MS – thesis January 2015

Genome-wide association study of muscle traits in Icelandic sheep

Ólöf Ósk Guðmundsdóttir

Faculty of Land and Animal Resources

MS – thesis January 2015

Genome-wide association study of muscle traits in Icelandic sheep

Ólöf Ósk Guðmundsdóttir

90 ECTS thesis submitted in partial fulfillment of a *Magister Scientiarum* degree in Agricultural Sciences

Academic advisors: Jón Hallsteinn Hallsson & Emma Eyþórsdóttir.

Agricultural University of Iceland
Faculty of Land and Animal Resources

Clarification of contribution

I hereby declare that the work of DNA isolation, SNP data analysis, genome-wide association analysis, primer design, PCR amplification, analysis of sequencing results, and writing of this thesis is my work under the supervision and assistance of my advisors, Dr. Jón Hallsteinn Hallsson and Emma Eyþórsdóttir MSc.

_____

Ólöf Ósk Guðmundsdóttir

Abstract

The study presented here investigated the possible association of genome-wide SNPs to a BLUP score for muscularity in the Icelandic sheep breed. Genomic DNA was extracted from 96 blood samples and genotyped for 606,006 SNPs. The selected samples were divided in five groups; sheep from Hestur experimental farm, sheep from the artificial insemination station and polled sheep from the North-west of Iceland, all of which have high muscularity and sheep from Stafholtsveggir and leadersheep both with low muscularity. The 'high muscle' and 'low muscle' sheep have significantly different BLUP scores for muscularity. Icelandic leadersheep that belonged to the 'low muscle' controls differed from the other groups in a multidimensional scaling analysis based on genomic kinship. A genome-wide association analysis was performed using both a continuous trait (BLUP score) and a binary trait with samples divided into cases and control (high vs. low-muscle). Few statistically significant SNPs were detected, but the SNPs that scored highest for association were further analyzed. Close to the highest ranking SNPs 13 genes were identified as possible candidate genes for muscularity of the Icelandic sheep. They are *CSF3R, ADAM17, GADD45B, GRID2, SPG11, DAB2, FREM3, GAB1, KLF13, AKAP6, PNN, DOCK1* and *TRRAP*. *GADD45B*, a gene involved in regulation of growth and apoptosis located on chromosome 5, was sequenced in 11 samples which were aligned and mapped to the reference gene. No variants causing a functional change of the protein were detected. More genotyped samples are needed to increase statistical significance of results and sequencing of more candidate genes is needed to locate causal variants.

Ágrip

Rannsókn þessi leitaði að mögulegum tengslum á milli BLUP einkunnar fyrir vöðvasöfnun og einbasabreytileika í erfðamengi íslensku sauðkindarinnar. DNA var einangrað úr blóðsýnum úr 96 sauðkindum og það arfgerðargreint með örflögu sem inniheldur sæti fyrir 606.006 einbasabreytileika. Sýnum var skipt í fimm hópa eftir uppruna; einn hópurinn inniheldur sýni frá tilraunabúinu á Hesti í Borgarfirði, annar sýni frá Sæðingastöðvum Vesturlands og Suðurlands og sá þriðji sýni úr kollóttu fé á Ströndum, í þessum hópum er fé sem hefur verið valið skipulega fyrir vöðvasöfnun. Hinir tveir hóparnir eru Höfðafé safnað á Stafholtsveggjum í Borgarfirði og svo forystufé safnað hér og þar um landið, í þessum hópum er fé sem hefur ekki verið valið fyrir vöðvasöfnun. Tölfræðilega marktækur munur er á BLUP einkunnum fyrir holdfyllingu skrokka á milli hópa eftir því hvort valið hefur verið fyrir vöðvasöfnun eða ekki. Við skoðun á stofngerð kom í ljós að forystuféð skar sig frá öðrum hópum á mynd sem sýnir breytileika reiknaðan út frá erfðafjarlægð milli allra gripa. Erfðamengis-tengslagreining (*e. genome-wide association study*) var framkvæmd annars vegar með BLUP einkunn sem eiginleika og hins vegar með öll sýnin skipt í tvo hópa; valið vöðvafé og óvalið fé sem tilfelli og viðmið (*e. case/control*). Fáir einbasabreytileikar voru tölfræðilega marktækt tengdir við eiginleikana en þeir sem sýndu háa fylgni voru teknir til skoðunar. Nálægt þeim einbasabreytileikum sem sýndu fylgni við vöðvastærð fundust 13 gen sem geta talist koma til greina sem gen sem hefur áhrif á vöðvastærð í íslensku sauðfé. Þessi gen eru *CSF3R, ADAM17, GADD45B, GRID2, SPG11, DAB2, FREM3, GAB1, KLF13, AKAP6, PNN, DOCK1* og *TRRAP*. *GADD45B* hefur virkni sem tengist stjórnun á vexti og frumudauða. Það var raðgreint í 11 sýnum, sýnin voru borin saman og borin við genið í viðmiðunarerfðamengi sauðfjár en enginn breytileiki sem hefur áhrif á virkni próteinsins fannst. Til þess að fá tölfræðilega marktækar niðurstöður á erfðamengis-tengslagreiningunni þyrfti að endurtaka rannsóknina með fleiri arfgerðargreindum sýnum. Til þess að finna breytingu sem gæti orsakað aukna vöðvasöfnun þyrfti að raðgreina öll genin sem teljast koma til greina.

Acknowledgements

I would like to thank my advisors, Dr. Jón Hallsteinn Hallsson and Emma Eyþórsdóttir sincerely for all their help, guidance and patience during the planning and working process of this project.

I would also like to thank Prof. Juha Kantanen from the University of Eastern Finland who assisted me with the analysis of the data. He provided me with assisting material and helped me to resolve the problems that arose during the analysis. His Ph.D. students Kisun Pokharel, Melak Weldenegodguad and Xiaoju Hu also receive my thanks for assisting with all computational work and other help.

Eyjólfur Ingvi Bjarnason and Þorsteinn Ólafsson receive great thanks for collecting blood samples with me all around Iceland for the project.

I wish to thank all the farmers that contributed samples from their sheep, without them this project would not have been possible.

My family and friends receive my thanks for support and encouragement when needed. Thanks to Hrannar Smári Hilmarsson for endless patience during my laboratory work, for sharing knowledge on practical work and for keeping me as positive as he is. Herborg Árnadóttir also receives my thanks for assistance with picture work and Jóna Björg Hlöðversdóttir and Hafþór Finnbogason for proofreading the thesis.

The Icelandic Sheep Research and Development Fund is gratefully acknowledged for financial support.

Table of Contents

List of tables

List of figures

List of abbreviations

AFLPs: Amplified fragment length polymorphism

CNV: Copy number variants

BLUP: Best linear unbiased prediction

dbSNP: Single nucleotide polymorphism database

DNA: Deoxyribonucleic acid

EBV: Estimated breeding value

EST: Expressed sequence tag

GS: Genomic selection

GWAS: Genome-wide association study

MAF: Minor allele frequency

MAS: Marker assisted selection

MDS: Multidimensional scaling

MNA: Mean number of alleles

MSTN: Myostatin

PCA: Principal component analysis

PCR: Polymerase chain reaction

QTL: Quantitative trait loci

RAPD: Random amplified polymorphic DNA

SSR: Simple sequence repeats

SNP: Single nucleotide polymorphism

1. Introduction

1.1. Sheep

Sheep (*Ovis aries*) have been under strong artificial selection by humans ever since domestication approximately 11,000 years ago (Chessa et al., 2009). They were one of the first livestock species to be domesticated along with goat and are believed to have undergone several domestication events (Meadows, Cemal, Karaca, Gootwine & Kijas, 2007; Pedrosa et al., 2005). At first they were mostly kept for meat, but later also for wool and milk (Chessa et al., 2009) and have therefore been selected for meat, fiber and milk production (The International Sheep Genomics et al., 2010), as well as for other characteristics such as color (Leymaster, 2002). The result is a multitude of sheep breeds all over the world, with a total number of 2,502 sheep breeds when counted for each country separately. In Europe alone there are 1,262 breeds although it should be noted that some breeds exist in many countries and are therefore counted more than once (Domestic Animal Diversity Information System DAD-IS, 2014). Different breeds are bred for different purposes such as reproductive traits, milk production, carcass and easy care or for many production purposes simultaneously (Leymaster, 2002). Crossbreeding systems are used to exploit the breed diversity, leading to increased productivity in the crossbred offspring. It is for example possible to use specialized sire breeds that complement characteristics of ewes from another breed, this can make the resulting offspring more productive compared to a purebred flock (Leymaster, 2002).

1.1.1 Icelandic sheep

Sheep have always been important farm animals in Iceland. The sheep population has been larger than the human population of Iceland for a long time and although the number of sheep in Iceland has decreased since it peaked in 1978-79 (Figure 1), they are still more numerous than people. Today around 2,600 registered sheep owners in Iceland keep ca. 475,000 sheep, with 92% of the population registered in the national recording system of the Farmers Association of Iceland, which contributes to the central breeding system of the breed (Landssamtök sauðfjárbænda, 2013). In 2012 the industry produced 9,900 tons of sheep meat and 1,000 tons of wool. The average meat production was 27.3 kg per ewe that year, a record high at that time (Landssamtök sauðfjárbænda, 2013).

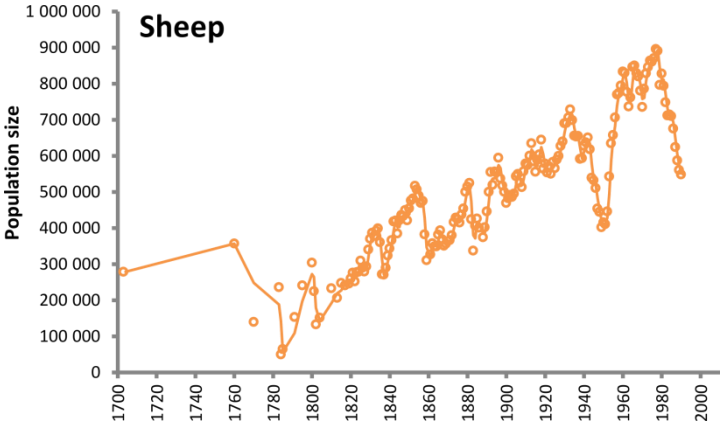

Figure 1 Population size of the Iceland sheep from 1700 to 2000 (Hagstofa Íslands, 1997).

In Iceland there is only one sheep breed, the Icelandic landrace, which is believed to have been brought there with settlers from Norway during the settling period (874-930) (Adalsteinsson, 1981). Sheep were imported to Iceland in the 18th, 19th and 20th century but in many cases the imported animals introduced new diseases which led to eradication of sheep in the infected area (Adalsteinsson, 1981). Therefore, the effect of the imported sheep on the Icelandic breed is considered to be limited (Adalsteinsson, 1981). In fact the Icelandic sheep are genetically distant to other European sheep breeds, except to the Faroe Islands sheep, Norwegian Spael; and Greenland sheep which are mostly derived from the Icelandic breed (Tapio et al., 2005). The Icelandic sheep breed belongs to a Northern-European group of short tailed sheep breeds along with other Scandinavian landraces. The old Scandinavian breeds are hardy and fertile but have less muscle mass compared to other European breeds (Eythórsdóttir, Tapio, Olsaker & Kantanen, 2002). The main breeding goal for the Icelandic sheep, ever since the start of organized breeding in the 1940s, has been to improve meat growth and carcass quality (Þorgeirsson & Þorsteinsson, 1991). Selection of sheep has mostly been based on carcass quality traits and from 1999 selection has been based on a genetic evaluation score calculated by the best linear unbiased prediction (BLUP) method (Árnason & Jónmundsson, 2008). The breed shows a great variety of color and horn forms with horned and polled sheep, very small horns and even four horned sheep (Adalsteinsson, 1981).

Among the Icelandic sheep breed there is a phenotypically unique breeding line called leadersheep (Dýrmundsson, 2002). They are usually non-white, horned and with a slender body conformation. Many believe that they are particularly intelligent sheep and that they walk or run in front of their sheep flock and lead them when the flocks are being moved or

even when the weather is bad and the flock needs to seek shelter (Adalsteinsson, 1981; Dýrmundsson, 2002). Little is known about the genetics of these sheep but they have been included in some genetic studies, sometimes as a subpopulation of the Icelandic sheep (Tapio et al., 2005). Leadersheep can be found within sheep flocks all around Iceland, but the breeding population is relatively small. There were only around 1,000 purebreds and about 500 crossbred animals in 2002 (Dýrmundsson, 2002) and around 1,300 leadersheep were reported in Iceland in 2008 (Dýrmundsson, 2011).

Sheep farmers in Iceland tend to keep their sheep in separate flocks of horned and polled individuals (Bjarnason & Kristjánsson, 2012; Valsdóttir, Jónmundsson & Eyþórsdóttir, 2012) and breed them separately, with the same breeding goal. A difference between body conformation and body fat composition in horned and polled sheep has been detected in a study about fat in lamb meat (Þorgeirsson, 1988) and at the artificial insemination stations there are both horned and polled rams, the polled rams having a higher average BLUP score for ewe productivity (estimated from lamb carcass weight) than the horned rams (Árnason & Jónmundsson, 2008).

1.2. Genetic diversity

Genetic diversity is important for adaptation to changing environmental conditions and therefore the long-term survival of populations (Talle et al., 2005). It is one of three levels of biodiversity that the World Conservation Union (IUCN) has recommended for conservation, along with species diversity and ecosystem diversity (Frankham, Briscoe & Ballou, 2002). Genetic diversity can both be observed as phenotypic variation among individuals within breeds and among different breeds (Talle et al., 2005). Variation among individuals within breeds is essential for selection in animal breeding (Meuwissen, 2009; Talle et al., 2005) and genetic diversity among breeds is useful for future breeding as it provides alternatives if a commercial breed cannot respond to changes in the production systems (Meuwissen, 2009).

Genetic diversity of a population can for example be characterized as observed or expected heterozygosity or as mean number of alleles (MNA) (Barreta et al., 2012; Tapio et al., 2005; Tolone, Mastrangelo, Rosa & Portolano, 2012). Differences between observed and expected heterozygosity in a specific loci indicate whether there is random mating or inbreeding in the population (Meuwissen, 2009). It is closely related to inbreeding as inbreeding can be described by loss of heterozygosity (Reed & Frankham, 2003). Effective

population size can also be used as an indicator of genetic diversity; endangered populations normally have a small effective population size (Talle et al., 2005).

1.2.1 Genetic diversity in domestic animals

Genetic diversity of populations is influenced by genetic drift, mutation, migration and selection (Talle et al., 2005). Natural selection favors animals best suited for their environment, but artificial selection by humans favors animals with traits that are profitable for production, both resulting in different effects on genetic diversity of breeds (Talle et al., 2005). During the several last decades selection programs have become very efficient and increased genetic improvement in a number of breeds along with environmental factors such as feed technology (Groeneveld et al., 2010). The highly productive breeds resulting from the selection programs have replaced local breeds across the world and this development has led to increased concern about the reduction of genetic resources of livestock (Groeneveld et al., 2010).

Effective population size (N_e) has decreased in many production breeds, for example in all Northern-European sheep breeds studied by Tapio et al. (2005). In the same study, within breed genetic diversity (h_k) varied a lot, ranging from 0.38 for Swedish Roslag sheep to 0.76 for Finnsheep. Heterozygosity has also been used to describe genetic diversity of sheep breeds in a recent study including 74 breeds from around the world. The estimates of heterozygosity ranged from 0.24 to 0.38 within breeds. The inbreeding coefficient F was also evaluated in the same study. It varied greatly between breeds with resulting values from 0.07 up to 0.42 (Kijas et al., 2012).

Genetic diversity of the Icelandic sheep population has not been studied specifically but samples of Icelandic sheep have been included in European research. In the study by Tapio et al. (2005), 30 samples of Icelandic sheep and 35 of Icelandic leadersheep were included. The average genetic diversity of the Icelandic sheep samples was 0.71 which was relatively high in this study. They were in the same group as a few Baltic breeds that showed above average contributions to molecular diversity of all breeds included. The leadersheep did not contribute to the above average variation and had slightly lower within breed diversity (0.65) (Tapio et al., 2005). In another study the Icelandic sheep were classified in a Nordic cluster within Northern European sheep breeds (Tapio et al., 2010). They were again defined as having relatively high genetic diversity and described as a unique pool of heterogeneous sheep populations in North Europe. When compared with breeds from

Southern Europe however, they have shown lower diversity (H_o=0.537, with the mean of Southern breeds H_o=0.641) (Handley et al., 2007). Inbreeding coefficient for the Icelandic sheep breed has been calculated using pedigree data and the average inbreeding is low compared to other breeds, ranging from less than 1% (0.01) up to about 5% (0.05) (Figure 2).

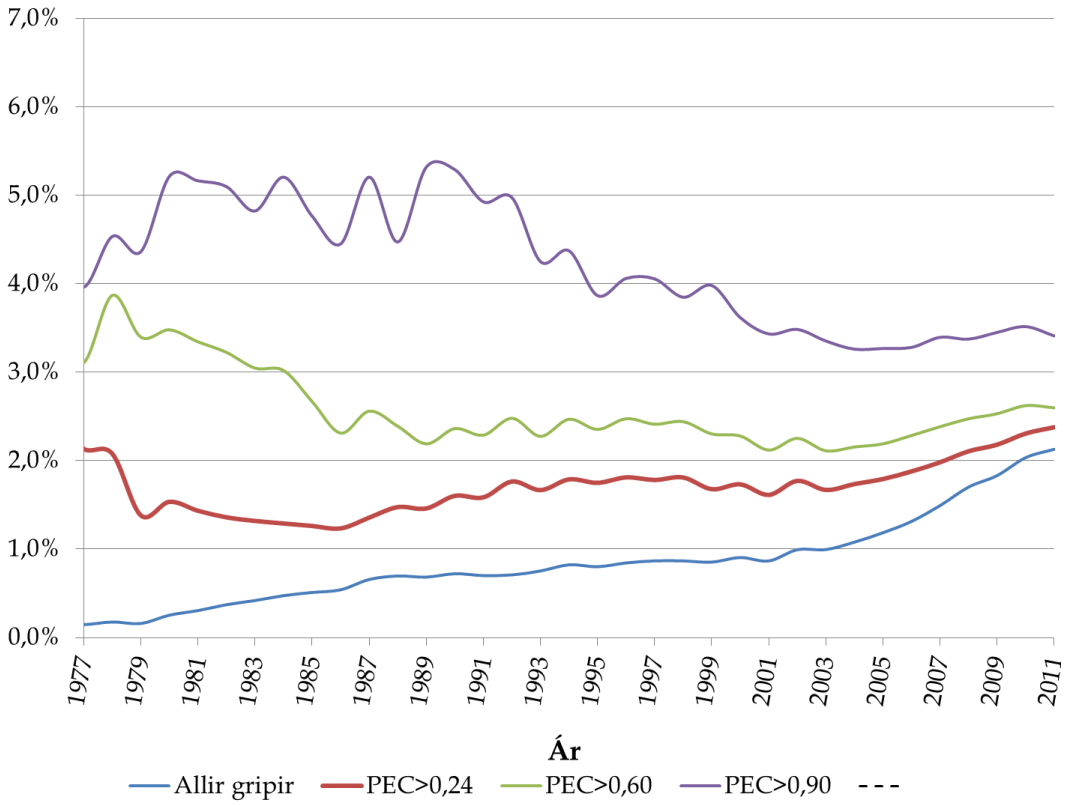

Figure 2 Average inbreeding coefficient per birth year (*Ár*) based on pedigree data of Icelandic sheep born from 1977 to 2011. The lines represent different PEC values (pedigree completeness index); with the blue line (*Allir gripir*) including all sheep regardless of PEC value (Jónmundsson & Eyþórsdóttir, 2014).

1.3. Population structure

A population is defined as a group of individuals from the same species that live in the same geographical area (Hartl, 1994). Another definition is that individuals within a population are able to exchange gametes and create offspring. Populations can be divided into subpopulations which are more or less distinct breeding groups in limited areas (Hartl, 1994). Allelic frequency is one of the factors used to study populations. The allelic frequency of a population should be representative for the whole population. If the population is divided into subpopulations then the allelic frequencies can be different

between the subgroups (Hartl, 1994). Therefore it is important to reveal all possible subpopulations before a genetic analysis of a population is performed.

