $f(AB)$ frequency of haplotype made of common alleles at the loci of interest. LD was measured among all syntenic marker loci (loci only on the same chromosome).

For the chromosomes having genome-wide significant SNPs, LD structures (LD heatmap) of those regions with high number of significant hits were plotted based on squared correlation between pairs of loci (r^2)using 'LDHeatmap' package provided by R environment (Shin et al., 2006).

Furthermore, LD decay examination and further LD structure analysis including plots showing clumping of significant SNPs was carried out with 'cgmisc' package based on R environment (Kierczak et al., 2015).

2.2. Genome-wide association analyses

2.2.1. Discovery of SNPs in co-segregation with traits

For both production traits and clinical mastitis a regression-based score test was utilized, where genotypes for each SNP markers were coded as 0, 1 and 2 based on the present number of the rare allele. Observations of clinical mastitis were coded as 0 and 1, which are control and case respectively, so that a linear regression-based analysis could be applied also to this trait. Dependent variables for the analyses of production traits were de-regressed EBVs.

The score test developed by (Chen and Abecasis, 2007) was implemented on GenABEL R package via the function 'mmscore' for the analyses. It is a mixed animal model approach to family based association tests which can accommodate a genomic kinship matrix weighted with allele frequencies, to account for population structure and relatedness.

It is a two-step approach, where in the first step, SNP effects were set to "0" and fixed effects other than SNP effects together with polygenic background effects were fitted in the model. After variance components estimation via maximum likelihood method, adjusted residuals of the first model were used in the second step with a score test to detect whether the effects of SNPs were actually "0" as proposed by the null hypothesis. The main assumption behind calculating variance components in the first step and keeping them as actual parameters of the analysis for the second step comes from the "infinitesimal" model, which suggests that quantitative trait inheritance is determined by infinite number of unlinked loci with small effects (Fisher, 1930). Therefore, incorporating those SNPs in the second step one at a time does not change the estimates of variance components. By this, enormous amount of time will be saved as parameters are not going to be re-estimated for each SNP inclusion to the model.

The general model used in the analyses was (Chen and Abecasis, 2007):

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

where $\mathbf{y}$ is the vector of observations; μ the population meanfor the observed trait; $\boldsymbol{\beta}$ is the vector of fixed effects, which is divided into two parts during the analysis as $\boldsymbol{\beta_x}$ (vector of fixed effects other than SNP effects) and $\boldsymbol{\beta_g}$(vector of fixed SNP effects);$\mathbf{u}$ vector of random polygenic background effects of each animal and assumed to be drawn from a multivariate normal distribution (MVN)~(0,$\sigma_u^2\mathbf{G}$), where **G** is the genomic relationship matrix obtained as described further in the paper; $\mathbf{e}$ is the vector of random residual effects assumed to be drawn from a MVN~(0, $\sigma_e^2\mathbf{I}$) where **I** is the identity matrix;

$\boldsymbol{X}$ (divided into two parts during analysis as$\boldsymbol{X_x}$ and $\boldsymbol{X_g}$) and $\mathbf{Z}$ are design matrices mapping vectors of fixed and random effects respectively, to the observations. Thus, variance-covariance structure can be summarized as:

$$Var\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}\sigma_u^2 & 0 \\ 0 & \boldsymbol{I}\sigma_e^2 \end{bmatrix}$$

Here, σ_u^2 and σ_e^2 represent the maximum likelihood estimates (MLE) of variance components from the first step of the analyses, where SNP effects, denoted as $\boldsymbol{\beta_g}$, are set to "0" and therefore the model used:

$$\boldsymbol{y} = \mu + \boldsymbol{X_x\beta_x} + \mathbf{Zu} + \mathbf{e}$$

In the second step, heritability and variance components estimates are used to estimate all the marker effects one at a time, by fitting residuals obtained from the first analysis:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{g}} = (\boldsymbol{X_g^T}\ \boldsymbol{V^{-1}_{\hat{\sigma}^2,\hat{h}^2}}\ \boldsymbol{X_g})^{-1}\boldsymbol{X_g^T}\ \boldsymbol{V^{-1}_{\hat{\sigma}^2,\hat{h}^2}}\ \boldsymbol{R_{\beta_x}}\ \ ,$$

where $\boldsymbol{V^{-1}_{\hat{\sigma}^2,\hat{h}^2}}$is the inverse of the variance-covariance matrix used for MLE estimates of aforementioned parameters and $\boldsymbol{R_{\beta_x}}$is the residuals from the model in the first step.

The score test statistic for a particular marker can be denoted as below:

$$\boldsymbol{T^2} = \frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{g}}}{\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{g}})} = \boldsymbol{N}\frac{\boldsymbol{Cov}(\boldsymbol{y},\boldsymbol{g})^2}{\boldsymbol{Var}(\boldsymbol{y}).\boldsymbol{Var}(\boldsymbol{g})}$$

Here, $\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{g}})$ is the variance of SNP effects; **y** is the vector of observations (N x 1), whose number of rows is equal to the number of subjects in our analysis, since all subjects have only one observation (de-regressed EBVs for milk production traits and 0, 1 coded observations for mastitis); **g** is the vector of genotypes at a particular locus taking the values 0, 1 or 2 for N observations. Here "0" represent the homozygous genotype for one variant, "1" heterozygous genotype and "2" is homozygous genotype for the other variant.

As it was suggested by (Eu-ahsunthornwattana et al., 2014), the choice of method among linear mixed models cannot be based on type I error rate since all show similar results, instead the choice of method was done regarding speed of computation, ease of use and most importantly method of dealing with missing genotypes. However, it is clear that imputing missing genotypes with a probabilistic genotype score allows incorporating individuals with missing genotypes into the analyses and therefore indirectly increases the statistical power for the detection of true associations.

Finally, inflation of test statistics was further controlled and corrected via genomic control method (Devlin and Roeder, 1999). Conceptually, since it is expected to have only a small number of tested SNPs to be associated with the trait of interest, it can be suggested that the test statistics of rest of the SNPs can be used as "null loci" to quantify the inflation factor lambda (λ). However, since cut-off for detection of "null loci" is still not obvious, practically all loci are used for estimation. As before, the test statistics under genomic inflation (T_i^2) are distributed as $\lambda\boldsymbol{X}^2$**.** Under this circumstance, the mean of random variables coming from$\lambda\boldsymbol{X}^2$isλ, whereas it is equal to 1 when those variables are coming from $\boldsymbol{X}^2$**.** Since, truly associated SNPs will increase the mean of test statistics, using the mean could be over-conservative and may lead to increased proportion of false negatives. Therefore, in practice, the median of the test statistics is preferred. Thus, $\hat{\lambda}$can be formulated as:

$$\hat{\lambda} = \frac{Median\left(T_i^2\right)}{0.4549}$$

where, 0.4549 is the significance threshold value of the test statistic distributed as $\boldsymbol{X}_1^2$ under a= 0.50 significance level. The obtained value of lambda is then used to divide each test statistic value from the association analysis to obtain corrected test statistic values, i.e. p-values, for genomic inflation.

2.2.2. Imputation of missing genotypes

Imputation of missing genotypes was done automatically by the function 'mmscore' via an expected genotype score calculated using the flanking marker information obtained from subsets of genotyped individuals at the loci of interest from each observed 'families'. Estimation of expected genotype score ($\bar{g}_{ijm}$) at a particular locus was undertaken using the following approach:

$$E\left(g_{ijm} \middle| \boldsymbol{G}i, \theta, \boldsymbol{F}\right) = 2\text{P}\ (G_{ijm}=\text{A/A}\middle|\boldsymbol{G}i, \theta, \boldsymbol{F}) + \text{P}\ (G_{ijm}=\text{A/B}\middle|\boldsymbol{G}i, \theta, \boldsymbol{F})$$

$$\bar{g}_{ijm} = E\left(g_{ijm} \middle| \boldsymbol{G}i, \theta, \boldsymbol{F}\right)$$

Here, $\boldsymbol{G}i$ denotes the matrix of all the observed genotypes for i'th family; θ is the vector of inter-marker recombination fractions and finally $\boldsymbol{F}$is the vector of allele frequencies. Further details regarding estimation of genotype scores and incorporating into the model is available (Chen and Abecasis, 2007).

2.2.3. Estimation of kinship matrix from genomic data

In the analyses, the genomic **relationship** matrix (**G**) (i.e., twice the **genomic** kinship matrix) was used to estimate model effects with the purpose of correcting for population structure and familial relatedness. Since it provides realised kinship between individuals via capturing Mendelian sampling terms, usage of genomic kinship matrix preferred over pedigree based estimations.

Computation of genomic kinship matrix was conducted via "ibs" function of GenABEL package in R, following the method described by (Astle and Balding, 2009). Here, *Kij* represents the correlation coefficient between a particular locus of *i*

th and *j* th individuals, while *xi* and *xj* are allelic counts for that spesific locus of the same individuals represented as 0, 1 or 2 depending upon the numbers of referred allele. Then the covariance between the two allelic phases can be written as:

$$Cov(x_i , x_j) = 4p(1 - p)K_{ij}$$

where p is the frequency of the same allele. If *x* is written as a column vector over individuals and L is the number of loci to be compared, estimator of kinship matrix can be formulated as:

$$\widehat{\boldsymbol{K}} = \frac{1}{L}\sum_{l=1}^{L}\frac{(xl - 2pl\mathbf{1})(xl - 2pl\mathbf{1})^{\boldsymbol{T}}}{4pl(1 - pl)}$$

2.2.4. Significance thresholds

Two different significance thresholds were used in the analyses. Bonferroni correction was used to obtain thresholds with the purpose of accounting for inflation due to multiple testing. A SNP was recognised as 'significant ' at genome-wide level, when the obtained p-value was smaller than 0.05/N, where N was the number of total markers that passed quality control criteria since it represents the number of independent tests applied in the study. Furthermore for each chromosome, the significance was assessed via the threshold 0.05/n where n was the number of markers on a given chromosome. Representative thresholds on Manhattan plots are $-\log_{10}(0.05/N)$ and $-\log_{10}((0.05/N) \times 29)$ for genome-wide significance and chromosome-wide significance respectively.

2.2.5.Q-Q plots

Quantile- Quantile (Q-Q) plots were utilized to visualise the extent of concordance between expected (null hypothesis of "no association") and observed test statistics. Essentially, quantiles of two probability distributions are plotted against each other and any deviation between the two distributions is demonstrated. Mass deviation of test statistics can be taken as a reporter of inflation due to population structure, relatedness or other kinds of systematic bias, such as genotyping error.

2.2.6.Determining possible QTL regions and effect sizeestimations

Manhattan plots were used to visualise significant SNP on all chromosomes at the same time (Gibson, 2010). Subsequently for all genome and chromosome wide significant hits, 'regional LD heatmaps' were used for analysing regional linkage structure and minor allele frequency for SNP markers. SNP effect sizes were estimated and presented by the GenABEL software. All effect sizes indicated in the study are additive effects. Output for genotypes 0,1 and 2 was converted to AA, AB and BBcoding, where AA was homozygous for one vairant (A allele), AB heterozygous genotype and BB was homozygous genotype for another variant (B allele). Effect sizes estimated here represent specified unit of increase or decrease in the values of de-regressed EBVs for production traits. For CM, they indicate odds ratio concerning the odds of getting the disease having a particular allele compared to another.

NCBI and Ensembl genome browser (Yates et al., 2016) was used for recovering further information on the regions of interest, position of variant in the genome and functions of suggested gene regions in homologous regions of cattle and sheep genomes.

3. RESULTS & DISCUSSION

To date, there have not been many studies regarding the genetics and genomics of traits of economic interest in goats. However, the initial publication of the goat reference genome (Dong et al., 2013) and design of 50K SNP chip for goats (Tosser-Klopp et al., 2014) led to genetic improvement aimed GWAS studies to be applicable for goats. Before that, due to the lack of well-established microsatellite markers; studies on milk production and disease related traits had rarely been conducted on genome-wide scale, thus leading to poorly reported QTL regions.

To our knowledge, this is the first genome-wide association study of clinical mastitis in dairy goats as well as being the most powerful study conducted on milk production traits of dairy goats in the UK, regarding the sample size.

In this study, GWAS for 4 milk production traits -milk yield (MY), milk fat yield (MFY), milk protein yield (MPY), milk lactose yield (MLY) - and clinical mastitis were performed. Approximately ~7500 animals were used for milk production traits whereas this number for clinical mastitis was 3968 animals. For milk production traits de-regressed EBVs used as pseudo-phenotypes. It encapsulates all the information for an animal in a single value that is inherently corrected for systematic environmental effects and appropriately weighted for removing the bias of EBVs sourced from varying offspring numbers of animals. For clinical mastitis, binary observations -all recorded in the same lactation- were used.

3.1. Characterisation of phenotype and genotype data

3.1.1. Traits and phenotype data

For the analysed traits summary statistics were obtained as represented below in the table 2. Clinical mastitis was excluded from the table since it was a binary observed trait. The mean values for the de-regressed EBVs were ranged between 5.009 and 370.641 corresponding to milk protein yield and milk yield respectively.

