A.D. MDLXII

UNIVERSITÀ DEGLI STUDI DI SASSARI

CORSO DI DOTTORATO DI RICERCA

Scienze Agrarie

Curriculum Scienze e Tecnologie Zootecniche

Ciclo XXXI

APPLICATION OF GENOMIC TOOLS TO BREEDING

AND TO GENETIC STRUCTURE STUDIES

IN LIVESTOCK POPULATIONS

dr. Alberto Cesarani

Direttore della Scuola	prof. Ignazio Floris
Referente di Indirizzo	prof. Corrado Dimauro
Docente Guida	prof. Nicolò Pietro Paolo Macciotta
Tutor	dr. Giustino Gaspa

Anno Accademico 2017-2018

Acknowledgements

I want to acknowledge Prof. Nicolò Pietro Paolo Macciotta for his constant openness and for the several great opportunities that he offered me during my PhD course.

I want to express my gratitude to Dr. Giustino Gaspa and Dott.ssa Silvia Sorbolini for helping me during my work.

Thanks to genetic research group and all my colleagues of the "*Scienze Zootecniche*" unit of the department of Agriculture of the University of Sassari.

I would also thank Prof. Ignacy Misztal and Prof. Daniela Lourenco for the kind hospitality and helpfulness to teach me their work, as well as all people I met at the Animal Breeding and Genetics Group of the University of Georgia.

Index

CHAPTER 1 .... **1**

General introduction **1**

Introduction *2*
a. Animal breeding and advances in molecular genetics 3
b. Genomic selection methods 9

Genetic variability in animal populations *14*
a. Tool for studying of genetic diversity and application to livestock conservations 14
c. Linkage disequilibrium in animal population 19
d. Runs of homozygosity 23

Objective of the Thesis *25*

References *26*

CHAPTER 2 .... **45**

Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies **45**

Abstract *46*

Introduction *47*

Materials and methods *49*
Data 49
Models and analysis 53

Results *54*

Discussion *62*

Conclusions *66*

Acknowledgements *67*

References *68*

Appendix Chapter 2 .... *75*

CHAPTER 3 .... **82**

Genomic selection of milk fatty acid composition in Sarda dairy sheep: effect of different phenotypes and relationship matrices on heritability and breeding values accuracy **82**

Abstract *83*

Introduction *85*

Materials and methods *87*

Data 87
Variance component estimation 89
Breeding Value Predictions 91
Results *92*
Genetic Parameters of Milk Fatty Acid profile 93
Accuracy of EBV and GEBV predictions 98
Discussion *103*
Genetic Parameters of Milk Fatty Acid profile 104
Accuracy of EBV and GEBV predictions 107
Conclusions *108*
Acknowledgements *109*
References *110*
Appendix Chapter 3 .... *120*

CHAPTER 4 123

Genome-wide variability and selection signatures in Italian island cattle breeds 123

Abstract *124*
Introduction *126*
Materials and methods *129*
Animals and genotypic data 129
Heterozygosity, Minor allele frequency and Linkage Disequilibrium 130
Multi-dimensional scaling and admixture analysis 130
Wright Fixation Index and LOWESS 130
Runs of homozygosity 131
Gene discovery 132
Results *132*
Discussion *142*
Conclusion *148*
Acknowledgments *149*
References *150*
Appendix Chapter 4 .... *159*

CHAPTER 5 163

The use of different statistical approaches highlights consensus signatures of selection harboring diversity among sheep breeds 163

Abstract *164*
Introduction *166*

Materials and methods *169*
Breeds and animals 169
Genotyping 170
Population stratification and admixture 170
Fixation Index (F_{ST}) 170
Canonical Discriminant Analysis (CDA) 171
Runs of Homozygosity Analysis (ROH) 172
Genes Annotation, QTLs and Functional Analysis 172
Results *174*
Detection of significant SNPs with F_{ST}, CDA and ROH approaches 174
Detection of signatures of selection 177
Selection signatures shared among procedures 178
Candidate gene discovery, functional annotation and QTLs 182
Discussion *186*
Conclusions *192*
References *194*
Appendix Chapter 5 .... *207*
CHAPTER 6 241
General conclusion 241

CHAPTER 1

General introduction

Introduction

Modern genetics is usually considered to have started with the rediscovery of Mendel's paper in 1900. Genetic selection for enhance productivity or fitness traits gave impressive results in the last 50 year due to the implementation of the Fisher's infinitesimal model – that assume infinite number of loci with infinitesimal effects underlying the observed phenotypes – in the mixed models statistics framework.

The expected genetic progress (ΔG) of a breeding scheme is generally evaluated through the breeders' equation developed by Rendel and Robertson (1950): $\Delta G = \frac{i \cdot r}{T} \sigma_A$ that positively associates ΔG to intensity of selection (*i*), accuracy of breeding values (r) and genetic variation (σ_A) and negatively to generation interval (T). Moreover, the introduction of best unbiased linear predictor (BLUP) methodology for predicting the genetic merit of a selection candidate represented a breakthrough in the animal quantitative genetics in the late-seventies of the past century.

More recently, genomic selection – based on extensive genotypization for thousands of single nucleotide polymorphisms (SNPs) along animal genomes – has transformed breeding programs worldwide providing more accurate estimated breeding values (EBVs) of selection candidates, especially of those without own records or daughters' information. Genomic selection has had a big impact on the accuracy and generation interval terms of the breeders equation. Under some preconditions, mainly about the size of genotyped population and the SNP density, genomic selection allowed to give to an animal a more reliable EBV at birth, with a great advantage over traditional quantitative breeding, especially for those traits that are recorded late in life, sex limited or expensive to measure.

If on the one hand the extensive use of advanced selection strategies has doubled production performances of farm animals, on the other hand a general reduction of biodiversity (*e.g*. number of breeds raised) and, with some exceptions, a general increase of within breed homogeneity have been observed worldwide. Despite of that, the current and large availability of genomic tools at affordable price might allow to cope with these emerging issues. In particular, the SNP marker scan be used both to estimate the genetic diversity among breeds and the within breed degree of homogeneity. The use of genomic tools may be helpful to safeguard endangered population or to control inbreeding in small population where no pedigree records are available.

a. Animal breeding and advances in molecular genetics

The application of genetics to the evaluation of animals and to the improvement of their productions revolutionized this industry and laid the foundations for the modern animal breeding and genetics' disciplines. Later on, the animal genetic divided into two main branches: quantitative genetics and molecular genetics. The main steps of quantitative genetics were: the proposal of selection index as tool for choosing the parents of the next generation (Hazel, 1943); the application on large dataset of best linear unbiased prediction animal model (BLUP, Henderson, 1975) in order to predict animal breeding values (Gianola *et al*., 1988); the use of complex statistical models for genetic evaluation, such as *animal model*, *test day model* or *random regression model* (Kennedy *et al*., 1988; Jensen, 2001; Schaeffer, 2004). The combination of use of genetic markers, linkage analysis and measured phenotypic trait, allowed to identify quantitative trait loci (QTLs) for farm animal species (Dekkers and Hospital, 2002;

Dekkers, 2004). QTLs are DNA regions associated with quantitative traits of economic interest. After several years of study about QTL, public datasets with information about productive characters were created (http://www.animalgenome.org/cgi-bin/QTLdb/index).

However, the development of affordable *genome sequencing* techniques represented the most important insight of molecular genetics both for animal breeding and genetics field. The DNA sequencing technique dated 1975 when this concept was introduced in the scientific community (Sanger, 1975). Thirty years later, in the early 2000th, novel DNA sequencing method open new frontiers: the high-throughput next generation sequencing (*HT-NGS*) completely replaced the Sanger based sequencing methods (Pareek *et al*., 2011). The main advantages of NGS over the Sanger based method are the highly automatization and parallelization of workflow, both reflecting in a dramatic reduction of the time needed to sequence a DNA sample. The main consequence of the introduction of NGS has been the drastic reduction of cost of a complete genome sequence (Altman *et al*. 2012). Just as example, in the nineties of the past century about $3-billion project were spent to sequence the human genome producing a "rough draft" of the genome in 2001 (Weller *et al*., 2016). In the 2008-2012 period, while a project aimed at sequencing the genome of at least 1000 of human individuals was completed (1000 genome project, 2012) the first complete genome sequence of a cow was also obtained (Bovine Genome Sequencing and Analysis Consortium *et al*., 2009). Three years later (2012) the 1000 bull genomes project was launched in cattle (Daetwyler *et al*.2014). Since the discovery of microsatellites in 1990 costs for obtaining individual genotype were dramatically reduced from $10 per each marker to $0.002/marker typed (Weller *et al*. 2016).

This sequencing process allowed to find several variants in the genome that can be used as molecular markers such as mitochondrial DNA (mtDNA), microsatellites, SNPs, insertion or deletion (INDEL) (Yamey, 2000). Mitochondrial DNA are small plasmids that can be found only in mitochondrial organelle that have some inherent characteristics: it shows a strong variability within species, has an almost exclusive maternal inheritance and it is not affected by recombination. The mtDNA has been the reference molecular markers for domestication studies since their maternal inheritance only (Bruford *et al*., 2003; Naderi *et al*., 2008; Di Lorenzo *et al*., 2015). Microsatellite markers, known also as simple sequence repeats (SSRs), identified short repeated sequences (usually less than 6 base pairs ex poly(TG) sites) that can be found several times along the genome: their origin can be ascribable to the DNA polymerase sliding during DNA replication. SNPs identify a variation of a single nucleotide across the genome (Figure 1), that is a single base change in DNA (Vignal *et al*., 2002). In animal genome, SNPs markers are usually bi-allelic.

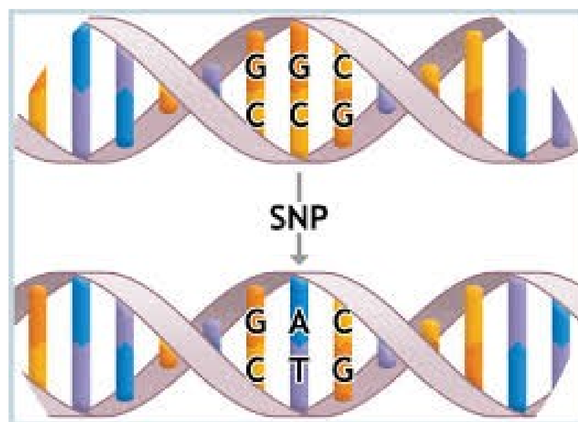


Figure 1. Example of Single Nucleotide Polymorphism (SNP) (http://www.socmucimm.org/).

In the first draft of human genome, a total of 1.4 Millions of SNPs has been discovered, with an average distribution of 1.9 marker each kb (Sachidanandam *et al.*, 2001, Abdel-Mawgood, 2012).

In animal genetics, first genome-wide linkage maps had 150-200 markers (15-20 centimorgan between each one on bovine genome) and were used to find QTL with whole genome scan, with high confidence interval though. The biggest problem of this type of map was that in the interval between two adjacent markers can harbor up to 50-60 genes. Only in 2009, the Bovine Genome Sequencing and Analysis Consortium (www.bovinegenome.org) produced the reference sequence of taurine genome. Fan *et al*. (2010) reported the number of discovered SNPs in some animal genomes. In cattle, The Bovine HapMap Consortium (2009) detected around 2.2 million SNPs (with an average distance between two SNPs of 1 kb). In 2012, the International Sheep Genomics Consortium (ISGC) started to assemble the ovine reference genome using both Sanger and NGS sequencing (www.sheephapmap.org): this project allowed to discover 2.8 million of SNPs. In horse (Wade *et al*., 2009) and chicken (International Chicken Polymorphism Map Consortium, 2004), around 1.1 million SNPs and 2.8 million SNPs were highlighted, respectively. In 2014, Daetwyler *et al*. analyzing several bull genomes of different breeds (*1000 genome Project*), sequenced using NGS platform, found about 28 millions of variants (26.7 Million SNPs and 1.6 Million INDEL).

Starting from the genome sequencing and knowledge about SNPs, genome maps with medium and high density (with an interval between two markers < 1 cM) of these polymorphisms were created. The subsequent step was the commercialization of *SNP beadchips* (Figure 2), that allows to automatically genotype tens of DNA

samples with hundreds of thousands of SNPs starting from blood, saliva or male semen (Matukumalli *et al*., 2009).

Figure 2. BovineSNP50 v3 DNA Analysis BeadChip (www.illumina.com).

In Table 1, the most common commercial SNP beadchips are reported. SNPs are very useful also because are markers that divide the animal genome and circumscribe the information of these small DNA regions. A plus point of SNPs is that they simplify genomic studies because the total amount of DNA information (billions of base pairs) is conveyed in a smaller sample of base pairs (~ 50 thousand bps). Due the high number of SNPs and their uniform distribution along the animal genome, some of them could be near a QTL and, therefore, they could explain a fraction of the trait variance allowing to map QTL more precisely.

Table 1. Commercial SNP chip for most common domesticated species (*via* http://bioinformatics.tecnoparco.org/SNPchimp/).

Species	Producer	Chip name	SNPs
Cattle	Illumina	*Bovine 3k BeadChip*	2,900
		Bovine LD BeadChip	6,909
		BovineLD v1.1	6,912
		BovineLD v.2	7,931
		BovineSNP50v1 BeadChip	54,001
		BovineSNP50v2 BeadChip	54,609
		BovineHD BeadChip	777,962
	GeneSeek	*Dairy Ultra LD v2*	7,049
		Genomic Profiler LD v1	8,610
		Genomic Profiler LD v2	19,721
		Genomic Profiler LD v3	26,151
		Genomic Profiler HD	76,879
		Genomic Profiler HD v2	139,480
	ICBF	*International Dairy and Beef v2*	17,807
		International Dairy and Beef v3	53,262
	Affymetrix	*Axiom Bovine*	648,875
Sheep	Illumina	*Infinium Ovine SNP50 v1 BeadChip*	54,241
	AgResearch	*OvineHD BeadChip*	606,006
Goat	IGGC	*CaprineSNP50 BeadChip*	53,347
Pig	Illumina	*Infinium PorcineSNP60 v1 BeadChip*	62,163
		Infinium PorcineSNP60 v2 BeadChip	61,565
	GeneSeek-Neogen	*Genomic Profiler 10k BeadChip*	10,241
		PorcineSNP80 BeadChip	68,528
	Affymetrix	*Axiom PorcineHD*	658,692
Horse	Illumina	*Infinium EquineSNP50 BeadChip*	54,602
	GeneSeek	*EquineSNP65 BeadChip*	65,157
	Affymetrix	*Axiom EquineHD*	670,796
Chicken	Affymetrix	*Axiom Chicken*	580,961

b. Genomic selection methods

One of the main applications of SNP beadchips in animal breeding has been defined "genomic selection". The use of thousands of SNP genotypes in the genetic evaluation of animals deeply changed the breeding schemes and their response. At the beginning, selection programs were based only on phenotypes recording and relationship among animals. These breeding programs took advantages from pedigree and the related pedigree relationship matrix (**A**). The **A** matrix is constructed according to the relationship coefficients among animals: e.g. sire/dam and offspring 0.5, grandparents and grandchild 0.25, half-sibs 0.25 *etc*. Pedigrees are used to identify the additive relationship coefficients that is the proportion of genes shared because inherited from a common ancestor.

The possibility of incorporate information deriving from molecular markers in the design of new breeding schemes was initially termed marker assisted selection (MAS) (Fernando and Grossman, 1989; Lande and Thompson, 1990) and was based on the combination of breeding values and DNA microsatellite information. Theoretically, MAS should have been replaced by genes assisted selection (GAS), once the causative mutations affecting economic traits had been discovered (Ron and Robertson, 2007). However, the major revolution in the animal breeding has been represented by genomic selection (GS) as formalized in influential papers by Meuwissen *et al*. (2001) and Schaeffer (2006).

The first step prior to start with a breeding program, either traditional or genomic assisted, is the calculation of variance components (Hofer, 1988; Misztal, 2008). Variance components are used to indicate variation in animal populations; they are usually identified using σ^2 and they are divided in phenotypic, genetic and

environmental. According to Falconer and Mackay (1996) the genetic variance is further scattered in additive, dominance and interaction. The additive variance is the variance portion which account for breeding values variation. Two main methods are used to calculate variance components: restricted maximum likelihood (REML) and Bayesian (Thompson *et al*., 2005; Misztal, 2008). The majority of animal traits of human interest (e.g. milk production, meat production, sport performance in horse, etc.) is heritable and it means that related animals show more similar phenotypic expressions than unrelated: this implication is used to estimate variance components and to separate the portion due to the genetic from the one due to the environment. Variance components are used to calculate the heritability (h^2) of one trait, usually defined as the ratio between the additive animal variance and the total phenotypic variance, i.e. as the fraction of phenotype determined by the individual genome (Falconer and Mackay, 1996). Another involvement that must be considered in order to engineer a breeding program is the involved traits, specifically if qualitative or quantitative. Quantitative traits (e.g. milk production, average daily gain, etc.), representing the major traits of livestock, show continuous variability, can be measured using a metric scale, and are under polygenic control. According to the infinitesimal model, all genes affecting a quantitative trait have an infinitesimal effect. Most of the quantitative traits present a frequency distribution similar to the normal distribution. On the other hand, qualitative traits (e.g. coat color, disease resistance, etc.) show discontinuous variability and they are controlled only be one or few genes.