There are many algorithms and programs designed to study the structure of populations. Common methods are principal component analysis (PCA) (Lee, Abdool & Huang, 2009) and multidimensional scaling (MDS) (Purcell et al., 2007) which are both multivariate statistical techniques. In PCA the principal components are constructed from a linear combination of the genotypes of genetic markers such as SNPs. Each principal component should maximize the variance between the samples used in the analysis. MDS is used to analyze genetic distance matrices and places samples on a graph so that the distances between them represent their true genetic distances (Wang, Zöllner & Rosenberg, 2012). Population structure based on genetic markers is often viewed on a two-dimensional graph plotting two components of MDS or PCA. The results from these two methods are generally quite similar to each other (Wang et al., 2012). Other methods are also available for analysis of population structure, including programs such as STRUCTURE or fastSTRUCTURE based on a model-based clustering method to infer population structure and identify possible subpopulations (Pritchard, Stephens & Donnelly, 2000; Raj, Stephens & Pritchard, 2014).

1.4. Molecular markers

Molecular markers or genetic markers are small sequences of DNA that reveal polymorphism in genomes (Tanksley, 1983). They are powerful tools to detect genetic uniqueness of individuals and the diversity of populations (Chauhan & Rajiv, 2010). Molecular genetic markers that have been used in genetic analyses are allozymes, restriction fragment length polymorphisms (RFLP), randomly amplified fragment polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), expressed sequence tags (EST), microsatellites (or simple sequence repeats, SSR) and single nucleotide polymorphisms (SNP) (Vignal, Milan, SanCristobal & Eggen, 2002).

Allozymes are enzyme coding genes which used to be popular genetic markers. The genes have different alleles, resulting in different protein isoforms which can be detected by comparing migration rate of the enzymes in gel electrophoresis (Chauhan & Rajiv, 2010). RFLPs are markers that are based on a nucleotide change which creates or destroys a restriction endonuclease recognition site so the DNA sequence acquires or loses the ability to be cleaved by a particular restriction endonuclease. The result is either one long

fragment or two shorter fragments (Beuzen, Stear & Chang, 2000). It is identified by using electrophoresis and allows identification of only two alleles per locus. Inbreeding in domestic animals makes many RFLPs sites non-informative (Montaldo & Meza-Herrera, 1998). RAPD markers are amplified products of less functional parts of the genome that do not respond strongly to phenotypic selection. RAPDs can detect high levels of DNA polymorphisms by amplification of genomic DNA by PCR with arbitrary nucleotide sequence primers. A large number of loci and individuals can be screened simultaneously using RAPDs. However it is not possible to distinguish between homozygotes and heterozygotes (Chauhan & Rajiv, 2010). AFLPs are based on selective PCR amplification of restriction fragments from a total digest of genomic DNA. Sets of restriction fragments can be visualized by PCR without knowledge of nucleotide sequence. The method allows co-amplification of high numbers of restriction fragments but the number of fragments that can be analyzed simultaneously is dependent of the resolution of the detection system (Vos et al., 1995). ESTs are single-pass sequences made from random sequencing of complementary DNA clones generated from mRNA. They are used to identify genes and rapid analysis of genes expressed in specific physiological conditions. They are also useful for linkage mapping and physical mapping in animal genomics (Chauhan & Rajiv, 2010). The disadvantages of using ESTs are that it is very difficult to isolate mRNA from some tissues and cell types. Also, important gene regulatory sequences that may be found within introns are removed because ESTs are made from mRNA from which the introns have been removed (National Center for Biotechnology Information, n.d.). Microsatellites are repeats of two- to six-nucleotides that are interspersed throughout the genome. They are abundant, highly polymorphic and the mutation rate is considered to be high. Microsatellite length variation is detected by polymerase chain reaction (PCR) using appropriate primers. Microsatellite markers are a powerful way of mapping genes that control economic traits. Once a marker is identified, specific primers can de designed for PCR for genotyping other samples (Beuzen et al., 2000). The drawback of using microsatellite markers is the need of a large amount of up-front investment and effort. Each microsatellite locus has to be identified and its flanking region sequenced to design the PCR primers (Chauhan & Rajiv, 2010).

Single nucleotide polymorphisms (SNPs) are single base-pair differences which have been growing in popularity as genetic markers and have been studied in many species. SNPs have many advantages over use of allozymes, microsatellites and other molecular markers.

The advantages are availability in high numbers, presence in both coding and non-coding regions, fewer errors, it is easy to compare results between studies and to change them to simple models of mutation (Haynes & Latch, 2012). They also represent the most abundant polymorphism in any organism's genome and reveal polymorphism not detected with other markers and methods (Chauhan & Rajiv, 2010).

One way to discover SNPs is to use commercially available SNP chips for a related, well studied model species (species with fully sequenced genomes). These chips are microarrays that are specially produced for genotyping known SNP loci and allow thousands of such loci to be scored simultaneously for two alleles. Recently SNP chips from livestock species have been used to identify SNPs in closely related, non-model species (Haynes & Latch, 2012).

Another type of variation involves structural variants such as deletions, duplications and complex rearrangements of genomic segments. A subset of these structural variants is known as copy number variants (CNV) or copy number polymorphisms (CNP) (Beckmann, Estivill & Antonarakis, 2007). CNVs are not as numerous as single nucleotide polymorphisms in most genomes. However, they are larger and can affect from one kilo-base of DNA up to several mega-bases per event (Beckmann et al., 2007). Genetic studies based on CNVs have increased in recent years and advances have been made in characterization of these variants. They are considered to be important contributors to phenotypic variation and have been associated with both Mendelian and complex disease traits (Clop, Vidal & Amills, 2012). They have been studied in humans, animals and plants. In humans they have been found to influence gene expression, phenotypic adaptation and expression by changing gene dosages and disrupting genes. Association to disease susceptibility has also been reported and CNVs have also been considered as an important source of genetic variation (Liu, Zhang, Liu & Arendt, 2013). In domestic animals they have been associated with pigmentation and morphological traits along with susceptibility to various diseases (Clop et al., 2012). CNVs have been studied in sheep by Fontanesi et al. (2011) who used an aCGH platform with 385,000 probes designed based on the bovine genome to analyze DNA samples of 11 ewes. They found 135 CNV regions covering ~10.5 Mb of the virtual sheep genome including many genes with important biological functions. Liu et al. (2013) identified a total of 238 CNV regions in an analysis of three sheep breeds using the Ovine SNP50 BeadChip array.

DNA-markers can be used to assist in selection (MAS, marker assisted selection) of domestic animals and other species. Mapping markers associated with phenotypic traits can be done for instance with linkage studies or quantitative trait loci (QTL) analyses which identify markers in linkage disequilibrium with one or more causal genes. The identified markers can then be used to select for preferred phenotypes (Tellam, Cockett, Vuocolo & Bidwell, 2012). Genomic selection (GS) is also based on markers but uses markers covering the whole genome so there should be a marker in linkage disequilibrium with all known loci that are useful for breeding. MAS and GS increase the rate of genetic gain but would be even more effective if the causal variants were selected for directly (Tellam et al., 2012).

An example of the use of molecular markers in sheep research is a study where a single nucleotide polymorphism in the Ho region of chromosome 10 in Australian Merino sheep has been found to be highly predictive for the polled phenotype (Dominik, Henshall & Hayes, 2012). An experimental population of sheep was genotyped with the Illumina ovine SNP chip for 54,977 SNPs. The genotype data was studied by building haplotypes and carrying out a linkage disequilibrium study to detect the association between the SNPs and the phenotype polled status (Dominik et al., 2012).

1.5. Genome-wide association studies

When a dense set of polymorphic markers across a genome is genotyped in samples it is possible to look for common genetic variants associated with a specific phenotype. These are so called genome-wide association studies (GWAS) and are mainly used to identify genetic risk factors associated with diseases in humans (Bush & Moore, 2012) and economic traits of animals (Zhang et al., 2013). GWAS compare the frequency of alleles or genotypes of many genetic markers between different phenotypes. They are considered relatively powerful and fast compared to other methods used to identify genetic effects (Hirschhorn & Daly, 2005). For example QTL studies have a long confidence interval of the associated loci so the causal genes can be hard to locate within them (Zhang et al., 2013).

To be able to conduct a powerful genome-wide association study a large set of polymorphic markers that captures the common variation across genomes is needed (Hirschhorn & Daly, 2005). This condition is met for example by the use of high density SNP chips that are now commonly used for genotyping samples for GWAS. For most

GWAS, two primary platforms have been used; those are the Illumina platform (San Diego, California, USA) and the Affymetrix platform (Santa Clara, California) (Hirschhorn & Daly, 2005). Information about phenotypes must also be available in order to find association between genotypes and phenotypes. Phenotypes can be either categorical or quantitative. The design of the study is based on what kind of phenotype is used. If the phenotype is categorical with information about disease status (for example affected or unaffected) then the GWAS is called a case/control study, even if the disease status is based on many underlying factors. However, if the phenotypic information is quantitative, some kind of measurement for example, then the study design will be quantitative. The quantitative design might seem more powerful but the case/control design has also resulted in many successful results (Bush & Moore, 2012). It is also important to have a large set of samples. Variants that contribute to complex phenotypic traits usually have small effects and therefore many samples are needed to get accurate results (Hirschhorn & Daly, 2005). When using a quantitative study design the power of the results is dependent on the effect size of the associated allele affecting the phenotypic trait under study (Table 1). To generate results with power above 70%, 300 samples are needed when the effect size of the SNP is 0.1, but when the effect size is 0.2, then 200 samples is enough and 100 samples is enough when the effect size is 0.3 or higher.

Table 1 Sample size needed to generate statistical significance of association when using 500,000 SNPs and full linkage disequilibrium is between the associated SNP and the causative SNP, with different values of effect size of the associated allele and difference in resulting power of the study. The required sample sizes were calculated using GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits (Feng, Wang, Chen & Lan, 2011).

Effect size	Power	Sample size
0.1	2.9%	100
0.1	31%	200
0.1	71%	300
0.1	92%	400
0.2	40%	100
0.2	97%	200
0.3	90%	100
0.4	99.8%	100
0.5	100%	100

In case/control study designs, the number of cases and controls needed depends on the frequency of the high risk allele (Table 2). A number of 2212 case/control samples has

been suggested to gain a statistical power of 81.8% when using a 610K Illumina chip in a human GWAS (Spencer, Su, Donnelly & Marchini, 2009) and a study of 6,000 cases and 6,000 controls could result in 94% power (if minor allele frequency, MAF, of the trait-susceptible allele is 0.1) and if the MAF is less than 0.1 and the effect of the allele is very small then sample sizes of more than 10,000 cases and 10,000 controls are required to achieve statistically significant results (Wang, Barratt, Clayton & Todd, 2005). This size of sample set is unrealistically large and good study designs using SNPs with higher MAF do not require so many samples (Wang et al., 2005). To generate statistically significant results of association with a study power of 80% it is necessary to include at least 903 cases and 903 controls when the effect of the allele is low and its frequency high (Table 2). Fewer cases and controls are needed when the effect of the allele is bigger and when its frequency is lower.

Table 2 Number of cases and controls needed to detect a dominant allele with statistically significant association (p=0.05) with a phenotypic trait, with a required 80% study power. The high risk allele frequency is the frequency of the allele causing the phenotype; Aa is the effect of heterozygosity and AA the effect of homozygosity for causative allele. The numbers of cases/controls were calculated using Genetic Power Calculator, available on http://pngu.mgh.harvard.edu/~purcell/gpc/ (Purcell, Cherny & Sham, 2003).

High risk allele frq	Aa	AA	no cases / controls
0.1	1.5	2	405
0.1	2	3	130
0.1	3	4	50
0.2	1.5	2	599
0.2	3	4	106
0.2	6	7	52
0.3	1.5	2	903
0.3	2	3	357
0.3	6	7	127

Statistical tests for association need to be adjusted for factors that can possibly have impact on the result of the analysis. These are factors such as; age, sex and study site and they should be included as fixed effects in the analysis. Another important factor is genetic structure of the populations under study. In association studies allelic differences are assumed to be related only to the trait of interest. However, if the individuals in the association study are of different subpopulations it is possible that the allelic difference is related to the background of the individuals (Liu et al., 2013). This is called population stratification and can produce false positives or overlook true associations when it is not

accounted for. Therefore it is important to study the structure of populations before performing association studies and adjust the study according to stratification if it is present (Helgason, Yngvadóttir, Hrafnkelsson, Gulcher & Stefánsson, 2004; Liu et al., 2013).

1.5.1 Genome-wide association studies in sheep

In recent years genomes of several domesticated animals have been sequenced, partially or completely. Information on whole genomes of animals in production is becoming more interesting for researchers and breeders with the possibility to identify genetic variation causing different performance (Bai, Sartor & Cavalcoli, 2012). It could increase opportunities for resisting pathogens that challenge animal production and will provide valuable information for production of lean, healthy and economic animal protein for human consumption (Bai et al., 2012). Chicken were the first species to be sequenced (Burt, 2005) followed by pig (Archibald et al., 2010), cow (Zimin et al., 2009), horse (Wade et al., 2009) and sheep (The International Sheep Genomics et al., 2010) which have all been partially or completely sequenced (Bai et al., 2012).

The sheep genome was generated by sequencing the DNA of a single Texel ewe and a single Texel ram using Illumina technology. The latest assembly of the sheep genome (Oar_v3.1) is based on the dataset of the Texel ewe (Jiang et al., 2014). The coverage of the reference genome is ~150 fold with a contig length of ~40kb and a total assembled length of 2.61 Gb (Jiang et al., 2014). Before the release of the sheep genome, there were only about 700 genes known in sheep (Zhang et al., 2013) but the current gene build by Ensembl counts 20,921 coding genes in the sheep genome and 43,449 Genscan gene predictions (Flicek et al., 2013).

Few GWA studies have been carried out for sheep data due to limited information about the sheep genome. With the recently released assembly of the whole sheep genome the number of GWAS on sheep is growing (Zhang et al., 2013). A Chinese GWA study on 329 sheep of different breeds looked for association to 11 traits related to muscle growth. The study identified 5 candidate genes for growth and meat production traits (Zhang et al., 2013). Milk production traits in Spanish Churra sheep have been studied and association found with a QTL on chromosome 3 and with the *LALBA* gene (García-Gámez et al., 2012). An association of genomic regions to susceptibility and control of Ovine lentivirus has been studied and a few candidate genes found (White et al., 2012). A genome-wide

scan of Finnsheep identified a single nucleotide substitution in the *ASIP* gene associated with coat color variation in sheep (white vs. non-white) (Li, Tiirikka & Kantanen, 2013).

1.5.2 Estimated breeding value as phenotype

The most common phenotypes used in genome-wide association studies and other genomic based analyses are individual measurements. Another possibility is to use estimated breeding values (EBV) of the individuals. This has been done in several GWAS. Estimated breeding values are usually based on information about individuals, their offspring and their relatives (depending on available information and the model used to calculate the value). EBVs based on the BLUP animal model have been used to detect SNPs related to calving ease in cattle, where the two most significant SNPs explained 10% of the EBV variation (Pausch et al., 2011). Another study also used EBVs in a GWAS for 9 traits in cattle but they calculated a so-called deregressed EBV by removing all information about relatives to reduce risk of finding SNPs only showing association based on relative information (Bolormaa, Pryce, Hayes & Goddard, 2010). A GWAS for fertility traits in Finnish Ayrshire cattle using EBVs found several significant QTL regions related to female fertility (Schulman et al., 2011).

In Iceland the BLUP animal model is used for estimation of breeding values. The BLUP score is based on the Best Linear Unbiased Prediction of breeding value for a specific trait and uses information about individual measurements and measures of relatives and offspring. It is used in animal breeding for estimating genetic quality of individual traits or a summary of traits (Henderson, 1975). Breeders can select animals based on their BLUP score to increase genetic improvement. The BLUP method has been used for genetic evaluation of carcass quality traits (muscle and fat) in Icelandic sheep since 1999 (Árnason and Jónmundsson, 2008).

1.6. Muscle growth

Understanding the control of growth and development of skeletal muscle is one of the most important goals in animal breeding and animal science. Muscles are mostly made of muscle fibers so muscle mass is determined by the number and size of muscle fibers. Current research shows that animals with greater number of moderate size muscle fibers produce more meat (Rehfelt, Fiedler & Stickland, 2004). Skeletal muscle in mammals is composed of several types of fibers, classified by the predominant type of myosin heavy-

chain isoform they contain. The major types are designated I, IIa, IIb and IIx (Klover, Chen, Zhu & Hennighausen, 2009). The number of muscle fibers is determined by the extent of multiplication of muscle cells in myogenesis. Genetic and environmental factors that affect prenatal myogenesis therefore determine the number of muscle fibers. Postnatal growth of skeletal muscle does not increase number of fibers but does increase their length and girth (Rehfelt et al., 2004). Complicated interactions of extrinsic and intrinsic regulatory mechanisms control myogenesis both prenatal and postnatal (Bentzinger, Wang & Rudnicki, 2012). Growth and function of all muscle fiber types are influenced by hormones. The growth hormone (GH) and IGF-1 appear to play a big role in postnatal growth of skeletal muscle (Klover et al., 2009; Liu, Baker, Perkins, Robertson & Efstratiadis, 1993) and the Leptin hormone is a major regulator of energy intake and expenditure and has been shown to positively regulate muscle mass by suppressing the *FOXO3A* gene (Braun & Gautel, 2011).

1.6.1 Muscle growth genes

Few genetic variants with relatively large effects on muscling traits have been discovered. Myogenic differentiation genes (*MYOD*) play an important role in growth and muscle development. They are involved in muscle fiber formation and proliferation during embryonic development along with maturation and function of fibers postnatal (Bhuiyan et al., 2009). The *MYOD* genes are considered as candidate genes for meat production traits because of their roles in muscle fiber development. SNPs within the *MYOD* genes have for example been associated with live weight in cattle and live and carcass weight in Korean cattle specifically (Bhuiyan et al., 2009). Skeletal muscle differentiation is also regulated by transcriptional mechanisms where myogenic regulatory factors (MRFs) play a role in muscle development along with other transcription factors and epigenic effects (Braun & Gautel, 2011). MRFs such as myogenic factor 5 (MYF5), muscle specific regulatory factor 4 (MYF6), myoblast determination protein and myogenin activate many downstream genes to begin muscle cell differentiation (Braun & Gautel, 2011). Variation in these factors or total absence of them can affect muscle development or adult muscle regeneration (Bismuth & Relaix, 2010).

Mutations of several major genes influence muscle fiber number and/or muscle fiber size in skeletal muscle. These mutations are associated with extreme muscular hypertrophy and sometimes changes in meat quality (Rehfelt et al., 2004). A single mutation in the *RYR1*

gene in pigs can result in leanness and muscle hypertrophy in heterozygotes. Mutations at the *IGF2* and *PRKAG3* loci are also known in pigs and can cause increased muscle mass (Gordon, Gordish Dressman & Hoffman, 2005). Deletions and missense mutations in the *MSTN* gene that codes for the myostatin protein can cause reduced expression or loss-of-function. These variations can cause double muscling in animals (Gordon et al., 2005; Lee, 2007) because myostatin (also known as growth differentiation factor 8 or GDF8) regulates generation of muscle fibers during development and growth of muscle fibers postnatal (Lee, 2007). Other known genes that have been associated with muscle growth traits are for instance the *NEB* gene that codes for a cytoskeletal matrix protein and is associated with weight in cattle. *DGAT2* has also been associated with weight in cattle; it has been annotated as a gene that influences fat deposition in animals (Dunner et al., 2013). MicroRNAs (miRNAs) have been connected to muscle hypertrophy by repressing the muscle specific expression of *miR-1* gene cluster in mice (McCarthy & Esser, 2007). A cysteine and glycine-rich protein 3 (*CSRP3*) plays an important role in myofiber differentiation and four SNPs within the gene have been associated with growth and carcass traits (He, Zhang, Li, Liu & Liu, 2014). It has been suggested as a candidate gene for selection programs to improve growth and carcass traits of cattle by selecting specific genotypes (He et al., 2014).