Table 3. Descriptive statistics for the analysed traits

Trait	Acronym	N	Mean	SD	Min	Max
Milk yield	MY	7967	370.641	174.829	-189	1039
Milk fat yield	MFY	7341	6.375	4.559	-12.590	24.890
Milk protein yield	MPY	7341	5.009	3.149	-8.080	16.380
Milk lactose yield	MLY	7341	6.998	4.370	-11.620	22.490

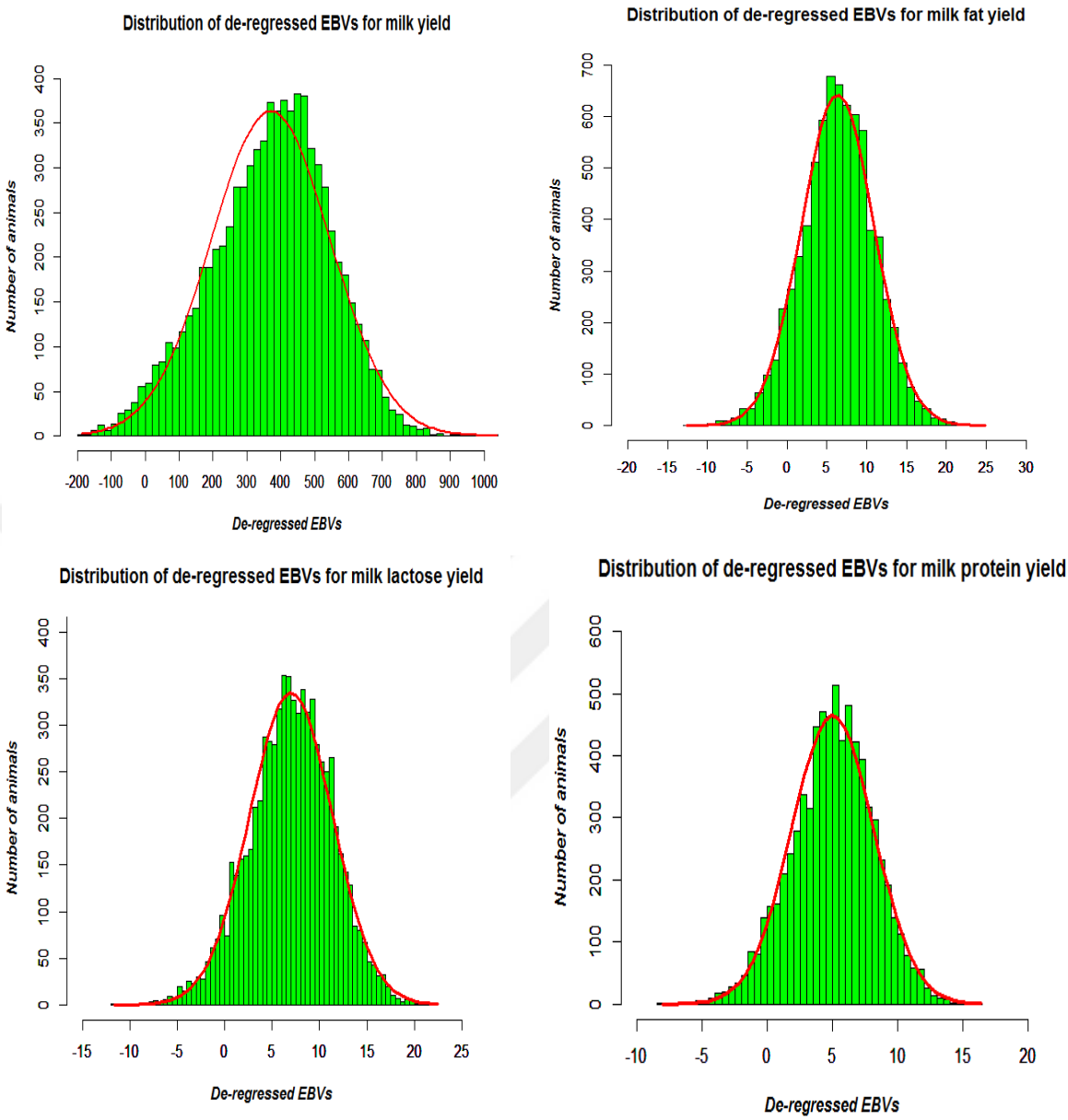
N: Number of animals used in the analysis after quality control

SD: Standard deviation

For clinical mastitis (CM) 3968 animals passed imposed quality control checks, of which 140 were cases and 3828 controls.

Distributions of de-regressed EBVs for each trait were examined via histograms and no major deviation from normality was observed. Results below were obtained where drawn curve represents normal distribution:

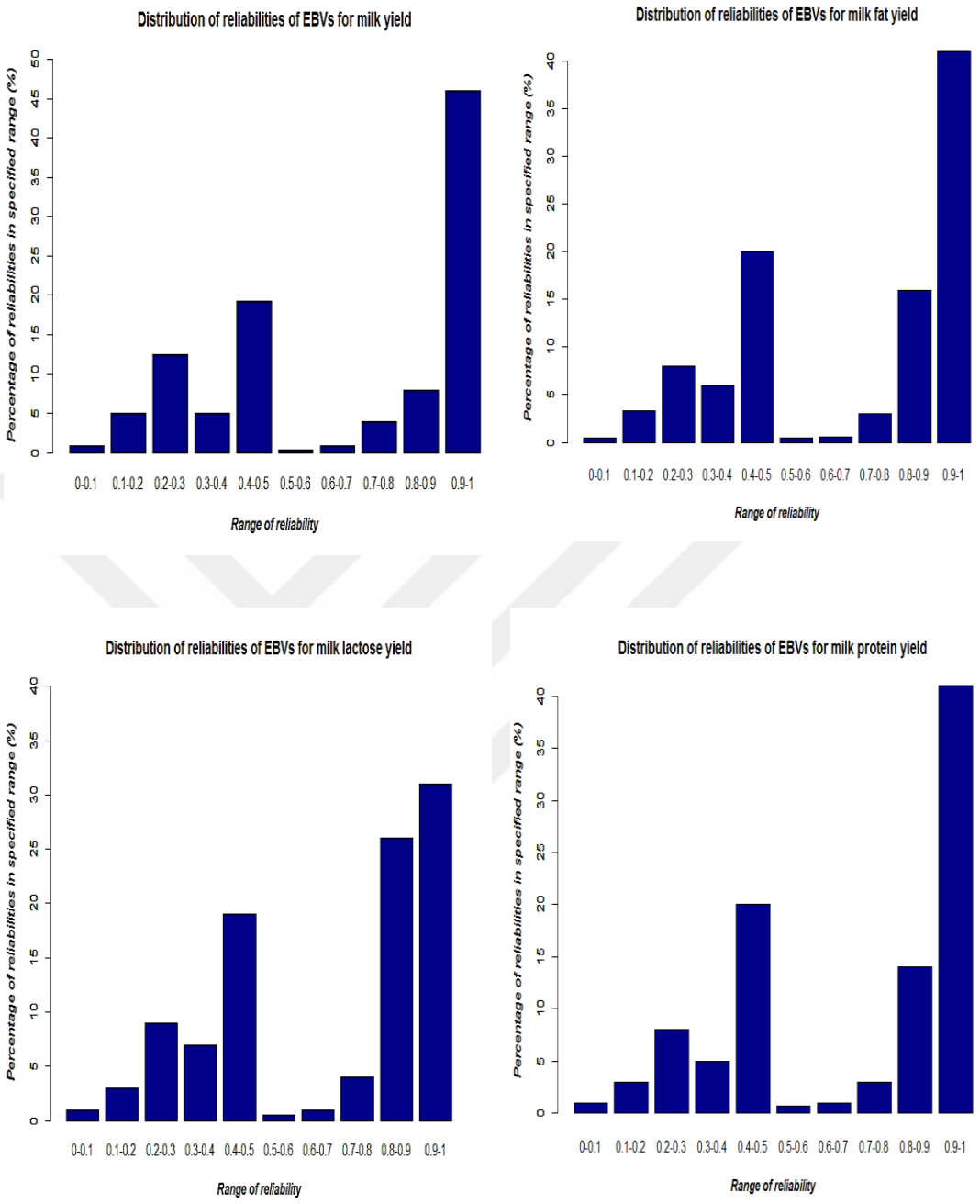
Figure 1: Distributions of de-regressed EBVs



All production traits showed normal distibution characteristics. Therefore, the de-regressed EBVs were used without any transformation applied.

Since reliabilities of EBVs may have an effect on obtained results with de-regressed EBVs (Garrick et al., 2009), their distribution was also obtained and presented below (Figure 2):

Figure 2: Distribution of reliabilities for EBVs



One of the main assumptions for GWAS is that homogenous variance (equal variance for all response variables) exists for the trait analysed (Bush and Moore, 2012). Having a wide range of reliabilities for the de-regressed EBVs, undermines this assumption. One important step for using de-regressed information as a

response variable for GWAS analysis is to weight the de-regressed EBVs with a proper weighting as suggested by (Garrick et al., 2009), to deal with heterogeneous variance sourced by varying r^2 values of EBVs. If the range of reliabilities used in the analysis is narrow, then the effect of not weighting the de-regressed values on the analysis would be a minor one, however if it is otherwise then ignoring weighting may lead spurious associations and power lose (John Woolliams, pers. comm.). Therefore, the range of reliabilities of EBVs used in the analyses was determined. Even if the majority of EBVs have quite high reliabilities, small values of reliabilities observed. However, to our knowledge de-regression procedure undertaken in MIX99 at least partially accounts for the differences in reliabilities by weighting with EOC (effective offspring contribution). However, to compare and validate the obtained results a further step can be taken and GWAS with weightings imposed to the de-regressed breeding values can be implemented. .

3.1.2. Genotype data characterisation and quality control (QC)

QC has resulted in different number of markers for milk yield, three other milk production traits and clinical mastitis since the number of animals used in studies were different.

For milk yield, 129 markers which had call rate less than 95 % were eliminated as well as 270 SNPs with MAF less than 0.05 and 169 SNPs out of Hardy- Weinberg equilibrium (HWE) ($1.11x10^{-6}$). Furthermore, 26 individuals were excluded from analysis due to IBS > 0.95 since it might represent the same sample repeated during the genotyping. Therefore, analysis had ended up 7,967 animals with 44,335 markers, with mean heterozygosity for a SNP being 0.398 (SD= 0.014 and 0.399 (SD= 0.018) for an animal.

For MFY, MLY and MPY the same thresholds were applied and as a result 269 markers with MAF< 0.05 excluded as well as 128 due to low call rates and 178 due to deviation from HWE. Subsequently 26 animals were removed because of having high IBS sharing as since they might be duplicates. Thus after the QC, milk production traits ended up with 7,341 animals and 44,346 markers, where the mean

heterozygosity for a SNP was 0.398 (SD= 0.105) and for animals it was 0.399 (SD= 0.018).

For CM, 3968 animals with 44,471 markers were obtained from the same thresholds, where 157 markers removed due to low call rates, 186 due to MAF < 0.05 and 91 due to imposed p-value threshold from HWE. The number of animals excluded due to high IBS was 12 for CM Mean heterozygosity for a SNP was 0.398 (SD(0.102 similarly for an animal it was obtained as 0.399 with SD of 0.018.

Distribution of MAF obtained after quality control presented below:

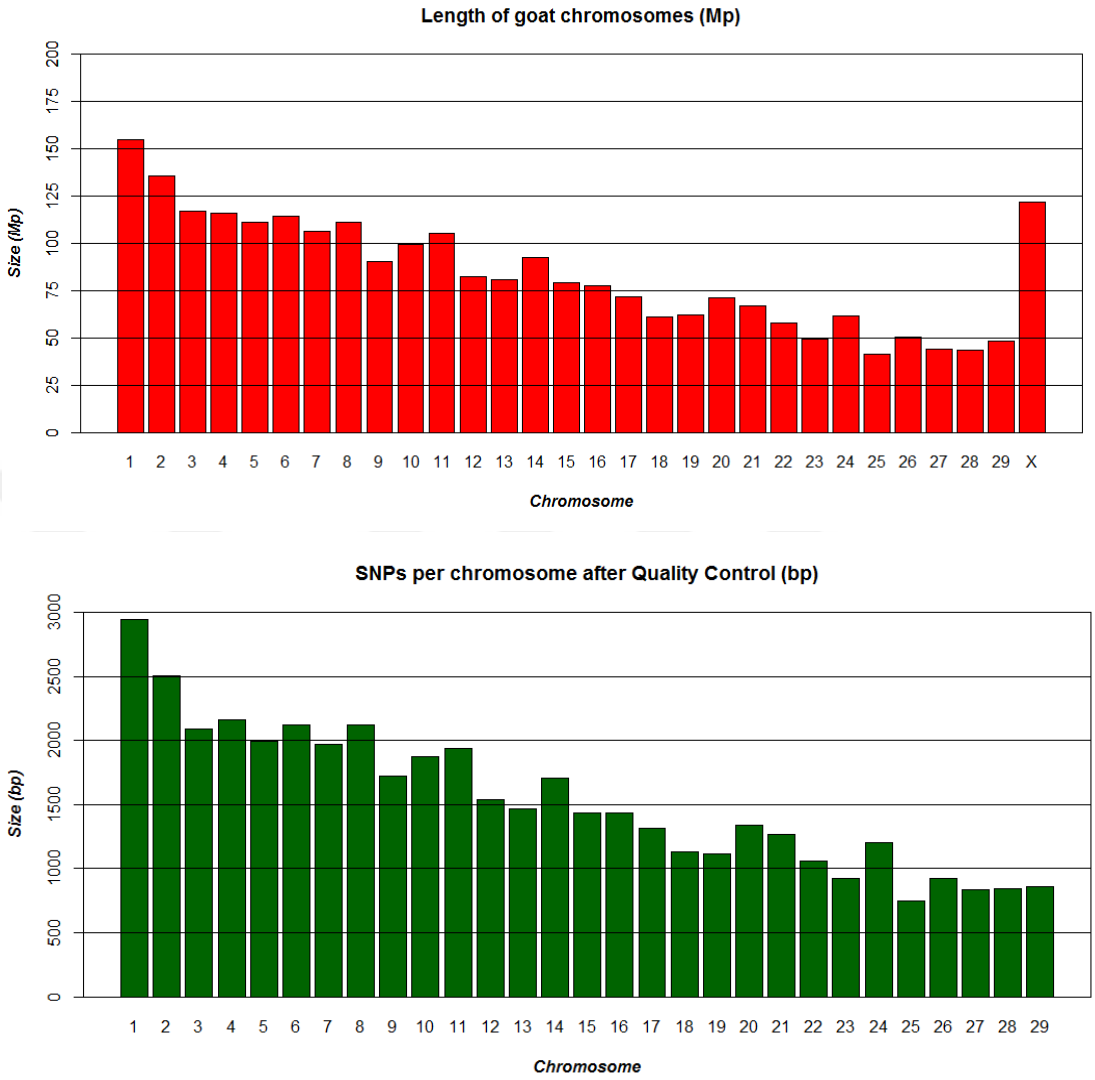
Table 4. Distribution of SNPs regarding MAF after quality control

MAF	Number of SNPs	Proportion
0.05- 0.1	2,492	0.056
0.1-0.2	7,225	0.163
> 0.2	34,629	0.781
Total	44,346	1

The same proportions were obtained for the other two traits (MY and CM), although number of SNPs within defined groups changed slightly.

SNPs were distributed on every chromosome with approximately 50 kb spacing. They were distributed to each chromosome in proportion with their length, as it is also clearly suggested by the figures below (Figure 3):

Figure 3: Length of goat chromosomes and average number of SNPs per chromosome



There are various approaches towards implementing quality control on genotype data of GWAS. Some studies do not set any minor allele frequency (MAF) threshold while others do not remove any alleles regarding deviation from Hardy-Weinberg Equilibrium. However, a study by The Wellcome Trust Case Control Consortium proved that certain thresholds needs to be applied to raw genotype data to obtain reliable results regarding association studies (The Wellcome Trust Case Control Consortium, 2007).