After the first QTL detection experiments, MAS selection took hold (Boichard *et al*., 2002; 2006; Bennewitz *et al*., 2004) in French and German cattle breeds. MAS was mainly applied in dairy cattle industry because the trait of interest was sex-limited

(milk could be registered only in females) and the generation interval was long. MAS was implemented with the goal to increase the reliability of breeding values and select for new functional traits (for example fertility). A simulation on the application of MAS program demonstrated that the increase of the breeding value reliability was rather limited (larger for sons of sires with genotyped progeny daughters with records), nonetheless the high number of genotyped animals (Guillaume *et al*., 2008)

As previously told, GS is the latest proposed methodology to evaluate and to select candidate animals. GS can be performed using different model and methods: SNP-BLUP, genomic BLUP (GBLUP), Bayesian (BayesA, B, C, Cπ-R *etc.*) or multivariate techniques (VanRaden *et al*., 2008; Gianola *et al*., 2009; Moser *et al*., 2009; Macciotta *et al*., 2010). Regardless of the statistical method, GS considerably reduced generation interval because it allowed to estimate breeding value at candidate birth (VanRaden *et al*., 2009).

Most of the GS application involved so called two-step GBLUP models. In GBLUP **A** matrix has been replaced by the genomic relationship matrix (**GRM** or **G**), that use genetic information to establish parental relationship among animals (Goddard and Hayes, 2007). The cornerstone of GS is the genomic breeding values (**GEBV**) that combine, with different weight, direct genomic value (**DGV**) estimated from individual SNP effects and traditional **EBV** (VanRanden *et al*., 2008).

Recently, a new methodology for GS has been developed: the single step genomic evaluation that combine pedigree and genomic information at the same time. Single-step GBLUP (ssGBLUP) exploits contemporarily phenotypic, pedigree and genomic information, whereas the previous methods were planned in subsequent different steps (Aguilar *et al*., 2010). In this method, the two relationship matrices

integrated in traditional or GBLUP animal model (i.e. **A** and **G**) are blended in one unique matrix identified with the **H** letter. The main advantages of ssGBLUP are the simplicity of implementation, without extensive change in the mixed model equation settings.

GS is strongly applied in dairy cattle breeding schemes (Hayes *et al*., 2009; VanRaden *et al*., 2009; Pryce and Daetwyler, 2012) and pigs (Ostersen *et al*., 2011; Wellman *et al*., 2013), but with less emphasis in other farm animal species, even if GS in sheep is gaining popularity (Van der Werf, 2009; Daetwyler *et al*., 2010; Duchemin *et al*., 2012).

Objectives of selection programs are highly varied over the years: next to the traditional improvement of the yield (milk, meat, eggs…) composition (e.g. fat and protein content in the milk), conformation or the production and disease resistance (e.g. mastitis or scrapie), novel traits (e.g. milk coagulation properties, milk and meat fatty acids content, residual feed intake, methane emissions, *etc*.) are coming onstage or could be integrated in the future breeding schemes.

The huge increase of livestock production was mainly due to better farm management, functional nutrition system and genetic improvement: thanks to increasingly precise breeding programs, livestock production drastically increased in the last 60 years (World Livestock 2011; Alexandratos and Bruinsma, 2012).

Regarding disease resistance, an example can be the *scrapie* in the sheep. It is a disease already known about 250 years ago; Dickinson (1968), studying Britannic sheep breeds, showed the individual variability for the disease sensibility. This disease was spread in France, United Kingdom, Italy (Macciotta *et al*., 2005). The prion protein gene was identified as the responsible. The prion protein is a membrane protein

with a not well-known role in the organism. The altered structure is accumulated in the nervous tissue of the affected animals. Mutations of the prion protein have been associated with different resistance levels: the ARR allele was the most resistance (with dominant effect), while the VRQ allele was the most sensitive. The scrapie disease was eradicated by choosing only rams with ARR allele. Several studies analyzed this phenomenon (Barillet *et al*., 2002; Hurtado *et al*., 2002; Macciotta *et al*., 2005; Ligios *et al*., 2006; Palhiere *et al*., 2008).

Moreover, with the increased availability of records about the abovementioned new phenotypes, several researches are being conducting in order to establish the possibility to implement this kind of data in the breeding schemes. The genetic background of fatty acid profile has been investigated both for milk (Soyeurt *et al.*, 2007; Mele et al., 2009, Hein et al., 2018) and meat (Cecchinato *et al*., 2012; Chiaia *et al*., 2017) in cattle. Recently, some studies, even if without a homogeneity of goals and techniques, have been carried out also in dairy (Moioli *et al*., 2012) and meat (Rovadoscki *et al*., 2018) sheep. Few researches on beef cattle were directly oriented to genomic selection for FA composition (Uemoto *et al*., 2011; Chen *et al*., 2015; Zhu *et al*., 2017).

Still, to restrict the farm costs and pollution, future breeding programs will be aimed at improving residual feed intake (RFI) and methane emissions. Luckily, these two traits are strongly related: studies about possible selection schemes aimed to reduce the RFI demonstrated that could also bring to a drop of methane emitted by animals (Alford *et al*., 2006). Finally, some researchers analyzed the role of enteric methane emission in the farm management and breeding decisions (Negussie *et al*., 2017).

Genetic variability in animal populations

a. Tool for studying of genetic diversity and application to livestock conservations

Although the use of SNP marker has been applied for improving livestock performance through genomic selection (Schaeffer, 2006; Hayes *et al.*, 2009; VanRaden *et al.*, 2009; 2011) in recent times genomics is widely applied also on genetics and biodiversity (Peter *et al.*, 2007; Bovine HapMap Consortium, 2009; Kijas et al., 2009; 2012; Ciani *et al.*, 2014; Sharma *et al.*, 2016; Upadhyay *et al.*, 2017; Colli *et al.*, 2018; Mastrangelo *et al.*, 2018; Stella *et al.*, 2018), conservation genomics (Allendorf *et al.*, 2010), linkage disequilibrium estimation (Grossi *et al.*, 2017; Prieur *et al.*, 2017; Snelling *et al.*, 2017), runs of homozygosity analysis (Ferenčaković *et al.*, 2011; Marras *et al.*, 2015), genome-wide association studies GWAS (Hayes et al., 2010; Mai *et al.*, 2010; Peñagaricano *et al.*, 2012; Zhang *et al.*, 2013; Macciotta *et al.*, 2017; Martinez-Royo *et al.*, 2017), relationship between environment and animal genome (Manel *et al.*, 2003; Finocchiaro *et al.*, 2005) and the occurrence and spread of animal diseases (Becker *et al.*, 2010; Testoni *et al.*, 2012; Zhao *et al.*, 2012).

The study of genetic biodiversity of farm animals and its conservation and safeguard is becoming increasingly relevant. The main issue is that a big portion of this genetic diversity has been unknowingly already lost: lots of plant varieties or animal breeds are disappearing and without precise figures about the exact amount of this loss. Moving in this scenario, Scherf (2000) published for the Food and Agriculture Organization of the United Nations (FAO) his “World watch list for domestic animal diversity”. The aim of this book was to monitor the animal genetic biodiversity about mammalian and avian species, trying to classify animal breeds in

extinct, at risk and safe breeds. Moreover, in 2007 FAO organization published the World's Animal Genetic Resources that debated about the animal history and domestication, the actual status of animal genetic resources, and the future challenges and management of genetic resources (www.fao.org).

The domestication of animals doubtless symbolized one of the most important events for humanity: this occurrence sealed the passage from hunting to cultivators and breeder mankind and therefore was the first step of human sedentary settling (Larson and Fuller, 2014). Through domestication, mankind caused significant morphological and physiological changes in animal and plants: the new characteristics of domesticated individuals discriminated between tamed animals and their ancestors (Ryder, 1983). After dog, sheep was the first domesticated species occurred in the Fertile Crescent at ca. 10,000-9,500 (Clutton-Brock, 1987; Zeder, 2008) and some studies allowed to trace this process (Bruford *et al*., 2003; Chessa *et al*., 2009). Sheep was earlier chosen for domestication thanks to its several advantageous characteristics, i.e. small size and versatile productions (meat, milk and wool). Ryder (1983) highlighted that domestication of sheep lead to a reduction of body size and of the horn length if compared with its wild ancestor. Simultaneously, also the domestication of cattle and goat, occurred nearly at the same time in the Fertile Crescent while pigs and buffalo domestication somewhere in the southwest Asia; on the contrary, horse seems to be the last major species that has been domesticated in several independent occasions and places (Bruford *et al*., 2003).

An overview of modern domestic species reveals a variety of different breeds within each species that are the results of years of crossbreeding and different selection ways. The concept of cattle breeds emerged in the 18th century in Britain primarily

based upon the experiences of Robert Blackwell (Porter, 1991). At a later time, Felius (1995) in his “Cattle breeds: an encyclopedia” described more than 700 different cattle breeds. Instead, at the moment, the known sheep breeds that are raised worldwide are approximately 200 pure and 400 composite ones (Rasali *et al*., 2005).

The domestication shaped the genome of farm animals: this human action became the first unaware genetic application in animal science because, choosing animal to domesticate, men also chose genes controlling those phenotypes. Indeed, domestication can be defined as the animals and plants genetic adaptation to better suit the human needs (Bruford *et al*., 2003). Conversely, the aware genetics application on plant science started on 1923 when Karl Sax showed the first evidence that genes control both qualitative and qualitative traits.

The domestication history and the modern breeds can be studied using genetics population approach. The bedrocks of this discipline are Darwin, Mendelian and Hardy-Weinberg laws. The latter affirms that in an infinitely large population distinguished by random matings (with no selection, mutation or migration) there is an equilibrium of allelic and genotype frequencies, unless something from outside happens (Hartwell *et al*., 2004). Populations showing these constant gene and genotype frequencies through generations are defined to be in *Hardy-Weinberg equilibrium* (Falconer and Mackay, 1996).

In summary, the process of animal domestication and the consequent selection and crossbreeding, left trace in the animal genomes. These traces are identified with the locution *selection sweeps* or *selection signatures*. Selection sweeps can have two main origins: natural or artificial. Selection signatures can be ascribed to natural

adaptation of animal to the environment (e.g. climate, orography) or due to the selective choice of farmers; in both cases this force leave trace in the farm animal genomes. The appearance of first artificial selection signatures dates back to the farm animal domestication, when an increase in the frequency of genes making the animals more adaptable to human needs was observed (Fan *et al*., 2010). Population genetics theory and genomic analysis are able to identify *selection sweeps* due to more or less recent selection events. In the following paragraphs some techniques that are able to identify these signatures will be examined.

b. Method based on allelic frequency: Wright's F statistics

The genome-wide technique using SNP panel with the aim to identify selection signature or quantify diversity among population, truly used the population genetics theory developed in the past century by the founder of this discipline.

The allelic frequencies and the proportion of polymorphic loci were initially proposed as genetic diversity measurement. However, the use of polymorphic loci proportion was problematic because of the selection criteria of these loci. In order to solve this problem, new indices were created: expected and observed heterozygosity, and observed homozygosity. In a population with N individuals, the *observed homozygosity* is the proportion of homozygotes on the total of individuals (Ho/N) and the *observed heterozygosity* is the proportion between the number of heterozygotes on the total number of individuals in the population (He/N). The *expected heterozygosity* is the possibility that taking randomly one individual from a population it is heterozygote under the hypothesis of HWE.

Another important aspect of population genetics are the Wright statistics, developed during the fifties of the last century. These three indices (F_{IS}, F_{IT} and F_{ST}) allow to represent the genetic structure of a population and the potential presence of sub-populations (Hartl and Clark, 2007). F_{IS} coefficient is used to measure a possible surplus of homozygotes in the population, for example because of matings between relatives (*inbreeding*). F_{ST} parameter represents the probability of extract from one sub-population two different alleles compared to the same probability in the entire population. F_{ST} is used to estimate genetic distance among populations and the genes flow: the more different are the populations, the higher is the F_{ST} value; conversely, the stronger is the flow between the populations the lower is the value.

The F_{IS} and F_{IT} coefficients can assume both positive and negative values: positive if there is a lack of heterozygotes and negative when there is an excess of heterozygotes. Wright F_{ST}, instead, can be only positive, ranging from 0 to 1. Following the Wright (1951) formulation of F statistics, a relationship that combine different source of reduction in expected heterozygosity occurs: $(1 - F_{IT}) = (1 - F_{ST}) \cdot (1 - F_{IS})$ where in the right-hand term the products of the deviations from the expected heterozygosity due to the subpopulation differentiations and the deviations due to the within population inbreeding are equated the overall deviation from hardy-Weinberg expectation (Falconer and Mackay, 1996)

According to Holsinger and Weir (2009), F_{ST} is related to variation in allele frequency among different populations; instead, within population, is an evaluator of the resemblance degree among individuals. They argued that high value of F_{ST} for one locus could be determined by natural selection that concern this locus over others. From pioneer work of Wright (1955), various estimators of F_{ST} have been proposed in

literature (*e.g.* Nei at al., 1977; Weir and Cockerham, 1984). For example, another useful definition of F_{ST} is the ratio between allele variance among populations and the genetic diversity in the entire population according to Weir and Cockerham (1984).

Typically, to estimate the degree of diversity among animal populations – usually belonging to different groups or breeds – F_{ST} is a suitable measure (Holsinger and Weir, 2009) and it became, among the others, one of the most commonly used parameters (Laval *et al.*, 2000; Peter *et al.*, 2007; Mastrangelo *et al*., 2014; Pintus *et al*., 2014; Kelleher *et al*., 2017) within the same species.

c. Linkage disequilibrium in animal population

The genome of an individual can be separated in intervals harboring genes that control the external phenotypic expression. Association among genes is strongly related with their distance on the genome and it can be identified through segregation analysis. With independent assortment, the dissimilarity between the observed and expected segregation ratios is called linkage disequilibrium (LD). The LD level decreases with the succession of generations: after some generations, no LD can be found also between two close genes. LD is related with recombination frequency: the smaller is the LD value, the higher is this frequency. Haldane (1919) derived a relationship aimed at converting recombination frequencies into genetic map units (Morgan, M) assuming a random distribution of recombination events along the chromosome (Weller *et al*., 2016). The genetic distance among genes is measured with centiMorgan (cM), unit that account for crossing-over: two genes are separated by 2 cM if one crossing-over event happens each 200 meiosis. Linkage Disequilibrium reveals a statistically significant association between alleles of two or more loci. The

LD level is higher in homogeneous or closed populations because of their individuals inherited loci from common ancestors.

Several measures (e.g. D', r^2 and $4N_er$) have been identified in order to measure linkage disequilibrium (Zapata, 2000; Ardie *et al*., 2002). If we define A_i and B_i the i-th allele of A and B genes, respectively and with $f(A_iB_i)$ the frequency of haplotype A_iB_i in the population, the simplest LD measure are D and D′:

$$D = [f(A_1B_1)f(A_2B_2)] - [f(A_1B_2)f(A_2B_1)]$$

D is the difference between parental haplotype frequency and the recombinant frequency, while *D′* is calculate dividing D by its maximum value. When *D′* is equal to 1, meaning that two markers have not been separated by any event of recombination, there is a complete, or perfect, linkage disequilibrium; lower values of *D′* indicate that this initial complete ancestral LD has been disrupted during the generations. *D′* is influenced by the sample size: using a small sample LD is overestimated because of a biased D′ (McRae *et al*., 2002). Moreover, *D′* is strongly affected by allele frequency variation (Bohmanova *et al*., 2010).

A better LD estimator is r^2 that is the ratio between D^2 and the product of the four possible allele frequencies; it indicates the alleles correlation of the two considered loci. Furthermore, r^2 allows to estimate the proportion of variance captured by the genetic marker in LD with an eventual causative mutation which cause major changes in the observable phenotype.

$$r^2 = \frac{D^2}{f(A_1)\,f(A_2)\,f(B_1)\,f(B_2)}$$

As though to $D' = 1$, also r^2 equal to 1 is identified as the perfect linkage disequilibrium level; moreover, r^2 is less biased than *D′* because it is less sensitive to

small sample (Ardlie *et al*., 2002; Bohmanova *et al*., 2010). On the other hand, LD calculated using r^2 is underestimated when calculated between two loci with a low minor allele frequency (Espigolan *et al*., 2013).

Another LD measure is the population recombination parameter, $4N_e r$, indicating the four times multiplication of r, recombination rate in the investigated genome region, and N_e (effective population size). Effective population size is identified with the breeding population size, that is an idealized population that reflect effective number of males and females that can be mated rather than actual census.

LD has been largely studied in humans (Pritchard and Przeworski, 2001; Reich *et al*., 2001) and in several domesticated animal species, both of economic and affective interest: such as cattle (Porto-Neto *et al*., 2014; Biegelmeyer *et al*., 2016), buffalo (Nagarajan *et al*., 2009), goat (Brito *et al*., 2015; Mdladla *et al*., 2016), sheep (Kijas *et al*., 2012, 2014; Al-Mamum *et al*., 2015), pigs (Du *et al*., 2007; Amaral *et al*., 2008; Grossi *et al*., 2017), chicken (Andreescu *et al*., 2007; Fu *et al*., 2015), and horse (Corbin *et al*., 2010); or cat (Alhaddad *et al*., 2013) and dog (Boyko, 2011; Stern *et al*., 2013), respectively. Studies about LD are useful to evaluate genetic diversity among breeds, identify genome regions that have been subjected to selection and mapping quantitative trait loci (QTL) on the genome (McRae *et al*., 2002; McKay *et al*., 2007).

In livestock, the average LD shows a typical decay with the increases of distance between loci: in Figure 3 a comparison between LD level in cattle and sheep was reported.