1.6.2 Muscle growth genes in sheep

A trait is heritable when phenotypic variance is affected by differences in genes. Heritability of a trait is defined as the ratio between genetic variance and phenotypic variance and measures the proportion of phenotypic variation of a trait that is due to genetic differences (Griffiths, Miller, Suzuki, Lewontin & Gelbart, 2000). Heritability of muscling traits in sheep are moderate; 0.22-0.54 on average for various sheep breeds, with traits like muscle weight and meat yield on the lower half; 0.22-0.35 for Merino and Border-Leicester (Mortimer et al., 2010) but up to 0.38-0.54 for muscle depth in Texel, Suffolk and Charollais (Tellam et al., 2012). The heritability of carcass weight of Icelandic sheep is lower, with reported values ranging from 0.11-0.18 (Eythórsdóttir, 2012), recent calculations of heritability of cold carcass weight was 0.18 and heritability of lean meat yield estimated as lean weight in major cuts ranged from 0.17-0.21 (Einarsson, Eythórsdóttir, Smith & Jónmundsson, 2014).

There are many genes involved in the development of skeletal muscle of sheep. A transcriptional profiling experiment of back muscle of sheep revealed changes in a great number of genes during skeletal muscle development (Byrne et al., 2010). The transcription of a large number of genes changed substantially during late skeletal development of sheep. These changes happened between an interval of late fetal stage and few days postnatal and are likely to affect adaptation of muscle to new physiological demands in the postnatal environment of the sheep (Byrne et al., 2010).

Table 3 Known genes related to muscle growth and development in sheep (Flicek et al., 2013).

Function	Genes	Chromosome
Muscle growth	*USMG5*	13
Muscle development	*IFRD1*	4
	MSC	9
	PPP2R3A	1
	PITX1	5
	TCF21	8
	CACNA1S	12
	PITX2	6
Muscle fiber development	*MYOG*	12
	MYOD1	15
	1)	Various

1) There are 38 more genes related to muscle fiber development, located on different chromosomes.

To date, about 610 genes have been annotated with an association with muscle in sheep. This is based on information from the sheep genome (Oar_v3.1), EntrezGene record, HGNC etc. (Flicek et al., 2013). A search for 'muscle growth' results in one gene; a gene called *USMG5* which is a predicted gene on chromosome 13 in sheep; it is annotated by the Ensembl gene build (Table 3). It was reported as up-regulated during skeletal muscle growth in mice but has not yet been annotated in sheep (Flicek et al., 2013). Eight genes are registered with function related to 'muscle development'. They are; *IFRD1* which codes for interferon-related development regulator 1, *MSC* or Musculin that plays a role in skeletal muscle development, *PPP2R3A* codes for a protein phosphatase with a record in somatic muscle development, *PITX1* a paired-like homeodomain recorded in skeletal muscle development, *TCF21* codes for a transcription factor connected to skeletal muscle, *CACNA1S* codes for a calcium channel and *PITX2* paired-like homeodomain both recorded in skeletal muscle development and one novel gene which is still uncharacterized but is recorded in smooth muscle development (Flicek et al., 2013). Even more genes are found recorded for 'muscle fiber development' (40 genes). These genes are for example *MYOG* myogenic factor and *MYOD1* myoblast determination protein which are both part of the

MYOD genes that have been suggested as candidate genes for muscle growth etc. (Table 3) (Flicek et al., 2013).

Genetic variants affecting muscle growth and meat quality have been discovered in several sheep breeds (Table 4). The callipyge effect is a major increase in hindquarter muscling caused by myofiber hypertrophy along with a change in myofiber type. The causative mutation is a point mutation (A/G) that is located on the distal end of the ovine chromosome 18 and was discovered in an American Dorset ram. This locus is known as the *CLPG* locus and lies in an intragenic region between *DLK1* gene and *GTL2* gene that belong to a cluster of imprinted genes (White et al., 2008). Callipyge muscle hypertrophy only appears in heterozygous animals that inherit the mutation from their father and a wild-type allele from their mother (Tellam et al., 2012). The Carwell effect is described by larger loin muscle area and increased loin muscle weight. It is due to a quantitative trait loci (QTL) mapped at the telomeric region of chromosome 18. The locus was identified in Australian Poll Dorset rams and it overlaps with the site of the callipyge mutation (Tellam et al., 2012). Another QTL that increases weight of the loin muscle area has been located on chromosome 18. This QTL is called TM-QTL or Texel muscling because it was found in Suffolk and Texel sheep breeds (Hopkins, Fogarty & Mortimer, 2011; Walling et al., 2004).

Table 4 Known genetic variants affecting muscle traits in sheep including information about effect, chromosome (Chr) and sheep breeds that the effect has been reported in.

Gene/variation	Effect	Chr	Sheep breeds	Reference
CLPG	Muscle hypertrophy	18	American Dorset	White et al., 2008
Carwell	Increased loin muscle area	18	Australian Poll Dorset	Tellam et al., 2012
TM-QTL	Increased weight of loin muscle	18	Suffolk & Texel	Hopkins et al., 2011; Walling et al., 2004
MSTN	Muscle hypertrophy	2	Australian White Suffolk, Poll Dorset, Lincoln, Charollais, Texel, Romney & Norwegian White-sheep	Kijas et al., 2007; Hadjipavlou, Matika, Clop & Bishop, 2008; Hickford et al., 2010; Boman et al., 2009

A mutation in the *MSTN* gene causing increased muscle growth was discovered in cattle in the late 1990s (McPherron, Lawler & Lee, 1997). Few years later a search for a similar mutation was started in sheep, leading to the identification of variation in the *MSTN* gene on chromosome 2 found to affect muscle and fat growth in sheep (Hopkins et al., 2011).

Polymorphisms in *MSTN* have been reported in many sheep breeds since, including Australian White Suffolk, Poll Dorset and Lincoln (Kijas et al., 2007), Charollais (Hadjipavlou, Matika, Clop & Bishop, 2008), Texel and Romney (Hickford et al., 2010) and Norwegian White-Sheep (Boman, Klemetsdal, Blichfeldt, Nafstad & Vage, 2009).

Few genetic studies based on genomic data have been conducted on the Icelandic sheep population. Some Icelandic sheep have been included in international studies comparing many breeds, one using endogenous retroviruses to study domestication and differentiation between 133 breeds (Chessa et al., 2009) and the other using microsatellite markers to study molecular variation of northern European sheep (Tapio et al., 2005). Polymorphism in the *PrP* gene and its effect on scrapie has also been studied in Icelandic sheep using RFLP analysis to identify a *PrP* allelic variant associated with scrapie status (Thorgeirsdottir, Sigurdarson, Thorisson, Georgsson & Palsdottir, 1999). Variation in the known areas affecting muscle growth has not yet been studied in the Icelandic sheep breed.

2. Aims of study

The main objectives of this study are the following.

To analyze genetic diversity within the Icelandic sheep breed and within subgroups of sheep samples collected in different places in Iceland, and to compare the groups. To analyze the genetic difference of the groups and do a multidimensional scaling analysis of genetic distance to find out if some of the groups can be considered subpopulations.

To analyze variation found related to differences in muscularity in Icelandic sheep grouped together based on high or low muscularity. Genome-wide association studies will be used to search for association of genomic area to muscle traits in the samples.

All genotyped SNPs that are associated with muscularity in the Icelandic sheep breed will be studied. Variation in muscle growth gene areas/loci that have already been identified in other sheep breeds will be searched for and candidate genes identified.

Selected candidate genes will be sequenced to identify causal variants.

The results will lay the foundation for further studies on the genetics of the Icelandic sheep, possibly concerning association to leadership behavior and color variation and is the first step in creating a tool for marker assisted selection to increase muscularity in Icelandic sheep.

3. Materials and methods

3.1. Sample selection and phenotypic data

Blood was collected from 231 sheep around Iceland and DNA isolated (Figure 3). Additionally, 69 DNA samples previously isolated for other studies (16 thereof for the sheep hapmap project; http://www.sheephapmap.org/hapmap.php) were added to the data collection. Of the 300 samples, 96 were chosen for genotyping. Samples from flocks selected for 'high muscle' were 56 and 40 were from sheep not selected for this trait. The sheep selected for 'high muscle' were from the Hestur experimental flock that has been strongly selected for conformation and carcass quality (all horned); from the breeding rams (horned and polled) at the artificial insemination stations (Saed) and from polled flocks in the North West of Iceland, also with a selection history for increased muscling. The samples from unselected sheep were from a flock of sheep in the West of Iceland (Stafh) and from leadersheep from several farms, mostly in East-Iceland. A list of samples used, including origin and phenotypes can be found in Appendix 1.

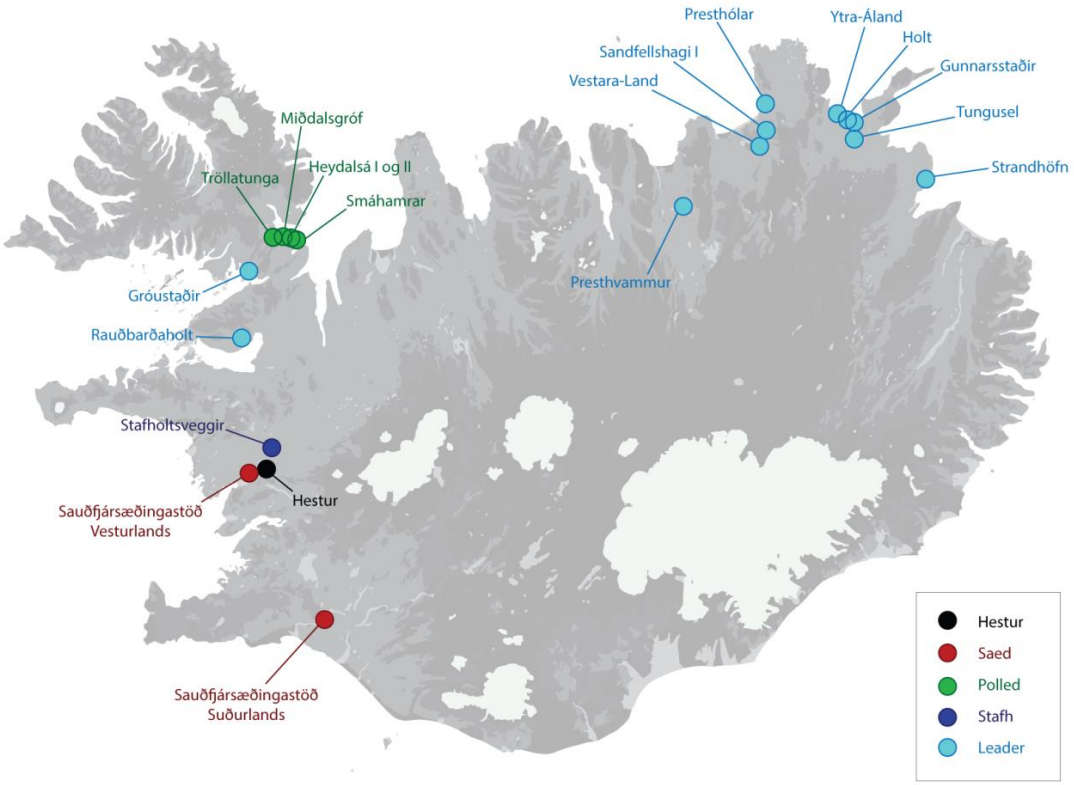

Figure 3 Blood samples were collected from different locations in Iceland. Leadersheep (light blue) were collected in the Northwest and Northeast of Iceland, the polled sheep (green) in the Northwest, artificial insemination rams (red) from the insemination stations in the West and South, sheep from Stafholtsveggir (dark blue) in the West and sheep from Hestur (black) also from the West.

Phenotypes for all samples were obtained from the national recording system of the Farmers Association of Iceland. The information included is age, sex, origin, color, horned/polled and BLUP scores for muscularity; based on evaluation of muscularity of carcasses of offspring and relatives. For some of the leadersheep samples there were no BLUP scores in the central database (individuals Oa218, Oa220, Oa231, Oa232, Oa244, Oa246, Oa247, Oa250 and Oa251). These samples got an estimated score based on the average score of the other leadersheep (BLUP=65).

3.2. DNA extraction, genotyping and quality control

DNA was extracted from blood samples using MasterPure™ Complete DNA Purification Kit (Epicentre Biotechnologies) following DNA purification protocol for whole blood samples. The DNA was genotyped by FIMM (Institute for Molecular Medicine Finland, Helsinki) with a high density SNP chip (SheepHD_AgResearch_Cons_iSelect beadchip) with 606,006 SNPs selected by scientists at AgResearch (Hamilton, New Zealand).

Quality control was performed using PLINK (Purcell et al, 2007) and GenABEL package in R (Aulchenko, Ripke, Isaacs & Van Duijn, 2007) to generate different datasets. In PLINK markers were excluded if minor allele frequency (MAF) was 0, missing rate per sample (mind) was higher than 0.1, missing rate per SNP (geno) was higher than 0.1 and if they failed Hardy-Weinberg equilibrium (hwe) at 0.001 threshold. Also, markers with unknown position and markers on X chromosome were removed. This dataset is called Dataset 1. In GenABEL the function "check.marker" with the following thresholds (default values) was used; SNP call rate = 0.95, sample call rate (perid.call) = 0.95, hardy Weinberg equilibrium (p.level) = 0.001 and minor allele frequency (MAF) = 0. This generated Dataset 2.

3.3. Data analysis

The samples were divided in two groups, 'high muscle' and 'low muscle'. R statistical program (R Core Team, 2014) was used in basic data handling, to compute descriptive statistics and conduct t-tests for difference between mean BLUP score for carcass traits in the two groups.

3.3.1 Population structure

The structure of the population was examined by performing multidimensional scaling (MDS) in PLINK based on identity by state (IBS) in Dataset 1 (Purcell et al, 2007) and a MDS plot was made in R. The population substructure was further studied using the program fastSTRUCTURE (Pritchard et al, 2000). fastSTRUCTURE uses an algorithm for inferring population structure from large SNP genotype data (Raj et al, 2013). It is a faster version of the program STRUCTURE (Pritchard et al, 2000). fastSTRUCTURE was run on Dataset 1 with K from 1 to 6 (command for K=1: python structure.py –K 1 –input=data –output k1 –full –seed=100).

The inbreeding coefficient was calculated in PLINK and GenABEL separately. In PLINK Dataset 1 was pruned by identifying pairs of SNPs in linkage disequilibrium using a sliding window method (PLINK command: --indep-pairwise 50 5 0.5). Pruned dataset was generated by extracting one SNP of each of the pairs (PLINK command: --extract data.prune.in). The pruned dataset was then used for estimation of the inbreeding coefficient (PLINK command: --het). In GenABEL the data was used directly to compute the inbreeding coefficient with hom(data) command.

Outliers were detected in GenABEL using the MDS picture made from Dataset 2.

Genomic kinship coefficient was calculated in R using the ibs function (ibs(data[, autosomal(data)], weight = “freq”) in GenABEL on Dataset 2. The command calculates the covariance between the vectors of individual genotypes and returns a number that can be lower than zero. The resulting matrix can be used to draw a histogram of genomic kinship coefficients as well as calculating and drawing a multidimensional scaling plot showing relationships between samples. Linkage disequilibrium was calculated between all SNPs in PLINK using –r2 0.2 command (generates a list with all SNPs with LD above 0.2). LD was also calculated within groups of different origin.

3.3.2 GWAS

Dataset 2 was used for genome-wide association analysis which was run in R using the GenABEL package. Both case-control models and continuous models were used. The continuous models used the individual BLUP score as a phenotype and for the case-control analysis the two groups of different muscle size were used, with the ‘high muscle’ group as the case and the ‘low muscle’ as the control.

In GenABEL there are a few models available for association testing. There is a fast score test for association between a trait and genetic polymorphism; qtscore(), a score test for association in samples of related individuals; mmscore() and a fast score test for association adjusted for possible stratification by principal components; egscore(). All tests require a formula with trait and fixed effects and data. The mmscore requires a previously determined formula, calculated by a polygenic model for example and the egscore requires information about genomic kinship. Sex was used as a fixed effect and coefficients of genetic distance as a covariate in one model (data88.pca). The commands that were used can be found in Appendix 2.

3.3.3 Annotation of associated SNPs

The online genomic databases Ensembl (www.ensembl.org), the UCSC genome browser (http://genome.ucsc.edu/) and NCBI (http://www.ncbi.nlm.nih.gov/) were used to explore the regions surrounding the top 25 SNPs from the results of the GWAS. The nearest genes of each SNP were located and relations to muscle size, growth, development etc. were searched in published literature.

3.3.4 Candidate gene sequencing

Eleven DNA samples were selected, two from each origin group except three from Hestur, for amplifying and sequencing of exon 2 of candidate gene *KLF13*, exon 9 of *PNN* and the whole *GADD45B* gene. The DNA samples were diluted to 4 ng/µl concentration and the exons and gene desired for sequencing were magnified using the PCR method (Saiki et al., 1988). One Taq® 2X Master Mix with GC Buffer (NEW ENGLAND BioLabs, M0483L) was used, including all necessary ingredients for PCR except DNA, primers and water. The PCR mix consisted of 12.5 µl of master mix, 0.5 µl of forward primer and 0.5 µl of reverse primer, 1.0 µl of DNA template, 0.5 µl of $MgSO_4$ and 10.0 µl of water. The primers were designed using Primer 3 (http://primer3.sourceforge.net-/releases.php/) in Geneious 7.1 (Kearse et al., 2012); two forward primers upstream of the exons and two reverse primers downstream of the exons.

The PCR was done with the following programs; 94°C for 1 minute to denature the DNA, 35 cycles of 94°C for 30 seconds followed by annealing for 30 seconds at temperatures 57°C, 57°C and 58°C followed by extension at 68°C for 40 seconds, 2 minutes and 1.5 minutes for *KLF13*, *GADD45B* and *PNN* respectively; followed by final extension at 68°C for 7 minutes and then cooling down to 4°C.

The PCR products were then mixed with 5 µl of 6x loading dye (NEW ENGLAND BioLabs,) and loaded on a 1.8% agarose gel made with TAE buffer and 1 µl of SYBR® Safe DNA Gel stain (Invitrogen, S33102) for visualization of DNA, along with 2µl of 1 kb DNA ladder (NEW ENGLAND BioLabs, N3232L). The DNA in the samples and the ladder were separated by size using electrophoresis (Johansson, 1972) on the loaded gel for 45 minutes at 90 Volts. The bands were viewed with UV light exposure (ImageQuant 300). Clear bands of correct size were excised from the gel and the DNA purified using NucleoSpin® Gel and PCR clean-up (Macherey-Nagel GmbH & Co KG, Düren, Germany). The purified PCR product was diluted to 5 ng/µl and 15 µl were mixed with 2 µl of primer (10 pmol/ µl), the same one as used in the PCR. Two samples were sequenced for each PCR product, one with the forward primer and the other with the reverse primer. The resulting tubes, with a total volume of 17 µl, were sequenced by Eurofins Genomics (Sequencing Department, Edersberg, Germany). The resulting sequences were aligned and analyzed using Geneious 7.1 (Kearse et al., 2012).

4. Results

4.1. Genotyping and quality control

Before quality control there were 96 samples of sheep DNA with 606,006 SNPs genotyped with the AgResearch HD SNP chip. Samples were genotyped with a call rate ranging from 94.7-98.4%. Dataset 1 was processed in PLINK. After the quality control there were 94 samples left and 547,892 SNPs (after removal of markers with unknown positions, markers on chromosome X and markers with MAF=0). Dataset 1 was also pruned in PLINK, leaving 185,517 SNPs in the pruned dataset that were used for calculation of inbreeding coefficient (F). Dataset 2 was processed in R using GenABEL package. The quality control in GenABEL was stricter and left 93 samples and 467,103 SNPs. The average call rate of SNPs after quality control in GenABEL was 0.998 and >0.99 after quality control in PLINK. There were 31,689 SNPs with maf from 0.01-0.05; 51,445 with maf from 0.05-0.1; 102,158 with maf from 0.1-0.2 and 281,811 with maf above 0.2.

4.2. Population structure

Dataset 1 was used to do a multidimensional scaling (MDS) of an identity by state (IBS) matrix of the samples. The matrix was used to draw a MDS picture to view the genetic distance of the samples. Figure 4 shows substructure in the population. The first component (C1) explains most of the variation between individuals in the population (0.135) and shows that the group of leadersheep differs most from the other groups. The second component separates the samples from Hestur and the rest and explains 0.105 of the variation. Samples from the artificial insemination stations (Saed) are however both among the sheep from Hestur and among the polled and Stafholt. The third component separates the samples from Stafholt from the other groups and explains 0.0709 of the variation.