MAF threshold was set to 0.05 to eliminate markers with rare variants since they are highly prone to genotype misclassifications. Furthermore, potential power loss as a result of increased number of tests and false positive rate increase due to informative missingness are other strong-holding arguments that encourages such a threshold (Weale, 2010).

Since the discovery of relationship between alleles and genotypes, known as Hardy-Weinberg equilibrium (HWE) (Hardy, 1908; Weinberg, 1908), it has become a powerful tool of use for both theoretical and applied research on population and quantitative genetics (Crow, 1988). Essentially, under the conditions of no mutation, no migration, no genetic drift, no selection and given that subjects are randomly mating; both allele and genotype frequencies suggested to remain constant over generations. Thus, regarding the Hardy-Weinberg equation, there is a stable relationship between genotypes and allele frequencies. Consequently, departure from HWE can be used as an indicator of genotype calling problems with care since the source of departure may be selection or other above-mentioned forces (Ziegler, 2008). Different studies have used different pre-defined thresholds for deviation tests to eliminate highly deviating SNPs (Samani et al., 2007; The Wellcome Trust Case Control Consortium, 2007). Here we used a Bonferroni corrected p-value threshold for removing high deviants. However to our knowledge our population was selected for best performing animals in every generation regarding higher milk yield. Therefore to make sure that we were not removing SNPs desired to be discovered in association with milk yield and production traits, a further series of analyses were undertaken with having no threshold for HWE and no significant changes has been observed regarding associated SNPs.

SNP call rate and animal call rate thresholds were imposed, to eliminate problems arising from poor quality clustering of genotypes for each SNP and bad quality DNA for animals respectively.

Gender checks were done in the previous study conducted by (Mucha et al., 2016) and X chromosome was removed after quality control checks for the favour of

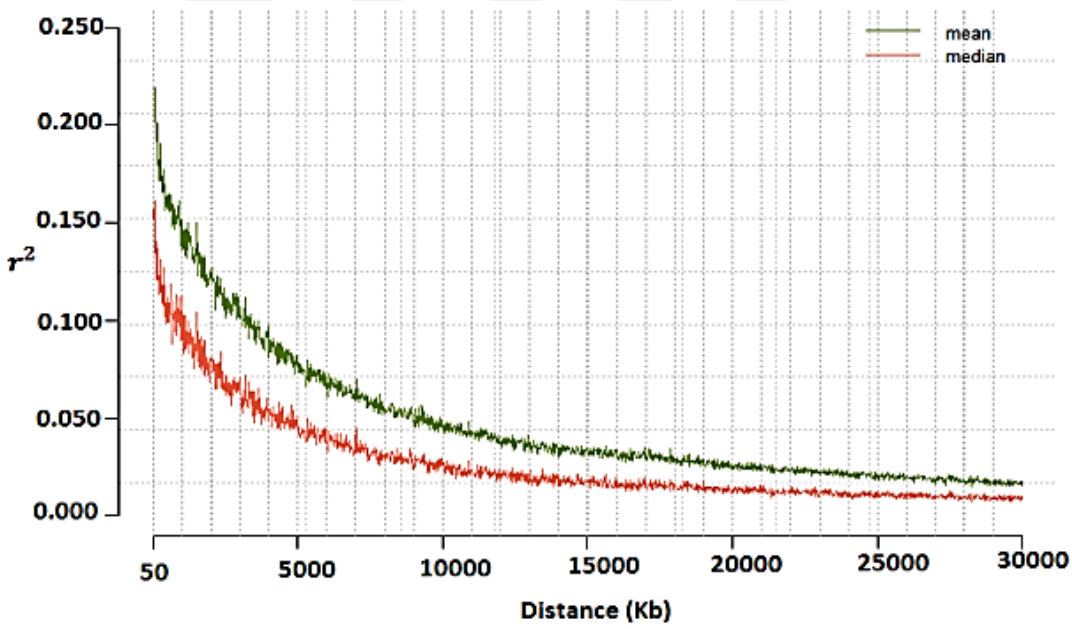
further analysis since it causes problems with kinship estimations and principal component analysis. Therefore in this study X-related SNP markers were not involved in the obtained initial genotypes.

3.1.3. Linkage Disequilibrium analysis and visualisation

Mean pairwise LD for all syntenic markers from 29 autosomes was estimated to be 0.034 (SD= 0.005). LD per chromosome revealed an opposite trend with the length of chromosome as it was expected. Mean LD between adjacent markers (within average 50 Kb distance) found to be 0.221 (SD= 0.043).

LD decay plot was obtained for 50 Kb frames and presented below:

Figure 4: LD decay plot



In the **Figure 4,** the mean LD (upper line) showed expectedly an exponential downward trend with increasing distance. Largest decline happened for distances below 100 kb, which was from 0.221 for 50 kb to 0.185 for 100 kb. It was followed by 0.159 for markers in 500 kb distance and 0.120 for 2000 kb distance. Finally average LD for markers within 5000 kb distance was estimated to be 0.078. Results

overall are higher than those estimated by (Mucha et al., 2015) via using the same population, possibly due to having an increased sample size, with more related animals included in the analysis from subsequent generations.

Since chromosome 19 had significant hits for all four milk production traits (MY, MFY, MLY, MPY) and chromosome 6 for fat yield; LD structures of the interested regions were further investigated, plotted and given in the Figure 5.

In Figure 5, bold blue SNP (6^{th} SNP from the left-hand side) is the top associated SNP while others are the SNPs that also exceed the genome-wide significance threshold. Total mapped region is 8 Mb fragment covering the genome-wide associated SNPs. The big triangle encapsulates a 3 Mb region (from 24^{th} to 27^{th} Mega base pairs of the chromosome 19) that is characterised with high LD and high number of protein coding genes in the genome. The small triangle covers the region where significance hits get intensive, possibly due to high LD between SNPs typed in the region. The sequence in this particular region codes fora number of protein coding genes such as RNASEK -where the top associated SNP is – and ALOX12, ASGR2, KIF1C, DNAH2; each at least having one genome-wide associated SNP either as intragenic or closely located to the coding sequence.

Figure 5: LD heatmap of the region on chromosome 19 with significant SNPs

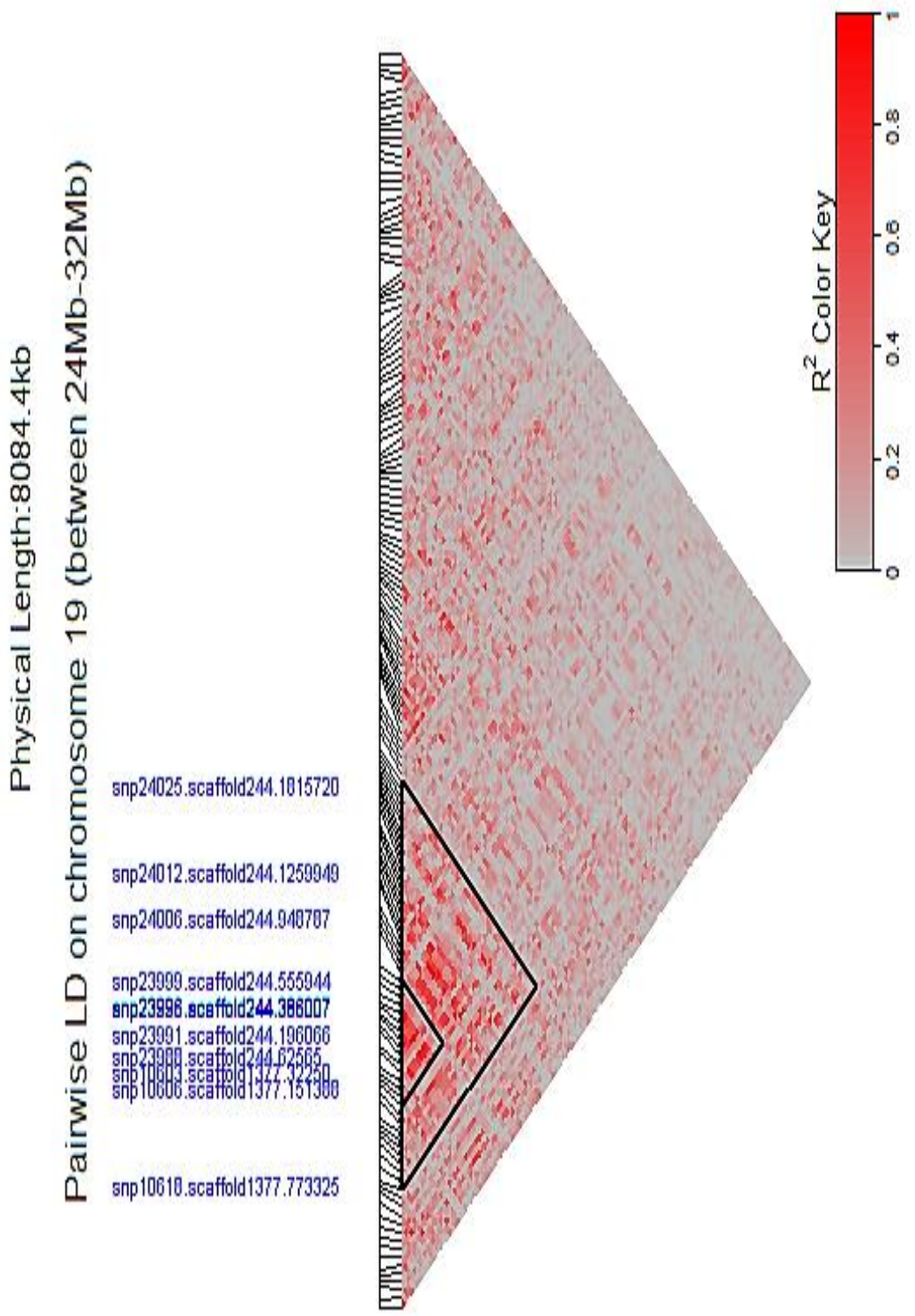
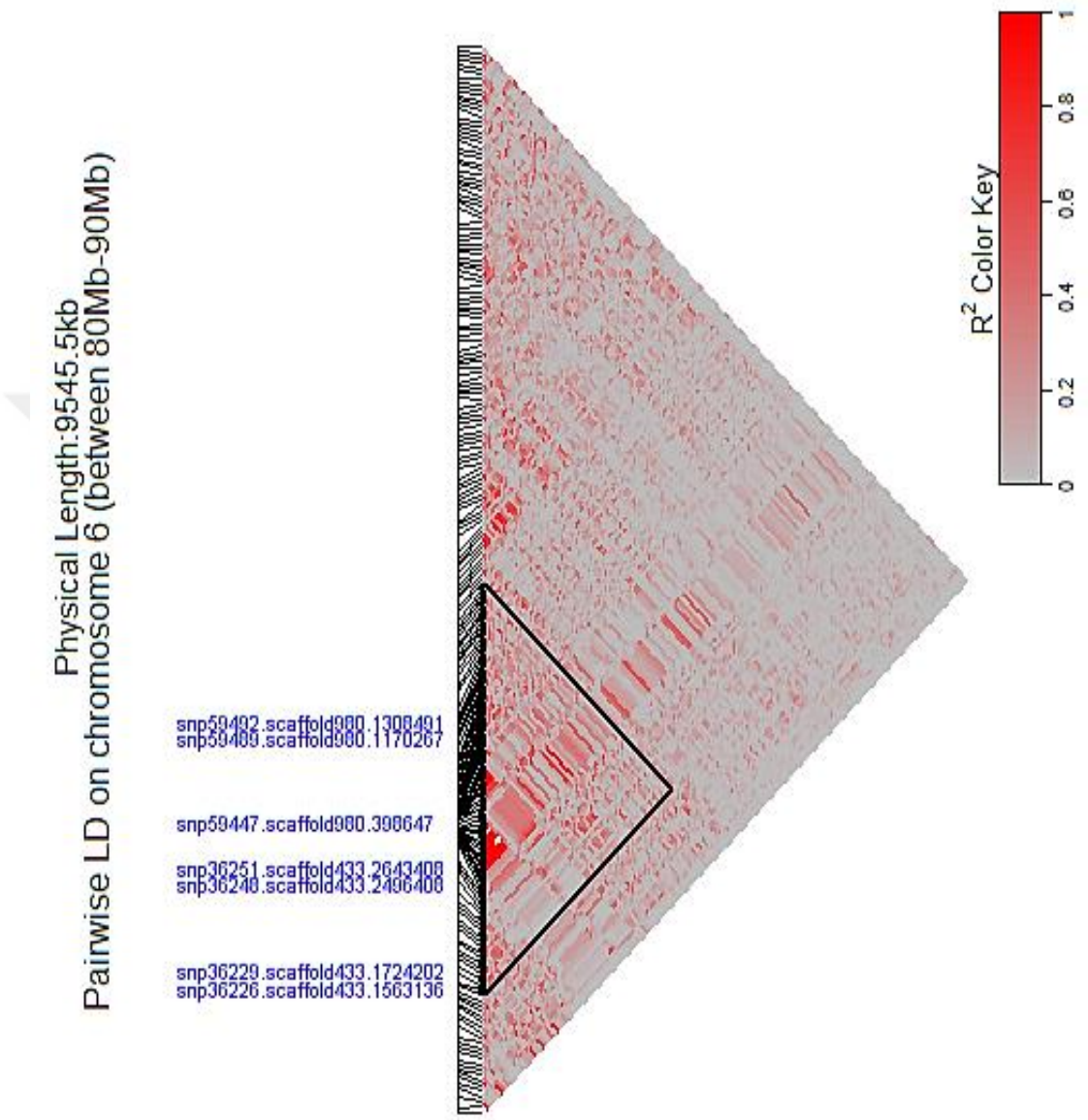


Figure 6: LD heatmap of the region on chromosome 6 with significant SNPs



Here the full red triangle regions represent sequence with high number of protein coding genes, including caseine coding genes, discussed later. Top genome-wide associated SNP arepresented on the left-hand side of the plot, whereas other denoted SNPs have at least chromosome-wide significance. Pairwise LD between

SNPs are coloured according to the "R^2 Color Key" regarding strength of LD. Moderately high LD was also observed in this region.