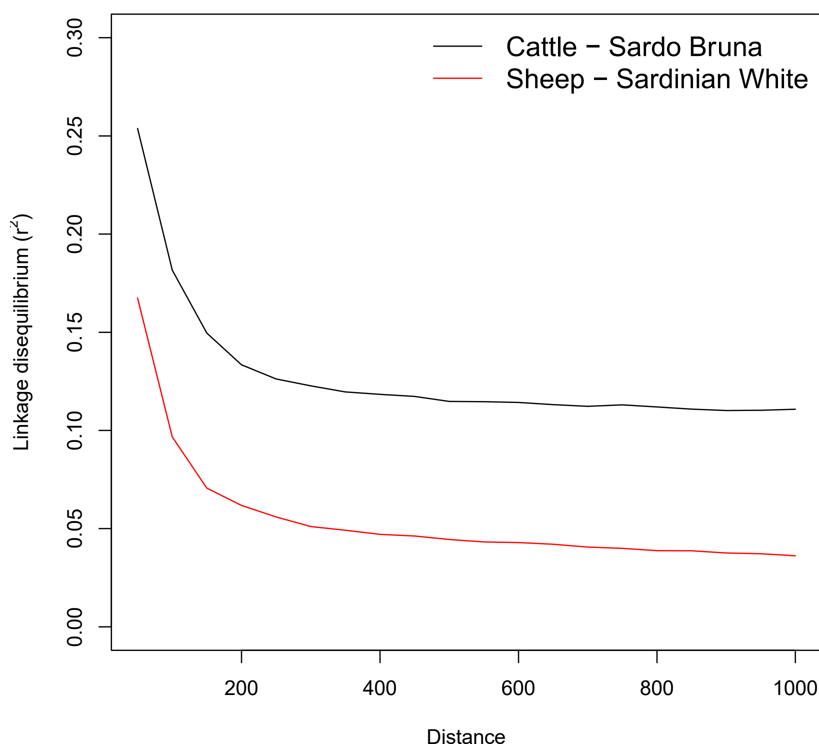


Figure 3. Linkage disequilibrium trend in cattle and sheep breeds with the characteristic decay.

Kijas *et al*. (2014) reported that sheep shows a lower level of LD, probably due to their past domestication history, if compared to pigs or cattle.

The knowledge about linkage disequilibrium is essential for performing genome-wide association studies (GWAS), that are modern genetic tools used to establish relationships between observed phenotypes and genotypes. In order to apply successfully GWAS, SNPs and QTLs must be in linkage disequilibrium at population level (Hayes, 2013, Kijas *et al*., 2014). GWAS are the reference model to map candidate genes affecting trait of economic interest in livestock (Goddard and Hayes, 2009). Moreover, LD knowledge is a requirement for successfully apply genomic selection methodology.

d. Runs of homozygosity

Single Nucleotide polymorphism beadchips can be also used to identify runs of homozygosity (ROH), that are genome segments without heterozygosis in the diploid state (Figure 4). This region is a long string containing only couples of identical alleles (Gibson *et al*., 2006). The most important ROH feature is the length, that can be measured in base pairs (*bp*) or consecutive homozygote SNPs.

Animal 1
AGGATCGAGCTTAGGCAACTGTAGCATGCATATCGATGCATATCG
ATTGACTAGCTTAGGCAACTGTAGCATGCATCATCGTAGTTAGCT

Animal 2
AGGATCGAGC TTAGGCAACTGTAGCATGCATATCGAT GCATATCG
ATCGTAGCTG TTAGGCAACTGTAGCATGCATATCGAT ATCGATGC

Figure 4. Example of runs of homozygosity (ROH) (Marras, PhD thesis).

Runs of homozygosity gained a huge interest in animal science: Peripolli *et al*. (2017) gathered several researches about this topic in a useful review. Interesting ROH applications are the study of inbreeding (Marras *et al*., 2015), population structure and demographic (Curik *et al*., 2014) and selective pressure (Kim *et al*., 2013). Moreover, ROHs allow to simply compute genomic inbreeding coefficients (i.e. F_{ROH}) through the ratio between genome length covered by runs of homozygosity on the total genome length. This index is used to calculate individual inbreeding and its correlation with pedigree-based inbreeding (Kim *et al*., 2015; Marras *et al*., 2015). Inbreeding arose when mating between parents sharing one or more ancestors occur (Curik *et al*., 2014): this occurrence is more frequent in small, close and isolated populations where also with a random choice of parents there is the possibility of mating between relatives. In

particular, inbreeding is defined as the probability that two alleles (or genes) are identical by descent (IBD), meaning that they were inherited from the same ancestral haplotype (Falconer and Mackay, 1996). Alleles are identified as IBD when compared to the base population where alleles are independent. Inbreeding is a typical characteristic of selected population, where mating between relatives occurred more frequently. A higher inbreeding level could lead to a negative phenomenon kwon as *inbreeding depression*, with some negative effects on animal fitness. When an individual inherits from the two related parents the same DNA segment, in that genome region it can be harbor a ROH. The ROH length is a good index to establish how far in the time the mating between relatives (i.e. inbreeding) occurred (Curik *et al*., 2014). Long ROHs suggest recent inbreeding, while small ROH ancient inbreeding: as time passes, some events (e.g. recombination, crossing-over) have more opportunity to cut the IBD region in more smaller segments. Long ROH segments and recombination rate are, indeed, inversely correlated (Bosse *et al*., 2012).

Objective of the Thesis

This work is structured into a general introduction, four chapters of experimental contributions and general conclusions. The first Chapter contains the general introduction, aiming to provide a synopsis about genetic application in animal science. Two main topics were debated in this thesis, modern genomic tools applied in animal breeding and animal biodiversity.

The first two experimental contributions (Chapters 2 and 3) concerned animal breeding programs, particularly the first and the last step: variance components and breeding values estimation, respectively. The contribution about variance components estimation was the result of my abroad period at the Animal Breeding and Genetics Group of the University of Georgia (USA). The third chapter dealt with the estimation of heritability and genomic breeding values for milk fatty acid profile.

Regarding the animal biodiversity, two researches were carried out analyzing both cattle and sheep breeds farmed in Italy. Several statistical approaches were used in order to identify *selection signatures* that can be distinguish among different breeds or productive aptitude within the same species.

Third contribution (Chapter 4) analyzed genetic diversity among five cattle breeds: two purebreds and three crossbreeds raised in Sardinia, the second biggest island of Mediterranean Sea. In the Chapter 5, in order to highlight genome regions that can differentiate livestock production, selection signatures were investigated among different productive aptitudes in the Italian ovine stock.

Finally, the general conclusions reported in Chapter 6 offered a short overview of the main results obtained during the PhD period.

References

1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56.

Abdel-Mawgood A. L. (2012). DNA based techniques for studying genetic diversity. In Genetic Diversity in Microorganisms. InTech.

Aguilar I., Misztal I., Johnson D. L., Legarra A., Tsuruta S., Lawlor T. J. (2010) Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of dDairy Science, 93, 743–752.

Al-Mamun H. A., Clark S., Kwan P., & Gondro C. (2015) Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. Genetics Selection Evolution, 47(1), 90.

Alexandratos N., & Bruinsma J. (2012) World agriculture towards 2030/2050: the 2012 revision (Vol. 12, No. 3). FAO, Rome: ESA Working paper.

Alford A. R., Hegarty R. S., Parnell P. F., Cacho O. J., Herd R. M., & Griffith G. R. (2006) The impact of breeding to reduce residual feed intake on enteric methane emissions from the Australian beef industry. Australian Journal of Experimental Agriculture, 46(7), 813-820.

Alhaddad H., Khan R., Grahn R. A., Gandolfi B., Mullikin J. C., Cole S. A., *et al.* (2013) Extent of linkage disequilibrium in the domestic cat, Felis silvestris catus, and its breeds. PLoS One, 8(1), e53537.

Allendorf F. W., Hohenlohe P. A., & Luikart G. (2010) Genomics and the future of conservation genetics. Nature Reviews Genetics, 11(10), 697-709.

Altmann A., Weber P., Bader D., Preuß M., Binder E. B., & Müller-Myhsok B. (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. Human Genetics, 131(10), 1541-1554.

Amaral A. J., Megens H. J., Crooijmans R. P., Heuven H. C., & Groenen M. A. (2008) Linkage disequilibrium decay and haplotype block structure in the pig. Genetics, 179(1), 569-579.

Andreescu C., Avendano S., Brown S. R., Hassen A., Lamont S. J., & Dekkers J. C. (2007) Linkage disequilibrium in related breeding lines of chickens. Genetics, 177(4), 2161-2169.

Ardlie K. G., Kruglyak L., & Seielstad M. (2002) Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics, 3, 299–309.

Barillet F., Andreoletti O., Palhiere I., Aguerre X., Arranz J. M., Minery S., et al. (2002) Breeding for scrapie resistance using PrP genotyping in the French dairy sheep breeds. In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production (Vol. 31, pp. 683-6).

Becker D., Tetens J., Brunner A., Bürstel D., Ganter M., Kijas J., *et al.* (2010) Microphthalmia in Texel sheep is associated with a missense mutation in the paired-like homeodomain 3 (PITX3) gene. PLoS One, 5(1), e8689.

Bennewitz J., Reinsch N., Reinhardt F., Liu Z., & Kalm E. (2004) Top down preselection using marker assisted estimates of breeding values in dairy cattle. Journal of Animal Breeding and Genetics, 121(5), 307-318.

Biegelmeyer P., Gulias-Gomes C. C., Caetano A. R., Steibel J. P., & Cardoso F. F. (2016) Linkage disequilibrium, persistence of phase and effective population size estimates in Hereford and Braford cattle. BMC Genetics, 17(1), 32.

Bohmanova J., Sargolzaei M., & Schenkel F. S. (2010) Characteristics of linkage disequilibrium in North American Holsteins. BMC Genomics, 11(1), 421.

Boichard D., Fritz S., Rossignol M. N., Boscher M. Y., Malafosse A., & Colleau J. J. (2002) Implementation of marker-assisted selection in French dairy cattle. In Proceedings of the 7th world congress on genetics applied to livestock production (pp. 18-23).

Boichard D., Fritz S., Rossignol M. N., Guillaume F., Colleau J. J., & Druet T. (2006) Implementation of marker-assisted selection: practical lessons from dairy cattle. Proceedings of 8th World Congress on Genetics Applied to Livestock Production, Commun, 22-11.

Bosse M., Megens H. J., Madsen O., Paudel Y., Frantz L. A., Schook L. B., Crooijmans R. P., & Groenen M. A. (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genetics 8, e1003100.

Bovine HapMap Consortium. (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science, 324(5926), 528-532.

Boyko A. R. (2011) The domestic dog: man's best friend in the genomic era. Genome Biology, 12(2), 216.

Brito L. F., Jafarikia M., Grossi D. A., Kijas J. W., Porto-Neto L. R., Ventura R. V., *et al.* (2015) Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats. BMC Genetics, 16(1), 67.

Bruford M. W., Bradley D. G., Luikart G. (2003) DNA markers reveal the complexity of livestock domestication. Nature Reviews Genetics, 4(11), 900-910.

Cecchinato A., De Marchi M., Penasa M., Casellas J., Schiavon S., & Bittante G. (2012) Genetic analysis of beef fatty acid composition predicted by near-infrared spectroscopy. Journal of Animal Science, 90(2), 429-438.

Chen L., Ekine-Dzivenu C., Vinsky M., Basarab J., Aalhus J., Dugan M. E., *et al*. (2015) Genome-wide association and genomic prediction of breeding values for fatty acid composition in subcutaneous adipose and longissimus lumborum muscle of beef cattle. BMC Genetics, 16(1), 135.

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

Chiaia H. L. J., Peripoli E., de Oliveira Silva R. M., Aboujaoude C., Feitosa F. L. B., de Lemos M. V. A., *et al.* (2017) Genomic prediction for beef fatty acid profile in Nellore cattle. Meat Science, 128, 60-67.

Ciani E., Crepaldi P., Nicoloso L., Lasagna E., Sarti F. M., Moioli B., *et al.* (2014) Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. Animal Genetics, 45(2), 256-266.

Clutton-Brock, J. (1999). A natural history of domesticated mammals. Cambridge University Press.

Colli L., Milanesi M., Talenti A., Bertolini F., Chen M., Crisà A., Daly K. G., Del Corvo M., Guldbrandtsen B., Lenstra J. A., Rosen B. D., Vajana E., Catillo G., Joost S., Nicolazzi E. L., Rochat E., Rothschild M. F., Servin B., Sonstegard T. S., Steri R., Van Tassell C. P., Ajmone-Marsan P., Crepaldi P., Stella A., & the AdaptMap Consortium (2018) Genome-wide SNP profiling of worldwide goat

populations reveals strong partitioning of diversity and highlights post-domestication migration routes. Genomic Selection Evolution, 50(1), 58.

Corbin L. J., Blott S. C., Swinburne J. E., Vaudin M., Bishop S. C., & Woolliams J. A. (2010) Linkage disequilibrium and historical effective population size in the Thoroughbred horse. Animal Genetics, 41(s2), 8-15.

Curik I., Ferenčaković M., & Sölkner J. (2014) Inbreeding and runs of homozygosity: a possible solution to an old problem. Livestock Science, 166, 26-34.

Daetwyler H. D., Capitan A., Pausch H., Stothard P., Van Binsbergen R., Brøndum, R. F., *et al.* (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics, 46(8), 858.

Daetwyler H. D., Hickey J. M., Henshall J. M., Dominik S., Gredler B., Van Der Werf J. H. J., & Hayes B. J. (2010) Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Animal Production Science, 50(12), 1004-1010.

Dekkers J. C. M., & Hospital F. (2002) Multifactorial genetics: The use of molecular genetics in the improvement of agricultural populations. Nature Reviews Genetics, 3, 22-32.

Dekkers, J. C. (2004) Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. Journal of Animal Science, 82(13), 313-328.

Di Lorenzo P., Ceccobelli S., Panella F., Attard G., & Lasagna E. (2015) The role of mitochondrial DNA to determine the origin of domestic chicken. World's Poultry Science Journal, 71(2), 311-318.

Dickinson A. G., Stamp J. T., Renwick C. C., & Rennie J. C. (1968) Some factors controlling the incidence of scrapie in Cheviot sheep injected with a Cheviot-passaged scrapie agent. Journal of Comparative Pathology, 78(3), 313-321.

Dimauro C., Nicoloso L., Cellesi M., Macciotta N. P. P., Ciani E., Moioli B., *et al.* (2015) Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. Small Ruminant Research, 128, 27-33.

Du F. X., Clutter A. C., & Lohuis M. M. (2007) Characterizing linkage disequilibrium in pig populations. International Journal of Biological Sciences, 3(3), 166.

Duchemin S. I., Colombani C., Legarra A., Baloche G., Larroque H., Astruc J. M., *et al*. (2012) Genomic selection in the French Lacaune dairy sheep breed. Journal of Dairy Science, 95(5), 2723-2733.

Espigolan R., Baldi F., Boligon A. A., Souza F. R. P., Gordo D. G. M., *et al*. (2013) Study of whole genome linkage disequilibrium in Nellore cattle. BMC Genomics, 14, 305.

Falconer D. S., & Mackay T. F. C. (1996) Introduction to quantitative genetics, 4.

Fan B., Du Z. Q., Gorbach D. M., & Rothschild M. F. (2010). Development and application of high-density SNP arrays in genomic studies of domestic animals. Asian-Australasian Journal of Animal Sciences, 23(7), 833-847.

FAO (2011) World Livestock 2011 – Livestock in Food Security. Rome, FAO.

Farnir F., Coppieters W., Arranz J. J., Berzi P., Cambisano N., Grisart B., *et al*. (2000) Extensive genome-wide linkage disequilibrium in cattle. Genome Research, 10(2), 220-227.

Felius M. (1995) Cattle Breeds: an encyclopedia (No. Ed. 1). C Misset bv.

Ferenčaković M., Hamzic E., Gredler B., Curik I. & Sölkner J. (2011) Runs of homozygosity reveal genome-wide autozygosity in the Austrian Fleckvieh cattle. Agriculturae Conspectus Scientificus, 76, 325–8.

Fernando R. L., & Grossman M. (1989) Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution, 21(4), 467.

Finocchiaro R., Van Kaam J. B. C. H. M., Portolano B., & Misztal I. (2005) Effect of heat stress on production of Mediterranean dairy sheep. Journal of Dairy Science, 88(5), 1855-1864.

Fu W., Dekkers J. C., Lee W. R., & Abasht B. (2015) Linkage disequilibrium in crossbred and pure line chickens. Genetics Selection Evolution, 47(1), 11.

Gianola D., Gustavo A., Hill W. G., Manfredi E., & Fernando R. L. (2009) Additive genetic variability and the Bayesian alphabet. Genetics.

Gianola D., Im S., and Fernando R. L. (1988) Prediction of breeding values under Henderson's selection model: A revisitation. Journal of Dairy Science, 71:2790-2798.

Gibson J., Newton E. M. & Collins A. (2006) Extended tracts of homozygosity in outbred human populations. Human Molecular Genetics, 15, 789–95.

Goddard M. E., & Hayes B. J. (2007). Genomic selection. Journal of Animal breeding and Genetics, 124(6), 323-330.

Goddard M. E., & Hayes B. J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics, 10(6), 381.

Grossi D. A., Jafarikia M., Brito L. F., Buzanskas M. E., Sargolzaei, M., & Schenkel F. S. (2017) Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in Canadian pigs. BMC Genetics, 18(1), 6.

Guillaume F., Fritz S., Boichard D., & Druet T. (2008) Correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls. Journal of Dairy Science, 91(6), 2520-2522.