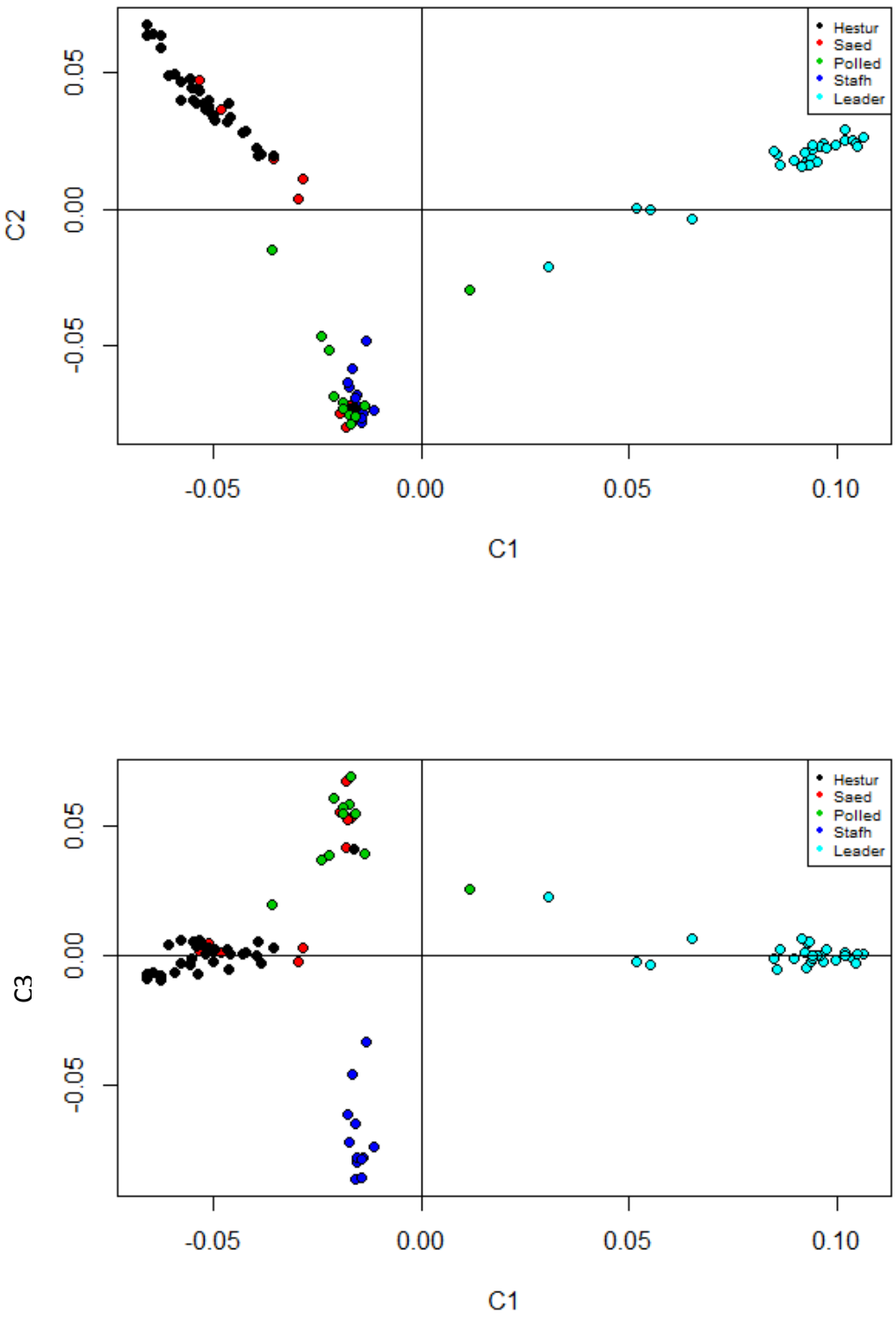

Figure 4 Multidimensional scaling of calculations of genetic distance between samples based on identity by state (IBS) method. Component 1 explains 0.135 of the variation, component 2 explains 0.105 of the variation and component 3 explains 0.0709 of the variation.

Results from fastSTRUCTURE confirm substructure in the samples. A model complexity of 3 maximized the marginal likelihood with 4 components used to explain the structure in the data. It was therefore concluded that there are 3 subgroups present in the data and according to the MDS picture the groups are; Leadersheep, Hestur sheep and all other sheep.

4.2.1 Genetic diversity and inbreeding

Results for inbreeding coefficient (F) in PLINK and GenABEL were similar. The average inbreeding coefficient for the samples was 0.06 in PLINK and 0.07 in GenABEL. Average homozygosity of samples was 0.68 in PLINK and 0.67 in GenABEL. The mean heterozygosity for a SNP was 0.341 and 0.320 per sample, calculated in GenABEL. The average minor allele frequency (MAF) was 0.23 in PLINK and 0.25 in GenABEL. The groups were also compared and leadersheep had the highest inbreeding coefficient (F) in both programs (0.124 in PLINK and 0.122 in GenABEL). The average of Hestur group was F=0.0573 (PLINK) and 0.0583 (GenABEL). 89,584 of all SNPs had a minor allele frequency of 0 so 516,458 (85.2%) were polymorphic.

When looking at the genomic kinship coefficient it seems like most of the sheep in the whole group are unrelated (Figure 5). Comparison of genomic kinship coefficients between groups shows that the Stafh sheep and leadersheep are most related to each other (Figure 6).

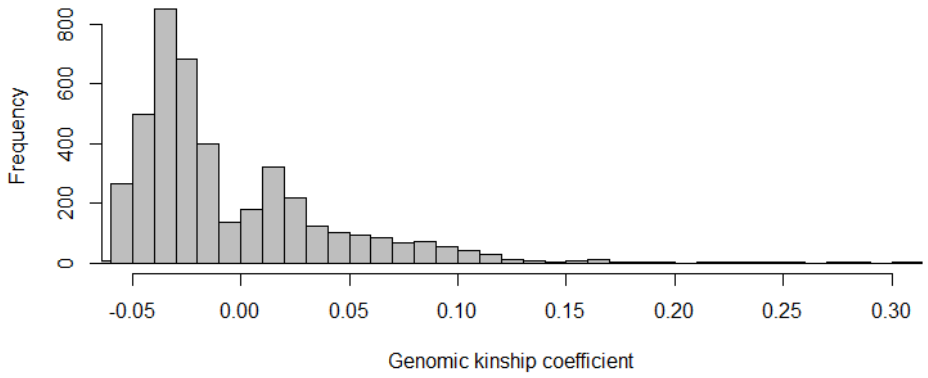

Figure 5 Histogram showing frequency of different genomic kinship coefficients between all animals.

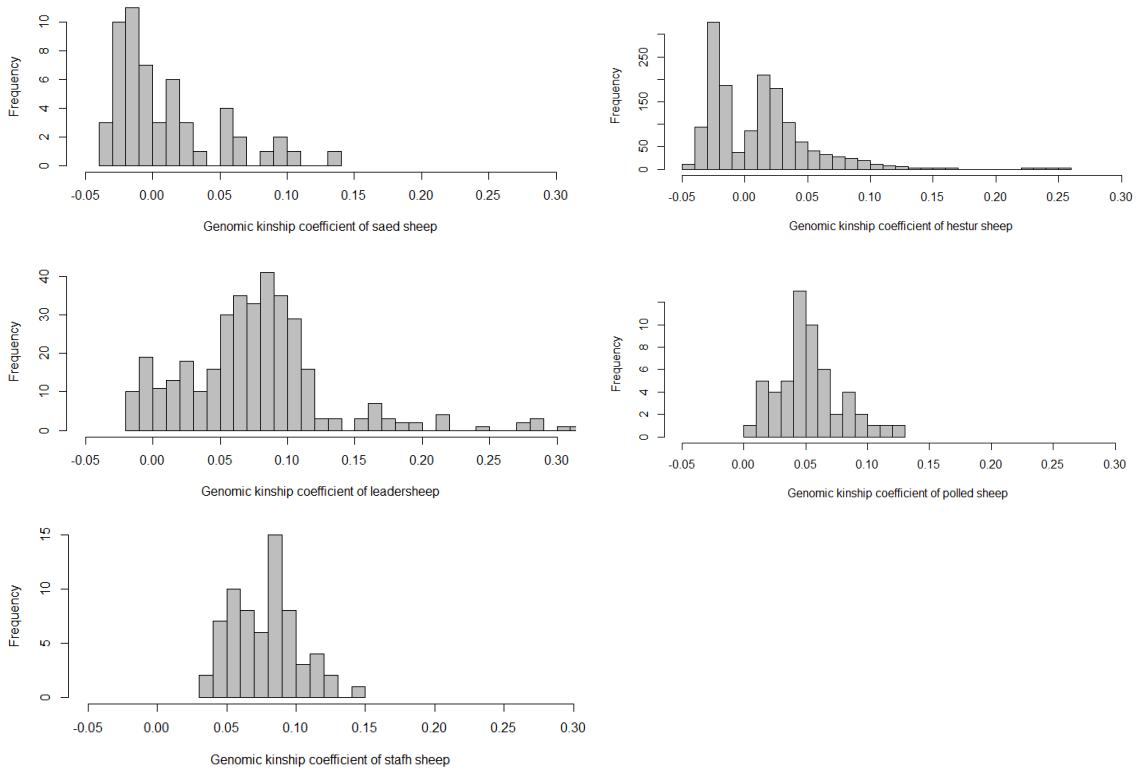

Figure 6 Histograms showing the frequency of different genomic kinship coefficients of animals of each origin group. Leadersheep and Stafh sheep have higher genomic kinship coefficients more frequently than the other groups.

4.2.2 Outliers

Five outliers were identified from results of MDS in R (individuals Oa211, Oa212, Oa247, Oa250 and Oa251) (Figure 7). They were removed from the dataset before the genome-wide association analysis. These individuals are all leadersheep, Oa211 and Oa212 from a farm in the Northeast and Oa247, Oa250 and Oa251 from a farm in the Northwest.

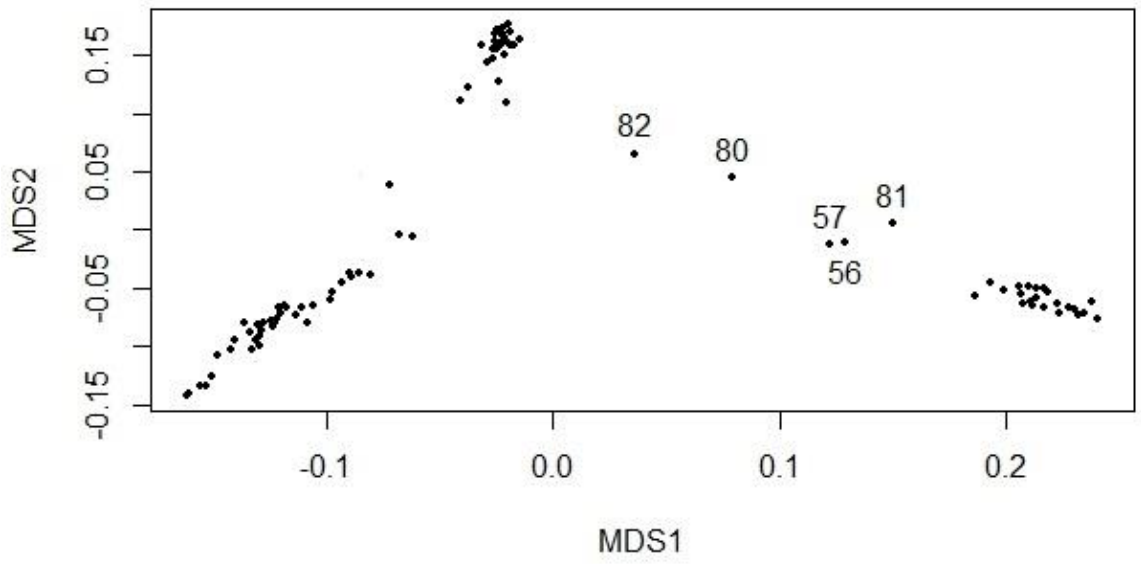

Figure 7 Sample distribution based on multidimensional scaling of genetic distances was used to identify outliers. The outliers are the numbered samples; they are the ones that lie far from the clusters when looking at the horizontal axis (MDS1).

4.2.3 Linkage disequilibrium

The average value of linkage disequilibrium (LD) for the SNPs included in the analysis was r^2=0.544. Only SNPs with r^2=0.2 or higher were included in LD calculations between SNPs. LD between samples decays fastest from 0.70 to 0.55 when distance between SNPs increases from 0 kb to 10 kb (Figure 8).

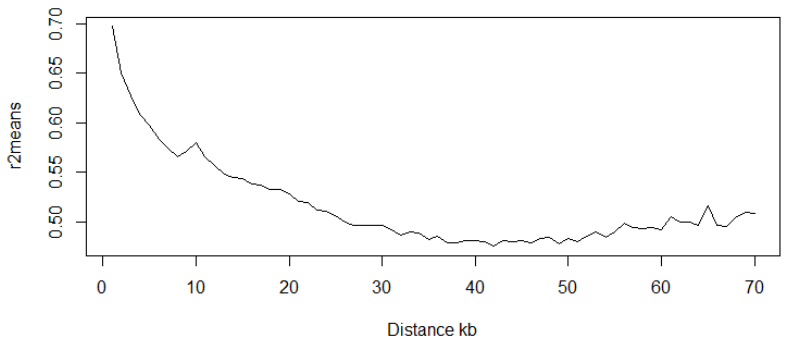

Figure 8 Linkage disequilibrium (R^2) decay relative to distance (kb) between SNPs.

Average value of LD between SNPs of samples within different groups of origin was similar. The leadersheep had the highest average value; r^2=0.602. Sheep from Hestur had r^2=0.581, polled sheep had r^2=0.560, Stafholt sheep r^2=0.570 and artificial insemination rams r^2=0.564.

4.3. Genome-wide association analysis

The 'low muscle' group has significantly lower BLUP scores for muscularity than the 'high muscle' group ($p<2.2x10^{-16}$). Therefore these groups were used as cases and controls for the GWAS; the sheep with higher BLUP scores were defined as cases and the ones with lower BLUP scores as controls.

The results of the models with the best value for the genomic inflation factor (λ) were used for further analysis. If the λ value is equal to 1 then there should be no false positives among the results. Results from four models of association were considered most reliable, two with the continuous trait and two with binary trait. The genome-wide Manhattan plot for each model is shown, displaying the resulting p-values with respect to genomic position. A Q-Q plot that presents the deviation from expectation in each model is also shown. For the continuous trait (BLUP-score) the model that gave the most reliable results was the fast score association test including MDS coefficients and sex as covariates; qtscore (Figure 9) and a score test using polygenic model including the genomic kinship as formula; mmscore (Figure 10).

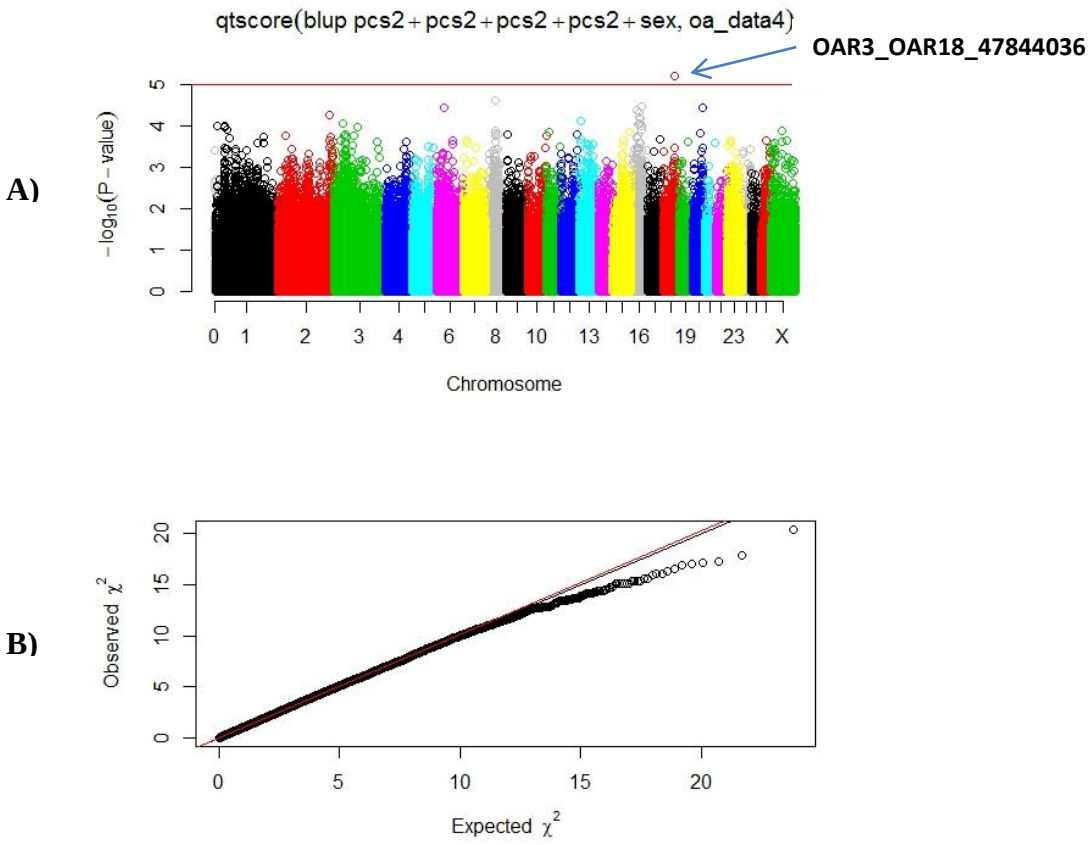

Figure 9 (A) Manhattan plot showing scores of SNPs calculated using a fast score test including coefficients from MDS analysis, with respect to chromosomes. The dots represent the SNPs and their association to the BLUP score, showing the negative $\log_{10}$ of the p-value of association. (B) QQ-plot showing relationship of observed and expected results from the association test.

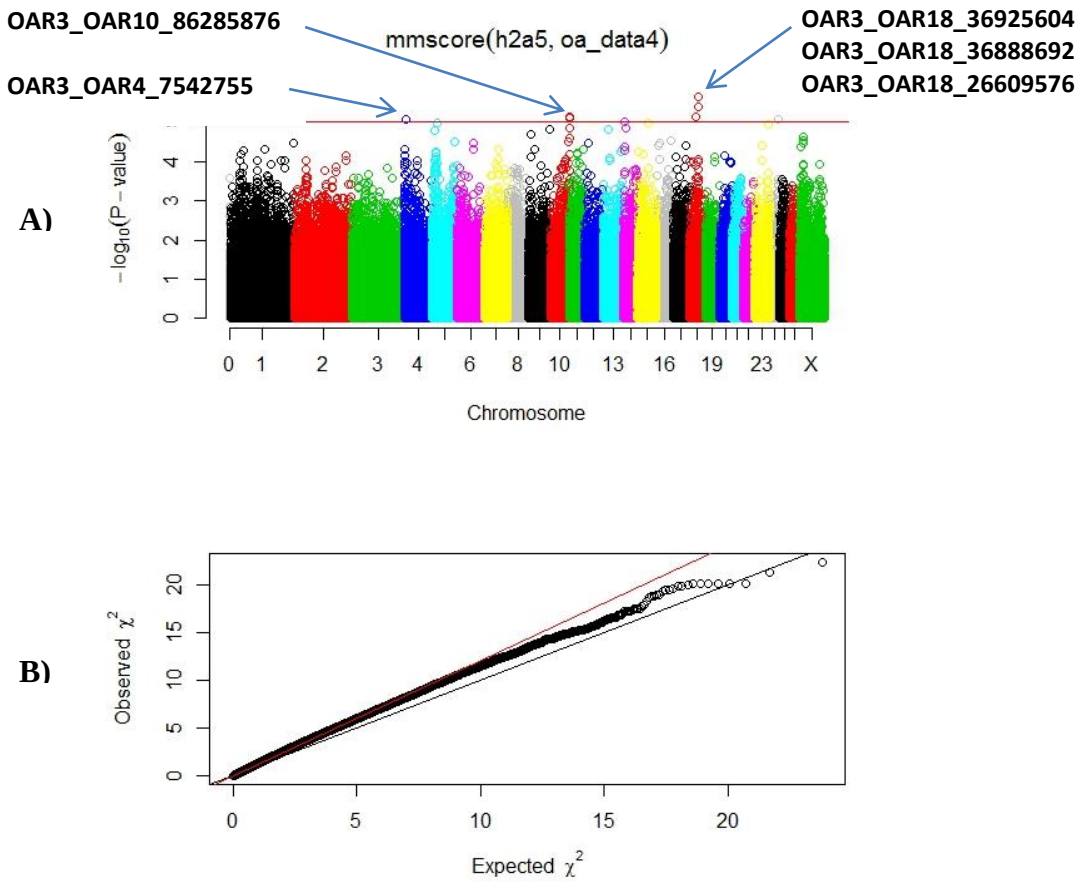

Figure 10 (A) Manhattan plot showing scores of SNPs with respect to chromosomes, each dot represents a SNP and its association to the BLUP score, which was calculated using a score test with a mixed model. (B) QQ-plot showing relationship of observed and expected results from the association test.