LD structure was examined because since typed SNP markers used in GWA analyses were only covering $1.6x10^{-5}$ of the analysed goat genome, they possibly do not capture all the existing genetic variance lying in the genome. Therefore association studies rely on linkage disequilibrium between marker locus and the causal variant. Perceiving to which distance LD is effective, gives a safe ground to speculate about the relative position of causal variant to the marker locus. Mean LD (r^2) between adjacent markers (average 50 Kb spacing) found to be 0.221, showing relatively high LD compared to the literature for Alpine and Saanen; slightly low LD compared to Toggenburg (Brito et al., 2015).

LD decay was found to be quick in short distances (below ~100 Kb), sincelargest fall happened from 50 kb ($\overline{r^2}$=0.221)distance to 100 kb ($\overline{r^2}$= 0.185), which was followed by 0.159 for markers in 500 kb and 0.120 for 2000 kb.

Results demonstrate that the population under study is characterised with relatively high overall LD between syntenic markers and adjacent SNPs compared to the results obtained by (Brito et al., 2015) in Table 2 and by (Mucha et al., 2015). Furthermore, low LD decay speed with distance compared to (Carillier et al., 2013) was also observed.

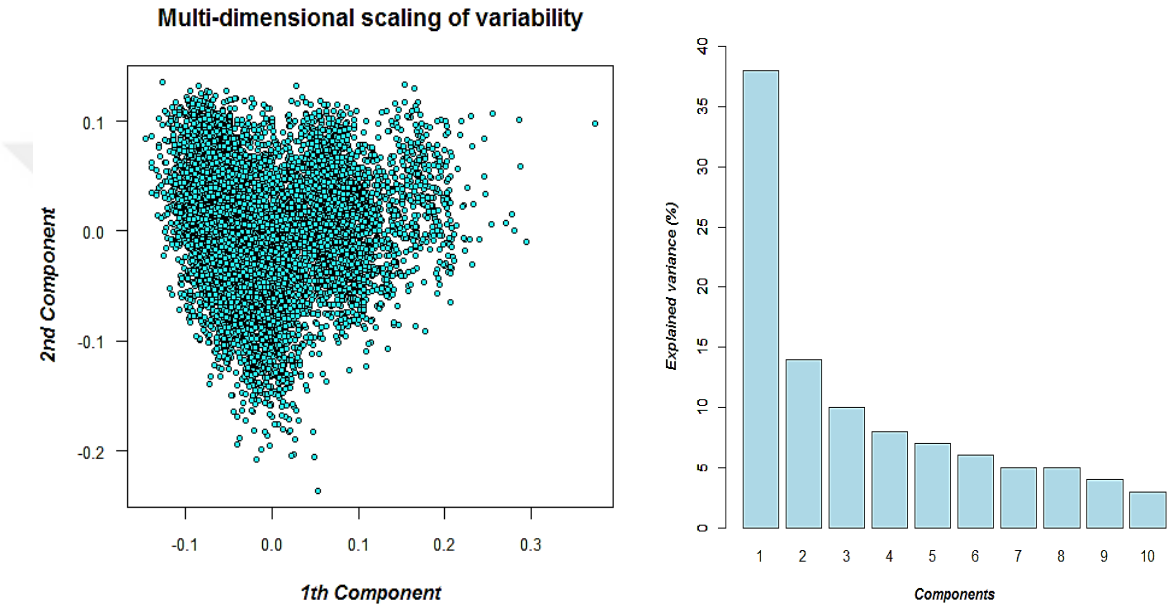
Consequently it can be inferred that any variant reached genome or chromosome-wide significance can be in reasonably high LD with the causal variant (if not already itself) up until relatively remote distances. Thus LD heatmaps were utilised to visualise regional LD structure of the genome-wide significant hits.

High LD throughout the region visualised on chromosome 19 (LD heatmap) might be as a result of intensive selection. Since we know that our population was bred regarding best performing animals for higher milk yield in every generation, rise of regions like this in the genome is highly possible.

3.1.4 Multidimensional scaling analysis

Since breed composition of the animals was not recorded and therefore could not be taken in the analyses, to prevent GWAS results to be confounded by possible unaccounted stratification, a multidimensional scaling analysis was done initially with 2 components. Later, proportion of variance explained by each of 10 principal components was obtained via the same function. Results are plotted in the following figures respectively:

Figure 7: Plots of principal components



Multi-dimensional scaling did not show any major clustered groups, however as it can be seen that there is a grouping of animals in the background which possibly captures the breed composition. Moreover, the variance explained by the first principal component was 38 percent of total genotypic variance, which is remarkably higher than the other components as it was illustrated by thesecond plot (**Figure 7**).The first three components in total explain almost 70 percent of variation in the data, which suggests some degree of stratification may exist.

3.2. Genome wide association analyses

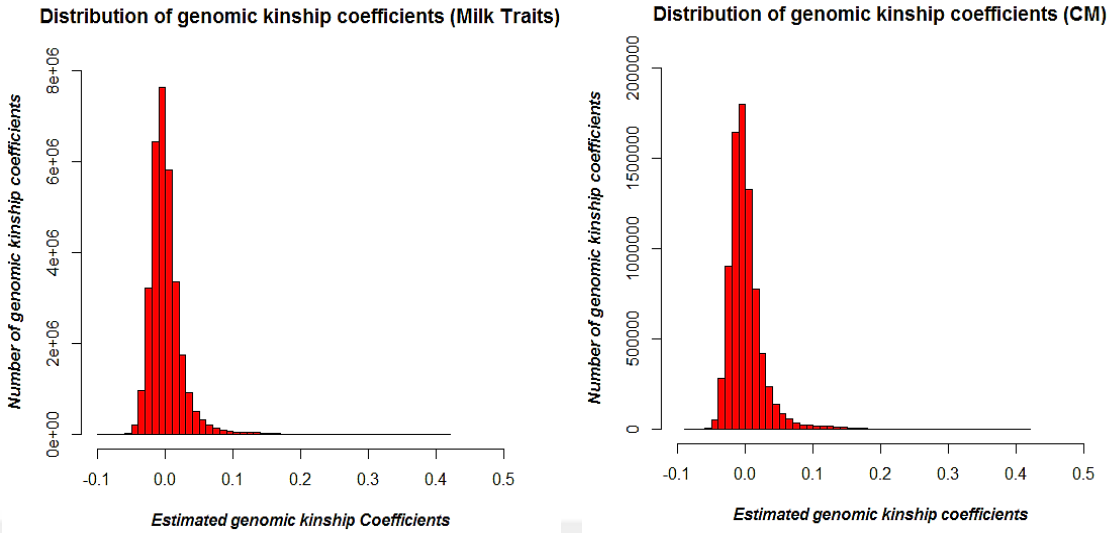
3.2.1. Estimation of kinship matrix from genomic data

The mean estimated kinship between two individuals was 0.016 ±0.016 with a minimum of 0.0001 and a maximum of 0.417.Results obtained were approximately the same for all milk yield and other milk production traits although the number of individuals among studies were slightly different. Proportion of kinship coefficients that is compatible with relationship of at least first degree cousin (and/or closer) was 2 % of the all estimated coefficients, among which 15% were characterised as half-sib (and/or closer) relationships. In general, 0.01 % all estimated coefficients were found to be higher than 0.25, which is relationship of 0.5 (full-sib and higher relationship coefficients possibly due to inbreeding).

For CM, estimated kinship coefficients among relatively smaller number of animals (3968) had a mean of 0.017 ±0.017 with a minimum of 0.0001 and maximum of 0.419, which is relatively similar to results from milk production traits. 2.5 % of pairwise kinship coefficients were compatible with first degree cousin and/or closer relationships, among which 19 % were characterised as half sib and/or closer relationships. Among all coefficients estimates, only 0.006 % were found to be representing full sib or higher relationship coefficients.

Distribution of kinship coefficients obtained for milk production traits and CM shown separately in the figures given since the number of animals among these studies varied significantly:

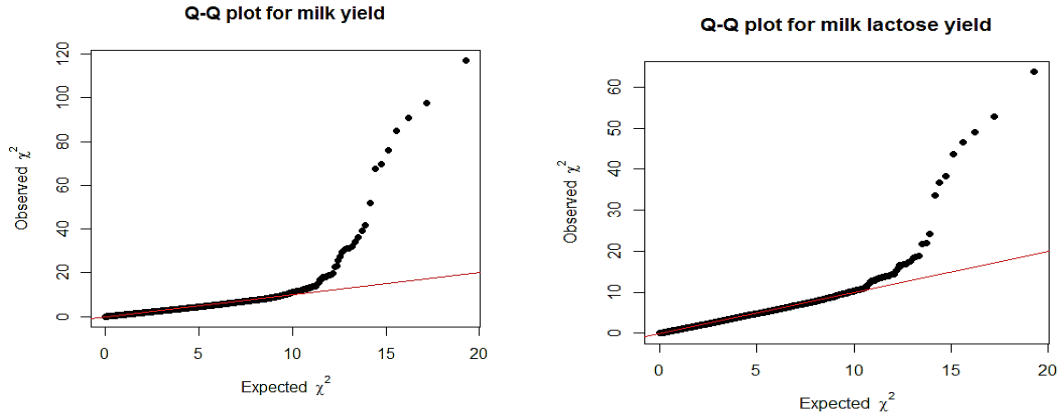
Figure 8: Histograms of estimated genomic kinship coefficients

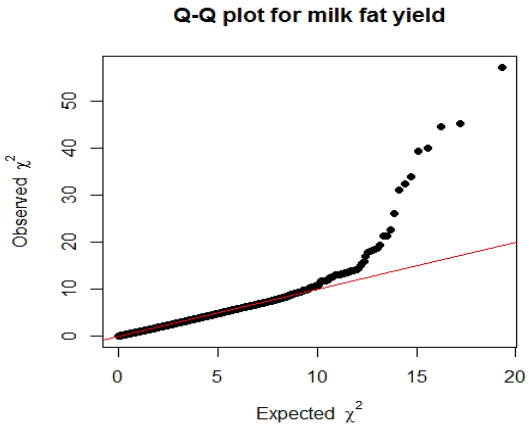


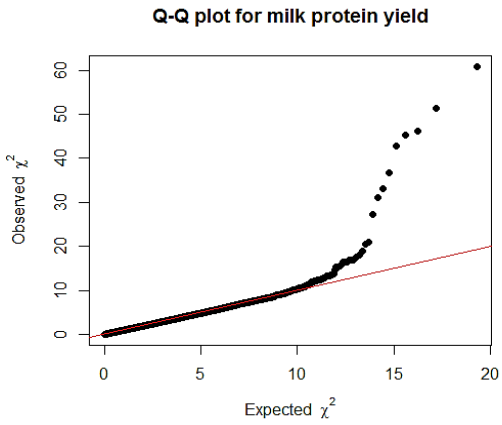
3.2.2. Q-Q plots

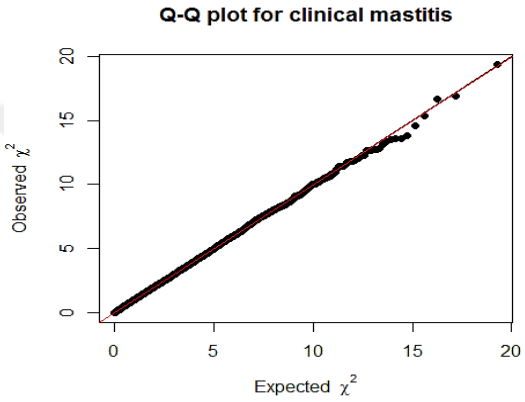
Quantiles of observed and expected test statistics for SNPs were plotted with corrected test statistics via genomic control. Although inflation before genomic control was varying among analysed traits (between 1 and 1.16), with genomic control correction all traits were set to unity (1) inflation. Drawn (red) line indicates identity between observed and expected test statistics, which is the null hypothesis of 'no association'.

Figure 9: Q-Q plots of tested SNPs for all analyses









As it was expected, analysed milk production traits showed similar trends of deviation from the null hypothesis; though regarding MY, the test statistic values are remarkably higher than the other three traits.

Here inflation values of the observed test statistics are given as before and after genomic control (GC) correction:

Table 5. Inflation factors for the analysed traits

Trait	Before GC	After GC
Milk yield	1.160 ± 0.003	1 ± 0.002
Milk fat yield	1.095 ± 0.001	1 ± 0.001
Milk lactose yield	1.091 ± 0.001	1 ± 0.001
Milk protein yield	1.089 ± 0.001	1± 0.001
Clinical mastitis	1 ± 0.000	1± 0.000

Since genomic inflation (ʎ) of all traits was unity, deviations observed in Q-Q plots are highly likely that are representing true associations.

3.2.3.Identification of associated regions

Genome- and chromosome-wide significance thresholds were calculated as described above in methods. Table 6 shows chromosome-wide significance thresholdsseparately for each chromosome:

Table 6. Chromosome-wide significance thresholds

Chr1	**Chr2**	**Chr3**	**Chr4**	**Chr5**	**Chr6**	**Chr7**	**Chr8**	**Chr9**	**Chr10**
$1.7x10^{-5}$	$2x10^{-5}$	$2.4x10^{-5}$	$2.3x10^{-5}$	$2.5x10^{-5}$	$2.4x10^{-5}$	$2.5x10^{-5}$	$2.4x10^{-5}$	$2.9x10^{-5}$	$2.7x10^{-5}$
Chr11	**Chr12**	**Chr13**	**Chr14**	**Chr15**	**Chr16**	**Chr17**	**Chr18**	**Chr19**	
$2.6x10^{-5}$	$3.2x10^{-5}$	$3.4x10^{-5}$	$2.9x10^{-5}$	$3.5x10^{-5}$	$3.5x10^{-5}$	$3.8x10^{-5}$	$4.4x10^{-5}$	$4.5x10^{-5}$	
Chr20	**Chr21**	**Chr22**	**Chr23**	**Chr24**	**Chr25**	**Chr26**	**Chr27**	**Chr28**	**Chr29**
$3.7x10^{-5}$	$3.9x10^{-5}$	$4.7x10^{-5}$	$5.4x10^{-5}$	$4.1x10^{-5}$	$6.7x10^{-5}$	$5.4x10^{-5}$	$6x10^{-5}$	$5.9x10^{-5}$	$5.8x10^{-5}$

Chr: Chromosome

Results of multi-dimensional scaling analysis proposed the presence of a certain degree of stratification in the analysed population. Furthermore, relationships inferred from estimated genomic kinship coefficients also suggested the existence of a wide range of close relationships between individuals. Therefore a mixed linear model based association test was preferred over other existing methods. With this purpose, ‘mmscore’ (FASTA method described by Chen and Abecasis) function of GenABEL was used since it was proved that unbiased estimations for SNP effect sizes can be obtained via this particular method, in the existence of relatedness (Do et al., 2013).