Hartl, D. L., & Clark, A. G. (2007) Principles of Population Genetics. Sinauer Associates.

Hartwell L. H., Hood L., Goldberg M. L., Reynolds A. E., Silver L. M., & Veres R. C. (2004) Genetics – from gene to genomes, Third edition, McGraw-Hill

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

Hayes B. (2013) Overview of statistical methods for genome-wide association studies (GWAS). In Genome-wide association studies and genomic prediction (pp. 149-169). Humana Press, Totowa, NJ.

Hayes B. J., Bowman P. J., Chamberlain A. J., & Goddard, M. E. (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science, 92(2), 433-443.

Hazel L. N. (1943). The genetic basis for constructing selection indexes. Genetics, 28(6), 476-490.

Hein L., Sørensen L. P., Kargo M., & Buitenhuis A. J. (2018) Genetic analysis of predicted fatty acid profiles of milk from Danish Holstein and Danish Jersey cattle populations. Journal of Dairy Science, 101(3), 2148-2157.

Henderson C. R. (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics, 423-447.

Hofer A. (1998) Variance component estimation in animal breeding: a review. Journal of Animal Breeding and Genetics, 115(1-6), 247-265.

Holsinger K. E., & Weir B. S. (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics, 10(9), 639-650.

Hurtado A., García-Pérez A. L., de Heredia I. B., Barandika J., Sanz-Parra A., Berriatua E., & Juste R. A. (2002) Genetic susceptibility to scrapie in a population of Latxa breed sheep in the Basque Country, Spain. Small Ruminant Research, 45(3), 255-259.

International Chicken Polymorphism Map Consortium (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature, 432(7018), 717.

Jensen J. (2001) Genetic evaluation of dairy cattle using test day models. Journal of Dairy Science, 84:2803–2812.

Kelleher M. M., Berry D. P., Kearney J. F., McParland S., Buckley F., & Purfield D. C. (2017) Inference of population structure of purebred dairy and beef cattle using high-density genotype data. Animal, 11(1), 15-23.

Kennedy B.W., Schaeffer L. R., and Sorensen D. A. (1988) Genetic proprieties of animal models. Journal of Dairy Science, 71:17-26.

Kijas, J. W., Townley, D., Dalrymple, B. P., Heaton, M. P., Maddox, J. F., McGrath, A., ... & Tang, D. (2009). A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PloS one, 4(3), e4668.

Kijas J. W., Lenstra J. A., Hayes B., Boitard S., Neto L. R. P., 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), e1001258.

Kijas J. W., Porto-Neto L., Dominik S., Reverter A., Bunch R., McCulloch R., *et al.* (2014) Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. Animal Genetics, 45(5), 754-757.

Kim E. S., Sonstegard T. S., Van Tassell C. P., Wiggans G., & Rothschild M. F. (2015) The relationship between runs of homozygosity and inbreeding in Jersey cattle under selection. PloS One, 10(7), e0129967.

Kim E.S., Cole J. B., Huson H., Wiggans G. R., Van Tassell C. P., Crooker B. A., Liu G., Da Y., & Sonstegard T. S. (2013) Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle. PLoS One 8, e80813.

Lande R., & Thompson R. (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124(3), 743-756.

Larson G., & Fuller D. Q. (2014) The evolution of animal domestication. Annual Review of Ecology, Evolution, and Systematics, 45, 115-136.

Laval G., Iannuccelli N., Legault C., Milan D., Groenen M. A., Giuffra E., Andersson L., Nissen P. H., Jørgensen C. B., Beeckmann P., Geldermann H., Foulley J., Chevalet C., & Olivier L. (2000). Genetic diversity of eleven European pig breeds. Genetics Selection Evolution, 32(2), 187.

Ligios C., Cancedda M. G., Madau L., Santucciu C., Maestrale C., Agrimi U., *et al.* (2006) PrP Sc deposition in nervous tissues without lymphoid tissue involvement is frequently found in ARQ/ARQ Sarda breed sheep preclinically affected with natural scrapie. Archives of Virology, 151(10), 2007-2020.

Macciotta N. P. P., Mele M., Cappio-Borlino A., & Secchiari P. (2005) Issues and perspectives in dairy sheep breeding. Italian Journal of Animal Science, 4(1), 5-23.

Macciotta N. P. P., Gaspa G., Steri R., Nicolazzi E. L., Dimauro C., Pieramati C., and Cappio-Borlino A. (2010) Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. Journal of Dairy Science, 93:2765-2774.

Macciotta N. P. P., Biffani S., Bernabucci U., Lacetera N., Vitali A., Ajmone-Marsan P., & Nardone A. (2017) Derivation and genome-wide association study of a principal component-based measure of heat tolerance in dairy cattle. Journal of Dairy Science, 100(6), 4683-4697.

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

Manel S., Schwartz M. K., Luikart G., & Taberlet P. (2003) Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology & Evolution, 18(4), 189-197.

Marras G., Gaspa G., Sorbolini S., Dimauro C., Ajmone-Marsan P., Valentini A., *et al.* (2015) Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Animal Genetics, 46(2), 110-121.

Martinez-Royo A., Alabart J. L., Sarto P., Serrano M., Lahoz B., Folch J., & Calvo, J. H. (2017) Genome-wide association studies for reproductive seasonality traits in Rasa Aragonesa sheep breed. Theriogenology, 99, 21-29.

Mastrangelo S., Di Gerlando R., Tolone M., Tortorici L., Sardina M. T., & Portolano B. (2014) Genome wide linkage disequilibrium and genetic structure in Sicilian dairy sheep breeds. BMC Genetics, 15(1), 108.

Mastrangelo, S., Ciani, E., Sardina, M. T., Sottile, G., Pilla, F., Portolano, B., & Bi. Ov. Ita Consortium. (2018). Runs of homozygosity reveal genome-wide autozygosity in Italian sheep breeds. Animal genetics, 49(1), 71-81.

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

McKay S. D., Schnabel R. D., Murdoch B. M., Matukumalli L. K., Aerts J., Coppieters W., *et al.* (2007) Whole genome linkage disequilibrium maps in cattle. BMC Genetics, 8(1), 74.

McRae A. F., McEwan J. C., Dodds K. G., Wilson T., Crawford A. M., & Slate J. (2002) Linkage disequilibrium in domestic sheep. Genetics, 160(3), 1113-1122.

Mdladla K., Dzomba E. F., Huson H. J., & Muchadeyi F. C. (2016) Population genomic structure and linkage disequilibrium analysis of South African goat breeds using genome-wide SNP data. Animal Genetics, 47(4), 471-482.

Mele M., Dal Zotto R., Cassandro M., Conte G., Serra A., Buccioni A., *et al.* (2009). Genetic parameters for conjugated linoleic acid, selected milk fatty acids, and milk fatty acid unsaturation of Italian Holstein-Friesian cows. Journal of dairy Science, 92(1), 392-400.

Meuwissen T. H. E., Hayes B. J., & Goddard M. E. (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics, 157:1819-1829.

Misztal I. (2008) Reliable computing in estimation of variance components. Journal of Animal Breeding and Genetics, 125(6), 363–370.

Moioli B., Contarini G., Pariset L., Marchitelli C., Crisà A., Catillo G., *et al.* (2012) Genetic variation of C18: 1 and C18: 2 isomers in sheep milk fat. Small ruminant research, 103(2-3), 187-193.

Moser G., Tier B., Crump R. E., Khatkar M. S., & Raadsma H. W. (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution, 41:56.

Naderi S., Rezaei H. R., Pompanon F., Blum M. G., Negrini R., Naghash H. R., *et al.* (2008) The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. Proceedings of the National Academy of Sciences, 105(46), 17659-17664.

Nagarajan M., Kumar N., Nishanth G., Haribaskar R., Paranthaman K., Gupta J., *et al.* (2009) Microsatellite markers of water buffalo, Bubalus bubalis-development, characterisation and linkage disequilibrium studies. BMC Genetics, 10(1), 68.

Negussie E., de Haas Y., Dehareng F., Dewhurst R. J., Dijkstra J., Gengler N., *et al.* (2017) Invited review: Large-scale indirect measurements for enteric methane emissions in dairy cattle: A review of proxies and their potential for use in

management and breeding decisions. Journal of Dairy Science, 100(4), 2433-2453.

Nei M. (1977). F-statistics and analysis of gene diversity in subdivided populations. Annals of human genetics, 41(2), 225-233.

Ostersen T., Christensen O. F., Henryon M., Nielsen B., Su G., & Madsen P. (2011) Deregressed EBV as response variable yield reliable genomic predictions for pigs. Genetics Selection Evolution, 43, 38.

Palhiere I., Brochard M., Moazami-Goudarzi K., Laloë D., Amigues Y., Bed'Hom B., *et al*. (2008) Impact of strong selection for the PrP major gene on genetic variability of four French sheep breeds (Open Access publication). Genetics Selection Evolution, 40(6), 663.

Pareek C. S., Smoczynski R., & Tretyn A. (2011). Sequencing technologies and genome sequencing. Journal of Applied Genetics, 52(4), 413-435.

Peripolli E., Munari D. P., Silva M. V. G. B., Lima A. L. F., Irgang R., & Baldi F. (2017). Runs of homozygosity: current knowledge and applications in livestock. Animal Genetics, 48(3), 255-271.

Peñagaricano, F., Weigel, K. A., & Khatib, H. (2012). Genome-wide association study identifies candidate markers for bull fertility in Holstein dairy cattle. Animal genetics, 43, 65-71.

Peter, C., Bruford, M., Perez, T., Dalamitra, S., Hewitt, G., Erhardt, G., & Econogene Consortium (2007). Genetic diversity and subdivision of 57 European and Middle-Eastern sheep breeds. Animal genetics, 38(1), 37-44.

Pintus E., Sorbolini S., Albera A., Gaspa G., Dimauro C., Steri R., *et al.* (2014) Use of locally weighted scatterplot smoothing (LOWESS) regression to study

selection signatures in Piedmontese and Italian Brown cattle breeds. Animal Genetics, 45(1), 1-11.

Porter V. (1991) Cattle, A Handbook to the Breeds of the World. A & C Black: London. CABI.

Porto-Neto L. R., Kijas J. W., & Reverter A. (2014) The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genetics Selection Evolution, 46(1), 22.

Prieur V., Clarke S. M., Brito L. F., McEwan J. C., Lee M. A., Brauning R., *et al.* (2017). Estimation of linkage disequilibrium and effective population size in New Zealand sheep using three different methods to create genetic maps. BMC Genetics, 18(1), 68.

Pritchard J. K., & Przeworski M. (2001) Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics, 69(1), 1-14.

Pryce J. E., & Daetwyler H. D. (2012) Designing dairy cattle breeding schemes under genomic selection: a review of international research. Animal Production Science, 52(3), 107-114.

Rasali D. P., Shrestha J. N. B., Crow G. H. (2005) Development of composite sheep breeds in the world: a review. Canadian Journal of Animal Science, 86, 1-24.

Reich D. E., Cargill M., Bolk S., Ireland J., Sabeti P. C., Richter D. J., *et al.* (2001) Linkage disequilibrium in the human genome. Nature, 411(6834), 199.

Rendel J. M., & Robertson A. (1950) Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. Journal of Genetics, 50(1), 1-8.

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

Rovadoscki G. A., Pertile S. F. N., Alvarenga A. B., Cesar A. S. M., Pértille F., Petrini J., *et al.* (2018). Estimates of genomic heritability and genome-wide association study for fatty acids profile in Santa Inês sheep. BMC genomics, 19(1), 375.

Ryder M. L. (1983) Sheep and man. Gerald Duckworth & Co. Ltd.

Sachidanandam R., Weissman D., Schmidt S. C., Kakol J. M., Stein L. D., Marth G., Sherry S., Mullikin J. C., Mortimore B. J., Willey D. L., Hunt S. E., *et al* (2001) A map of human genome sequence variation containing 1.42 milions single nucleotide polymorphisms. Nature, 409, 928-933

Sanger F. (1975) The Croonian Lecture, 1975: Nucleotide Sequences in DNA, B191. Proceedings of the Royal Society, London, 317

Schaeffer L. R. (2004) Application of random regression models in animal breeding. Livestock Production Science, 86(1), 35-45.

Schaeffer L. R. (2006) Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics, 123(4), 218-223.

Scherf B. D. (2000) World watch list for domestic animal diversity (No. Ed. 3). Food and Agriculture Organization (FAO).

Sharma A., Lee S. H., Lim D., Chai H. H., Choi B. H., & Cho Y. (2016) A genome-wide assessment of genetic diversity and population structure of Korean native cattle breeds. BMC Genetics, 17(1), 139.

Snelling W. M., Kuehn L. A., Keel B. N., Thallman R. M., & Bennett G. L. (2017) Linkage disequilibrium among commonly genotyped SNP variants detected from bull sequence. Animal Genetics, 48(5), 516-522.

Soyeurt H., Gillon A., Vanderick S., Mayeres P., Bertozzi C., & Gengler N. (2007) Estimation of heritability and genetic correlations for the major fatty acids in bovine milk. Journal of Dairy Science, 90(9), 4435-4442.

Stella A., Nicolazzi E. L., Van Tassell C. P., Rothschild M. F., Colli L., Rosen B. D., Sonstegard T. S., Crepaldi P., Tosser-Klopp G., Joost S., & the AdaptMap Consortium (2018). AdaptMap: exploring goat diversity and adaptation.

Stern J. A., White S. N., & Meurs, K. M. (2013) Extent of linkage disequilibrium in large-breed dogs: chromosomal and breed variation. Mammalian Genome, 24(9-10), 409-415.

Testoni S., Bartolone E., Rossi M., Patrignani A., Bruggmann R., Lichtner P., *et al.* (2012). KDM2B is implicated in bovine lethal multi-organic developmental dysplasia. PloS One, 7(9), e45634.

The Bovine HapMap Consortium. (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science, 324:528-532.

Thompson R., Brotherstone S., White I.M.S. (2005) Estimation of quantitative genetic parameters. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 360, 1469–1477.

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

Upadhyay M. R., Chen W., Lenstra J. A., Goderie C. R. J., MacHugh D. E., Park S. D. E., *et al.* (2017) Genetic origin, admixture and population history of aurochs (Bosprimigenius) and primitive European cattle. Heredity, 118(2), 169.

Van der Werf J. H. J. (2009) Potential benefit of genomic selection in sheep. In Proceedings of the Association for the Advancement of Animal Breeding and Genetics (Vol. 18, pp. 38-41).

VanRaden P. M. (2008) Efficient methods to compute genomic predictions. Journal of Dairy Science, 91(11), 4414-4423.

VanRaden P. M., O'Connell J. R., Wiggans G. R., & Weigel K. A. (2011) Genomic evaluations with many more genotypes. Genetics Selection Evolution, 43(1), 10.

VanRaden P. M., Van Tassell C. P., Wiggans G. R., Sonstegard T. S., Schnabel R. D., Taylor J. F., & Schenkel F. S. (2009) Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science, 92(1), 16-24.

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.

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

Weir B. S., & Cockerham C. C. (1984). Estimating F-statistics for the analysis of population structure. evolution, 38(6), 1358-1370.

Weller J. (2016) Genomic selection in animals. John Wiley & Sons.

Wellmann R., Preuß S., Tholen E., Heinkel J., Wimmers K., and Bennewitz J. (2013) Genomic selection using low density marker panels with application to a sire line in pigs. Genetics Selection Evolution, 45:28.

Yamey G. (2000) Scientists unveil first draft of human genome. BMJ: British Medical Journal, 321(7252), 7.

Zapata C. (2000) The D′ measure of overall gametic disequilibrium between pairs of multiallelic loci. Evolution, 54: 1809–1812.

Zeder M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proceedings of the National Academy of Sciences, 105(33), 11597-11604.

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).

Zhao X., Onteru S. K., Piripi S., Thompson K. G., Blair H. T., Garrick D. J., Rothschild M. F. (2012). In a shake of a lamb's tail: using genomics to unravel a cause of chondrodysplasia in Texel sheep, Animal Genetics, 43, 9-18.

Zhu B., Niu H., Zhang W., Wang Z., Liang Y., Guan L., *et al* (2017) Genome wide association study and genomic prediction for fatty acid composition in Chinese Simmental beef cattle using high density SNP array. BMC Genomics, 18(1), 464.

CHAPTER 2

Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies

This is the published version in **Journal of Animal Breeding and Genetics** of:

Cesarani A, Pocrnic I, Macciotta NPP, Fragomeni BO, Misztal I, Lourenco DAL. Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies. J Anim Breed Genet. 2019; 136:40–50. https://doi.org/10.1111/jbg.12367

Abstract

We investigated the effects of different strategies for genotyping populations on variance components and heritabilities estimated with an animal model under restricted maximum likelihood (REML), genomic REML (GREML), and single-step GREML (ssGREML). A population with 10 generations was simulated. Animals from the last one, two, or three generations were genotyped with 45,116 SNP evenly distributed on 27 chromosomes. Animals to be genotyped were chosen randomly or based on EBV. Each scenario was replicated five times. A single trait was simulated with three heritability levels (low, moderate, high). Phenotypes were simulated for only females to mimic dairy sheep and also for both sexes to mimic meat sheep. Variance component estimates from genomic data and phenotypes for one or two generations were more biased than from three generations. Estimates in the scenario without selection were the most accurate across heritability levels and methods. When selection was present in the simulations, the best option was to use genotypes of randomly selected animals. For selective genotyping, heritabilities from GREML were more biased compared to those estimated by ssGREML, because ssGREML was less affected by selective or limited genotyping.