For the case-control association (binary trait) the models with λ closest to 1 were both polygenic models including the kinship matrix to generate the formula of association. The score test was used for data both excluding and including additional information about stratification (Figures 11 and 12).

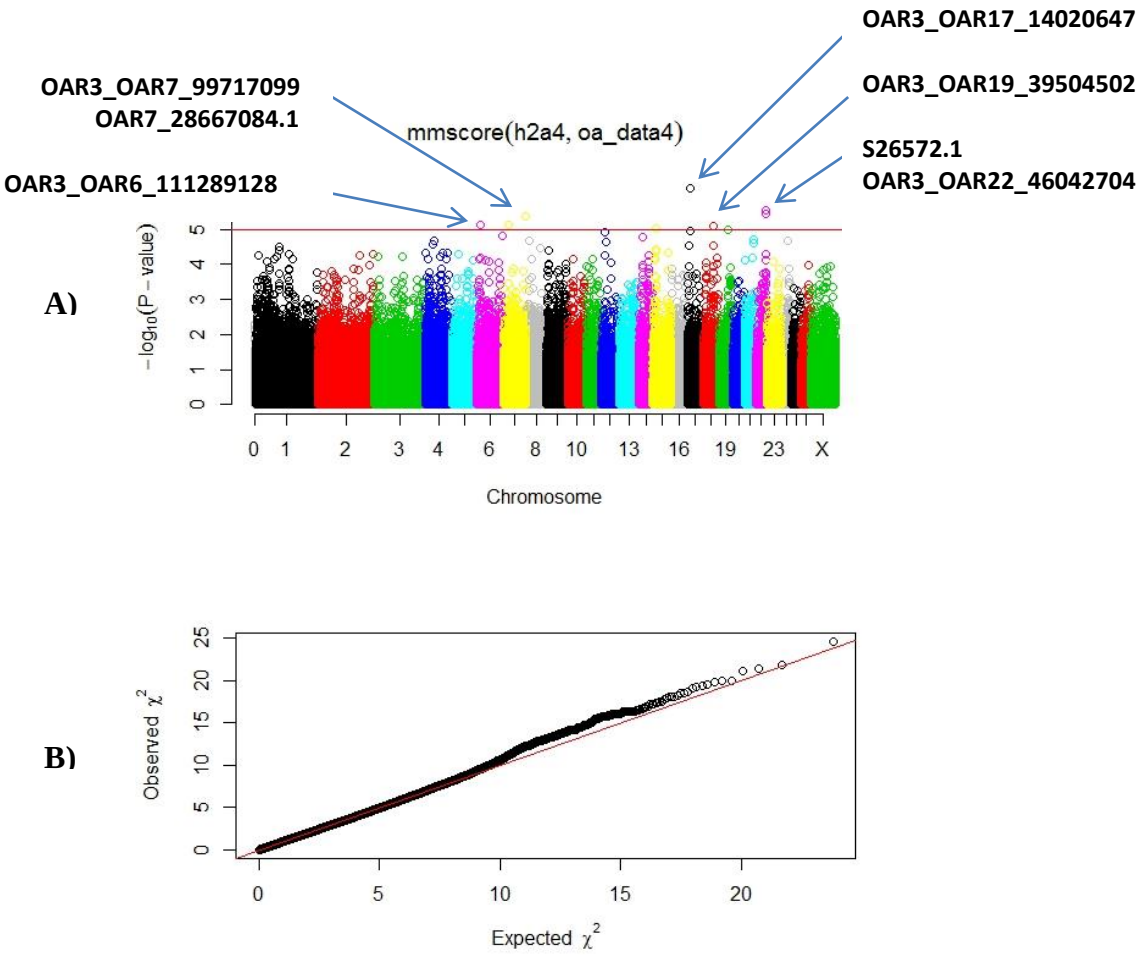

Figure 11 (A) Manhattan plot showing scores of SNPs with respect to chromosomes with a case/control study design. Each dot represents a SNP and its association to muscularity ('high muscle' or 'low muscle') calculated with a score test using a mixed model. (B) QQ-plot showing relationship of observed and expected results from the association test.

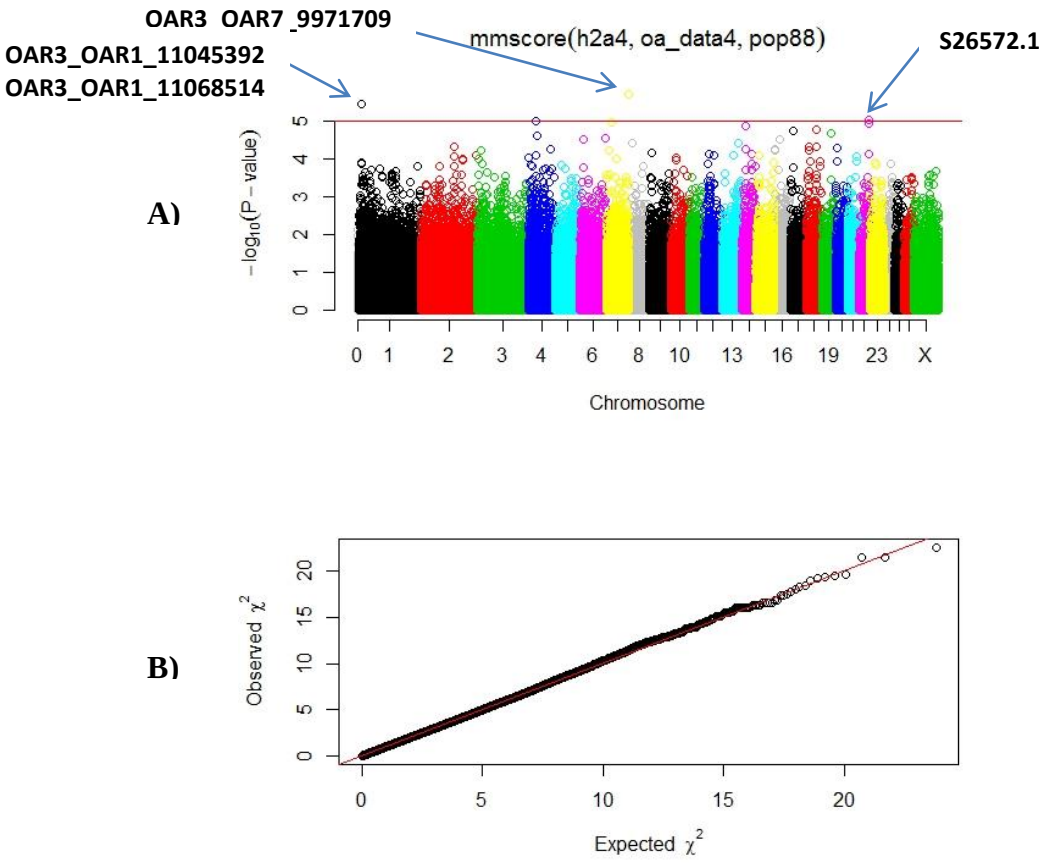

Figure 12 (A) Manhattan plot showing scores of SNPs with respect to chromosomes in a case/control study design. Each dot represents a SNP and it's association to muscularity ('high muscle' or 'low muscle'), calculated with a score test using a mixed model and accounting for stratification. (B) QQ-plot showing relationship of observed and expected results from the association test.

Other models were also tested but had less reliable results ($\lambda > 1$) and will not be considered in following analysis.

The top 25 SNPs, with the lowest corrected p-value for association to muscle size trait were considered for each model. All 25 SNPs have a corrected p-value lower than 1.32 x 10^{-4}. No SNP reached genome-wide significance, but many reached a commonly used threshold of p-values lower than 10^{-5}. Some SNPs appeared in the top 25 hits for more than one model, most often only twice, in two score tests using the same trait (continuous or binary). Some of the SNPs are located within known genes, most of them within introns. Three are located in exons; OAR3_OAR1_11068514 is a missense mutation located in a gene called *CSF3R* that changes C>T which alters the amino acid sequence; Arginine becomes Glutamine, but the length of the protein is preserved (Flicek et al., 2013). OAR3_OAR3_18747575 is a synonymous mutation in a gene known to be expressed in

skeletal muscle and OAR3_OAR3_18756234 is a missense mutation located in the same gene that changes the base T>C and alters the amino acid sequence; Lycine becomes Glutamic acid but the length of the protein is preserved (Flicek et al., 2013). Other SNPs are located in intragenic regions of the sheep genome, some close to known genes and other relatively far from the next gene.

The SNPs for all four models are listed in Appendix 3 (Table A-D) with information about position, alleles, corrected p-value for association and distance to the nearest gene. The nearest genes were annotated using online genetic databases and published literature. The genes that have functions relevant to muscle growth or development, or are known to be expressed in muscle were considered as possible candidate genes.

4.4. Candidate genes

In Table 5 all possible candidate genes are listed. They have some connection to muscle development and are relevant when the muscularity of the Icelandic sheep is considered. A few genes were selected for further analysis. They were selected for PCR amplification and sequencing of their exons. The genes that were selected are the possible candidate genes that are of small size and could be analyzed with few steps. The selected genes were *KLF13*, *PNN* and *GADD45B*. Exon 2 of *KLF13*, exon 9 of *PNN* and all exons of *GADD45B* were magnified with PCR and *GADD45B* was sequenced with Sanger sequencing.

Table 5 Genes that lie close to the top 25 SNPs of all four models and are annotated as having functions related to muscle growth or development. The references are publications where the muscle related function is explained. The SNP column shows the SNPs close to the gene that were associated with muscle traits in the GWAS.

Gene	Chr	Function	Reference	SNPs	Distance from gene
CSF3R	1	Associated with number of regenerating myocytes in the regenerating skeletal muscle (significantly decreased in csf3r −/− mice).	Hara et al, 2011	OAR3_OAR1_11045392 OAR3_OAR1_11068514	20kb 0 kb (exon variant, missense)
ADAM17	3	Cell-cell and cell-matrix interactions, including fertilization, muscle development, and neurogenesis). Widely expressed, for example in skeletal muscle.	Gooz, 2010	OAR3_OAR3_18716983 OAR3_OAR3_18747575 OAR3_OAR3_18753039	8 kb 0 kb (exon variant, synonymous) 0 kb (intron variant)

				OAR3_OAR3_18756234	0 kb (exon variant, missense)
				OAR3_OAR3_18780585	12kb
				S62291.1	10kb
				OAR3_OAR3_18782837	14kb
GADD45B	5	Regulation of growth and apoptosis. *GADD45B* is a paralog to *GADD45A* which is associated with muscular atrophy.	Ebert et al, 2012	OAR3_OAR5_18640578	50kb
GRID2	6	Contains a SNP associated with carcass weight in Hanwoo cattle.	Lee et al, 2012	OAR3_OAR6_31353857	0 kb (intron variant)
SPG11	7	Associated with ALS (Amyotrophic lateral sclerosis)-dysfunction of muscles.	Daoud et al, 2012	OAR3_OAR7_99717099	80kb
DAB2	16	*Dab2* plays an essential role in the early development of skeletal muscle.	Shang et al, 2011	OAR3_OAR16_34117164	900kb
				OAR16_37082988.1	800kb
				OAR3_OAR16_34144581	800kb
FREM3	17	Contains a SNP associated with muscle mass in a previous study.	Kärst et al, 2011	OAR3_OAR17_14020647	0kb (intron variant)
GAB1	17	Plays a role in the migration of muscle progenitor cells.	Vasyutina et al, 2005; Sachs et al, 2000	OAR3_OAR17_14020647	80kb
KLF13	18	Expressed in skeletal muscle but role not yet known.	Haldar et al, 2007	OAR3_OAR18_26609576	150kb
				OAR3_OAR18_26644248	100kb
				OAR3_OAR18_26646267	120kb
				OAR3_OAR18_26649516	120kb
				OAR3_OAR18_26649945	120kb
AKAP6	18	A kinase (PRKA) anchor protein. Function is f.ex. muscle differentiation.	Vargas et al, 2012	OAR3_OAR18_41905736	0 kb (intron variant)
PNN	18	*Pnn* mutant mice exhibited reduced body mass and impaired muscle function during development.	Wu et al, 2014	OAR3_OAR18_47844036	300kb
DOCK1	22	Essential role in embryonic development. A dramatic reduction of all skeletal muscle tissues is observed in *Dock1*-null embryos.	Laurin et al, 2008	S26572.1	200kb
				OAR3_OAR22_46042704	150kb
				OAR3_OAR22_46133225	300kb
				OAR3_OAR22_46154824	300kb
TRRAP	24	Negative regulation of skeletal muscle development.	Ren et al, 2011	OAR3_OAR24_37189272	0 kb (intron variant)

4.5. Candidate gene sequencing

Primers were designed for the three selected genes, *KLF13, PNN* and *GADD45B*, but the PCR was only successfully optimized for one primer pair, for the *GADD45B* gene;

F2 sequence (5' to 3'): TCTCACGGGTTGGGTTGTTG.

R2 sequence (5' to 3'): TTTTGGGGGTGGATTTCGCT.

The gene is 1,572 base-pairs but the primers were designed to cover the coding sequence which is 1,240 base-pairs (Figure 13).

Figure 13 Sequence view of the *GADD45B* gene. The yellow lines are coding DNA sequence (CDS), the red lines are the mRNA and the short green lines are the primers (F2 and R2) (Geneious version 7.1 created by Biomatters. Available from http://www.geneious.com).

The gene was amplified using PCR in 11 samples and all samples were sequenced (Table 6). The sequence reads are all shorter than the gene and the read quality is rather low. After trimming of ends the quality increased slightly, ranging from 31.8% up to 95.8% (Table 6). The sequenced reads were aligned with the *GADD45B* gene from the reference genome (Oar_v3.1). The alignment showed that the reads cover both ends of the gene, but there is a piece missing in the middle for all but one sample (Oa070).

Table 6 Sequencing results of the *GADD45B* gene after trimming, including information about the desired product from the PCR and resulting product size and percentage of high quality base calls (HQ %) in all sequenced samples.

Sample	F2 Size bp	Quality HQ %	R2 Size bp	Quality HQ %
Oa014	271	78.6	357	51.3
Oa070	921	75.7	503	95.8
Oa073	450	60.4	506	52.4
Oa136	304	70.7	518	89.6
Oa142	300	72.3	544	66.4
Oa175	273	60.1	368	68.8
Oa189	275	69.7	369	59.6
Oa220	0	0	149	43.0
Oa227	402	67.7	445	64.0
Oa259	176	31.8	146	51.4
Oa264	266	62.0	326	63.5

The reads cover three exons on the gene and they seem to be mostly conserved between all samples and the reference gene. There is a one base deletion at base-pair 1,243 which is in an intron found in 9 samples (Appendix 4). An insertion of 16 base-pairs was found between base-pairs 226 and 227 in 9 samples (Figure 14). The sequence reads from the other samples did not cover the regions.

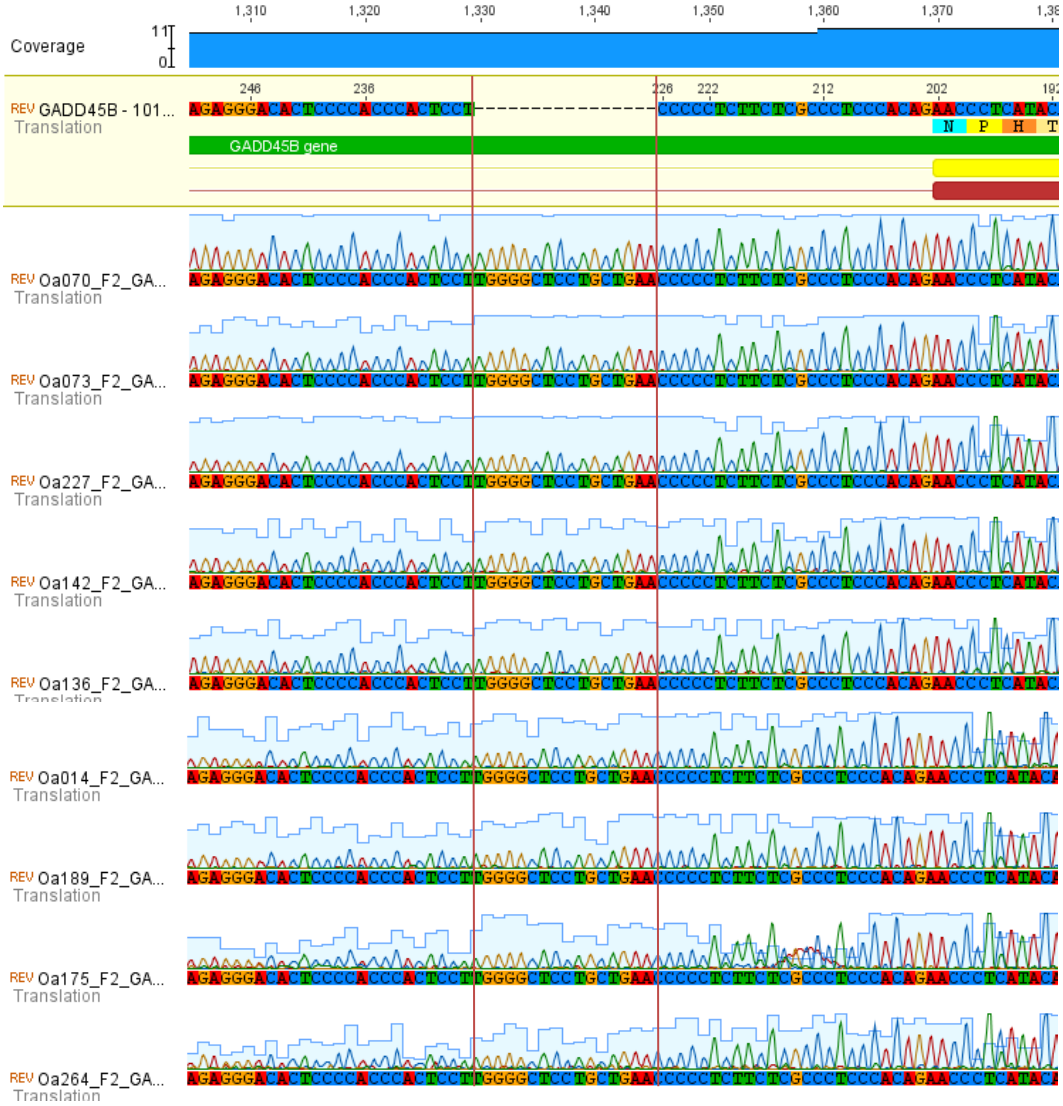

Figure 14 Alignment of sequenced samples of *GADD45B* gene to reference gene revealed a 16 base-pair insertion in 9 samples between the 226 and 227 base-pairs in the reference gene (Geneious version 7.1 created by Biomatters. Available from http://www.geneious.com).

5. Discussion

It is known that domestic animals are useful for exploring genotype-phenotype relationships. This is because they have a history of thousands of generations which is long enough to allow an evolution of phenotypes. At the same time it is not too old to allow a lot of dissection of phenotypic diversity. Therefore they are ideal for these studies (Andersson, 2009). One method to study this relationship is the genome-wide association approach used here. This is the first genome-wide association study conducted using samples of Icelandic sheep. It creates a foundation of genomic information about the Icelandic sheep breed. The SNPs that were genotyped are more than 600,000 and spread evenly across the sheep genome so the possibilities for analysis of the data are many. Phenotypic information about color and horned/polled status exists for all genotyped samples and for some samples there are even more information about body composition and about occurrence of yellow fat.