Inflations obtained for the analysed traits before genomic control correction ranged between 1.000 and 1.160, meaning that used FASTA method has done relatively a good job for dealing with stratification and familial relatedness. Considering large sample size of the analysed population (any subtle inflation might cause spurious

associations) and for the sake of unbiasedness, genomic control applied over raw p-values, which removed existing inflation and brought all inflation factors to unity (1).

In general, the variants found in common for the traits were in the majority and showing similar effects on the traits regarding size and direction. This supports the estimated genetic correlations among the traits (presented in the introduction) to be strong and proposes similar genetic background for the analysed traits. Moreover, the majority of genes carrying associated variants were found to be involving in processes such as fatty acid metabolism, vesicle transport (secretion) and ion transport; whose effects are integral to milk production.

Genome and chromosome-wide significant SNPs with their minimum observed p-values (among traits associated), effect allele, allele frequency andeffect size, proportion of genetic variance explained and various other information were provided as tables and presented in table 7 (Page 59) and table 8 (Page 60).

It should be noted that associations obtained here are only indicating that allelic phase of these regions are correlated with the existence or absence of binary traits and upward or downward trend in continuous traits. Therefore, to be able to talk about causality further biological validation is required. Here, to some degree, possible candidates or 'step up' points to biological processes were discussed and are not representative of causality and biological validation

3.2.3.1. Milk yield

All SNP effects reported here are additive effects. Genome-wide association analysis of milk yield resulted in a top genome-wide significant SNP, "snp23996.scaffold244.386007 (rs268292132)" on chromosome 19, with a p-value of 2.868×10^{-27} and a number of SNPs within 3 Mb surrounding, which are in high LD with the associated SNP (demonstrated in "Linkage Disequilibrium analysis and visualisation" section.The top associated variant is within a gene called RNASEK (Ribonuclease Kappa). However as it was suggested by regional LD heatmap, the

region has high LD in general. A close look on NCBI pointed out that a dozen of other genes and uncharacterised variants are quite closely located at the same 3 Mb region. Therefore all genome-wide significant SNPs had been taken into consideration with their minor allele frequencies and LD patterns. As a result, our focus was drawn by also a number of other candidate genes such as ALOX12 (~3 Kb downstream of top associated SNP), ALOXE3 (~900 Kb upstream associated SNP), ASGR2 (~ 50 Kb upstream) andKIF1C (~ 420 Kb downstream). Each of these gene regions have been found to have at least one genome-wide significant hits either within themselves or in 20 Kb vicinity. Here, the biological processes where identified genes involved are presented with their Gene Ontology (GO) terms.

Figure 10: Manhattan plot of milk yield association analysis

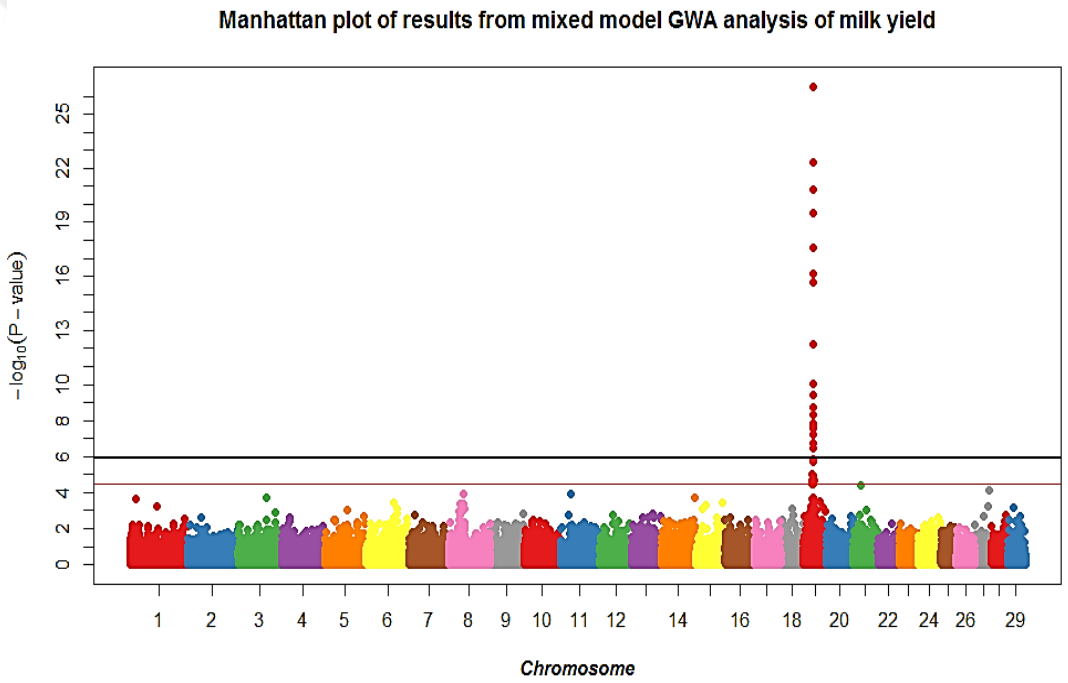


Fig. 10: $-\log_{10}$ (p-value) of each tested SNP markers are plotted against their position in the goat genome. Black line (upper line) represents genome-wide significance threshold, whereas red line (lower line) is overall chromosome-wide significance threshold. It is valid for all Manhattan plots presented in this paper.

Figure 11: Regional significance, LD structure and MAF of SNPs on chromosome19.

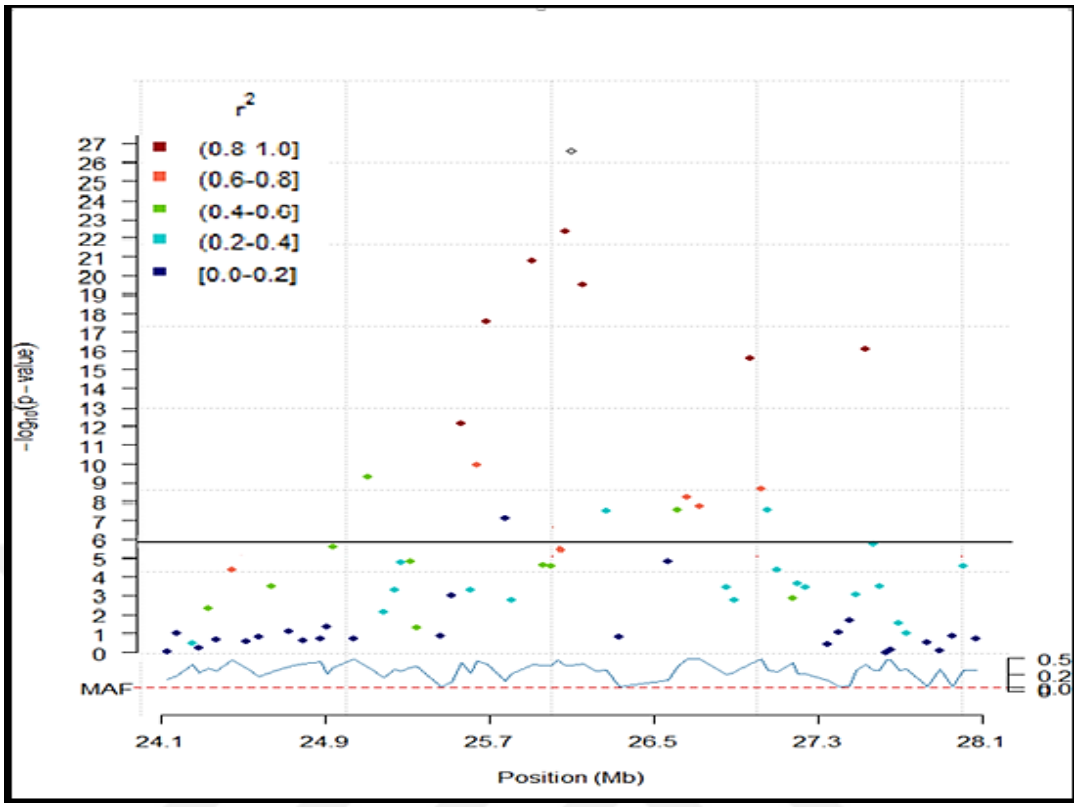


Fig. 11: Above plotted ~4 Mb region was centred on top-associated SNP. Points represent SNP markers, with the ones above marked threshold (genome-wide significance threshold) being SNPs reached genome-wide significance as a result of GWA analysis. Colour of a point shows its LD, measured as r^2, with the top associated SNP in accordance with legend present on top left of the figure. Mini-line chart below shows MAF (minor allele frequency) of SNP markers in the analysed population with cut-off being 0.05. This plot was used as a regional LD and minor allele frequency indicator for both chromosome 19 and further mentioned chromosome 6.

The SNP "snp23995.scaffold244.354162 (rs268256403)" is located at ~15 Kb downstream of ALOX12 gene region with a p-value of 4.768×10^{-23}. Homologous ALOX12 (Arachidonate 12-lipoxygenase) gene is positioned on the 19th chromosome of the cow genome (again closely located with a variant of RNASEK) and already associated with biological processes such as; arachidonic acid metabolic process (GO: 0019369), fatty acid metabolism (GO:00119395), linoleic acid metabolic process (GO: 0043651) and lipoxygenase pathway (GO: 0019372) (Ensembl, 2016; NCBI). It is located on the 11th chromosome of sheep genome and

in a similar way majorly associated with fatty acid metabolism, which nominates the gene region as a strong candidate for manipulating milk composition and therefore yield itself. In a study with dairy cattle, milk fat globules were targeted via RNA-seq technology to detect differentially expressed genes between the cows that are remarkably differing regarding milk production performance (Yang et al., 2016). ALOX12 was detected as differentially expressed.

The SNP "snp23997.scaffold244.438286 (rs268256404)" was found to be positioned within ASGR2 (asialoglycoprotein receptor 2) gene with a p-value of 2.829×10^{-20}. Homologs of this gene are also located on 19^{th} and 11^{th} chromosomes of cow and sheep genomes respectively and characterised by involving in regulation of protein stability (GO: 0031647), glycoprotein metabolic process (GO: 0009100) and lipid homeostasis (GO: 0055088) by Ensembl. As it is known, milk yield is strongly correlated with milk protein yield and lipid metabolism. Furthermore glycoproteins comprises a significant share of total milk protein, therefore this gene can be seen as another strong candidate to the observed variation in milk yield and can be suggested for further analysis.

The SNP "snp10603.scaffold1377.32250 (rs268243458)" reached genome-wide significance with a p-value of 2.710×10^{-18}. It is an intron variant of KIF1C gene which is characterised for vesicle-mediated transport (GO: 0016192) and wide range of metabolic processes (GO: 0008152) in cattle (19^{th} chromosome) and sheep (11^{th} chromosome). Furthermore another genome wide significant (nearly the same p-values and other characteristics of KIF1C gene; for all four production traits) SNP, "snp10606.scaffold1377.151388 (rs268243460)", was located within RABEP1 gene which also involves in vesicle mediated transport. However since the two regions have only ~100 Kb distance in between, it was not further discussed here. Milk components such as lipids, proteins and lactose are produced within cell and transported to the apical cell membrane with secretory vesicles (Mcmanaman and Neville, 2003). Therefore, variation regardingthese two genes may suggest another candidate region for further investigation.

Another variant reached genome-wide significance for milk yield was

"snp24012.scaffold244.1259949 (rs268256419)"with a p-value of $2.124x10^{-16}$. It is an intragenic variant of ALOXE3 gene on the 19th chromosome. Just as ALOX12 gene protein, it involves in fatty acid metabolism with similar activities on to 19th and 11thchromosomes of cow and sheep genomes respectively.

The last variant associated with milk yield was a chromosome-wide significant SNP "snp32593.scaffold374.339266 (rs268264775)". It is an inter-genic variantwith a p-value of $3.879x\ 10^{-5}$, which is just under the defined p-value threshold for chromosome 21. Nearest gene regions are MEX3B (~40 Kb upstream) and EFTUD1 (~ 120 Kb upstream). In cattle genome, 10 Mb closeby homologous region on 19thchromosome has been associated to milk yield with both association studies and linkage mapping studies however majorly with different genes (Wang et al., 2011; Oikonomou et al., 2011). In the sheep genome, linkage mapping studies resulted in a QTL region between 13.6-38.1 Mbp, covering also significant region in our study, on chromosome 11 (Jonas et al., 2011).

3.2.3.2. Milk fat yield

Top SNP obtained from milk yield was kept still as significant with a higher p-value ($3.898x10^{-15}$) along with SNPs in high LD with itself. The effect of top associated SNP was showing almost 4-fold increase in the same direction, comparing the results for milk yield and milk fat yield. This has a significant implication for implementing genomic selection. Regarding the market demand for milk fat composition, a selection aiming higher or lower fat content of milk with integrating this SNP to the breeding programme will have almost 4 times less effect on milk yield itself. For example, if lower fat content was aimed, fat content of milk will decrease four times quicker than milk yield.

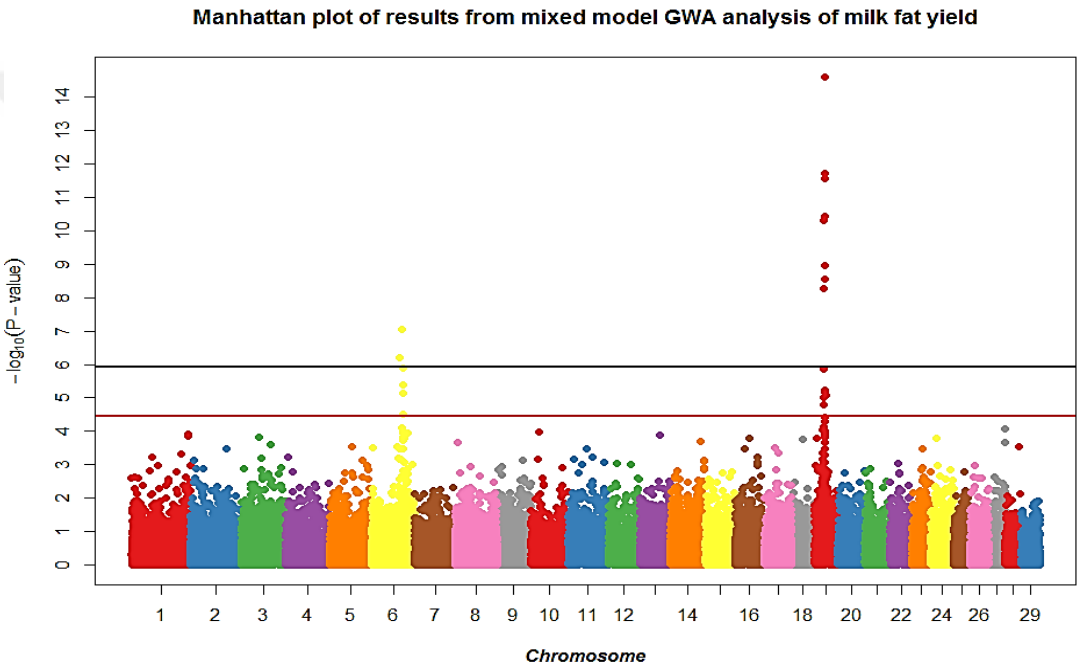
The estimated effect size (0.896 ± 0.126) of ALOX12 variant was threefold higher than that observed for milk yield as well as being in the same direction (p-value= $1.750x\ 10^{-11}$.