Keywords

genotyping scheme, selective genotyping, single-step genomic BLUP, REML, variance

Introduction

The most popular method for estimation of variance components is the restricted maximum likelihood (REML; Patterson and Thompson, 1971). The REML theory is based on the BLUP mixed model equations (Henderson, 1975), which provide unbiased predictions if all data used for selection is used in the analysis and the model is correct. When genomic information is available, variance components can be estimated using genomic REML (GREML) if only phenotypes of genotyped animals are considered, or single-step GREML (ssGREML) if phenotypes of all available including ungenotyped animals are considered. While initial costs of REML with genomic data was high due to dense blocks of mixed model equations generated by the genomic information, updated sparse-matrix techniques allow the use of a large number of genotyped individuals in the estimation of variance components (Masuda *et al*., 2015).

In general, the transition from REML using all available data to ssGREML should be straightforward as the only difference is availability of more data. Forni *et al*. (2011) and Veerkamp *et al*. (2011) found that estimates by REML and ssGREML are similar, however, standard errors with the latter are smaller due to more data available. The advantage of GREML over REML depends on whether important data on ungenotyped individuals can be discarded. Yang *et al*. (2010) applied GREML to a human dataset to find out what fraction of the genetic variance for human height is accounted for by SNP. In human populations, pedigrees are disconnected and genomic relationships can capture distant relationships among genotyped individuals. In a simulation study mimicking livestock, Hayes and Goddard (2008) stated that estimates by GREML can

be more accurate than by REML when pedigrees have errors as the genomic information is not subject to pedigree errors.

While the main goal of using genomic information in animal breeding is to increase the accuracy of selection, estimating variance components can be a quality control check and provide more accurate parameters for the model. Cost-effectiveness in genomic selection (GS) usually requires selective genotyping. While early genotyping in dairy cattle focus on high-reliability and young bulls, the number of such bulls is limited and subsequently most of current genotyping is for females (Cooper *et al*., 2015). Inclusion of genotypes from cows with phenotypic records could increase GEBV accuracy in small populations (Pryce *et al*., 2012; Jenko *et al*., 2017). In small populations like dairy sheep the artificial insemination is not used and rams have small progeny groups. For such populations, a feasible strategy for genomic selection is the use of female genomic information, and a breeding program based on both males and females.

In extensive sheep farming systems, it is difficult to collect accurate relationship information due to the simultaneous presence of more rams in the same group (Hayes and Goddard, 2008). Missing or incomplete pedigree information, especially on the side of the pedigree with larger progeny size, can severely bias variance components estimation (Israel and Weller, 2000; Visscher *et al*., 2002). However, the genomic information may compensate for pedigree problems (Hayes and Goddard, 2008).

In small populations, genotyped animals tend to be highly selected and do not truly represent the population structure. This could bias estimation of variance components because selection can be viewed as a missing-data process (Im *et al*., 1989). Also, the number of genotyped animals increased over times as genotyping become less

expensive, which made early genotyping highly selective. In conjunction with possibly incomplete pedigrees, the main information in a population could be phenotypes and genotypes of highly selective animals. The purpose of this study was to investigate the effect of artificial selection and selective genotyping on heritability estimates when genomic information is included in GREML and ssGREML. The study focuses on simulated populations that lack deep pedigree and have genotyped animals only from a few recent generations.

Materials and methods

Data

Sheep populations were simulated using QMSim (Sargolzaei and Schenkel, 2009). The simulated genome consisted of 27 chromosomes with 45,116 evenly allocated biallelic SNP markers and 1,038 biallelic and randomly distributed QTL with allelic effects sampled from a gamma distribution with a shape parameter of 0.4. All genetic variance was assigned to QTL. The recurrent mutation rate of SNP and QTL was assumed to be 2.5×10^{-5} per locus per generation (Solberg *et al*., 2008). First, 2,000 generations of a historical population were generated with a gradual decrease from 10,000 animals in generation 0 to 5,000 animals in generation 500 and then with a gradual increase to 10,000 animals in generation 2,000. This first step was performed to create initial linkage disequilibrium (LD) and establish mutation-drift balance in the population. Ten recent, overlapping generations were simulated for each scenario with 40 males and 2,000 females as founders (ratio of 1 male:50 females), which corresponds to an effective population size about 150 when calculated from classical formula based on number of breeding males and females (Wright, 1931). The number

of breeding males and females was kept constant throughout the recent generations. Phenotypes were simulated with overall mean (μ = 3) as the fixed effect, and with three different heritability levels: low (h^2=10%), moderate (h^2=30%), and high (h^2=50%). Phenotypes were simulated for all 10 recent generations, whereas genotypes were simulated for only the last 3 generations.

Different scenarios considered artificial selection as a process causing missing data. In the first scenario (NF), no artificial selection was applied and phenotypes were simulated only for females (i.e., sex-limited trait). In the second (SF), selection was based on estimated breeding values (EBV) calculated by QMSim using BLUP, and phenotypes were simulated only on females (i.e., sex-limited trait). In the third (SMF), both males and females were selected also based on EBV, and both sexes had phenotypes. In all scenarios the mating was random. The NF and SF scenarios mimicked a dairy sheep population, whereas SMF mimicked a meat sheep population. For all scenarios, twinning rate was 40%, proportion of male progeny was 50%, and sire and dam replacement rate were 60 and 30%, respectively. Therefore, selection intensity was the same for all selection scenarios. Number of animals in pedigrees and phenotypes are shown in Table 1 for each scenario. Three different subscenarios based on the number of generations (1, 2, or 3) with genotypic and phenotypic records were also considered for each of the three simulated scenarios.

Table 1. Numbers of phenotype and pedigree records in simulated data.

	Simulated scenario		
Data category	No selection, female phenotypes	EBV-based selection, female phenotypes	EBV-based selection, male and female phenotypes
Phenotype			
One generation	1,405 ± 31	1,397 ± 26	2,795 ± 14
Two generations	2,810 ± 39	2,773 ± 39	5,582 ± 11
Three generations	4,194 ± 44	4,174 ± 40	8,391 ± 25
Pedigree			
REML/ssGREML	29,975 ± 81	30,048 ± 66	30,013 ± 80
Inbreeding coefficient	0.007 ± 0.0002	0.037 ± 0.003	0.030 ± 0.008

REML, restricted maximum likelihood; ssGREML, single-step genomic restricted maximum likelihood.

For all simulations, 1,000 animals were genotyped for each of the last 3 generations (from 8 to 10). Only females were genotyped for NF and SF; both males and females or only females were genotyped for SMF. Two genotyping strategies were tested: randomly chosen animals or the best animals ranked on EBV. Average minor allele frequency was close to 0.5 for both genotyping strategies and in all selection scenarios. After replicating the simulation five times, we observed very small standard errors (0.003 – 0.06); therefore, average of five replicates are reported. All possible scenarios are reported in Table 2.

Table 2. All possible scenario combinations tested for the three heritability levels.

Data category	Selection scheme: No selection, female phenotypes	EBV-based selection, female phenotypes	EBV-based selection, male and female phenotypes
Randomly genotyping			
Females	yes	yes	yes
Both males/females			yes
Best genotyping			
Females	yes	yes	yes
Both males/females			yes

Based on that, the missing data process in our simulations was represented by selection, amount of generations to be genotyped, and existence of selective genotyping. The complete information was represented by pedigree and phenotypes for all 10 recent generations and genotypes for all animals in the last 3 generations. According to Im *et al*. (1989) and confirmed by Cantet *et al*. (2000), if a population is undergone selection but all the data used to make decisions is considered, the selection process is ignorable. In this way, heritability estimates should be close to the simulated value when having the complete information. Contrarily, bias is expected when data is missing. Under the complete information we tested REML and ssGREML, and under the missing data scenarios, we tested the ability of REML, GREML, and ssGBLUP in estimating heritabilities, as described below.

Models and analysis

A mixed linear model was used to estimate variance components:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e},$$

where y is a vector of simulated phenotypes, μ is overall mean, u is a vector of additive animal effects, Z is an incidence matrix relating phenotypes in y to additive genetic effects in u, and e is a vector of random residuals. Heritability (h^2) was estimated as

$$h^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2),$$

where σ_a^2 is the additive genetic variance and σ_e^2 is the residual variance.

To test the effect of different amounts of information on variance component estimation, three different genetic (co)variance structures were considered for **u**: (i) REML that included only pedigree information, $\mathbf{u} \sim \mathrm{N}(0, \mathbf{A}\sigma_a^2)$; (ii) GREML that included only genomic information, $\mathbf{u} \sim \mathrm{N}(0, \mathbf{G}\sigma_a^2)$; (iii) ssGREML that included combined pedigree and genomic information, $\mathbf{u} \sim \mathrm{N}(0, \mathbf{H}\sigma_a^2)$, where **H** is a matrix that combines **A** and **G** with its inverse as defined in Aguilar *et al*. (2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

where $\mathbf{A}_{22}^{-1}$ is the inverse of the pedigree relationship matrix among genotyped animals. Variance components were estimated considering genotypes for the last one, two, or three generations. For both random and selective genotyping, genotyped animals could be only females or both sexes. For REML and ssGBLUP, ancestors were allowed in the pedigree in addition to the last three generations. For GREML, data were restricted to the last three generations because of method limitations. For all computations, average-information REML was used as implemented in AIREMLF90 (Misztal *et al*., 2015).

In GREML and ssGREML, the genomic relationship matrix was constructed as in VanRaden (2008):

$$G = \frac{MM'}{2\sum p_i(1\text{-}p_i)}$$

where **M** is a matrix of genotypes centered by twice the current allele frequencies (p); i is the ith locus. To avoid singularity problems, **G** was blended with 5% of **A22**. This **G** matrix has been widely used in applications of genomic BLUP (GBLUP) and single-step GBLUP (ssGBLUP), although it is identity by state. Other ways to compute G that account for identity by descent relationships, and can be more accurate, have been proposed (Forneris *et al*., 2016; Odegard and Meuwissen, 2014); however, they are less trivial than VanRaden's **G**.

As pedigree and genomic relationships are combined in ssGREML, **G** was tuned to match **A22** as proposed by Vitezica *et al*. (2011). This procedure is done to account for the shift in the mean genetic value of genotyped animals because of selection, as usually only recent and highly selected generations are genotyped. Therefore, tuning helps to adjust **G** to a common base population.

Results

For sake of comparison and to confirm the theory presented by Im *et al*. (1989) and Cantet *et al*. (2000), when the complete information was used to estimate heritability under REML and ssGREML assuming selection, values were very close to the simulated ones (i.e., 10%, 30%, and 51% for REML and 10%, 29%, and 51% for ssGREML). As only the 3 last generations were genotyped, using GREML with complete information was not possible.

Results for the moderate heritability (h^2=30%) under missing information were reported in Figure 1 and Table S1. For the NF scenario with no selection (random mating) and only female phenotypes, heritability estimates (Figure 1a) from all methods except GREML were quite close to the simulated heritability of 30% and ranged from 28 to 31%. Heritability estimates were closest to 30% with three generations of genotypes and phenotypes. Random genotyping resulted in the best heritability estimates, whereas genotyping the best females led to overestimates with GREML. Heritability estimates were most accurate with REML.

When selection was based on EBV calculated from female phenotypes (SF scenario), heritability estimates from all methods (Figure 1b) were less accurate than for the NF scenario and ranged from 20 to 44%. Again, heritability estimates were best with three generations of data except for GREML, which overestimated heritability when best females were genotyped regardless of the number of generations. When best females were genotyped, heritability estimates were most accurate with ssGREML.

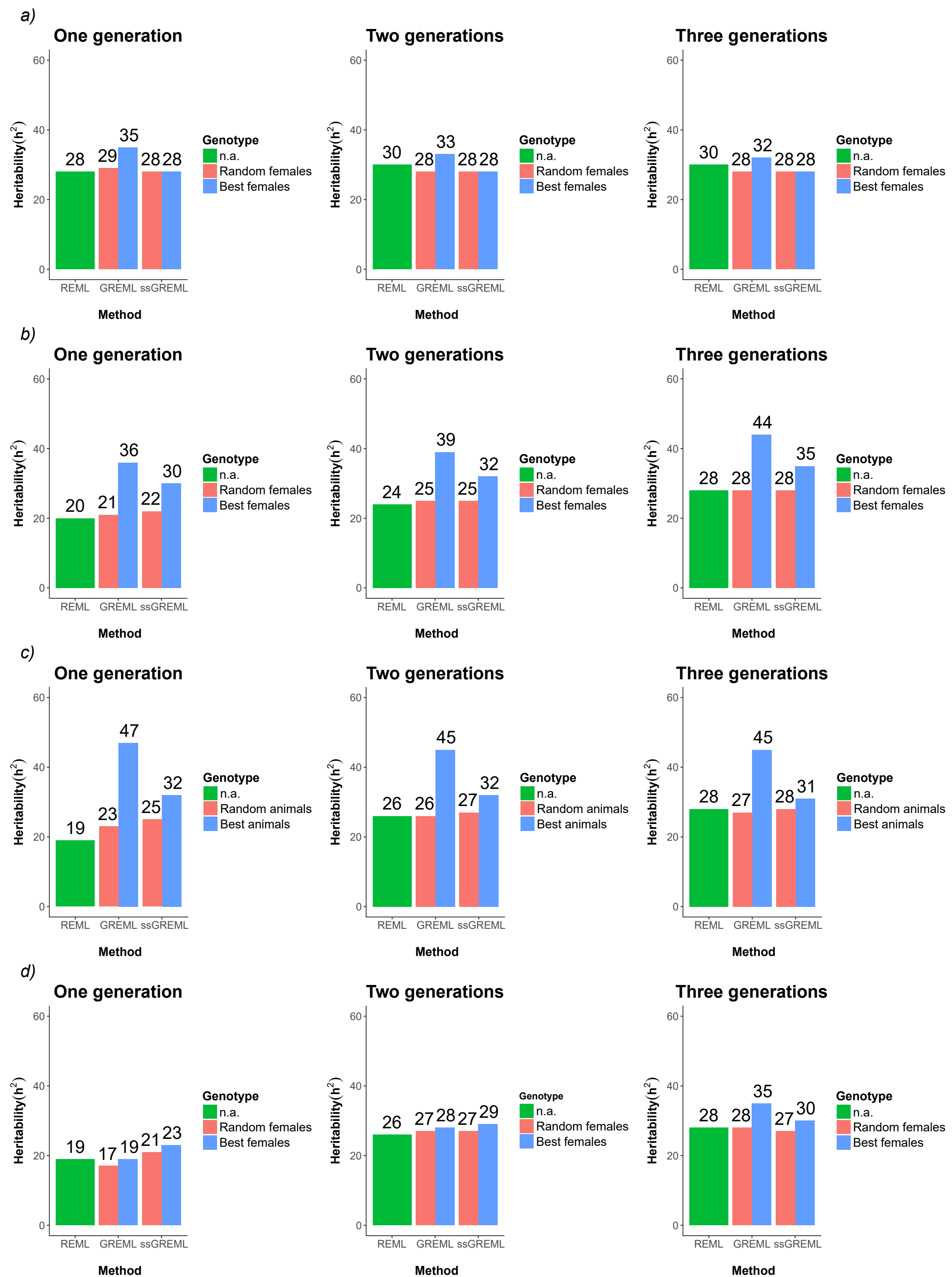


Figure 1. Heritability estimates (h^2=30%) from four simulated scenarios for genotyping strategies of randomly chosen animals or the best animals ranked on EBV: no selection and only female phenotypes (a), selection and only female phenotypes (b), selection and male and female phenotypes and genotypes (c), and selection, male and female phenotypes, and female genotypes (d). For one, two, or three generations, heritability was estimated using restricted maximum likelihood (REML) with only pedigree information, genomic REML (GREML) with only genomic information, and single-step GREML (ssGREML) with combined pedigree and genomic information. Simulated heritability was 0.3.

For the SMF scenario with both male and female genotypes, variation in heritability estimates was greater among methods and genotyping strategies (Figure 1c) than for NF and SF scenarios. Once again accuracy of heritability estimates increased with number of generations. The best method was ssGREML combined with genotyping the best animals. However, including genotypes of the best animals when using GREML resulted in a marked overestimation of heritability, especially when only one generation of animals was considered. For selective genotyping, heritabilities from GREML were, on average, overestimated by 49% compared with the simulated heritability of 30%. However, overestimation with ssGREML was only 3%.

Variation of heritability estimates in the SMF scenario was reduced when only female genotypes were included (Figure 1d). As before, heritability estimates were most accurate with three generations. The highest overestimate was found using GREML with only the best females genotyped and three generations of data.

A similar trend was observed in the simulation with low heritability, i.e. 10% (Figure 2 and Table S2). Best estimates were found when three generations were included in the analysis, with reduction in over/underestimation when gradually moving from one to three generations. As far as the genotyping strategy is concerned, the worst cases were observed when only best animal/females were genotyped. Scenarios without selection (NF) were less affected by the genotyping strategy (Figure 2).