5.1. Genetic diversity measures

Genetic diversity was estimated using a few different parameters. Average homozygosity of samples ranged from 0.67 to 0.68 which is higher than was seen in an older study which included Nordic and Southern breeds. Icelandic sheep had a H_o=0.537 in that study, with the mean of Southern breeds H_o=0.641) (Handley et al., 2007). The mean heterozygosity for a SNP was 0.341 which is within the range that was reported in a genome-wide study of 74 sheep breeds using a 50K SNP chip, 0.24 to 0.38 (Kijas et al., 2012). The average inbreeding coefficient, F, ranged from 0.06-0.07 per animal. No older results of inbreeding coefficients estimated from genomic data exist for the Icelandic sheep breed but an estimation based on pedigree data reported an inbreeding coefficient even lower (0.01-0.05) (Jónmundsson & Eyþórsdóttir, 2014). When compared to inbreeding coefficients of other sheep breeds the values seem to be normal. The inbreeding coefficient ranged from 0.07-0.42 in the study including 74 sheep breeds (Kijas et al., 2012). In a study of Sicilian sheep breeds the within population inbreeding estimate F_{is} was 0.032 and total inbreeding estimate F_{it} was 0.080 (Tolone et al., 2012). In a study on Bolivian alpacas the highest F_{is} values for an individual population was 0.114 and the lowest was 0.019 (Barreta et al, 2012). However, both these studies used microsatellite data for calculation of the inbreeding coefficients. It is possible that the inbreeding coefficient calculated in this study is slightly overestimated because of overrepresentation of leadersheep in the samples

compared to the breed as a whole. In this study the leadersheep are 28/96 samples = 29% but in the whole breed they are about 1,300/500,000 = 0.26%. The leadersheep had the highest inbreeding coefficient of the groups, the highest average value of linkage disequilibrium and had the second highest genomic kinship coefficient, so they seem to include more similar individuals then the other sheep groups.

5.2. Data substructure

The dataset used in this study shows substructure. The most obvious subgroup is the leadersheep, ending up furthest away from the other groups when looking at the first component of the multidimensional scaling picture. Population structure analysis resulted in three structural groups; leadersheep, Hestur sheep and the other sheep, with the greatest difference between leadersheep and other sheep. When looking at the second component of the MDS picture the sheep from Hestur seemed to diverge from the other sheep, however there were still some samples from the artificial insemination station among them. This was expected because the rams on the station include rams that come from or are directly related to sheep from the Hestur farm. The leadersheep are considered to be a subgroup because of their genetic difference as well as their phenotype. They have been included as a rare subtype of the Icelandic breed in a study about molecular variation in northern European sheep (Tapio et al., 2005). In that study the leadersheep had different results for estimations of molecular diversity. The difference of the leadersheep compared to the other sheep in this study supports these previous results. The Stafh sheep cluster with the ‘high muscle’ sheep when the first component is considered although they are phenotypically more similar to the leadersheep than the muscular sheep. They only diverge from the others as a special group when looking at the third component, which explains only a small part of the variation. This is a bit surprising since this is a flock that is outside the national recording system and has been considered unselected, while the muscular sheep are all strictly selected for muscularity.

The five outliers that were removed before the genome-wide association analysis were all leadersheep, coming from two different farms. It is possible that those leadersheep come from a different genetic origin than the others, or that they have either never been mixed with the other leadersheep or that they have some relation with other sheep than the leadersheep.

5.3. Possible candidate genes

The genes that lie nearest to the top scoring SNPs were all studied and relatedness to muscling traits was listed. If more than one gene was found close to the SNP relevant for muscles then they were all listed. Many of the genes are not very well known, at least not in sheep, but are listed as known by projection. This means that the gene is orthologous to a gene with known function in other species (Flicek et al., 2013). If the gene is known to have functions related to muscle traits in bovine species, mice or humans it is speculated that it has the same function in sheep and is therefore listed as well.

The *CSF3R* gene is associated with the number of myocytes in the regenerating skeletal muscle, with the number significantly decreased in *CSF3R*-null mutant mice (Hara et al., 2011). Two SNPs in one of the case-control models (mixed model including strata information) were associated with muscularity and one is located in an exon of *CSF3R* (see Table 5). The protein product of *CSF3R* has been detected at high or medium expression levels in 5 of 79 analyzed normal tissue cell types in the Human Protein Atlas (Uhlen et al., 2010). It was for example detected in placenta at high expression levels and in skin and heart but has not been detected in skeletal muscle cells (Uhlen et al., 2010). It would be interesting to further study the function of the SNP in the exon to find out if it affects muscle or even other related traits.

ADAM17 is a gene located on chromosome 3. It is widely expressed and has for example been reported in skeletal muscle but also in brain, heart and kidney (Gooz, 2010). *ADAM17* is one of many ADAM enzymes which are important contributors to many physiological and pathophysiological processes. *ADAM17*, in particular, has been described as a regulator of almost every cellular event from proliferation to migration (Gooz, 2010). There were 7 SNPs located within or close to *ADAM17* associated with muscularity in this study (Table 5) and therefore it was considered as a possible candidate gene. The gene has however not previously been associated with muscle size and it remains to be sequenced in the samples from this study.

GADD45B is a gene located on chromosome 5; with a role in regulation of growth and apoptosis. It is a paralog to *GADD45A* which has been associated with muscular atrophy (Ebert et al., 2012) and is therefore an interesting possible candidate gene. It is located around 50 kb from a SNP that was associated with muscularity (Table 5). It has been detected at high or medium expression levels in 54 of 82 analyzed normal human tissue

cell types in the Human Protein Atlas. It was for example detected at high levels in placenta and medium levels in skeletal muscle tissue cells (Uhlen et al., 2010).

GRID2 is a gene on chromosome 6, it contains 4 SNPs that have been associated with carcass weight in Hanwoo cattle (Lee, Lia & Kim, 2012) and one SNP that was associated with muscularity in this study (Table 5). It has been detected at medium expression levels in 7 of 81 analyzed normal tissue cell types, for example in gallbladder, small intestine and in liver. It has not been detected in skeletal muscle cells (Uhlen et al., 2010).

SPG11 on chromosome 7 has been associated with ALS (Amyotrophic lateral sclerosis) which can affect the function of muscles (Daoud et al., 2012). It has not been associated with muscle size, however. It has been detected at medium expression levels in 44 of 82 analyzed normal tissue cell types in humans but was only detected at very low expression levels in skeletal and smooth muscle cells (Uhlen et al., 2010). There was one SNP associated with muscularity located close to *SPG11* (Table 5).

DAB2 is located on chromosome 16, within one Mb from three SNPs that were associated with muscularity (Table 5). It is an intracellular adaptor protein as well as a potential tumor suppressor. It is involved in the MAKP signaling pathway which is important in muscle development and has therefore been suggested as a factor in the early development of skeletal muscle (Shang, Samuel, Zhao, & Chan, 2011). It has been detected at high or medium expression levels in 7 of 80 analyzed normal human tissue cell types, for example in placenta and epididymis, but has not been detected in skeletal or smooth muscle cells (Uhlen et al., 2010).

FREM3 is a gene on chromosome 17 that is known to contain a SNP that was associated with muscle mass (Kärst et al., 2011). In this study it contained one SNP that is associated with muscularity, in an intron (Table 5). It has been detected at high or medium expression levels in 9 of 79 analyzed normal human tissue cell types. It was detected at a high level in cerebellum and medium levels in lung and kidney for example. It has not been detected in skeletal or smooth muscle cells (Uhlen et al., 2010).

GAB1 is a gene that codes for a docking protein that binds phosphorylated c-Met receptor tyrosine kinase. It is important for migration of myogenetic precursor cells into the limb (Sachs et al., 2000). Reduced numbers of muscle progenitor cells reach the forelimb in *GAB1* mutant mice. Smaller size or even absence of limb muscles has also been seen in

GAB1 mutants (Vasyutina et al., 2005). There was one SNP with association to muscle located close to *GAB1* gene in both case-control models (Table 5).

The Krüppel like factor 13 (*KLF13*) gene is known to be expressed in skeletal muscle but its role remains unknown. It is a basic transcription element-binding protein that activates a minimal promoter in the *SM22α* gene which is specific for smooth muscle (Haldar, Ibrahim & Jain, 2007). *KLF13* and other *KLFs* have been identified in developing or mature skeletal muscle but few reports describe their possible role and regulation in the tissue. The KLF15 protein has been reported as a regulator of expression of the glucose transporter *GLUT4* gene and fasting induced transcription of mitochondrial acetyl-CoA sythetase-2 in skeletal muscle (Haldar et al., 2007). There were five SNPs associated with muscularity located within or close to *KLF13* in the GWAS (Table 5).

AKAP6, also known as *mAKAP* is a kinase (PRKA) anchor protein and is a regulatory factor for MEF2 which is a key element in induction of skeletal muscle differentiation. The interaction between *mAKAP* and MEF2 is required for differentiation of precursor cells in skeletal muscle (Vargas, Tirnauer, Glidden, Kapiloff & Dodge-Kafka, 2012). The *AKAP6* gene is located on chromosome 18 and is interesting because of its function in myogenesis and because of the already known TM-QTL locus on chromosome 18. It has been detected at medium expression levels in 2 of 81 analyzed normal tissue cell types, heart and skeletal muscle cells (Uhlen et al., 2010). One SNP associated with muscularity was located within an intron of the *AKAP6* gene (Table 5).

PNN or *Pinin* is a gene on chromosome 18 that has recently been connected with reduced body mass and impaired muscle function during development in mice (Wu, Hsu, Wu, Hu, & Ouyang, 2014). The study also concluded that down regulation of *PNN* in skeletal muscle can cause muscular dystrophy (Wu et al., 2014). Additionally it has been detected at high or medium expression levels in 73 of 75 analyzed normal human tissue cell types, for example in duodenum and skin and at medium levels in skeletal and smooth muscle tissue cell types (Uhlen et al., 2010). This gene is therefore an interesting possible candidate for muscle traits although there was only one SNP close to the gene that was associated with muscularity in one of the GWAS models (see Table 5).

Dock1 gene (also known as *Dock180*) codes for an atypical Rho GTPase activator and has an essential role in embryonic development. A dramatic reduction of all skeletal muscle

tissues was reported in *Dock1*-null mouse embryos (Laurin et al., 2008). This defect in the embryos was explained by deficiency in myoblast fusion and *Dock1* has been identified (along with the protein product of the *Dock5* gene) as an important regulator of the fusion step in myogenesis in mammals (Laurin et al., 2008). The *Dock1* gene is annotated in the sheep genome as a dedicator of cytokinesis 1 but has not been associated with muscle traits in sheep. It has been detected at high or medium levels in 31 of 81 analyzed normal human tissue cell types. It is predicted to be intracellular and has been detected in for example thyroid gland, breast and smooth muscle, but not in skeletal muscle cells (Uhlen et al., 2010). There were four associated SNPs located within 300 kb of the *Dock1* gene (Table 5).

TRRAP is a gene located on chromosome 24 and had one SNP in an intron that was associated with muscularity (Table 5). *TRRAP* is an adapter protein which participates in gene expression regulation and cell proliferation. It was one of differentially expressed genes in a study using the first specialized transcriptome-wide sheep oligo DNA microarray on fetal *longissimus* muscle in Texel sheep, which have high muscle proportion and low fat, and Ujumqin sheep, which have low muscle proportion and more fat. It was suggested that *TRRAP* negatively regulates skeletal development through canonical and Wnt/calcium pathways in sheep (Ren et al., 2011). It has been detected at high or medium expression levels in 71 of 80 analyzed tissue cell types, for example in liver and heart. It was detected at high levels in both skeletal and smooth muscle (Uhlen et al., 2010). One associated SNP was located in an intron of the *TRRAP* gene (Table 5).

5.3.1 Candidate gene sequencing

Three genes were selected for PCR amplification and sequencing but only one gene was successfully amplified and sequenced. To finish the amplification of the other genes, new primers need to be designed and the PCR optimized for the new primers. The results of sequencing of the *GADD45B* gene did not reveal any visible functional changes of the protein product of the gene. The deletion that was detected in the samples was located in an intron and the detected insertion as well. Introns are not translated to amino-acids, so functional changes resulting from the variations cannot be detected by comparing protein products of the samples and the reference gene. This does not mean that the variations are meaningless, only that more analysis is needed to determine if they have some sort of phenotypical effect. The variations were detected in almost all the sequenced samples and

no variation was found between the Icelandic samples in the gene. A part of the gene is missing in most sample reads, including exon three, because of incomplete sequence reads. New primers need to be designed so this area can be amplified and sequenced to reveal the whole sequence of the gene in those samples.

5.4. GWAS implications

The SNP chip was designed using the dbSNP build 140 which includes 35,439,092 SNPs and can be accessed on ftp://ftp.ncbi.nih.gov/snp/organisms/sheep_9940/. The dbSNP is developed and hosted by NCBI and contains all identified genetic variation in an organism. The SNP discovery process is usually derived from few individuals from selected populations (Albrechtsen, Nielsen & Nielsen, 2010). The sheep reference genome was for example generated by sequencing one Texel ewe and one Texel ram (Jiang et al., 2014). However, the dbSNP build is based on more resources than the two sheep sequenced for the reference genome.

Ascertainment bias in GWA studies is generated by SNP discovery process because of genetic differences between the breed of study and the breed used to select the SNPs for the SNP chip. The breed being studied can be heterozygous at different loci than those that are on the chip. This can result in a skewed assessment of genetic diversity (Albrechtsen et al., 2010). It can also lead to the study overlooking loci that are relevant for the trait under study but it is considered unlikely that it causes false-positives (Albrechtsen et al., 2010).

It is important to use SNPs that represent the whole genome of the sampled animals under study for the GWAS to be reliable. This can be difficult as 30% of common variants of cases and controls might remain undetected (Wang et al., 2005). However, this can be corrected by re-sequencing a larger set of genomes of unrelated individuals. Also, it is known that many SNPs have alleles that are in strong LD with nearby SNPs so the SNPs that are used on the chip can represent enough coverage of the region under study (Wang et al., 2005). When the true causative SNP is not on the genotyping chip there will typically be several SNPs on the chip which are correlated with it (Spencer et al., 2009). One or more of these could give a signal of significant association and hence allow detection of the locus (Spencer et al., 2009) and therefore all the top SNPs of the analysis were closely studied.

A large sample size is considered very important to maximize reliability of the results of a GWAS. It has been suggested that to achieve results with relatively high statistical power a sample size of more than 2,000 is required (Spencer et al., 2009) and sample sizes in GWAS of domestic animals vary from 329 sheep (Zhang et al., 2013) to ca. 1,000 sheep (White et al., 2012) and 2,000 cattle (Pausch et al., 2011). In this study there were only 96 samples of sheep genotyped so the power of the study is greatly affected by the sample size. According to Table 1, 96 samples can only generate power above 90% when the effect size of the associated SNP is 0.3 or higher in a study using around 500,000 SNPs and a Bonferroni correction to calculate p-value of significance. When using a case/control study design with 50 cases and 50 controls, like in this study, then the frequency of the high risk allele must be 0.1 and the effect of the allele must be big (Aa=3 and AA=4) (Table 2) to get a significant association with 80% power. If the frequency of the risk allele is higher (0.2) then its effect needs to be even greater to achieve 80% power. If the effect is lower, then more samples are needed to generate significant results with power above 80%.

A Bonferroni correction adjusts a p-value from a common threshold of 0.05 to 0.05/k, where k is the number of statistical tests conducted in a study. So, for a GWAS using 500,000 SNPs, statistical significance of a SNP association would be set at 1×10^{-7} (Bush & Moore, 2012) and for this study it would be $0.05/606{,}006 = 8.25 \times 10^{-8}$. Another widely used significance threshold is based on an effective number of statistical tests that need to be corrected for depending on numbers of independent genomic regions of a specific population. This is called genome-wide significance and for European-descent populations the threshold has been estimated to be 7.2×10^{-8} (Bush & Moore, 2012). The SNP with the highest score in the results of the GWAS in this study had a p-value of 5.26×10^{-7}. The others all had slightly higher p-values, ranging from 1.36×10^{-6} to 1.32×10^{-4}. The highest scoring SNP in this study does not reach the Bonferroni threshold or genome-wide significance and a likely explanation is the small sample size.

To generate more accurate results it is possible to carry out replication studies, make sure that the possible bias is accounted for in the estimation of association and use case/control samples that are similar in all way apart from the trait under study (Wang et al., 2005). In the case of this study it would be possible to do some replications in the future but it is more difficult to make sure that the samples are similar. The cases and controls in this study differ in other phenotypic traits apart from muscularity. The sheep with ‘low muscle’ usually have a less compact conformation and longer legs than the ‘high muscle’ animals

and even different colors. Another problem in this study is the form of the phenotype. The BLUP score does not have the same accuracy for all animals, because the score is based on information about its offspring and other relatives. The amount of information used to calculate the score varies between individuals. Using individual measurements of muscle thickness might generate more accurate results of association.

6. Conclusion

The results of this study demonstrated substructure in the dataset used. The so called leadersheep clearly differed from the other sheep sampled in the study. Genetic diversity parameters showed average diversity measures with an inbreeding coefficient F ranging from 0.07-0.08. Icelandic leadersheep differed from the other groups in a multidimensional scaling picture based on genomic kinship. Genetic diversity was measured using a few parameters; the results were similar to previous studies of genetic diversity that have included Icelandic sheep.

Genome-wide association analysis for muscularity resulted in few significant SNPs but many that scored high in association. Close to the highest ranking SNPs there were 13 genes identified as possible candidate genes for muscularity of the Icelandic sheep. Those genes are only possible candidates and should be studied further to find out if they are real candidates. Three of these genes were selected for PCR and sequencing to investigate variation between a few samples. The *GADD45B* gene was successfully magnified with PCR and sequenced in 11 samples but the other genes remain to be sequenced along with other possible candidate genes. Alignment of reads of the sequenced samples to the reference gene did not reveal any functional changes in the *GADD45B* gene. The middle part of the *GADD45B* gene, including exon three, was missing from all reads because of incomplete sequence reads and needs to be sequenced again to make sure if there is any variation causing a functional change in the resulting protein.

The results of the genome-wide association analysis should be carefully interpreted. The samples are too few to show genome-wide significance of the top scoring SNPs. With so many SNPs included in the study like here, the tests for association are many and there is a possibility of false results. However, the results can be used as indicators and suggestions for further studies. To improve the statistical power it would be possible to do a replication study, preferably including more genotyped samples.

7. References

Adalsteinsson, S. (1981). Origin and conservation of farm animal populations in Iceland. *Zeitschrift für Tierzüchtung und Züchtungsbiologie, 98*(1-4), 258-264.

Albrechtsen, A., Nielsen, F. C. & Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol, 27*(11), 2534-2547.

Andersson, L. (2009). Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. *Genetica, 136*(2), 341-349.

Archibald, A. L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M. A. M., Harlizius, B., et al. (2010). Pig genome sequence-analysis and publication strategy. *BMC Genomics, 11*(1), 438.

Aulchenko, Y. S., Ripke, S., Isaacs, A. & Van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics, 23*(10), 1294-1296.

Árnason, T. & Jónmundsson, J. V. (2008). Multiple trait genetic evaluation of ewe traits in Icelandic sheep. *Journal of Animal Breeding and Genetics, 125*(6), 390-396.

Bai, Y., Sartor, M. & Cavalcoli, J. (2012). Current status and future perspectives for sequencing livestock genomes. *Journal of Animal Science and Biotechnology, 3*(1), 1-6.

Barreta, J., Iniguez, V., Saavedra, V., Romero, F., Callisaya, A. M., Echalar, J., et al. (2012). Genetic diversity and population structure of Bolivian alpacas. *Small Ruminant Research, 105*(1), 97-104.

Beckmann, J. S., Estivill, X. & Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics, 8*(8), 639-646.

Bentzinger, C. F., Wang, Y. X. & Rudnicki, M. A. (2012). Building muscle: molecular regulation of myogenesis. *Cold Spring Harbor Perspectives in Biology, 4*(2), 1-16.

Beuzen, N. D., Stear, M. J. & Chang, K. C. (2000). Molecular markers and their use in animal breeding. *The Veterinary Journal, 160*(1), 42-52.

Bhuiyan, M. S. A., Kim, N. K., Cho, Y. M., Yoon, D., Kim, K. S., Jeon, J. T., et al. (2009). Identification of SNPs in MYOD gene family and their associations with carcass traits in cattle. *Livestock Science, 126*(1), 292-297.

Bismuth, K. & Relaix, F. (2010). Genetic regulation of skeletal muscle development. *Experimental Cell Research, 316*(18), 3081-3086.

Bjarnason, E. I. & Kristjánsson, Þ. (2012). Þróun skyldleikaræktar í íslenska sauðfjárstofninum (e. Evolution of inbreeding in the Icelandic sheep breed). *Búnaðarblaðið Freyja, 2*(2), 9-12. –In Icelandic.