The magnitude of the SNP effect size of the same variant found in ASGR2 gene was estimated to be 0.834 ± 0.119 (p-value= 2.394x 10^{-11}).

The variant on ALOXE3 gene was observed with an effect size of 0.739 ± 0.121 and a p-value of 5.725x10^{-09}.

The intron variant of KIF1C gene showed an effect size of 0.808 ± 0.123 with a p-value of 3.440 x 10^{-10}.

Figure 12:Manhattan plot of milk fat yield association analysis



Another region with a genome-wide significant SNP was found to be located on chromosome 6 with SNPs in high LD in the surrounding also exceeding chromosome-wide significance. Top SNP associated was "snp36226.scaffold433.1563136 (rs268268336)" with a p-value of 3.165x10^{-7}. The effect size (B allele) was obtained as-0.926 ± 0.173, which is in the opposite direction (decrease) with SNPs on chromosome 19. It is close located (~ 80 Kb downstream of the nearest gene in the described region) to a region that encapsulates a dozen of casein coding genes, signal transduction genes and uncharacterised transcript variants. It is a ~2.7 Mb region between 81.3 - 84 Mb that

was visualised in terms of significance, LD structure and MAF in Figure 13.The other SNPs exceeding chromosome-wide significance were found to be spread throughout the issued region.

Next genome-wide significant SNPon chromosome 6 was "snp59492.scaffold980.1308491 (268290910)" with a p-value of 1.708 x 10^{-6}. It is located just ~20 Kb upstream of SLC4A4 geneThis gene was annotated to be involving in ion transport process in both cattle and sheep genomes (Ensembl, 2016).

SNP "snp36248.scaffold433.2496408 (rs268268358)" was identified as being significant at chromosome-wide level with an effect size of (B allele) -0.723 ± 0.156 (decrease) and p-value of 1.034x 10^{-5}. It is an inter-genic variant that is closely surrounded by uncharacterised transcript variants. The nearest casein coding genes are CSN1S1, CSN2 and CSN1S2, which are at ~450 Kb, ~480 Kb and ~550 Kb respectively upstream of the SNP. Supportingly, a SNP "snp59447.scaffold980.398647 (rs268293091)" located within CSN1S2 has reached p-value of 2.992x 10^{-5}. It is above overall chromosome-wide threshold in figure 13, but not within the specified chromosome. Genes in this casein cluster (82.6 - 82.9 Mb of chromosome 6) have been repeatedly found to be significantly associated with a wide range of goat milk traits, including fat and protein content, and broadly reviewed in (Amills, 2014). Regarding the SNPs located within or closer vicinity of casein coding genes, their minor allele frequencies in our study population are remarkably low in comparison with other SNPs within ~4 Mb region presented in Figure 13. Therefore it could be interpreted that, those SNPs might not reach any significance threshold, due to lack of power sourced by low minor allele frequencies.

Figure 13: Regional significance, LD structure and MAF of SNPs on chromosome 6.

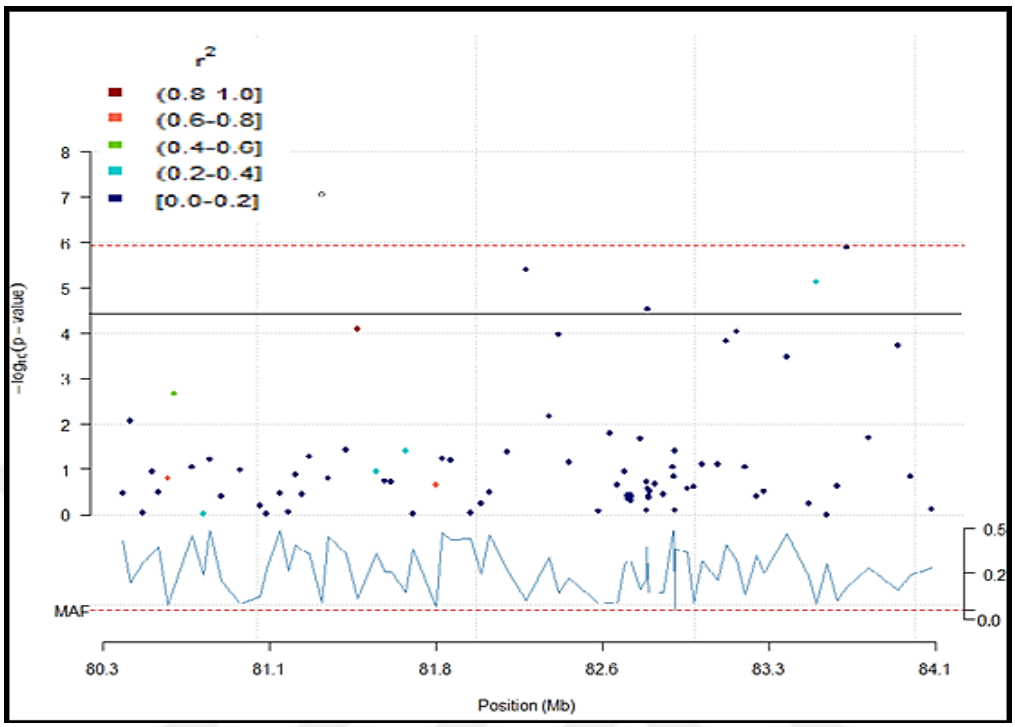


Fig. 13: Different from what is described above in Figure 12, black line represents chromosome-wide threshold for significance of association. The map is not centred on top associated SNP, however it is the only one exceeds genome-wide significance threshold (red dashed line). Area between 82.6 Mb and 82.9 Mb is where casein coding gene- rich regions located.

Last chromosome-wide significant SNP "snp9004.scaffold1328.341088 (rs268241909)" was on chromosome 19 and quite close located to a gene called SHISA6 with a p-value of 2.047×10^{-5}. The same SNP was found significant for only fat and lactose yield, whereas variants described until now were significant for all four production traits. This pottentially provides a closer connection regarding the pathways affecting milk fat and lactose yield. However, no literature entry that could reason this connection was found for the gene of interest.

3.2.3.3. Milk lactose yield

GWAanalysis of milk lactose yield resulted in the same region and top-associated SNP (p-value $1.450x10^{-15}$) on chromosome 19 with abovementioned traits as well as a nearly chromosome-wide significant SNP on chromosome 3.

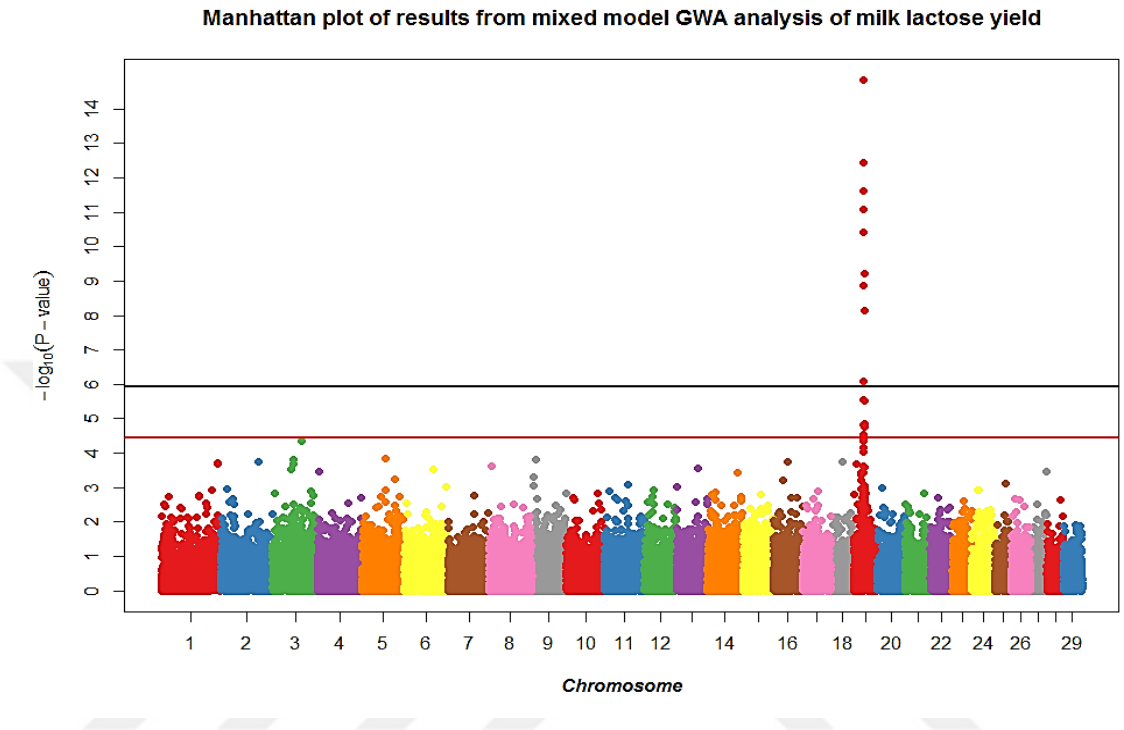
The estimated effect size of SNP variant close located to ALOX12 was the highest for the analysed traits as well as being in the same direction (P= $3.650 x10^{-13}$).The effect estimateof ASGR2 variant shows similar trend with milk fat yield with a p-value of $2.459 x 10^{-12}$. The variant from ALOXE3 gene reached to p-value of $5.941x10^{-10}$.KIF1C gene intron variant had a p-value of $3.883x10^{-11}$

In general, the SNP effectsobserved among composition traits (fat, lactose and protein) regarding common variants showed a pattern of being almost the same between lactose and fat yield, whereas considerably higher than those obtained with protein yield GWA analysis. Furthermore, those effects are also at least threefold higher than effects obtained with milk yield. Thus, selection for lower lactose and fat yield via incorporating these markers to the selection index would have a considerably less effect on average milk yield (almost three times slower decrease) and protein yield (~1.5 time slower decrease).

One SNP variant that is worthy of further discussion is “snp54293.scaffold83.702798 (rs268285914)” (P= $4.408x10^{-05}$). It is located within PDE4B (Phosphodiesterase 4B) gene coding region which is associated with signal transduction (GO: 0007165) via cAMP regulation, metabolic process (GO: 0008152) and some other key processes in cattle (chromosome 3) and sheep (chromosome 1) genome (Ensembl, 2016). A variant of this gene, PDE9A (phospodiesterase9A), has been validated to be strongly associated with milk production traits in Chinese Holstein cattle (Yang et al., 2015). And another variant PDE3A has been found to be associated with milk production traits in Buffaloes (Venturini et al., 2014). Therefore this gene region can

be suggested as being a candidate region that may have an effect on milk lactose yield in dairy goats.

Figure 14: Manhattan plot of milk lactose yield association analysis

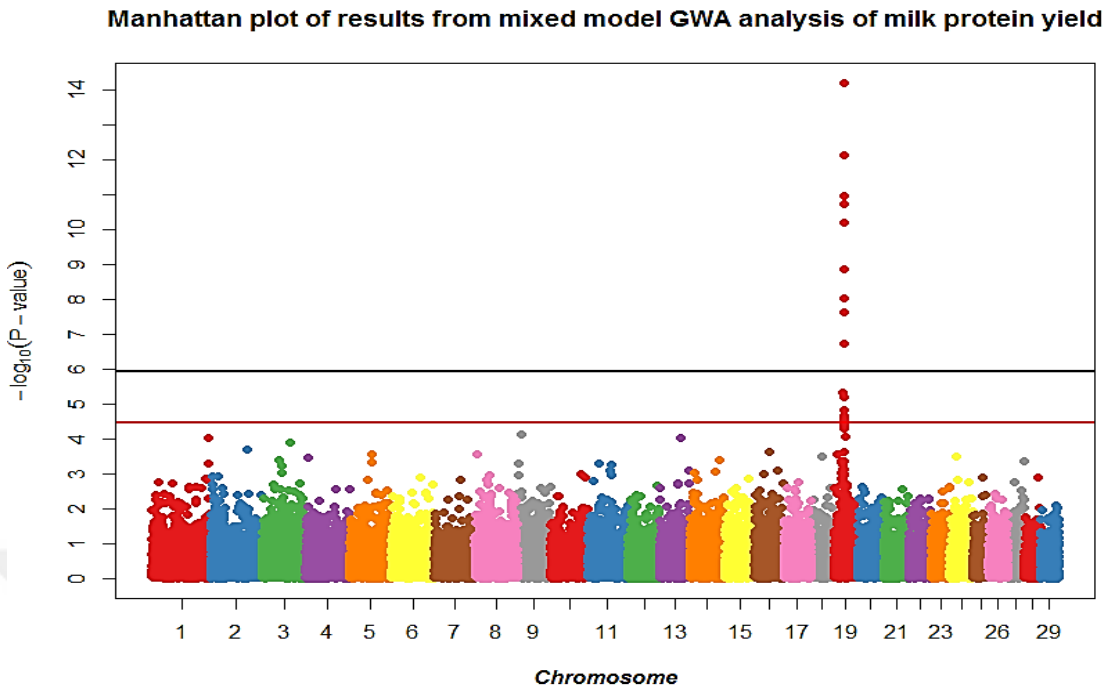


3.2.3.4. Milk protein yield

There was no significant association other than that observed region on chromosome 19 in milk yield. Most variants, as it was expected for all the analysed production traits (due to the high genetic correlation) kept their significance. However, two new variants located in the same gene within the pre-defined region on chromosome 19reached to the genome-wide significance.

The top associated variant had a p-value of $6.222x10^{-15}$. P-values of $7.443x10^{-13}$,$1.777x\ 10^{-11}$, $1.301x10^{-09}$ and $6.232x\ 10^{-11}$ were obtained for variants related with ALOX12, ASGR2, ALOXE3, KIF1C respectively.