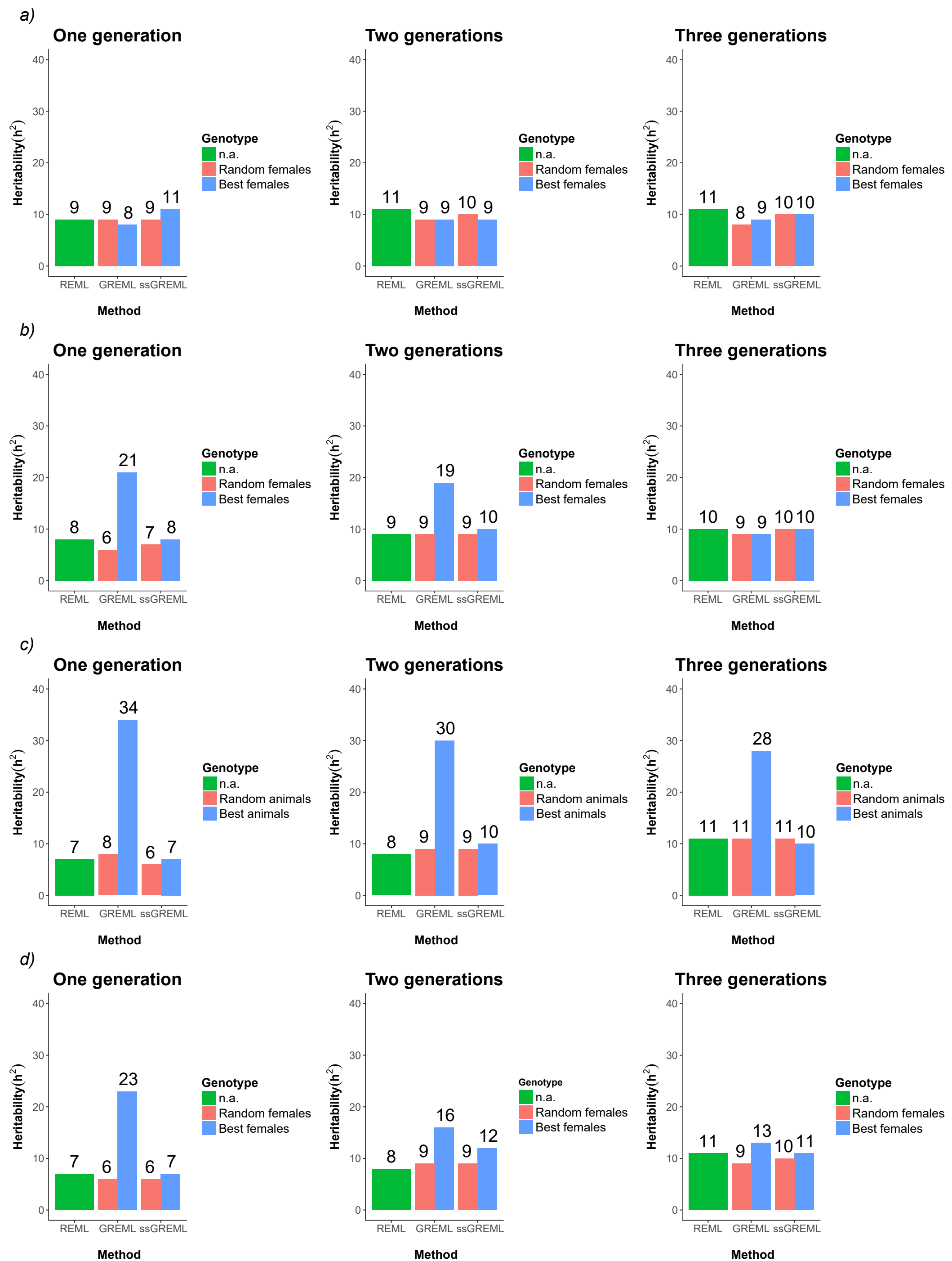


Figure 2. Heritability estimates (h^2=10%) from four simulated scenarios for genotyping strategies of randomly chosen animals or the best animals ranked on EBV: no selection and only female phenotypes (a), selection and only female phenotypes (b), selection and male and female phenotypes and genotypes (c), and selection, male and female phenotypes, and female genotypes (d). For one, two, or three generations, heritability was estimated using restricted maximum likelihood (REML) with only pedigree information, genomic REML (GREML) with only genomic information, and single-step GREML (ssGREML) with combined pedigree and genomic information. Simulated heritability was 0.1.

Regarding the estimation method, ssGREML was the most precise and less biased, whereas GREML showed estimate errors when extreme genotypes were considered. However, in these two methods, biases were lower when three generations were included in the analysis. The heritability estimate errors were greater when best animals were genotyped in the SMF scenario using GREML. When only one generation with phenotypes and genotypes was considered, the estimated value was three times higher than the simulated one. The inclusion of genotypes from randomly chosen females returned correct estimates independently of the method or scenario.

For the simulation with high heritability (50%), the estimated values fluctuated more (Figure 3 and Table S3). Also, in this simulation, the inclusion of phenotypes and genotypes from three generations resulted in the best estimates, except for SMF scenario with best females genotyped. In the scenario without selection (NF) all methods and genotyping strategies gave h^2 estimates very close to the simulated values. Regarding to the genotyping strategy, randomly genotyping females was best for obtaining h^2 estimates close to the true values. Heritability estimates were more biased when GREML was considered and only the best animals were genotyped. The combined use of selective genotyping and GREML method in SMF scenario showed an opposing trend, as choosing best animals strongly underestimated h^2 (especially with one or two generations where the estimated values were close to zero), whereas choosing best females resulted in overestimated values (Figure 2c and 2d). Overestimated h^2 values were observed using best female genotypes in SF scenario as well. The use of ssGREML method in the scenarios with selection (SF and SMF) resulted in good estimates and reduction of estimate errors even when best genotypes were included in the analysis (Table S3). For example, the h^2 value obtained using

ssGREML with three generations of best animals genotyped was very close to the simulated ones (0.52±0.02), whereas with GREML the value was 25% lower (0.27±0.03).

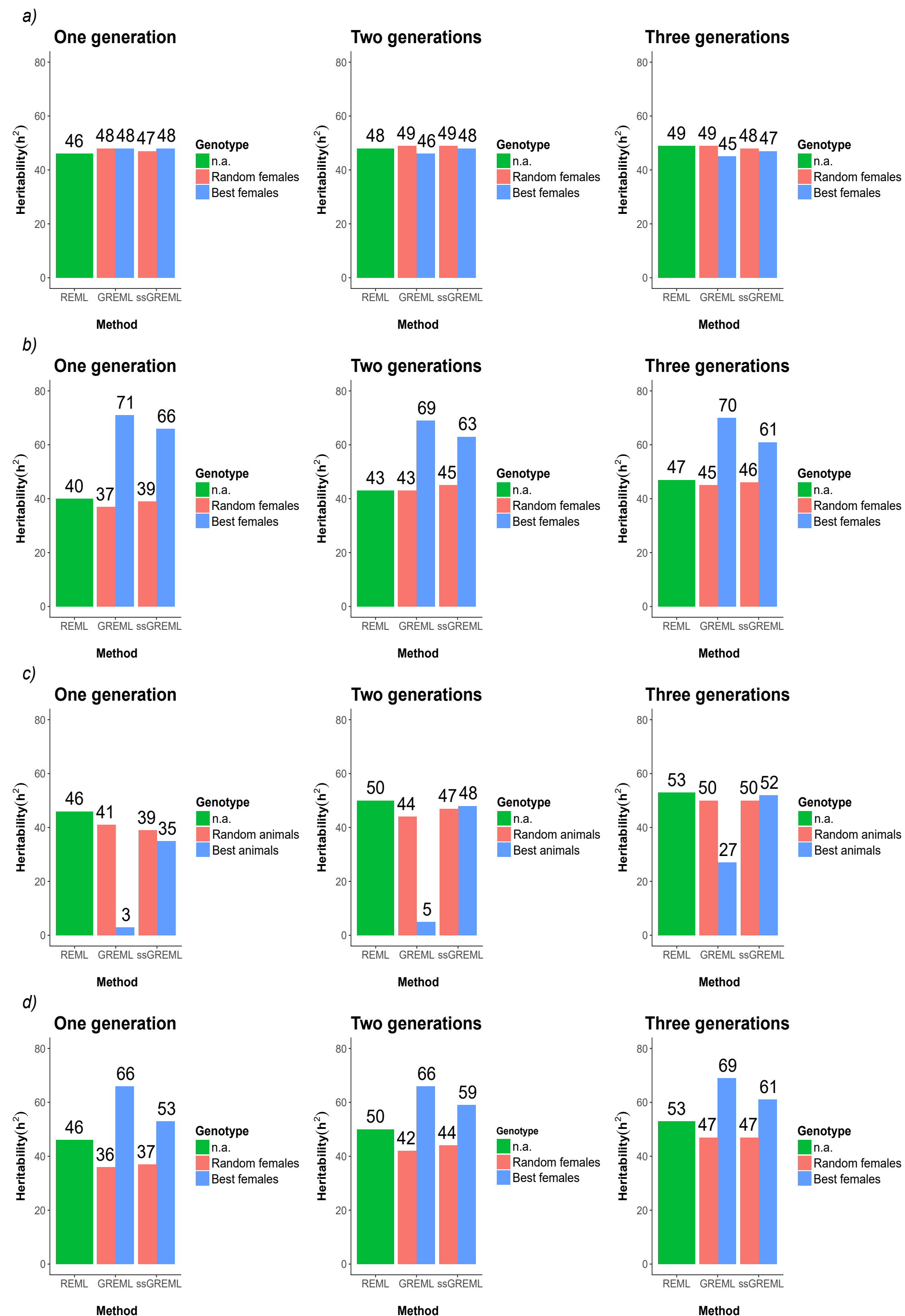


Figure 3. Heritability estimates (h^2=50%) from four simulated scenarios for genotyping strategies of randomly chosen animals or the best animals ranked on EBV: no selection and only female phenotypes (a), selection and only female phenotypes (b), selection and male and female phenotypes and genotypes (c), and selection, male and female phenotypes, and female genotypes (d). For one, two, or three generations, heritability was estimated using restricted maximum likelihood (REML) with only pedigree information, genomic REML (GREML) with only genomic information, and single-step GREML (ssGREML) with combined pedigree and genomic information. Simulated heritability was 0.5.

Discussion

The heritability of the trait is strongly related to the covariance among relatives. In fact, pedigree information is commonly incorporated in breeding schemes as a way to quantify these relationships. However, accurate relationship information is not always available. For instance, in an extensive sheep farming system it is difficult to collect accurate relationship information due to the simultaneous presence of more rams in the same group (Hayes and Goddard, 2008). Moreover, it is well known that pedigrees can contain several errors or partial relationships. Banos *et al*. (2001) showed that the paternity misidentification is common in several animal populations; Visscher *et al*. (2002) estimated 10% overall pedigree error rate in United Kingdom dairy populations; Legarra *et al*. (2014) reported unknown fatherhood of 50% and 20% for Latxa and Manech/Basco-Béarnaise sheep breeds, respectively. Missing or incomplete pedigree information, especially regarding the sire assignment, is a big problem in variance components estimation (Israel and Weller, 2000; Banos *et al*., 2001; Visscher *et al*., 2002). Additionally, in small or autochthonous populations pedigree is not even recorded (Mészáros *et al*., 2015). Starting a breeding program in such populations can be challenging because variance components may be overestimated, depending on the level of incompleteness of the pedigree. In this situation, genomic information could be useful to construct relationships among animals and improve breeds in the absence of complete pedigree (Thompson, 1975; Blouin *et al*., 1996; Ritland, 1996; Goodnight and Queller, 1999; Lynch and Ritland, 1999; Mészáros *et al*., 2015). Hayes and Goddard (2008) showed that heritability estimates using a sufficient number of markers can be more accurate than using pedigree information only for the last generation, because genomic information should

not be subject to pedigree errors. In this context, this paper investigated which method should be used to estimate heritability in populations with limited pedigree information and selective genotyping.

Phenotyping and genotyping strategies as well as statistical methodology affected heritability estimates when genomic information was available for populations that are not structured into large half-sib families. Genotyping randomly selected females, including more generations of genotyped and phenotyped animals, and using ssGREML was in almost all cases the best combination for improving accuracy of heritability estimates.

In the SFM scenario, in which both males and females were genotyped, GREML greatly overestimated heritability when the best animals were genotyped. This is because GREML only uses the sample of genotyped animals, and the highest ranked individuals for EBV are too different from the remaining population because of selection. Using GREML gave the worst estimates across the three different simulations because it does not account for selection. This explanation is supported by NF results; random genotyping or choosing the best females did not result in relevant differences for GREML heritability estimates in the absence of selection (Tables S1-S3). Jenko *et al*. (2017) found similar results for the effect of genotyping strategy on the accuracy of and bias in genomic predictions for Guernsey cattle: best predictions when all cows with phenotypes were genotyped, and worst predictions when only animals with the best phenotypes were genotyped. Pryce *et al*. (2012) suggested that the best strategy for including female genotypes in GS is to select them randomly, because females with the best phenotypes represent a biased sample of the whole population. Similarly, Gao *et al*. (2015) found that adding unselected females to a

reference population improved GEBV reliabilities for Nordic Jersey cattle and reduced prediction bias compared to adding genotypes for just the best animals. Although their results relate to the quality of genomic predictions, quality and unbiasedness of variance components in this study followed the same trend.

As expected, including more generations with genotypes and phenotypes resulted in the best estimates for scenarios that considered selection (SF and SMF). Including more generations reduced standard error and improved accuracy of heritability estimates. This was also observed by Van der Werf and de Boer (1990), who found unbiased additive genetic variance when all available generations were considered. Number of generations also affected genotyping strategy (random versus selective) results because of the different number of animals analyzed. The REML estimates were highly accurate across scenarios when three generations of animals were included in the analysis (Figures 1-3). Reduced heritability estimates from REML were observed across all h^2 levels for scenarios with selection (SF and SMF) and fewer generations, primarily because of reduced additive genetic variance (Tables S1-3). As expected, the number of generations included did not affect heritability estimates in the scenario with no selection (NF) regardless of methodology. When A is used to calculate variance components (REML), many generations can be incorporated in the analysis because pedigrees generally have good depth for most livestock, whereas it is uncommon for many generations to have been genotyped. Usually genotypes are available only for a few recent generations, especially for small breeds (e.g., breeding programs to enhance native breeds) and species with limited financial compensation. In such situations, the genotyped population is not representative of many previous generations, and GS application may be problematic. In this study, GREML estimation

issues can be attributed to this type of population structure (i.e., genotyped animals from only a few recent generations). When three generations where included in the analysis, ssGREML estimates were more accurate overall than those from GREML across all scenarios (Figures 1-3).

In this study, choosing the best animals regardless of sex resulted in equally good estimates using ssGREML but overestimates using GREML (especially for SMF with phenotypes for both males and females). The GREML method showed the worst estimates, especially when extreme genotypes were included in the analysis. When genotypes only from the best females were included in NF, heritability estimates were accurate from all methods, with almost no effect from number of generations included. When the best animals were used as the genotyped population, ssGREML outperformed GREML across all scenarios. Regardless of the number of generations included, GREML overestimated heritability for NF and SMF and underestimated it for SF. When selection was present (SF), ssGREML estimates were close to simulated heritability across generations, whereas using GREML were biased.

The ssGREML method was not as affected by genotyping strategy as the other methods. This is possibly because all generations of pedigree can be included and adjusting averages from **G** to match those from **A22** is a standard procedure in ssGBLUP (Vitezica *et al*., 2011). Such an adjustment can account for selection and also for the fact that animals in **A22** are more related than **G** can express (i.e., averages in **A22** usually are larger than in **G**). Veerkamp *et al*. (2011) showed that ssGREML can be successfully incorporated into variance component estimation. In their study, combination of **G** and **A** into **H** resulted in the most accurate estimates of variance

components and allowed inclusion of phenotypes from non-genotyped animals in the model.

In small populations where the missing data process is represented by selection, amount of generations genotyped, and selective genotyping, a method that accounts for all available pedigree, genotypes, and phenotypes is highly recommended. However, when pedigree is not available, GREML may be the only option for estimating variance components, but bias is expected.

Conclusions

Genomic information can be used for variance component estimation through the inclusion of G or H matrices into the well-known REML methodology. Selecting which method to use when genomic information is available is dependent on selection pressure, number of generations available, and most importantly the genotyping strategy. Worst results were found with selective genotyping and application of GREML methodology. In general, including more than two generations of phenotypes and genotypes improved estimates in most cases. Genotyping strategies affected GREML results more than those from ssGREML. Unlike GREML, ssGREML allows the use of phenotypes and pedigree information for genotyped and non-genotyped animals, which makes the method less biased. Consequently, variance components are less affected by selective or limited genotyping.

Acknowledgements

Editing by Suzanne Hubbard is gratefully acknowledged. This research was supported by the Regional Government of Sardinia (Grant no. CRP 61608 “Il latte Ovino della Sardegna”).

References

Aasmundstad T., Andersen-Ranberg I., Nordbø Ø., Meuwissen T., Vangen O., & Grindflek E. (2015). The effect of including genomic relationships in the estimation of genetic parameters of functional traits in pigs. Journal of Animal Breeding and Genetics, 132, 386–391.

Aguilar I., Misztal I., Johnson D. L., Legarra A., Tsuruta S., & Lawlor T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science, 93, 743–752

Baba T., Gotoh Y., Yamaguchi S., Nakagawa S., Abe H., Masuda Y., & Kawahara T. (2017). Application of single-step genomic best linear unbiased prediction with a multiple-lactation random regression test-day model for Japanese Holsteins. Animal Science Journal, 88, 1226–1231.

Baloche G., Legarra A., Sallé G., Larroque H., Astruc J. M., Robert-Granié C., & Barillet F. (2014). Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. Journal of Dairy Science, 97, 1107–1116.

Banos G., Wiggans G. R., & Powell R. L. (2001). Impact of paternity errors in cow identification on genetic evaluations and international comparisons. Journal of Dairy Science, 84, 2523-2529.