Bolormaa, S., Pryce, J. E., Hayes, B. J. & Goddard, M. E. (2010). Multivariate analysis of a genome-wide association study in dairy cattle. *Journal of Dairy Science, 93*(8), 3818-3833.

Boman, I. A., Klemetsdal, G., Blichfeldt, T., Nafstad, O. & Vage, D. I. (2009). A frameshift mutation in the coding region of the myostatin gene (MSTN) affects carcass conformation and fatness in Norwegian White Sheep (Ovis aries). *Animal Genetics, 40*(4), 418-422.

Braun, T. & Gautel, M. (2011). Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nature reviews Molecular cell biology, 12*(6), 349-361.

Burt, D. W. (2005). Chicken genome: current status and future opportunities. *Genome Research, 15*(12), 1692-1698.

Bush, W. S. & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput Biol, 8*(12), 1-11.

Byrne, K., Vuocolo, T., Gondro, C., White, J. D., Cockett, N. E., Hadfield, T., et al. (2010). A gene network switch enhances the oxidative capacity of ovine skeletal muscle during late fetal development. *BMC Genomics, 11*(378).

Chauhan, T. & Rajiv, K. (2010). Molecular markers and their applications in fisheries and aquaculture. *Advances in Bioscience and Biotechnology, 1*, 281-291.

Chessa, B., Pereira, F., Arnaud, F., Amorim, A., Goyache, F. l., Mainland, I., et al. (2009). Revealing the history of sheep domestication using retrovirus integrations. *Science, 324*(5926), 532-536.

Clop, A., Vidal, O. & Amills, M. (2012). Copy number variation in the genomes of domestic animals. *Animal Genetics, 43*(5), 503-517.

Daoud, H., Zhou, S., Noreau, A., Sabbagh, M., Belzil, V., Dionne-Laporte, A., et al. (2012). Exome sequencing reveals *SPG11* mutations causing juvenile ALS. *Neurobiology of Aging, 33*(4), 839.e5-839e9.

Domestic Animal Diversity Information System DAD-IS. (2014). Number of breeds by species and country. Retrieved 10.11, 2014, from http://dad.fao.org/

Dominik, S., Henshall, J. M. & Hayes, B. J. (2012). A single nucleotide polymorphism on chromosome 10 is highly predictive for the polled phenotype in Australian Merino sheep. *Animal Genetics, 43*(4), 468-470.

Dunner, S., Sevane, N., García, D., Cortés, O., Valentini, A., Williams, J. L., et al. (2013). Association of genes involved in carcass and meat quality traits in 15 European bovine breeds. *Livestock Science, 154*(1), 34-44.

Dýrmundsson, Ó. (2002). Leadersheep: the unique strain of Iceland sheep. *Animal Genetic Resources Information, 32*, 45-48.

Dýrmundsson, Ó. (2011). *The conservation and utiliztion of the genetically diverse, native Icelandic livestock breeds, with reference to selfsufficiency and national food security* Paper presented at the RBI 8th Global Conference on the Conservation of Animal Genetic Resources Tekirdag, Turkiye.

Ebert, S. M., Dyle, M. C., Kunkel, S. D., Bullard, S. A., Bongers, K. S., Fox, D. K., et al. (2012). Stress-induced skeletal muscle Gadd45a expression reprograms myonuclei and causes muscle atrophy. *Journal of Biological Chemistry, 287*(33), 27290-27301.

Einarsson, E., Eythórsdóttir, E., Smith, C. R. & Jónmundsson, J. V. (2014). Genetic parameters for lamb carcass traits assessed by video image analysis, EUROP classification and in vivo measurements. *Icelandic Agricultural Sciences* (accepted for publication).

Eythórsdóttir, E. (2012). Growth and carcass characteristics of Icelandic lambs - a review. *Icelandic Agricultural Sciences, 25*, 59-66.

Eythórsdóttir, E., Tapio, M., Olsaker, I. , Kantanen, J., Miceikiene, I., Holm, L.-E. et al. (2002). Uppruni og erfðabreytileiki norrænna sauðfjárkynja (e. Origin and genetic diversity of Northern sheep breeds). *Ráðunautafundur2002,* 313-315. –In Icelandic.

Feng, S., Wang, S., Chen, C. C., & Lan, L. (2011). GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits. *BMC Genetics, 12*(1), 12.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2013). Ensembl 2014. *Nucleic Acids Research*.

Fontanesi, L., Beretti, F., Martelli, P. L., Colombo, M., Dall'Olio, S., Occidente, M., et al. (2011). A first comparative map of copy number variations in the sheep genome. *Genomics, 97*(3), 158-165.

Frankham, R., Briscoe, D. A. & Ballou, J. D. (2002). *Introduction to conservation genetics*: Cambridge University Press.

García-Gámez, E., Gutiarrez-Gil, B., Sahana, G., Sánchez, J.-P., Bayón, Y. & Arranz, J.-J. (2012). GWA analysis for milk production traits in dairy sheep and genetic support for a QTN influencing milk protein percentage in the LALBA gene. *Plos One, 7*(10).

Gooz, M. (2010). ADAM-17: the enzyme that does it all. *Critical Reviews in Biochemistry and Molecular Biology, 45*(2), 146-169.

Gordon, E. S., Gordish Dressman, H. A. & Hoffman, E. P. (2005). The genetics of muscle atrophy and growth: the impact and implications of polymorphisms in animals and humans. *The International Journal of Biochemistry & Cell Biology, 37*(10), 2064-2074.

Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. (2000). Heritability of a trait. In *An introduction to genetic analysis* (7 ed.). New York: W. H. Freeman.

Groeneveld, L. F., Lenstra, J. A., Eding, H., Toro, M. A., Scherf, B., Pilling, D., et al. (2010). Genetic diversity in farm animals - a review. *Animal Genetics, 41*(s1), 6-31.

Hadjipavlou, G., Matika, O., Clop, A. & Bishop, S. C. (2008). Two single nucleotide polymorphisms in the myostatin (GDF8) gene have significant association with muscle depth of commercial Charollais sheep. *Animal Genetics, 39*(4), 346-353.

Hagstofa Íslands (1997). Hagskinna: Sögulegar hagtölur um Íslands (e. Icelandic historical statistics). G. Jonsson & M. M. S (Eds.). Reykjavík. -In Icelandic

Haldar, S. M., Ibrahim, O. A. & Jain, M. K. (2007). Kruppel-like Factors (KLFs) in muscle biology. *Journal of Molecular and Cellular Cardiology, 43*(1), 1-10.

Handley, L. J. L., Byrne, K., Santucci, F., Townsend, S., Taylor, M., Bruford, M. W., et al. (2007). Genetic structure of European sheep breeds. *Heredity, 99*(6), 620-631.

Hara, M., Yuasa, S., Shimoji, K., Onizuka, T., Hayashiji, N., Ohno, Y., et al. (2011). G-CSF influences mouse skeletal muscle development and regeneration by stimulating myoblast proliferation. *The Journal of Experimental Medicine, 208*(4), 715-727.

Hartl (1994). Population genetics and evolution. In *Genetics* (3 ed., pp. 197-224). Boston: Jones and Bartlett Publishers.

Haynes, G. D. & Latch, E. K. (2012). Identification of Novel Single Nucleotide Polymorphisms (SNPs) in Deer (Odocoileus spp.) Using the BovineSNP50 BeadChip. *Plos One, 7*(5), 11.

He, H., Zhang, H.-l., Li, Z.-x., Liu, Y. & Liu, X.-l. (2014). Expression, SNV identification, linkage disequilibrium, and combined genotype association analysis of the muscle-specific gene CSRP3 in Chinese cattle. *Gene, 535*(1), 17-23.

Helgason, A., Yngvadóttir, B., Hrafnkelsson, B., Gulcher, J. & Stefánsson, K. (2004). An Icelandic example of the impact of population structure on association studies. *Nature Genetics, 37*(1), 90-95.

Hickford, J. G. H., Forrest, R. H., Zhou, H., Fang, Q., Han, J., Frampton, C. M., et al. (2010). Polymorphisms in the ovine myostatin gene (MSTN) and their association with growth and carcass traits in New Zealand Romney sheep. *Animal Genetics, 41*(1), 64-72.

Hirschhorn, J. N. & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics, 6*(2), 95-108.

Hopkins, D. L., Fogarty, N. M. & Mortimer, S. I. (2011). Genetic related effects on sheep meat quality. *Small Ruminant Research, 101*(1-3), 160-172.

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., et al. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science, 344*(6188), 1168-1173.

Johansson, B. G. (1972). Agarose gel electrophoresis. *Scandinavian Journal of Clinical & Laboratory Investigation, 29*(S124), 7-19.

Jónmundsson, J. V. & Eyþórsdóttir, E. (2013). Erfðir og kynbætur sauðfjár (e. Genetics and breeding of sheep). In R. Sigurðardóttir (Ed.), *Sauðfjárrækt á Íslandi (e. Sheep-farming in Iceland)*. Reykjavík: Uppheimar. –In Icelandic.

Kärst, S., Cheng, R., Schmitt, A. O., Yang, H., de Villena, F. P. M., Palmer, A. A., et al. (2011). Genetic determinants for intramuscular fat content and water-holding capacity in mice selected for high muscle mass. *Mammalian Genome, 22*(9-10), 530-543.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28*(12), 1647-1649.

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M., et al. (2012). Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLoS Biology, 10*(2).

Kijas, J. W., McCulloch, R., Edwards, J. E. H., Oddy, V. H., Lee, S. H. & van der Werf, J. (2007). Evidence for multiple alleles effecting muscling and fatness at the Ovine GDF8 locus. *BMC Genetics, 8*, 11.

Klover, P., Chen, W., Zhu, B.-M. & Hennighausen, L. (2009). Skeletal muscle growth and fiber composition in mice are regulated through the transcription factors STAT5a/b: linking growth hormone to the androgen receptor. *FASEB Journal, 23*(9), 3140-3148.

Landssamtök sauðfjárbænda (2013). *Staðreyndir um íslenska sauðfjárrækt (e. Facts about Icelandic sheep farming)*. Retrieved 24.03.2014, from http://www.saudfe.is/images/stories/baeklingur_pdf_vefur_2013.pdf

Laurin, M., Fradet, N., Blangy, A., Hall, A., Vuori, K. & Cóté, J.-F. (2008). The atypical Rac activator Dock180 (Dock1) regulates myoblast fusion in vivo. *Proceedings of the National Academy of Sciences, 105*(40), 15446-15451.

Lee, C., Abdool, A. & Huang, C. H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC bioinformatics, 10*(Suppl 1), S73.

Lee, J. H., Lia, Y. & Kim, J. J. (2012). Detection of QTL for Carcass Quality on Chromosome 6 by Exploiting Linkage and Linkage Disequilibrium in Hanwoo. *Asian-Australasian Journal of Animal Sciences, 25*(1), 17-21.

Lee, S. J. (2007). Quadrupling Muscle Mass in Mice by Targeting TGF-ß Signaling Pathways. *Plos One, 2*(8).

Leymaster, K. A. (2002). Fundamental aspects of crossbreeding of sheep: Use of breed diversity to improve efficiency of meat production. *Sheep and Goat Research Journal, 17*(3), 50-59.

Li, M. H., Tiirikka, T. & Kantanen, J. (2013). A genome-wide scan study identifies a single nucleotide substitution in ASIP associated with white versus non-white coat-colour variation in sheep (Ovis aries). *Heredity, 112*(2), 122-131.

Liu, J.-P., Baker, J., Perkins, A. S., Robertson, E. J. & Efstratiadis, A. (1993). Mice carrying null mutations of the genes encoding insulin-like growth factor I (Igf-1) and type 1 IGF receptor (Igf1r). *Cell, 75*(1), 59-72.

Liu, L., Zhang, D., Liu, H. & Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. *BMC Bioinformatics, 14*(1), 132.

McCarthy, J. J. & Esser, K. A. (2007). MicroRNA-1 and microRNA-133a expression are decreased during skeletal muscle hypertrophy. *Journal of Applied Physiology, 102*(1), 306-313.

McPherron, A. C., Lawler, A. M. & Lee, S. J. (1997). Regulation of skeletal muscle mass in mice by a new TGF-beta superfamily member. *Nature, 387*, 83-90.

Meadows, J. R. S., Cemal, I., Karaca, O., Gootwine, E. & Kijas, J. W. (2007). Five ovine mitochondrial lineages identified from sheep breeds of the near East. *Genetics, 175*(3), 1371-1379.

Meuwissen, T. (2009). Genetic management of small populations: A review. *Acta Agriculturae Scand Section A, 59*(2), 71-79.

Montaldo, H. H. & Meza-Herrera, C. A. (1998). Use of molecular markers and major genes in the genetic improvement of livestock. *Molecular Biology and Genetics, 1*(2).

Mortimer, S. I., Van der Werf, J. H. J., Jacob, R. H., Pethick, D. W., Pearce, K. L., Warner, R. D., et al. (2010). Preliminary estimates of genetic parameters for carcass and meat quality traits in Australian sheep. *Animal Production Science, 50*(12), 1135-1144.

National Center for Biotechnology Information (n.d.). ESTs: Gene discovery made easier. Retrieved February 26, 2013, from NSBI: http://www.ncbi.nlm.nih.gov/About/primer/est.html

Pausch, H., Flisikowski, K., Jung, S., Emmerling, R., Edel, C., Götz, K.-U., et al. (2011). Genome-wide association study identifies two major loci affecting calving ease and growth-related traits in cattle. *Genetics, 187*(1), 289-297.

Pedrosa, S., Uzun, M., Arranz, J.-J., Gutiérrez-Gil, B., San Primitivo, F. & Bayón, Y. (2005). Evidence of three maternal lineages in near eastern sheep supporting multiple domestication events. *Proceedings of the Royal Society B: Biological Sciences, 272*(1577), 2211-2217.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics, 19*(1):149-150.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*(3), 559-575.

Raj, A., Stephens, M. & Pritchard, J. K. (2014). Variational Inference of Population Structure in Large SNP Datasets. *Genetics*, genetics. 114.164350.

Reed, D. H. & Frankham, R. (2003). Correlation between fitness and genetic diversity. *Conservation Biology, 17*(1), 230-237.

Rehfelt, C., Fiedler, I. & Stickland, N. C. (2004). Number and size of muscle fibres in relation to meat prudoction. In M. F. W. t. Pas, M. E. Everts & H. P. Haagsman (Eds.), *Muscle development of livestock animals - physiology, genetics and meat quality* (pp. 1-30). Cambridge: CABI Publishing.

Ren, H., Li, L., Su, H., Xu, L., Wei, C., Zhang, L., et al. (2011). Histological and transcriptome-wide level characteristics of fetal myofiber hyperplasia during the second half of gestation in Texel and Ujumqin sheep. *BMC Genomics, 12*(1), 411.

Sachs, M., Brohmann, H., Zechner, D., Müller, T., Hülsken, J., Walther, I., et al. (2000). Essential role of Gab1 for signaling by the c-Met receptor in vivo. *The Journal of Cell Biology, 150*(6), 1375-1384.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T. & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science, 239*(4839), 487-491.

Schulman, N. F., Sahana, G., Iso-Touru, T., McKay, S. D., Schnabel, R. D., Lund, M. S., et al. (2011). Mapping of fertility traits in Finnish Ayrshire by genome wide association analysis. *Animal Genetics, 42*(3), 263-269.

Shang, N., Samuel, M. C., Zhao, H. & Chan, W. Y. (2011). Dab2 in early skeletal muscle development. *The FASEB Journal, 25*, 874.872.

Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics, 5*(5).

Talle, S. B., Chenyabuga, W. S., Fimland, E., Syrstad, O., Meuwissen, T. & Klungland, H. (2005). Use of DNA technologies for the conservation of animal genetic resources: A review. *Acta Agriculturae Scandinavica, Section A-Animal Science, 55*(1), 1-8.

Tanksley, S. D. (1983). Molecular markers in plant breeding. *Plant Molecular Biology Reporter, 1*(1), 3-8.

Tapio, M., Ozerov, M., Tapio, I., Toro, M. A., Marzanov, N., Cinkulov, M., et al. (2010). Microsatellite-based genetic diversity and population structure of domestic sheep in northern Eurasia. *BMC Genetics, 11*(1), 76.

Tapio, M., Tapio, I., Grislis, Z., Holm, L. E., Jeppsson, S., Kantanen, J., et al. (2005). Native breeds demonstrate high contributions to the molecular variation in northern European sheep. *Molecular Ecology, 14*(13), 3951-3963.

Tellam, R. L., Cockett, N. E., Vuocolo, T. & Bidwell, C. A. (2012). Genes contributing to genetic variation of muscling in sheep. *Frontiers in Genetics, 3*.

The International Sheep Genomics, C., Archibald, A. L., Cockett, N. E., Dalrymple, B. P., Faraut, T., Kijas, J. W., et al. (2010). The sheep genome reference sequence: a work in progress. *Animal Genetics, 41*(5), 449-453.

Thorgeirsdottir, S., Sigurdarson, S., Thorisson, H. M., Georgsson, G. & Palsdottir, A. (1999). PrP gene polymorphism and natural scrapie in Icelandic sheep. *Journal of General Virology, 80*, 2527-2534.

Tolone, M., Mastrangelo, S., Rosa, A. J. M. & Portolano, B. (2012). Genetic diversity and population structure of Sicilian sheep breeds using microsatellite markers. *Small Ruminant Research, 102*(1), 18-25.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., et al. (2010). Towards a knowledge-based human protein atlas. *Nature Biotechnology, 28*(12), 1248-1250.

Valsdóttir, O. S., Jónmundsson, J. V. & Eyþórsdóttir, E. (2012). Blöndun á hyrndu og kollóttu fé : könnun á blendingsþrótti (e. Crossbreeding of horned and polled sheep : study of hybrid vigor). *Rit Lbhí, 42, 18p.*

Vargas, M. A. X., Tirnauer, J. S., Glidden, N., Kapiloff, M. S. & Dodge-Kafka, K. L. (2012). Myocyte enhancer factor 2 (MEF2) tethering to muscle selective A-kinase anchoring protein (mAKAP) is necessary for myogenic differentiation. *Cellular Signalling, 24*(8), 1496-1503.

Vasyutina, E., Stebler, J. r., Brand-Saberi, B., Schulz, S., Raz, E. & Birchmeier, C. (2005). CXCR4 and Gab1 cooperate to control the development of migrating muscle progenitor cells. *Genes & Development, 19*(18), 2187-2198.

Vignal, A., Milan, D., SanCristobal, M. & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution, 34*(3), 275-306.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. v. d., Hornes, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research, 23*(21), 4407-4414.

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science, 326*(5954), 865-867.

Walling, G. A., Visscher, P. M., Wilson, A. D., McTeir, B. L., Simm, G. & Bishop, S. C. (2004). Mapping of quantitative trait loci for growth and carcass traits in commercial sheep populations. *Journal of Animal Science, 82*(8), 2234-2245.

Wang, C., Zöllner, S. & Rosenberg, N. A. (2012). A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS Genetics, 8*(8).

Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics, 6*(2), 109-118.

White, J. D., Vuocolo, T., McDonagh, M., Grounds, M. D., Harper, G. S., Cockett, N. E., et al. (2008). Analysis of the callipyge phenotype through skeletal muscle development; association of Dlk1 with muscle precursor cells. *Differentiation, 76*(3), 283-298.

White, S. N., Mousel, M. R., Herrmann-Hoesing, L. M., Reynolds, J. O., Leymaster, K. A., Neibergs, H. L., et al. (2012). Genome-wide association identifies multiple genomic regions associated with susceptibility to and control of ovine lentivirus. *Plos One, 7*(10).

Wu, H.-P., Hsu, S.-Y., Wu, W.-A., Hu, J.-W. & Ouyang, P. (2014). Transgenic mice expressing mutant Pinin exhibit muscular dystrophy, nebulin deficiency and elevated expression of slow-type muscle fiber genes. *Biochemical and Biophysical Research Communications, 443*(1), 313-320.

Zhang, L., Liu, J., Zhao, F., Ren, H., Xu, L., Lu, J., et al. (2013). Genome-Wide Association Studies for Growth and Meat Production Traits in Sheep. *Plos One, 8*(6).

Zimin, A., Delcher, A., Florea, L., Kelley, D., Schatz, M., Puiu, D., et al. (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology, 10*(4), R42.