Figure 15: Manhattan plot of milk protein yield association analysis

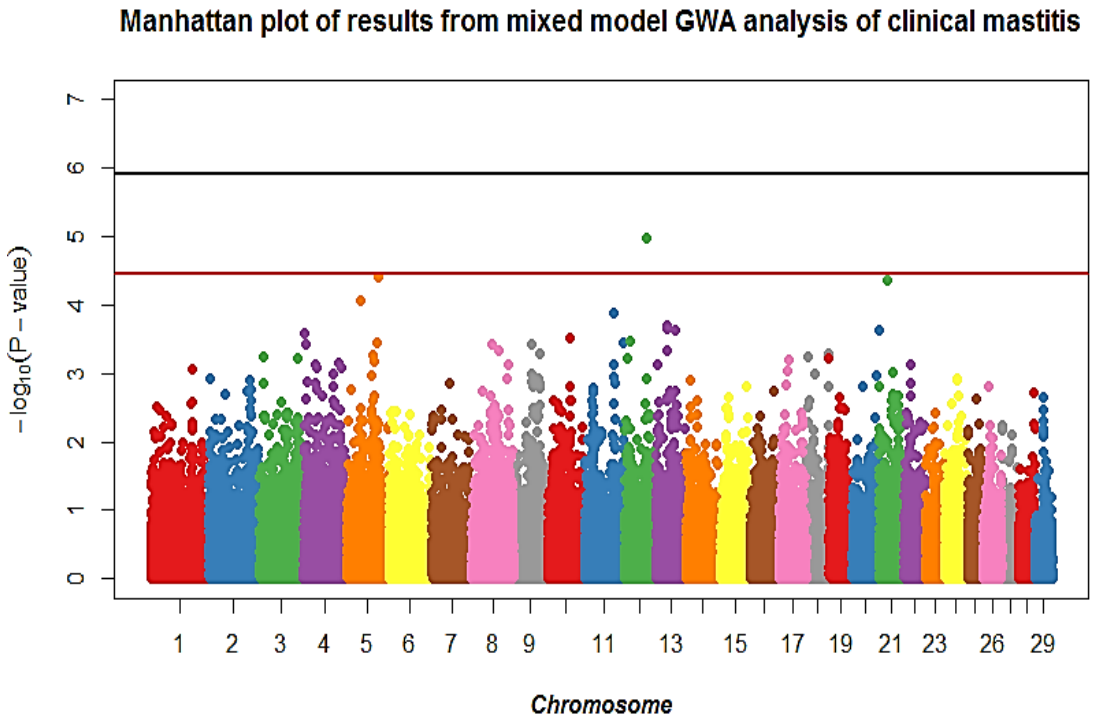


The GWA analysis withmilk protein yield resulted in two genome-wide significant variants different than for the other traits. Two of these (snp24008.scaffold244.1010180 (rs268256415), snp24006.scaffold244.948787 (rs268256413)) were located in the same gene with 5 Kb spacing, DNAH2 (Dynein Axonemal Heavy Chain 2). Effect sizes estimated were almost the same as well as p-values, which were 0.339 ± 0.017 for B allele and 2.205×10^{-05}.The fact that the two variants being in the same protein coding gene and arising genome-wide significance only in GWA analysis of milk protein yield prepares a safe ground for the significance of this gene on MPY related pathways. DNAH2 gene has been annotated to be involving in microtubule-based movement (GO: 0007018) and metabolic process (GO: 0008152) for both cattle (chromosome 19) and sheep genome (chromosome 11). However, no futher studies was observed regarding its significance for milk protein yield. It is suggested for further analysis (such as fine mapping).

3.2.3.5. Clinical Mastitis

Analysis of clinical mastitis resulted in one chromosome-wide significant and two suggestive region that is pretty close to the chromosome-wide significance. Top significant SNP "snp19179.scaffold193.159151 (rs268251725)" is an inter-genic region surrounded by SLITRK family protein coding genes (1.2 Mb downstream of SLITRK1 gene and 0.5 Mb upstream of SLITRK6 gene). The effect size of (B allele) 0.031 ± 0.007 and p-value of 1.065×10^{-5} was observed, meaning that B allele significantly increases risk of getting the disease. SLITRK1 takes place in homeostatic process (GO: 0042592) and together with SLTRK6 in nerve system development in both cattle (chromosome 12) and sheep genome (chromosome 10). As for some of the previous traits, there is no QTL region previously linked to this region.

Figure 16: Visualisaling GWA analysis with clinical mastitis



First suggestive SNP obtained was an intragenic variant ("snp2436.scaffold107.6510838 (rs268235499)")of an uncharacterised transcript variant (LOC106502131) on the chromosome 5 with a p-value $3.907x10^{-5}$(threshold P= $2.500x10^{-5}$). Furthermore, RASSF8 gene is located ~50 Kb upstream of the obtained region. In cattle genome regarding homologous region (~10 Mb surrounding), GWAS studies of clinical mastitis resulted in detected QTL on cattle chromosome 5 (Wu et al., 2015).

The last suggestive SNP, "snp32588.scaffold374.123703 (rs268264770)" had a p-value of $4.325x10^{-5}$ and is located on chromosome 21, where the chromosome specific p-value threshold for significance was $3.900x10^{-5}$.~400 Kb downstream and upstream of the suggestive SNP is occupied with HOMER2 and EFTUD1 genes respectively. In cattle chromosome 21, ~4 Mb downstream of the homologous region has been characterised with a QTL for clinical mastitis (Schulman et al.,2004). Interestingly, chromosome 21 was suggested to have a significant region for SCC in goats (Rupp et al., 2014). However, since the details regarding the region has not been published, further investigation could not be carried out.

It is also interesting to note that obtained region is closely located to the SNP (rs268264775) obtained as chromosome wide significant from milk yield GWAS analysis. Since an antagonistic relationship between milk yield and clinical mastitis is widely known and also mentioned earlier, especially this region might be taken as a bridge to unveil relationship between clinical mastitis and milk yield.

Table 7. Genome wide significant SNPs associated with the traits

Chr	SNP	Position (bp)	Trait	Highest p-value *	Effect Allele	EAF **	Effect Size	Variance Explained ***	Cand. gene	Distance
19	rs268292132	26098302	MY, MFY MLY,MPY	6.222×10^{-15}	A	0.394, 0.389 0.389, 0.389	0.291, 0.998 1.013, 0.704	0.090, 0.040, 0.046, 0.045	RNASEK	Exon (Syn.)
19	rs268256403	26066457	MY, MFY MLY, MPY	1.750×10^{-11}	A	0.393, 0.387 0.387, 0.387	0.269, 0.896 0.931, 0.653	0.075, 0.024 0.039, 0.037	ALOX12	~15 Kb d.stream
19	rs268256404	26150581	MY, MFY MLY, MPY	2.394×10^{-11}	B	0.418, 0.412 0.412, 0.412	0.235, 0.834 0.841, 0.574	0.061, 0.030 0.032, 0.030	ASGR2	Intron variant
19	rs268256419	26972244	MY, MFY MLY, MPY	5.725×10^{-09}	A	0.420, 0.415 0.415,0.415	0.213, 0.739 0.759, 0.526	0.022, 0.023 0.023, 0.025	ALOXE3	Intron variant
19	rs268243458	25679946	MY,MFY MLY, MPY	3.440×10^{-10}	B	0.417, 0.411 0.411, 0.411	0.229, 0.808 0.819, 0.576	0.056, 0.023 0.030, 0.030	KIF1C	Intron variant
6	rs268268336	81297166	MFY	3.165×10^{-7}	B	0.092	- 0.926	0.012	CENPC	~80 Kb d.stream
6	rs268290910	8369916	MFY	1.708×10^{-6}	A	0.172	-0.646	0.010	SLC4A4	~20 Kb d.stream
19	rs268256415	26722475	MPY	2.205×10^{-5}	B	0.490	0.339	0.010	DNAH4	Intron variant

* Regarding the p-values for different traits; ** Effect allele frequency ; *** Proportion of additive genetic variance explained by each SNP

Table 8. Chromosome wide associated and suggestive SNPs

Chr	SNP	Position (bp)	Trait	Highest p-value *	Effect Allele	EAF **	Effect Size	Variance Explained ***	Cand. gene	Distance
6	rs268268358	82230438	MFY	1.034×10^{-5}	B	0.103	- 0.723	0.008	CSN1S1	~400 Kb
6	rs268293091	82782072	MFY	2.992×10^{-5}	B	0.220	- 0.530	0.008	CSN1S2	Intron variant
21	rs268264775	21656080	MY	3.879×10^{-5}	A	0.092	0.140	0.007	MEX3B EFTUD1	~40 Kb ~120 Kb
3	rs268285914	76032437	MLY	4.408×10^{-5}	B	0.182	- 0.468	0.006	PDE4B	Intragenic variant
19	rs268241909	29181149	MFY, MLY	2.047×10^{-5}	A	0.250	0.461, 0.452	0.007, 0.007	SHISA6	~70 Kb upstream
12	rs268251725	57974104	CM	1.065×10^{-5}	B	0.119	0.031	0.134	SLITRK genes	~1.2 Mb, ~0.6 Mb
5	rs268235499	81537111	CM	3.907×10^{-5}	B	0.154	0.025	0.108	Unchar. transcript variant	Intron variant
21	rs268264770	21440517	CM	4.325×10^{-5}	A	0.081	0.033	0.108	HOMER2 EFTUD1	~400 Kb up and d.stream

* Regarding the p-values for different traits ** Effect allele frequency ; *** Proportion of additive genetic variance explained by each SNP

Overall,31 % of the total genetic variance for milk yield was explained by the associated SNPs. This figure was 18 % for milk fat yield, 15% for milk lactose yield, similarly 18%for milk protein yield and 35%for clinical mastitis, which shows there is still huge amount of variants segregating. Unexplained genetic variance might be due to SNPs with small effects or low minor allele frequencies. In this study population LD was relatively high, however there is still the chance of typed markers not capturing all the genetic variance.

All four production traits has shown a similar genetic background. Obtaining high number of SNPs in common among the milk production traits and to be in the same direction supports strong positive genetic correlation between the traits to be the case.

The effect sizes of the associated SNPs show a trend between fat yield and lactose yield, whereas they are overall higher than protein yield regarding the magnitude of the effect sizes.

In general variants found on chromosome 6 with negative effect sizes had low effect allele frequencies. The fact that they had low frequencies supports the hypothesis that most casein coding genes did not reach any significance level due to low minor allele frequencies. One explaination to this might be the fact that the analysed population is selected for best performing animals in terms of milk yield.

3.3. Possible further steps and considerations

All SNP effects estimated here assume an additive model for the analysed markers. Therefore a further step can be to test this obtained SNPs in terms of other genetic models such as dominance and reccesive.

Allelic coding used here was “A/B” coding. However taking another step and changing it into “A/C/T/G” could render allelic coding biologically more meaningful and possibly more helpful for further usage of the results (such as for genomic selection and RFLP).

Another useful approach would be to run an epistasis scan of whole genome for each associated SNPs and see the interaction web for each SNP in terms of genotypic states. This could be done easily with “cgmisc” package of R environment and p-values for epistatis between each SNP and the rest of the SNPs can be obtained. Furthermore interaction regarding genotypes of two particular SNPs can be visualised to see the direction of interaction with the same R package.

These particular regions can be sequenced for region-focused association analyses. Furthermore, functional properties of protein variants of these genes can be further investigated via proteomic techniques in terms of association with milk production pathways.

Finally, obtained regions can positionallybe amplified via appropriately designed primers and further studies with RFLP (restriction fragment length polymorphism) can be carried out with cheaper molecular biology techniques-for the sake of cost-effectiveness- on thousands of animals rather than genotyping all animals.

3.4. Applications of the obtained information

Once results are proved biologically, obtained genetic regions could be incorporated into the breeding programmes of potentially different goat breeds, since the analysed population was a mixed-breed. This could be done via marker assisted selection, aiming for goat genetic improvement in terms of disease resistance, increased milk yield and/or increased milk quality via manipulating milk fat, lactose and protein content regarding the demand from targeted market. Effectiveness of obtained regions on different goat breeds also depends on the allele frequencies of aimed population regarding delivered regions. Some populations might be fixed in terms of targeted allele, which in turn gives no genetic gain with further selection.

The variants found unique to one trait can show a great economic value regarding the market demand for milk composition. Once the SNP effects are ‘validated’ biologically, best way of utilizing the obtained results is to incorporate markers with

aimed effects to the breeding programmes via selection indices and implement selection with regards to economic values of the traits.

Cheaper molecular techniques could be easily applied to the obtained significant regions with the aim of understanding nature of milk composition traits and disease resistance.

4. CONCLUSIONS

On average ~7500 animals for milk production traits and 3968 animals for clinical mastitis were used in GWA analyses. In terms of statistical power, this study was considered to be one of the most powerful studies in the goat genomics field, regarding the sample size being analysed.

Here, genome-wide association studies for milk yield, milk fat yield, milk lactose yield, milk protein yield and clinical mastitis have been conducted. Despite the fact that conservative quality control and significance thresholds were applied, various number of genome and chromosome wide significant regions were obtained for the analysed traits. Regions obtained for milk production traits showed concordance among each other. Linkage disequilibrum analyses suggested high overall LD for population considered with LD decay being rapid with short distances but slow and persistent for long distances. LD was considered to be in a relatively effective margin, up until distances of ~ 2 Mega base pair between markers.

High regional LD could be leading to suggest that all significant SNPs in the two focused regions (chromosome 6 and 19) might be as a result of being dragged by truly associated SNP or SNPs. However, as it is well known high LD may arise from intensive selection on the same trait over generations. Furthermore, if the genes affecting the trait of interest are located quite closely to each other, this could lead the regional accumulation of high LD. Therefore, each regions were carefully inspected in terms of candidacy for further analysis.

ALOX12,, ASGR2, CSN1S1 (including CSN2, CSN1S2), PDE4B, DNAH2, HOMER2 and EFTUD1 along with other aforementioned genes regarding different traits were suggested as candidate genes for further analysis (such as fine mapping).

Finally, it should be noted that issued regions for all milk production traits are rich in terms of coding genes, therefore there might still be other genes or variants segregating but could not be detected by these analyses. All of these particular regions and genes are proposed by considering the position of associated variant, the biological processes being involved and their associations with related traits. Furthermore, it is worth remembering that regions issued under milk production traits are high in LD in general, therefore some associated SNPs might be significant due to being dragged by other SNPs. It is highly important to know that, there still might be variants segregating with these traits, however they might not be discovered due to low effect size, low minor allele frequency and/or genetic variance capturing capacity of 50 K SNP chip.