Blouin M. S., Parsons M., Lacaille V., & Lotz S. (1996). Use of microsatellite loci to classify individuals by relatedness. Molecular Ecology, 5, 393-401.

Carta A., Casu S., & Salaris S. (2009). Invited review: Current state of genetic improvement in dairy sheep. Journal of Dairy Science, 92, 5814–5833.

Chen C. Y., Misztal I., Aguilar I., Legarra A., & Muir W. M. (2011a). Effect of different genomic relationship matrices on accuracy and scale. Journal of Animal Science, 89, 2673–2679.

Chen C. Y., Misztal I., Aguilar I., Tsuruta S., Meuwissen T. H. E., Aggrey S. E., Wing T., Muir W. M. (2011b). Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. Journal of Animal Science, 89, 23–28.

Christensen O. F., & Lund M. S. (2010). Genomic prediction when some animals are not genotyped. Genetics Selection Evolution, 42, 2.

Cooper T. A., Wiggans G. R., & VanRaden P. M. (2015). Short communication: Analysis of genomic predictor population for Holstein dairy cattle in the United States—Effects of sex and age. Journal of Dairy Science, 98, 2785–2788.

Duchemin S. I., Colombani C., Legarra A., Baloche G., Larroque H., Astruc J. M., Barillet F., Robert-Granié C., & Manfredi E. (2012). Genomic selection in the French Lacaune dairy sheep breed. Journal of Dairy Science, 95, 2723–2733.

Forni S., Aguilar I., & Misztal I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution, 43, 1.

Gao H., Madsen P., Nielsen U. S., Aamand G. P., Su G., Byskov K., & Jensen J. (2015). Including different groups of genotyped females for genomic prediction in a Nordic Jersey population. Journal of Dairy Science, 98, 9051–9059.

Goddard M. E., & Hayes B. J. (2007). Genomic selection. Journal of Animal Breeding and Genetics, 124, 323–330.

Goodnight K. F., & Queller D. C. (1999). Computer software for performing likelihood tests of pedigree relationship using genetic markers. Molecular Ecology, 8, 1231–1234.

Haile-Mariam M., Nieuwhof G. J., Beard K. T., Konstatinov K. V., & Hayes B. J. (2013). Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. Journal of Animal Breeding and Genetics, 130, 20–31.

Hayes B. J., & Goddard M. E. (2008). Prediction of breeding values using marker-derived relationship matrices. Journal of Animal Science, 86, 2089-2092.

Hayes B. J., Bowman P. J., Chamberlain A. J., & Goddard M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science, 92, 433–443.

Israel C., & Weller J. I. (2000). Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. Journal of Dairy Science, 83, 181-187.

Jenko J., Wiggans G. R., Cooper T. A., Eaglen S. A. E., Luff W. G. de L., Bichard M., Pong-Wong R., Woolliams J. A. (2017). Cow genotyping strategies for genomic selection in a small dairy cattle population. Journal of Dairy Science, 100, 439–452.

Legarra A., Baloche G., Barillet F., Astruc J. M., Soulas C., Aguerre X., Arrese F., Mintegi L., Lasarte M., Maeztu F., Beltrán de Heredia I., & Ugarte E. (2014). Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. Journal of Dairy Science, 97, 3200-3212.

Loberg A., Dürr J. W., Fikse W. F., Jorjani H., & Crooks L. (2015). Estimates of genetic variance and variance of predicted genetic merits using pedigree or genomic relationship matrices in six Brown Swiss cattle populations for different traits. Journal of Animal Breeding and Genetics, 132, 376–385.

Lourenco D. A. L., Fragomeni B. O., Tsuruta S., Aguilar I., Zumbach B., Hawken R. J., Legarra A., Misztal I. (2015a). Accuracy of estimated breeding values with genomic information on males, females, or both: An example on broiler chicken. Genetics Selection Evolution, 47, 56.

Lourenco D. A. L., Tsuruta S., Fragomeni B. O., Masuda Y., Aguilar I., Legarra A., Betrand J. K., Amen T. S., Moser D. W., Misztal I. (2015b). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. Journal of Animal Science, 93, 2653–2662.

Lynch M., & Ritland K. (1999). Estimation of pairwise relatedness with molecular markers. Genetics, 152, 1753–1766.

Mäntysaari E. A., Evans R. D., & Strandén I. (2017). Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. Journal of Animal Science, 95, 4728–4737.

Mészáros G., Boison S. A., Pérez O'Brien A. M., Ferenčaković M., Curik I., Da Silva M. V. B., Utsunomiya Y.T., Garcia J.F., & Sölkner J. (2015). Genomic analysis for managing small and endangered populations: a case study in Tyrol Grey cattle. Frontiers in genetics, 6, 173.

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

Misztal I., Tsuruta S., Lourenco D., Aguilar I., Legarra A., & Vitezica Z. (2015, June 18). Manual for BLUPF90 family of programs. Retrieved from http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf

Moser G., Tier B., Crump R. E., Khatkar M. S., & Raadsma H. W. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution, 41, 56.

Mouresan E. F., Altarriba J., Moreno C., Munilla S., González-Rodríguez A., & Varona L. (2017). Performance of genomic selection under a single-step approach in autochthonous Spanish beef cattle populations. Journal of Animal Breeding and Genetics, 134, 289–299.

Powell J. E., Visscher P. M., & Goddard P. M. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics, 11, 800–805.

Pryce J. E., Hayes B. J., & Goddard M. E. (2012). Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. Retrieved from ICAR 38th session Cork Ireland: https://www.icar.org/wp-content/uploads/2015/09/Pryce-2.pdf

Pryce J. E., Nguyen T. T. T., Axford M., Nieuwhof G., & Shaffer M. (2018). Symposium review: Building a better cow—The Australian experience and future perspectives. Journal of Dairy Science, 101, 3702–3713.

Ritland K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical Research, 67, 175–185.

Sargolzaei M., & Schenkel F. S. (2009). QMSim: A large-scale genome simulator for livestock. Bioinformatics, 25, 680–681.

Simeone R., Misztal I., Aguilar I., & Vitezica Z. G. (2012). Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. Journal of Animal Breeding and Genetics, 129, 3–10.

Solberg T. R., Sonesson A. K., Woolliams J. A., & Meuwissen T. H. E. (2008). Genomic selection using different marker types and densities. Journal of Animal Science, 86, 2447–2454.

Thompson E. A. (1975). The estimation of pairwise relationships. Annals of Human Genetics, 39, 173–188.

Van der Werf J. H., & de Boer I. J. (1990). Estimation of additive genetic variance when base populations are selected. Journal of Animal Science, 68, 3124-3132.

VanRaden P. M. (2008). Efficient methods to compute genomic predictions. Journal of Dairy Science, 91, 4414–4423.

VanRaden P. M., Van Tassell C. P., Wiggans G. R., Sonstegard T. S., Schnabel R. D., Taylor J. F., & Schenkel F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science, 92, 16–24.

Veerkamp R. F., Mulder H. A., Thompson R., & Calus M. P. L. (2011). Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. Journal of Dairy Science, 94, 4189–4197.

Visscher P. M., Woolliams J. A., Smith D., & Williams J. L. (2002). Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. Journal of Dairy Science, 85, 2368–2375.

Vitezica Z. G., Aguilar I., Misztal I., & Legarra A. (2011). Bias in genomic predictions for populations under selection. Genetics Research, 93, 357–366.

Wellmann R., Preuß S., Tholen E., Heinkel J., Wimmers K., & Bennewitz J. (2013). Genomic selection using low density marker panels with application to a sire line in pigs. Genetics Selection Evolution, 45, 28.

Wright S. (1931). Evolution in Mendelian populations. Genetics, 16, 97–159.

Appendix Chapter 2

Genotyping structure is responsible for overestimation of heritability under genomic best linear unbiased predictor (GBLUP) models

Table S1. Variance components and heritability estimates across all scenarios for heritability 0.30

Table S2. Variance components and heritability estimates across all scenarios for heritability 0.10

Table S3. Variance components and heritability estimates across all scenarios for heritability 0.50

Table S1. Variance components and heritability estimates across all scenarios for heritability 0.30

Method	Scenario	Genotyping strategy	Gen	Estimation (mean ± s.e.)		
				σ_a^2	σ_e^2	h^2
REML	NF	n.a.	1	0.27±0.02	0.71±0.02	0.28±0.02
			2	0.30±0.01	0.71±0.02	0.30±0.01
			3	0.30±0.01	0.71±0.01	0.30±0.01
	SF		1	0.19±0.04	0.75±0.02	0.20±0.04
			2	0.23±0.02	0.73±0.01	0.24±0.02
			3	0.28±0.03	0.71±0.01	0.28±0.02
	SMF		1	0.18±0.02	0.75±0.02	0.19±0.02
			2	0.25±0.01	0.71±0.01	0.26±0.01
			3	0.27±0.01	0.71±0.01	0.28±0.01
GREML	NF	Random females	1	0.28±0.02	0.69±0.02	0.29±0.02
			2	0.28±0.01	0.72±0.01	0.28±0.01
			3	0.29±0.01	0.73±0.01	0.28±0.01
		Best females	1	0.32±0.01	0.60±0.02	0.35±0.01
			2	0.31±0.01	0.62±0.02	0.33±0.01
			3	0.30±0.01	0.64±0.01	0.32±0.01
	SF	Random females	1	0.20±0.03	0.72±0.02	0.21±0.03
			2	0.23±0.02	0.70±0.01	0.25±0.02
			3	0.27±0.02	0.70±0.01	0.28±0.02
		Best females	1	0.35±0.07	0.62±0.05	0.36±0.06
			2	0.40±0.03	0.63±0.02	0.39±0.03
			3	0.48±0.03	0.61±0.01	0.44±0.02
	SMF	Random animals	1	0.21±0.03	0.68±0.02	0.23±0.03
			2	0.25±0.02	0.68±0.01	0.26±0.02
			3	0.26±0.01	0.70±0.01	0.27±0.01
		Best animals	1	0.44±0.04	0.49±0.04	0.47±0.04
			2	0.44±0.03	0.54±0.02	0.45±0.03
			3	0.46±0.01	0.57±0.01	0.45±0.01
		Random females	1	0.16±0.01	0.73±0.01	0.17±0.02
			2	0.25±0.01	0.67±0.02	0.27±0.01
			3	0.27±0.01	0.69±0.01	0.28±0.01
		Best females	1	0.18±0.05	0.74±0.05	0.19±0.05
			2	0.26±0.01	0.69±0.01	0.28±0.01
			3	0.35±0.01	0.65±0.01	0.35±0.01

ssGREML	NF	Random females	1	0.27±0.01	0.71±0.02	0.28±0.01
			2	0.28±0.01	0.73±0.02	0.28±0.01
			3	0.28±0.02	0.72±0.01	0.28±0.01
		Best females	1	0.27±0.02	0.71±0.02	0.28±0.02
			2	0.28±0.01	0.72±0.02	0.28±0.01
			3	0.28±0.01	0.72±0.01	0.28±0.01
	SF	Random females	1	0.21±0.03	0.73±0.01	0.22±0.03
			2	0.25±0.02	0.72±0.01	0.25±0.02
			3	0.28±0.02	0.71±0.01	0.28±0.02
		Best females	1	0.30±0.04	0.68±0.02	0.30±0.04
			2	0.32±0.02	0.68±0.01	0.32±0.01
			3	0.36±0.02	0.67±0.01	0.35±0.01
	SMF	Random animals	1	0.24±0.01	0.71±0.02	0.25±0.01
			2	0.26±0.01	0.70±0.01	0.27±0.01
			3	0.28±0.01	0.70±0.01	0.28±0.01
		Best animals	1	0.31±0.03	0.67±0.03	0.32±0.03
			2	0.31±0.01	0.67±0.01	0.32±0.01
			3	0.30±0.01	0.67±0.01	0.31±0.01
		Random females	1	0.20±0.02	0.73±0.01	0.21±0.02
			2	0.27±0.01	0.73±0.01	0.27±0.01
			3	0.26±0.01	0.71±0.01	0.27±0.01
		Best females	1	0.21±0.03	0.71±0.02	0.23±0.03
			2	0.29±0.01	0.70±0.01	0.29±0.01
			3	0.30±0.01	0.69±0.01	0.30±0.01

Table S2. Variance components and heritability estimates across all scenarios for heritability 0.10

Method	Scenario	Genotyping strategy	Gen	Estimation (mean ± s.e.)		
				σ_a^2	σ_e^2	h²
REML	NF	n.a.	1	0.10±0.02	0.90±0.01	0.09±0.02
			2	0.11±0.01	0.89±0.01	0.11±0.01
			3	0.12±0.02	0.88±0.01	0.11±0.02
	SF		1	0.08±0.03	0.89±0.01	0.08±0.02
			2	0.09±0.02	0.87±0.01	0.09±0.02
			3	0.11±0.01	0.89±0.01	0.10±0.01
	SMF		1	0.07±0.01	0.93±0.02	0.07±0.01
			2	0.09±0.01	0.91±0.02	0.08±0.01
			3	0.11±0.01	0.89±0.02	0.11±0.01
GREML	NF	Random females	1	0.09±0.03	0.87±0.02	0.09±0.02
			2	0.09±0.02	0.90±0.01	0.09±0.02
			3	0.08±0.01	0.91±0.01	0.08±0.01
		Best females	1	0.08±0.04	0.89±0.03	0.08±0.03
			2	0.09±0.02	0.88±0.01	0.09±0.01
			3	0.09±0.01	0.90±0.01	0.09±0.01
	SF	Random females	1	0.06±0.02	0.91±0.02	0.06±0.02
			2	0.09±0.01	0.87±0.01	0.09±0.01
			3	0.09±0.01	0.89±0.01	0.09±0.01
		Best females	1	0.25±0.02	0.75±0.02	0.21±0.02
			2	0.21±0.01	0.80±0.01	0.19±0.01
			3	0.09±0.02	0.87±0.01	0.09±0.02
	SMF	Random animals	1	0.08±0.02	0.92±0.02	0.08±0.01
			2	0.09±0.02	0.87±0.02	0.09±0.02
			3	0.11±0.01	0.88±0.02	0.11±0.01
		Best animals	1	0.39±0.01	0.53±0.01	0.34±0.01
			2	0.36±0.02	0.58±0.01	0.30±0.02
			3	0.29±0.01	0.52±0.01	0.28±0.02
		Random females	1	0.06±0.02	0.91±0.02	0.06±0.02
			2	0.09±0.01	0.90±0.02	0.09±0.01
			3	0.09±0.01	0.90±0.01	0.09±0.01

		Best females	1	0.25±0.02	0.64±0.03	0.23±0.02
			2	0.16±0.01	0.69±0.02	0.16±0.01
			3	0.15±0.01	0.90±0.02	0.13±0.02
ssGREML	NF	Random females	1	0.09±0.03	0.87±0.02	0.09±0.02
			2	0.10±0.02	0.90±0.01	0.10±0.01
			3	0.10±0.01	0.90±0.02	0.10±0.01
		Best females	1	0.12±0.01	0.89±0.01	0.11±0.01
			2	0.09±0.02	0.87±0.01	0.09±0.02
			3	0.11±0.01	0.90±0.01	0.10±0.01
	SF	Random females	1	0.07±0.02	0.90±0.01	0.07±0.02
			2	0.09±0.01	0.88±0.01	0.09±0.01
			3	0.10±0.01	0.90±0.01	0.10±0.01
		Best females	1	0.08±0.03	0.90±0.01	0.08±0.03
			2	0.10±0.02	0.88±0.01	0.10±0.02
			3	0.11±0.01	0.89±0.01	0.10±0.01
	SMF	Random animals	1	0.06±0.01	0.94±0.02	0.06±0.01
			2	0.09±0.01	0.87±0.02	0.09±0.01
			3	0.11±0.01	0.89±0.01	0.11±0.01
		Best animals	1	0.07±0.02	0.93±0.03	0.07±0.02
			2	0.10±0.01	0.90±0.02	0.10±0.01
			3	0.11±0.01	0.89±0.02	0.10±0.01
		Random females	1	0.06±0.01	0.92±0.01	0.06±0.01
			2	0.09±0.01	0.90±0.01	0.09±0.01
			3	0.11±0.01	0.90±0.01	0.10±0.01
		Best females	1	0.07±0.01	0.92±0.02	0.07±0.01
			2	0.14±0.01	0.89±0.02	0.12±0.01
			3	0.12±0.01	0.89±0.01	0.11±0.01

Table S3. Variance components and heritability estimates across all scenarios for heritability 0.50