Þorgeirsson, S. (1988). Skýrsla um fitu í dilkakjöti (e. A report on fat in lamb meat). *Árbók landbúnaðarins, 1988*, 214-249.

Þorgeirsson, S. & Þorsteinsson, S. S. (1991). Sauðfjárkynbætur, markmið, viðhorf og leiðir (e. Sheepbreeding, goals, aspects and methods). *Ráðunautafundur, 1991, 200-226.* - In Icelandic.

Appendix 1

List of samples used in the analysis showing the year the sheep are born, their sex; male (1) or female (2). The status column explains if the samples were selected to represent muscular ('high muscle') or non-muscular ('low muscle') sheep. Origin explains where the sheep come from, the samples were collected on place of origin except the artificial insemination rams (labeled Saed); blood samples from them were collected at the insemination stations. BLUP is the BLUP score for muscle (gerð in Icelandic). Color explains if the sheep is white (1) or colored (2).

Sample	Born	Sex	Status	Status	Origin	Label	BLUP	Horn	Color
Oa004	2006	1	high muscle	1	Hestur	Hestur	103.5	Horned	1
Oa008	2006	2	high muscle	1	Hestur	Hestur	104.5	Horned	1
Oa011	2006	1	high muscle	1	Hestur	Hestur	104.5	Horned	1
Oa014	2006	2	high muscle	1	Hestur	Hestur	111.5	Horned	1
Oa019	2004	1	high muscle	1	Hestur	Hestur	107	Horned	1
Oa021	2006	1	high muscle	1	Hestur	Hestur	121	Horned	1
Oa024	2001	2	high muscle	1	Hestur	Hestur	104	Horned	1
Oa025	2002	2	high muscle	1	Hestur	Hestur	102	Horned	1
Oa028	2002	2	high muscle	1	Hestur	Hestur	105	Horned	1
Oa029	2002	2	high muscle	1	Hestur	Hestur	116	Horned	1
Oa031	2002	2	high muscle	1	Hestur	Hestur	101	Horned	1
Oa033	2003	2	high muscle	1	Hestur	Hestur	90	Horned	1
Oa069	U	2	low muscle	2	Stafhv 2012	Stafh	97	Horned	2
Oa070	U	2	low muscle	2	Stafhv 2012	Stafh	93	Horned	2
Oa071	U	2	low muscle	2	Stafhv 2012	Stafh	96	Horned	2
Oa073	U	2	low muscle	2	Stafhv 2012	Stafh	93	Horned	2
Oa075	U	2	low muscle	2	Stafhv 2012	Stafh	96	Horned	2
Oa077	U	2	low muscle	2	Stafhv 2012	Stafh	96	Horned	2
Oa078	U	2	low muscle	2	Stafhv 2012	Stafh	90	Horned	2
Oa080	U	2	low muscle	2	Stafhv 2012	Stafh	96	Horned	2
Oa083	U	2	low muscle	2	Stafhv 2012	Stafh	88	Polled	2
Oa087	U	2	low muscle	2	Stafhv 2012	Stafh	96	Horned	1
Oa089	U	2	low muscle	2	Stafhv 2012	Stafh	87	Polled	2
Oa096	U	2	low muscle	2	Stafhv 2012	Stafh	89	Horned	1
Oa117	2009	1	high muscle	1	Árbær, Reykhólahr.	Saed	116	Polled	1
Oa118	2008	1	high muscle	1	Melar, Árneshr	Saed	119	Polled	1
Oa120	2009	1	high muscle	1	Sauðadalsá, Vatnsn.	Saed	124	Polled	1
Oa121	2009	1	high muscle	1	Heydalsá, Ragnar	Saed	102	Polled	1
Oa123	2007	1	high muscle	1	Bær, Árneshr.	Saed	138	Polled	1
Oa136	2009	1	high muscle	1	Kirkjuból, Dýraf.	Saed	122	Horned	1
Oa137	2007	1	high muscle	1	Hagaland, Þistilf.	Saed	121	Horned	1
Oa142	2010	1	high muscle	1	Melar, Árneshr	Saed	134	Polled	1

Oa143	2010	1	high muscle	1	Bergsst., Vatnsn.	Saed	123	Horned	2
Oa149	2008	1	high muscle	1	Hestur	Saed	139	Horned	1
Oa153	2008	1	high muscle	1	Fremri Hlíð, Vopn.	Saed	132	Horned	1
Oa158	2009	1	high muscle	1	Skriða, Hörgárd.	Saed	130	Horned	1
Oa163	2008	2	high muscle	1	Hestur	Hestur	124	Horned	1
Oa165	2007	2	high muscle	1	Hestur	Hestur	128	Horned	1
Oa166	2008	2	high muscle	1	Hestur	Hestur	116	Horned	1
Oa171	2010	2	high muscle	1	Hestur	Hestur	130	Horned	1
Oa173	2010	2	high muscle	1	Hestur	Hestur	129	Horned	1
Oa174	2010	2	high muscle	1	Hestur	Hestur	120	Horned	1
Oa175	2006	2	high muscle	1	Hestur	Hestur	121	Horned	1
Oa181	2009	2	high muscle	1	Hestur	Hestur	120	Horned	1
Oa183	2009	2	high muscle	1	Hestur	Hestur	118	Horned	1
Oa186	2008	2	high muscle	1	Hestur	Hestur	120	Horned	1
Oa189	2011	2	high muscle	1	Hestur	Hestur	126	Horned	1
Oa190	2011	2	high muscle	1	Hestur	Hestur	115	Horned	1
Oa192	2011	2	high muscle	1	Hestur	Hestur	128	Horned	1
Oa193	2011	2	high muscle	1	Hestur	Hestur	129	Horned	1
Oa197	2009	2	high muscle	1	Hestur	Hestur	117	Horned	1
Oa199	2009	2	high muscle	1	Hestur	Hestur	131	Horned	1
Oa200	2007	2	high muscle	1	Hestur	Hestur	115	Horned	1
Oa201	2009	2	high muscle	1	Hestur	Hestur	130	Horned	1
Oa202	2010	2	high muscle	1	Hestur	Hestur	116	Horned	2
Oa204	2007	2	high muscle	1	Hestur	Hestur	114	Horned	2
Oa210	2007	2	high muscle	1	Hestur	Hestur	117	Horned	2
Oa211	2010	2	low muscle	2	Strandhöfn	Leader	90	Polled	2
Oa212	2010	2	low muscle	2	Strandhöfn	Leader	79	Horned	2
Oa213	2006	2	low muscle	2	Tungusel	Leader	60	Horned	2
Oa214	2010	2	low muscle	2	Tungusel	Leader	62	Horned	2
Oa215	2003	2	low muscle	2	Tungusel	Leader	51	Horned	2
Oa216	2012	2	low muscle	2	Tungusel	Leader	60	Horned	2
Oa218	U	2	low muscle	2	Gunnarsstaðir	Leader	65	Horned	2
Oa220	2006	2	low muscle	2	Gunnarsstaðir	Leader	65	Horned	2
Oa221	2011	2	low muscle	2	Gunnarsstaðir	Leader	78	Horned	2
Oa223	2006	2	low muscle	2	Gunnarsstaðir	Leader	77	Horned	2
Oa224	2006	2	low muscle	2	Gunnarsstaðir	Leader	52	Horned	2
Oa226	2012	2	low muscle	2	Holt	Leader	63	Horned	2
Oa227	2010	2	low muscle	2	Ytra-Áland	Leader	66	Horned	2
Oa228	2009	2	low muscle	2	Ytra-Áland	Leader	66	Horned	2
Oa231	2009	2	low muscle	2	Presthólar	Leader	65	Horned	2
Oa232	2009	2	low muscle	2	Presthólar	Leader	65	U	2
Oa233	2008	2	low muscle	2	Sandfellshagi 1	Leader	53	Horned	2
Oa234	2008	2	low muscle	2	Sandfellshagi 1	Leader	56	Horned	2
Oa235	2005	2	low muscle	2	Sandfellshagi 1	Leader	66	Horned	2
Oa237	2008	2	low muscle	2	Vestara-Land	Leader	53	Horned	2
Oa240	2010	1	low muscle	2	Vestara-Land	Leader	57	Horned	2

Oa241	2004	2	low muscle	2	Presthvammur	Leader	81	Horned	2
Oa243	2008	2	low muscle	2	Presthvammur	Leader	64	Horned	2
Oa244	2009	2	low muscle	2	Presthvammur	Leader	65	Horned	2
Oa246	2010	2	low muscle	2	Rauðbarðaholt	Leader	65	Horned	2
Oa247	2012	2	low muscle	2	Gróustaðir	Leader	65	Horned	2
Oa250	2007	2	low muscle	2	Gróustaðir	Leader	65	Horned	2
Oa251	2001	2	low muscle	2	Gróustaðir	Leader	65	Horned	2
Oa254	2009	2	high muscle	1	Smáhamrar	Polled	122	Polled	1
Oa259	2009	2	high muscle	1	Smáhamrar	Polled	132	Polled	1
Oa260	2010	2	high muscle	1	Smáhamrar	Polled	118	Polled	1
Oa264	2008	2	high muscle	1	Heydalsá II Guðjón	Polled	117	Polled	1
Oa268	2009	2	high muscle	1	Heydalsá II Guðjón	Polled	117	Polled	1
Oa270	2007	2	high muscle	1	Heydalsá II Guðjón	Polled	124	Polled	1
Oa276	2007	2	high muscle	1	Heydalsá I Ragnar	Polled	120	Polled	1
Oa281	2006	2	high muscle	1	Miðdalsgröf	Polled	119	Polled	1
Oa290	2010	2	high muscle	1	Miðdalsgröf	Polled	128	Polled	1
Oa292	2005	2	high muscle	1	Tröllatunga	Polled	120	Polled	1
Oa295	2008	2	high muscle	1	Tröllatunga	Polled	115	Polled	1

Appendix 2

GenABEL commands for genome-wide association analysis.

Continuous trait – BLUP score:

```
egscore(blup~sex, oa_data, kinship=oa_datagkin)
h2a5 <- polygenic_hglm(blup~sex, oa_data4, kin = oa_datagkin, trait="gaussian")
an.mm5<- mmscore(h2a5, oa_data4
pcs2 <- cmdscale(oa_datadist, k=10)
qtscore(blup~pcs2[,1]+pcs2[,2]+pcs2[,3]+pcs2[,4]+sex, oa_data4)
mmscore(h2a5, oa_data, strata=pop88)
```

Case-control, binary trait:

```
qtscore(cc ~ sex, oa_data, trait="binomial")
qtscore(cc ~ blup+sex, oa_data, strata=pop, trait="binomial")
egscore(cc ~ blup, oa_data, kinship=oa_datagkin3)
h2a3 <- polygenic_hglm(ph.x ~ blup+sex, oa_data,kin=oa_datagkin, trait="binomial")
an.mm3 <- mmscore(h2a3, oa_data)
mmscore(h2a4, oa_data)
mmscore(h2a4, oa_data, strata=pop)
```

Appendix 3

GWAS results from all four models. A1 and A2 represent the alleles, A2 being the causative allele. N is the number of samples of which the SNP was detected in, effB is the effect of allele A2 on the phenotype and Pc1df is the corrected p-value of the association.

Table A Case/control mixed model top 25 SNPs.

SNP	Chromosome	A1	A2	MAF	effB	Pc1df
OAR3_OAR17_14020647	17	G	A	0.114	-0.255	5.26e-07
S26572.1	22	A	G	0.0638	-0.309	1.36e-06
OAR3_OAR22_46042704	22	A	G	0.122	-0.210	2.93e-06
OAR3_OAR7_99717099	7	G	A	0.0479	-0.312	3.36e-06
OAR7_28667084.1	7	G	A	0.122	-0.219	5.76e-06
OAR3_OAR19_39504502	19	G	A	0.0372	-0.400	6.11e-06
OAR3_OAR15_14537738	15	A	G	0.0851	-0.287	7.08e-06
OAR3_OAR6_111289128	6	G	A	0.0798	-0.258	8.19e-06
OAR3_OAR18_41905736	18	A	G	0.0585	-0.303	8.92e-06
OAR3_OAR6_16095581	6	A	G	0.138	-0.192	9.50e-06
OAR3_OAR8_16566300	8	A	G	0.0426	-0.394	1.05e-05
OAR3_OAR14_16175371	14	G	A	0.0479	-0.311	1.11e-05
OAR3_OAR14_16184329	14	G	A	0.0479	-0.311	1.11e-05
OAR3_OAR4_43654975	4	A	G	0.0591	-0.341	1.49e-05
OAR3_OAR22_46133225	22	A	C	0.128	-0.179	1.62e-05
OAR3_OAR22_46154824	22	A	G	0.128	-0.179	1.62e-05
OAR3_OAR9_10822930	9	A	C	0.154	-0.188	2.30e-05
OAR3_OAR24_28793777	24	A	G	0.000	-0.283	2.35e-05
OAR3_OAR9_11339790	9	G	A	0.213	-0.151	2.45e-05
OAR3_OAR4_37180930	4	G	A	0.0426	-0.291	2.49e-05
OAR1_161578392.1	1	G	A	0.121	-0.187	2.69e-05
OAR3_OAR3_18716983	3	G	A	0.101	-0.238	2.73e-05
OAR3_OAR3_18747575	3	A	G	0.101	-0.238	2.73e-05
OAR3_OAR3_18753039	3	C	A	0.101	-0.238	2.73e-05
OAR3_OAR3_18756234	3	A	G	0.101	-0.238	2.73e-05

Table B Case/control mixed model with stratification accounted for top 25 SNPs.

SNP	Chromosome	A1	A2	MAF	effB	Pc1df
OAR3_OAR7_99717099	7	G	A	0.0479	-0.301	5.77e-06
OAR3_OAR1_11045392	1	A	G	0.0479	0.375	1.40e-05
OAR3_OAR1_11068514	1	G	A	0.0479	0.375	1.40e-05
S26572.1	22	A	G	0.0638	-0.278	1.49e-05
OAR3_OAR14_14883927	14	A	C	0.245	-0.140	2.77e-05
OAR3_OAR22_46042704	22	A	G	0.122	-0.190	2.80e-05
OAR7_28667084.1	7	G	A	0.122	-0.201	2.82e-05
OAR3_OAR4_37180930	4	G	A	0.0426	-0.284	2.97e-05
OAR3_OAR19_39504502	19	G	A	0.0372	-0.363	3.87e-05

OAR3_OAR6_111289128	6	G	A	0.0798	-0.237	3.93e-05
OAR3_OAR18_41905736	18	A	G	0.0585	-0.278	4.31e-05
OAR3_OAR4_43654975	4	A	G	0.0591	-0.318	4.62e-05
OAR3_OAR8_16566300	8	A	G	0.0426	-0.359	5.55e-05
OAR3_OAR3_18780585	3	G	A	0.0904	-0.227	7.87e-05
S62291.1	3	G	A	0.0904	-0.227	7.87e-05
OAR3_OAR3_18782837	3	A	G	0.0904	-0.227	7.87e-05
OAR3_OAR17_14020647	17	G	A	0.114	-0.210	8.01e-05
OAR3_OAR16_21809475	16	A	G	0.207	-0.143	8.72e-05
OAR3_OAR14_16175371	14	G	A	0.0479	-0.278	8.74e-05
OAR3_OAR14_16184329	14	G	A	0.0479	-0.278	8.74e-05
OAR3_OAR3_18716983	3	G	A	0.101	-0.221	8.75e-05
OAR3_OAR3_18747575	3	A	G	0.101	-0.221	8.75e-05
OAR3_OAR3_18753039	3	C	A	0.101	-0.221	8.75e-05
OAR3_OAR3_18756234	3	A	G	0.101	-0.221	8.75e-05
OAR3_OAR13_47270753	13	A	C	0.495	-0.111	9.13e-05

Table C Continuous trait – BLUP PCA model, top 25 SNPs.

SNP	**Chromosome**	**A1**	**A2**	**MAF**	**effB**	**Pc1df**
OAR3_OAR18_47844036	18	A	G	0.383	-5.84	7.30e-06
OAR3_OAR8_44085170	8	G	A	0.479	5.22	2.64e-05
OAR3_OAR16_49923680	16	G	A	0.367	-5.27	3.71e-05
S54536.1	20	G	A	0.303	-5.68	3.92e-05
OAR3_OAR6_31353857	6	A	G	0.468	5.21	4.09e-05
OAR3_OAR16_22562762	16	G	A	0.266	5.38	4.51e-05
OAR3_OAR16_34117164	16	G	A	0.457	-4.94	5.43e-05
S01424.1	2	A	G	0.335	-5.26	6.11e-05
OAR16_37082988.1	16	A	G	0.495	-4.77	6.86e-05
OAR3_OAR16_34144581	16	A	C	0.495	-4.77	6.86e-05
OAR3_OAR16_41683117	16	C	A	0.218	-5.41	7.26e-05
OAR3_OAR13_3436475	13	C	A	0.372	-4.87	8.68e-05
OAR3_OAR13_3438159	13	G	A	0.372	-4.87	8.68e-05
OAR3_OAR16_38201904	16	A	C	0.362	4.48	9.84e-05
OAR3_OAR16_38212677	16	G	A	0.362	4.48	9.84e-05
OAR3_OAR3_37847391	3	G	A	0.367	-5.04	9.87e-05
OAR3_OAR16_31124520	16	G	A	0.314	-5.08	1.01e-04
OAR3_OAR1_4968557	1	A	G	0.229	-5.40	1.10e-04
OAR3_OAR1_39946023	1	G	A	0.245	5.39	1.12e-04
OAR1_41306759.1	1	A	G	0.245	5.39	1.12e-04
OAR3_OAR16_48627603	16	G	A	0.165	-6.48	1.14e-04
OAR3_OAR1_39987538	1	G	A	0.250	5.38	1.16e-04
OAR3_OAR1_39989157	1	G	A	0.250	5.38	1.16e-04
OAR3_OAR3_101522485	3	A	G	0.335	-5.24	1.16e-04
OAR3_OAR16_48626512	16	A	G	0.161	-6.51	1.32e-04

Table D Continuous trait – BLUP mixed model, top 25 SNPs.

SNP	Chromosome	A1	A2	MAF	effB	Pc1df
OAR3_OAR18_36925604	18	A	G	0.115	-19.0	1.65e-05
OAR3_OAR18_36888692	18	G	A	0.112	-18.7	2.64e-05
OAR3_OAR18_26609576	18	A	G	0.223	-12.6	4.44e-05
OAR3_OAR18_26644248	18	A	C	0.223	-12.6	4.44e-05
OAR3_OAR18_26646267	18	A	C	0.223	-12.6	4.44e-05
OAR3_OAR18_26649516	18	A	C	0.223	-12.6	4.44e-05
OAR3_OAR18_26649945	18	G	A	0.223	-12.6	4.44e-05
OAR3_OAR10_86285876	10	A	G	0.156	-13.8	4.45e-05
OAR3_OAR10_86315861	10	A	G	0.158	-13.8	4.69e-05
OAR3_OAR4_7542755	4	C	A	0.495	-10.2	4.82e-05
OAR3_OAR24_37189272	24	A	G	0.00	-11.9	4.835e-05
OAR3_OAR14_7607475	14	C	A	0.393	-10.2	5.69e-05
OAR3_OAR15_49097869	15	G	A	0.176	-16.7	5.90e-05
OAR3_OAR5_18640578	5	G	A	0.102	-16.0	6.00e-05
OAR3_OAR23_61198865	23	G	A	0.143	-14.5	6.34e-05
OAR3_OAR14_10591551	14	A	G	0.138	-16.2	7.43e-05
OAR3_OAR10_86411658	10	G	A	0.147	-14.2	7.58e-05
OAR9_97825328.1	9	A	G	0.388	-10.5	7.85e-05
OAR3_OAR13_21064705	13	G	A	0.271	-11.1	7.90e-05
OAR3_OAR5_10964015	5	C	A	0.314	10.0	8.54e-05
S36767.1	9	A	G	0.269	-10.7	1.00e-04
OAR3_OARX_18785482	X	A	G	0.269	7.98	1.16e-04
OAR10_93987978.1	10	G	A	0.129	-14.7	1.24e-04
OAR3_OARX_18777960	X	A	G	0.452	7.56	1.37e-04
OAR3_OARX_18784666	X	A	G	0.452	7.56	1.37e-04

Appendix 4

A deletion in an intron at base-pair 1,243 was detected in 9 samples compared to the reference gene *GADD45B.*