5. REFERENCES

Abecasis, G.R., Cardon, L.R., Cookson, W.O.C. (2000). A general test of association for quantitative traits in nuclear families. Am J Hum Genet, 66, pp. 279–292

Amills M. (2014). The Application of Genomic Technologies to Investigate the Inheritance of Economically Important Traits in Goats. Advances in Biology, vol. 2014, Article ID 904281. doi:10.1155/2014/90428

Astle, William; Balding, David J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. Statist. Sci. 24, no. 4, 451--471.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics 23: 1294–129

Aulchenko, Y. S., D.-J. de Koning and C. Haley. (2007a). Genome-wide rapid association using mixed model and regression: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. Genetics 177: 577–585.

Aulchenko, Y.S., Karssen, L.C. (2015). The GenABEL project developers. The GenABEL Tutorial. Zenodo; 2015.

Bergonier, D., De Cr´emoux, R., Rupp, R., Lagriffoul, G., Berthelot, X. (2003). Mastitis of dairy small ruminants. Veterinary Research, BioMed Central, 34 (5), pp.689- 716.

Brito, L., Silva, F., Melo, A., Caetano, G., Torres, R., Rodrigues, M., & Menezes, G. (2011). Genetic and environmental factors that influence production and quality of milk of Alpine and Saanen goats.*Genetics and Molecular Research : GMR, 10*(4), 3794-802.

Brito, L. F., Jafarikia, M., Grossi, D. A., Kijas, J. W., Porto-Neto, L. R., Ventura, R. V., … Schenkel, F. S. (2015). Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats. *BMC Genetics*, *16*, 67.

Bush, S., Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies (F. Lewitter and M. Kann, Eds.). PLoS computational biology, 8(12): e1002822.

Carillier, C., Larroque, H., Palhière, I., Clément, V., Rupp, R., Robert-Granié C. (2013). A first step toward genomic selection in the multi-breed French dairy goat population. J. Dairy Sci. 96:7294–7305.

Castañeda-Bustos, Montaldo, Torres-Hernández, Pérez-Elizalde, Valencia-Posadas, Hernández-Mendo, & Shepard. (2014). Estimation of genetic parameters for productive life, reproduction, and milk-production traits in US. J. Dairy Sci; 97(4): 2462-73.

Chen and Abecasis. (2007). Family-Based Association Tests for Genomewide Association Scans. *The American Journal of Human Genetics, 81*(5), 913-926.

Chetroiu et al. (2013). *Worldwide trends and orientations of raising goats.* Agrarian Economy and Rural Development - Realities and Perspectives for Romania, Vol. 4, No. ISSN – 2285-6803; ISSN – L – 2285-6803 (21. November 2013): pp. 100-106

Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.***37**, 1243–1246

Contreras, A., Sierra, D., Corrales, J.C., Sánchez, A., Marco, J. (1996).Physiological threshold of somatic cell count and California mastitis test for diagnosis in caprine subclinical mastitis Small Rumin. Res., 21, pp. 259–264

Crow J. F. (1988). Eighty years ago: the beginnings of population genetics. Genetics, 119(3): 473–476

DEFRA. (2014). Farming Statistics Final crop areas, yields, livestock populations and agricultural workforce At June 2014 - United Kingdom

Devlin, B. and Roeder, K. (1999). Genomic Control for association studies. *Biometrics* 55, 997-1004
Do, D., Strathe, A., Ostersen, T., Jensen, J., Mark, T., & Kadarmideen, H. (2013). Genome-wide

association study reveals genetic architecture of eating behavior in pigs and its implications for humans obesity by comparative mapping. *PloS One, 8*(8), E71509.

Dong Y, Xie M, Jiang Y, Xiao N, Du X, et al. (2013). A reference genome of the domestic goat (Capra hircus) generated by Illumina sequencing and whole genome mapping. Nature Biotechnology 31: 135–141.

Eu-ahsunthornwattana, J., Miller, E. N., Fakiola, M., Wellcome Trust Case Control Consortium 2, Jeronimo, S. M. B., Blackwell, J. M., & Cordell, H. J. (2014). Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. *PLoS Genetics*, *10*(7), e1004445.

FAO. 2013. Enhancing animal welfare and farmer income through strategic animal feeding − Some case studies. Edited by Harinder P.S. Makkar. FAO Animal Production and Health Paper No. 175. Rome, Italy

FAOSTAT. (2014). (http://faostat3.fao.org/home/E).

Fikse, W.F. & Banos, G. (1997). Alternative weighing factors for daughter information in international genetic evaluations: progress report. Interbull bulletin, 16, 3-7.

Garrick, D. J., Taylor, J. F., & Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses.*Genetics, Selection, Evolution : GSE*, *41*(1), 55.

Gibson, Greg. (2010). Hints of hidden heritability in GWAS. *Nature Genetics*. **42** (7): 558–560.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., and Thompson, R. (2009). ASReml User Guide Release 3.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK

Gower JC. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338

Hao K, Di X, Cawley S. (2006). LdCompare: rapid computation of single- and multiple-marker r2 and genetic coverage. Bioinformatics, 23: 252-254.

Hardy G. H. (1908). Mendelian proportions in a mixed population. Science, 28(706): 49–50

Henderson, CR. (1975). "Best linear unbiased estimation and prediction under a selection model". Biometrics 31 (2): 423–447.

Henderson, C.R. (1984). Application of Linear Models in Animal Breeding (University of Guelph, Guelph, Canada.

Hill, W.G., Robertson A. (1968). Linkage disequilibrium in finite populations. Theor Appl Genet., 38: 226-31

Kierczak, M., Jabłońska, J., Forsberg, S. K. G., Bianchi, M., Tengvall, K., Pettersson, M., Lindblad-Toh, K. (2015). cgmisc: enhanced genome-wide association analyses and visualization. *Bioinformatics*, *31*(23), 3830–3831

Kirkpatrick, Lofsvold, & Bulmer. (1990). Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics, 124*(4), 979-93.

Kitchen B. (1981). Review of the progress of Dairy Science: Bovine mastitis: milk compositional changes and related diagnostic tests. *Journal of Dairy Researc* 48, 167-18

Koop, G. et al. (2010).The effect of subclinical mastitis on milk yield in dairy goats Journal of Dairy Science , Volume 93 , Issue 12 , 5809 – 5817

Jonas, E., Thomson, P. C., Hall, E. J., McGill, D., Lam, M. K., & Raadsma, H. W. (2011). Mapping quantitative trait loci (QTL) in sheep. IV. Analysis of lactation persistency and extended lactation traits in sheep. *Genetics, Selection, Evolution : GSE*, *43*(1), 22.

Leitner G., Merin U., Silanikove N. (2004). Changes in milk composition as affected by subclinical mastitis in goats. J Dairy Sci. 87, 1719-1726.

Leitner, G., Merin, U., Lavi, Y., Egber, A., Silanikove, N. (2007). Aetiology of intramammary infection and its effect on milk composition in goat flocks. J. Dairy Res. 74, 186–193

Lidauer, M., K. Matilainen, E. Mantysaari and I. Straden. (2011). MiX99 Manual. MTT, Jokioinen, Finland

Maatje, K., Rossing, W., Garssen, G. J., Pluygers, H.G. (1983). Automation of electrical conductivity measurements during milking. Page 89 in Proc. 3rd Symp. Automation in Dairying, Wageningen, Neth.

Manser, PA. (1986). Prevalence, causes, and laboratory diagnosis of subclinical mastitis in the goat. Vet. Rec. 118, 552-554.

Mcmanaman JL, Neville MC. (2003). Mammary physiology and milk secretion. Adv Drug Deliv Rev 55:629–64

Montaldo, H. H., and E. Manfredi. (2002). Organisation of selection programmes for dairy goats. Pages 1–8 in Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. Institut National de la Recherche Agronomique (INRA), Paris, France.

Mrode,R.A.,and Swanson,G.J.T. (1996).Genetic and statistical properties of somatic cell count and its suitability as an indirect means of reducing the incidence of mastitis in dairy cattle. *Anim.Breed.Abstr.* 64, 847–857.

Mucha, S., Mrode, R., Coffey, M., & Conington, J. (2014). Estimation of genetic parameters for milk yield across lactations in mixed-breed dairy goats. *Journal of Dairy Science, 97*(4), 2455-2461.

Mucha, S., Mrode, R., MacLaren-Lee, I., Coffey, M., & Conington, J. (2015). Estimation of genomic breeding values for milk yield in UK dairy goats. *Journal of Dairy Science*, *98*(11), 8201-8208. doi:10.3168/jds.2015-9682

Mucha S., Mrode R., Coffey M., Conington J. (2016). Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. PAG conference, San Diego, USA

Nielen, M., Schukken, Y.H., Brand, A., Haring, S., & Ferwerda-Van Zonneveld, R.T. (1995). Comparison of Analysis Techniques for On-Line Detection of Clinical Mastitis.*Journal of Dairy Science, 78*(5), 1050-1061.

Oikonomou G, Michailidis G, Kougioumtzis A, Avdi M (2011): Effect of polymorphisms at the *STAT5A* and *FGF2*gene loci on reproduction, milk yield and lameness of Holstein cows. Research in Veterinary Science, 91, 235–239. doi: 10.1016/j.rvsc.2011.01.009.

Ostersen, T., O.F. Christensen, M. Henryon, and B. Nielsen. (2011). Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet. Sel. Evol. 43:38.

Paape M.J., Capuco A.V. (1997). Cellular defense mechanisms in the udder and lactation of goats. J Anim Sci. 75 (2): 556-565.

Price, A.L. et al. (2006). Principal component analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904-909

Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170-181

R Development Core Team (2006). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/

Rannala, B. and Reeve, J. P. (2001). High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am. J. Hum. Genet. 69: 159-178.

Risch, N., & Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science, 273*(5281), 1516-1517.

Rupp, R. and Boichard, D. (2003).Genetics of resistance to mastitis in dairy cattle. *Vet*. *Res*. 34,671–688.doi:10.1051/vetres:2003020

Rupp, Clément, Piacere, Robert-Granié, & Manfredi. (2011). Genetic parameters for milk somatic cell score and relationship with production and udder type traits in dairy Alpine and Saanen primiparous goats. *Journal of Dairy Science, 94*(7), 3629-3634.

Rupp R., Palhière I., Maroteau C., Balloche G., Sallé G., Tircazes A., C. Moreno C., Foucras G. and Tosser-Klopp G. (2014). Mapping QTL Controlling Milk Somatic Cell Counts in Sheep and Goat Support the Polygenic Architecture of Mastitis Resistance. Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Schulman, N.F. et al. (2004). Quantitative Trait Loci for Health Traits in Finnish Ayrshire Cattle. Journal of Dairy Science, Volume 87, Issue 2, 443 – 449

Shin, J., Blay, S., Graham, J., & Mcneney, B. (2006). LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software, 16* (Code Snippet 3), Journal of Statistical Software, 2006, Vol.16 (Code Snippet 3).

Spielman, R.S. and Ewens, W.J. (1998). A sib-ship test for linkage in the presence of association: the sib transmission/disequilibrium test Am J Hum Genet, 62, pp. 450–458

Taylor, J. (2012). Current status of genomic selection in US Beef Cattle. XX Plant and Animal Genome Conference San Diego, USA, 14-18 January 2012

The 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*(7145), 661–678.

Tolone M, Riggio V, Portolano B. (2013). Estimation of genetic and phenotypic parameters for bacteriological status of the udder, somatic cell score, and milk yield in dairy sheep using a threshold animal model. Livestock Sci 151:134-139.

Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al. (2014). Design and characterization of a 52K SNP chip for goats. PLoS One.; 9(1): e86227. doi:10.1371/journal.pone.0086227.

UK House of Commons Library, Alex Bate. (2016). Dairy industry in the UK: Statistics. House of Commons Library Briefing Paper number 2721

VanRaden PM, Wiggans GR. (1991). Derivation, calculation, and use of national animal model information. J Dairy Sci. 74 (8): 2737-2746.

Venturini, G., Cardoso, D., Baldi, F., Freitas, A., Aspilcueta-Borquis, R., Santos, D., . . . Tonhati, H. (2014). Association between single-nucleotide polymorphisms and milk production traits in buffalo. *Genetics and Molecular Research: GMR, 13*(4), 10256-68.

Wang, Z., Wang, G., Huang, J. et al. (2011).Novel SNPs in the *ATP1B2* gene and their associations with milk yield, milk composition and heat-resistance traits in Chinese Holstein cows. Mol Biol Rep 38: 1749.

Weale, M. E. (2010). Quality control for genome-wide association studies. Methods in molecular biology (Clifton, N.J.), 628, 341-372.

Weinberg W. (1908). Über den Nachweis der Vererbung beim Menschen. Jahresh. Ver. Vaterl. Naturkd. Württemb, 64: 369–382. On the demonstration of heredity in man (Translated by S.H. Boyer, 1963)

Wu, X., Lund, M. S., Sahana, G., Guldbrandtsen, B., Sun, D., Zhang, Q., & Su, G. (2015). Association analysis for udder health based on SNP-panel and sequence data in Danish Holsteins. *Genetics, Selection, Evolution : GSE*,*47*(1), 50.

Yang, J., Jiang, J., Liu, X., Wang, H., Guo, G., Zhang, Q. and Jiang, L. (2016), Differential expression of genes in milk of dairy cattle during lactation. Anim Genet, 47: 174–180. doi:10.1111/age.12394

Yang, S.-H., Bi, X.-J., Xie, Y, Li, C., Zhang, S.-L., Zhang, Q., Sun, D.-X. (2015). Validation of *PDE9A* Gene Identified in GWAS Showing Strong Association with Milk Production Traits in Chinese Holstein. *Int. J. Mol. Sci.*, *16*, 26530-26542.

Yates, A., Akanni, W., Amode, M. R. et al. (2016). **Ensembl 2016.***Nucleic Acids Res.* 2016 44 Database issue: D710-6.PubMed PMID: 26687719; PubMed CentralPMCID: PMC4702834.

Yu, J.,G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doe-bley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness *Nat. Genet.* 38, 203-208

Zadoks, R. & Fitzpatrick, J. (2009). Changing trends in mastitis. *Irish Veterinary Journal*, *62*(Suppl 4), S59–S70

Ziegler A, König IR, Thompson JR. (2008). Biostatistical aspects of genome-wide association studies. Biom J.; 50:8–28.