Method	Scenario	Genotyping strategy	Gen	Estimation (mean ± s.e.)		
				σ_a^2	σ_e^2	h^2
REML	NF	n.a.	1	0.46±0.01	0.55±0.03	0.46±0.02
			2	0.48±0.01	0.52±0.01	0.48±0.01
			3	0.49±0.01	0.51±0.01	0.49±0.01
	SF		1	0.37±0.03	0.56±0.02	0.40±0.02
			2	0.40±0.02	0.53±0.01	0.43±0.01
			3	0.45±0.01	0.50±0.01	0.47±0.01
	SMF		1	0.45±0.01	0.54±0.03	0.46±0.01
			2	0.47±0.04	0.47±0.02	0.50±0.03
			3	0.50±0.02	0.45±0.01	0.53±0.02
GREML	NF	Random females	1	0.47±0.03	0.51±0.03	0.48±0.03
			2	0.49±0.02	0.51±0.02	0.49±0.02
			3	0.48±0.01	0.51±0.01	0.49±0.01
		Best females	1	0.43±0.01	0.47±0.02	0.48±0.01
			2	0.40±0.01	0.47±0.01	0.46±0.01
			3	0.39±0.01	0.47±0.01	0.45±0.01
	SF	Random females	1	0.31±0.03	0.53±0.02	0.37±0.03
			2	0.38±0.04	0.50±0.01	0.43±0.03
			3	0.41±0.02	0.50±0.01	0.45±0.02
		Best females	1	0.71±0.04	0.28±0.02	0.71±0.02
			2	0.75±0.04	0.33±0.01	0.69±0.02
			3	0.80±0.03	0.34±0.01	0.70±0.01
	SMF	Random animals	1	0.34±0.03	0.50±0.04	0.41±0.04
			2	0.39±0.02	0.49±0.01	0.44±0.02
			3	0.47±0.03	0.48±0.01	0.50±0.02
		Best animals	1	0.03±0.01	0.84±0.01	0.03±0.02
			2	0.05±0.01	0.83±0.01	0.05±0.01
			3	0.27±0.04	0.73±0.01	0.27±0.03
		Random females	1	0.31±0.03	0.54±0.02	0.36±0.03
			2	0.37±0.03	0.51±0.01	0.42±0.03
			3	0.44±0.03	0.50±0.01	0.47±0.02
		Best females	1	0.64±0.04	0.33±0.03	0.66±0.04
			2	0.69±0.03	0.36±0.01	0.66±0.02
			3	0.78±0.03	0.35±0.01	0.69±0.01

ssGREML	NF	Random females	1	0.48±0.03	0.53±0.03	0.47±0.03
			2	0.49±0.02	0.52±0.01	0.49±0.01
			3	0.48±0.01	0.52±0.01	0.48±0.01
		Best females	1	0.48±0.02	0.52±0.02	0.48±0.02
			2	0.48±0.02	0.52±0.01	0.48±0.01
			3	0.47±0.01	0.52±0.01	0.47±0.01
	SF	Random females	1	0.35±0.04	0.54±0.02	0.39±0.03
			2	0.42±0.04	0.51±0.01	0.45±0.03
			3	0.44±0.03	0.51±0.01	0.46±0.02
		Best females	1	0.70±0.03	0.37±0.02	0.66±0.02
			2	0.69±0.05	0.40±0.01	0.63±0.03
			3	0.67±0.03	0.43±0.01	0.61±0.01
	SMF	Random animals	1	0.34±0.04	0.54±0.02	0.39±0.04
			2	0.44±0.03	0.49±0.01	0.47±0.02
			3	0.47±0.02	0.48±0.01	0.50±0.02
		Best animals	1	0.31±0.06	0.57±0.03	0.35±0.05
			2	0.46±0.03	0.49±0.02	0.48±0.03
			3	0.52±0.03	0.47±0.01	0.52±0.02
		Random females	1	0.33±0.03	0.55±0.02	0.37±0.03
			2	0.41±0.03	0.51±0.01	0.44±0.02
			3	0.45±0.02	0.50±0.01	0.47±0.02
		Best females	1	0.52±0.06	0.45±0.03	0.53±0.04
			2	0.60±0.03	0.43±0.01	0.59±0.02
			3	0.65±0.03	0.42±0.01	0.61±0.01

CHAPTER 3

Genomic selection of milk fatty acid composition in Sarda dairy sheep: effect of different phenotypes and relationship matrices on heritability and breeding values accuracy

This is the accepted version in **Journal of Dairy Science** of:

Alberto Cesarani, Giustino Gaspa, Fabio Correddu, Massimo Cellesi, Corrado Dimauro, and Nicolo Macciotta. 2019. Genomic selection of milk fatty acid composition in Sarda dairy sheep: effect of different phenotypes and relationship matrices on heritability and breeding values accuracy. Journal of Dairy Science, 102, doi: https://doi.org/10.3168/jds.2018-15333.

This version is used by permission of the **Journal of Dairy Science**.

Abstract

Fatty acid (FA) composition is one of the most important aspects of milk nutritional quality. However, the inclusion of this trait as breeding goal for dairy species is hampered by the logistics and high costs of phenotype recording. Fourier transform Infrared Spectroscopy (FTIR) is a valid and cheap alternative to laboratory gas chromatography (GC) for predicting milk FA composition. Moreover, as for other novel phenotypes, the efficiency of selection for these traits can be enhanced by using genomic data. Objective of this research was to compare traditional versus genomic selection approaches for estimating genetic parameters and breeding values of milk fatty acid composition in dairy sheep using either GC measured or FTIR predicted FA as phenotypes. Milk FA profiles were available for a total of 923 Sarda breed ewes. The youngest 100 had their own phenotype masked to mimic selection candidates. Pedigree relationship information and genotypes were available for 923 and 769 ewes, respectively. Three statistical approaches were used: the classical pedigree based BLUP; the GBLUP that considers the genomic relationship matrix **G**; the single step GBLUP (ssGBLUP) where pedigree and genomic relationship matrices are blended into a single H matrix. Heritability estimates using pedigree were lower than ssGBLUP, and very similar between GC and FTIR regarding the statistical approach used. For some FA, mostly associated with animal diet (i.e. C18:2ω6, C18:3ω3), random effect of combination of flock and test date (FTD) explained a relevant quota of total variance, reducing accordingly h^2 estimates. Genomic approaches (GBLUP and ssGBLUP) outperformed the traditional pedigree method both for GC and FTIR FA. Prediction accuracies in older cohort were larger than young cohort. Genomic prediction accuracy (obtained using either **G** or **H** relationship matrix) in young cohort

of animals, where their own phenotype were masked, were similar for GC and FTIR. These results suggest that FTIR predicted milk FA composition could represent a valid option for the inclusion of this trait in breeding programs.

Keywords

Mid infrared spectra, REML, FTIR, genomic selection

Introduction

Dairy sheep breeding programs have been historically aimed at improving total milk yield per lactation (Carta *et al*., 2009). Although sheep milk is almost totally destined to cheese making (Pulina *et al*., 2018), selection for milk composition is carried out only in few breeds (Macciotta *et al*., 2005; Astruc *et al*., 2008). This is mostly because of the high recording costs compared to the income per ewe (Carta *et al*., 2009; Rupp *et al*., 2016). On the other hand, the increasing consumer interest on dairy product nutritional quality pushes toward the inclusion of fine milk composition traits among breeding goals of dairy species. An example is represented by the conjugated linoleic acid (CLA), known for its relationships with human health (Banni *et al*., 2003; Bhattacharya *et al*., 2006; Mele *et al*., 2011). Ruminant dairy products are among the most important sources of CLA in human diets (Nudda *et al*., 2014). Although animal feeding is considered the most important factor affecting milk fatty acid (FA) composition (Cabiddu *et al*., 2005; Sanchez *et al*., 2010), genetic variation for these traits has been reported in cattle (Stoop *et al*., 2008; Pegolo *et al*., 2016) and sheep (Sanchez *et al*. 2010; Correddu *et al*. 2018) suggesting the possibility for a genetic improvement.

The inclusion of milk FA composition as breeding goal for dairy sheep programs is constrained by logistics and costs of phenotype recording. The standard method for measuring milk FA composition is the gas chromatography (GC) analysis, that is expensive and time consuming. A population-scale recording of milk FA appears therefore rather unfeasible for species where also the routine phenotyping of milk components is economically unbearable. A valid alternative to GC is represented by Fourier transform Infrared (FTIR) spectroscopy. This technique, implemented in milk

lab equipment currently used for routine milk composition analysis, produces a spectrum of approximately one thousand variables that could be used for large scale prediction of novel phenotypes, including FA (e.g. Cecchinato *et al*., 2009; De Marchi *et al* 2011; McParland *et al*., 2011; Dehareng *et al*., 2012; Fleming *et al*., 2016). Good prediction accuracies of milk FA based on FTIR spectrum have been reported for dairy cattle (Arnould and Soyeurt, 2009; De Marchi *et al*., 2011). Similar results, even though with a certain degree of variability and in a limited number of studies, have been reported for dairy sheep (Ferrand-Calmels *et al*., 2014; Caredda *et al*. 2016; Correddu *et al*., 2018). Fatty acid predicted by FTIR exhibited genetic variation both in dairy cattle (e.g. Soyeurt *et al*., 2006; Bastin *et al*., 2013; Narayana *et al*., 2017) and sheep (Sanchez *et al*., 2010; Boichard *et al*., 2014). Moreover, genetic correlations ranging from 60% to 99% between FTIR predicted and GC measured milk FA have been reported both in cattle (Bonfatti *et al*., 2017) and sheep (Correddu *et al*., 2018).

Dairy sheep breeding programs are based on the classical quantitative genetic approach, with a pyramidal organization of the population, large scale registration of phenotypes and pedigree, and genetic evaluations of AI rams based on progeny testing (Carta *et al*., 2009; Baloche *et al*., 2014). The availability of high throughput SNP panel for sheep has opened the perspective of genomic selection (GS) also for this species. Researches have been carried out on dairy (Duchemin *et al*., 2012; Baloche *et al*., 2014), meat, and wool sheep (Daetwyler *et al*., 2012). An improvement of genomic breeding value (GEBV) accuracies over the traditional pedigree index has generally been observed, even though to a lesser extent compared to dairy cattle (Legarra *et al*., 2014).

Genomic studies on milk FA in cattle have focused mostly on the study of their genetic determinism (Stoop *et al*., 2009; Bouwman *et al*. 2011; Buitenhuis *et al*., 2014). In dairy sheep, the molecular basis of FA has been investigated by candidate gene (Crisà *et al*, 2010; Moioli *et al*., 2012), and QTL detection (Carta *et al*., 2008) approaches. Genomic selection studies for FA compositions are limited to beef cattle (Uemoto *et al*., 2011; Chen *et al*., 2015; Zhu *et al*. 2017) and meat sheep (Rovadoscki *et al*., 2018). One of the main advantange of GS over traditional selection is that, once a reference population with both phenotypic and genotypic records has been settled, breeding values of animals without their own phenotypes can be predicted with a reasonable accuracy (Meuwissen *et al*., 2001; Hayes *et al*., 2009). Therefore, GS seems to be an appealing option for novel traits that are difficult to measure routinely as milk FA composition (Boichard and Brochard, 2012; Daetwyler *et al*., 2012).

Aim of the present work is to explore the feasibility of breeding for milk FA composition in a dairy sheep breed by combining the use of FTIR predicted phenotypes and the genomic selection technology. At this purpose genetic parameters estimation and breeding values prediction were carried out running a pedigree based and two genomic models, using either FTIR predicted and GC measured FA as phenotypes.

Materials and methods

Data

A sample of 923 Sarda breed dairy ewes farmed in 47 flocks located in the island of Sardinia (Italy) were considered. Milk samples, one per animal, were collected from February to June 2015 (Table 1).

Table 1. Flock statistics and distribution of records for fixed effects considered in the analysis

Observations	n	%
Flocks	47	
Ewes/flock	19.6±7.2	
Parity		
1	186	20
2	123	13
3	151	16
4	164	18
5	116	13
6	95	10
7	68	7
>7	20	2
Lambing Month		
Jan	142	15
Feb-Mar	130	14
Oct-Nov	377	41
Dec	274	30
Altitude		
Mountain (>500 m)	135	15
Hill (200-500 m)	480	52
Plain (<200 m)	308	33

In this study 13 individuals FA (C4:0, C6:0, C8:0, C10:0, C12:0, C14:0, C16:0, C18:0, C18:1t11, C18:1c9, C18:2ω6, C18:3ω3, CLAc9,t11), 5 groups of FA and a ratio between groups of FA were analyzed. Groups of FA were calculated as follow (Table S1): SFA, sum of individual saturated fatty acids; MUFA, sum of individual monounsaturated fatty acids; PUFA, sum of individual polyunsaturated fatty acids; TFA-VA, sum of individual trans FA with the exclusion of C18:1t11 (vaccenic acid); Denovo, sum of individual FA that are de novo synthesized in the mammary gland; PUFA ω6:PUFA ω3, ratio between the sum of individual PUFA ω6 and the sum of all individual PUFA ω3. Milk FA (g FA/100 g total FA) composition was both measured

by gas chromatography (FA_GC) and predicted by partial least square regression (PLS) using the FTIR spectra (FA_FTIR) generated by milk analysis performed with Milkoscan FT6000 instrument (Foss, Hillerød, Denmark). PLS was carried out by extracting 18 latent factors. Prediction accuracies were tested by using a calibration data set of 700 ewes and a validation data set of 223 ewes, respectively. One-hundred replicates randomly assigning animals to the two data sets were performed. Details for GC analysis are reported in the work of Correddu *et al*. (2018).

Genotypes obtained with the Infinium Ovine SNP50 v1 BeadChip (Illumina Inc., San Diego, California) were available for 769 ewes out of 923. Quality control of SNP genotypes was carried out with PLINK software (Purcell *et al*., 2007). All genotyped ewes had a call rate greater than 0.95. A SNP was discharged if: the call rate was lower than 0.975 (867 markers removed), the minor allele frequency (MAF) was lower than 0.01 (1,309 markers removed), it deviated significantly from the Hardy Weinberg Equilibrium ($P < 0.01$, 1,264 markers removed), or it did not map to the OAR_v3.1 assembly (6,182 markers removed). After quality control, all genotyped ewes and 44,619 SNPs across 27 chromosomes were retained for the analysis. A pedigree with 633,317 animals was also available.

Variance component estimation

Variance components for FA_GC and FA_FTIR traits were estimated by restricted maximum likelihood (REML) using three mixed linear models that differed in the relationship matrix used.

The following mixed linear model was implemented:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Qf} + \mathbf{Za} + \mathbf{e}$$

where **y** is the vector of investigated FA; **X** is the incidence matrix linking records to fixed effects and **b** the related vector; **Q** is the incidence matrix for random flock test-date combination (**FTD**) effect and **f** the related vector (71 classes) distributed as N(0, Iσ^2_{FTD}) where **I** is an identity matrix and σ^2_{FTD} is the associated variance component; **Z** is the incidence matrix for random genetic effects, relating records to animals and **a** is the vector of breeding values (**a** distributed according to the relationship matrix used); **e** is the vector of random residuals distributed as **N**(0, **I**σ^2_e) where σ^2_e is the residual variance. The fixed effects (Table 1) considered in the model were: parity (8 classes), days in milk (5 classes), lambing month (4 classes), altitude of farm (3 classes).

The additive genetic effect was modelled using three genetic (co)variance structures. In the first model (**ABLUP**), the pedigree relationship matrix (**A**) was used and the animal effect was distributed as N(0, **A**σ^2_a) where σ^2_a is the additive genetic variance. The other two genomic models used the genomic relationship matrix (**G**) (**GBLUP**) or a blend of genomic and pedigree relationship matrices (**H**) in a single-step framework (**ssGBLUP**) with **a** distributed as N(0, **G**σ^2_a) and N(0, **H**σ^2_a), respectively. From whole pedigree, three generations were traced back from the phenotyped animals; the composition and number of animals of the different relationship matrices are reported in Table 2.

Table 2. Type of relationship matrices used and number of animals for the three (co)variance structures

	Matrix		
Animals	**A**	**G**	**H**
With genotypes and own phenotypes	769	769	769
Without genotypes and with own phenotypes	154	-	154
Other relatives without phenotype	3,924	-	3,924
Total number of animals	4,847	769	4,847

G and **H** matrices were computed according to VanRaden (2008) and Aguilar *et al.* (2010), respectively. AIREML algorithm implemented in blupf90 family software was used for estimating variance components (Mistzal *et al.*, 2015). Heritability (h^2) and intra-flock heritability (h^2_{IF}) were computed respectively as:

$$h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_{FTD}^2 + \sigma_e^2)$$

$$h_{IF}^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2);$$

moreover, variance explained by FTD (r_{FTD}^2) was computed as:

$$r_{FTD}^2 = \sigma_{FTD}^2/(\sigma_a^2 + \sigma_{FTD}^2 + \sigma_e^2)$$

Breeding Value Predictions

Breeding values were predicted using the above-mentioned model with the traditional (ABLUP) and the two GS (GBLUP and ssGBLUP) approaches, respectively. From the 769 animals with genotypes and own phenotypes, records of the 100 youngest ewes (born after November 2012) were masked in order to mimic the condition of candidate animals.

Accuracy of breeding values animals were estimated as:

$$accuracy = \sqrt{1 - SEP^2/\sigma_a^2}$$

where SEP is the standard error of prediction, derived from the diagonal element of the LHS inverse of the mixed model equations. In order to ensure a fair comparison among accuracies obtained in the three different methods, the same variance components (the ones estimated with ABLUP) were used in the three approaches for breeding values predictions and computation of accuracy.

Moreover, in order to reduce GEBV bias in the ssGBLUP, a weighing factor omega (ω) equal to 0.95 was applied in construction of the inverse of the **H** matrix (Tsuruta *et al*., 2013):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \omega\, \mathbf{A}_{22}^{-1} \end{bmatrix}$$

where $\mathbf{A}_{22}$ is the pedigree-based relationship matrix for genotyped animals

Results

Basic statistics (Table 3) of the milk FA_GC and FA_FTIR, and coefficients of determination of the regression between FA_GC and FA_FTIR (R2GC-FTIR) essentially confirm previous reports on dairy sheep (Ferrand-Calmels *et al*., 2014; Caredda *et al*., 2016; Correddu *et al*., 2018).

Table 3. Descriptive statistics of fatty acids measured using gas chromatography (FA_GC) or predicted using Fourier Transformed Infrared spectrum (FA_FTIR) and coefficients of determination ($R^2_{CG-FTIR}$).

Fatty Acid	Trait	FA_GC Mean	FA_GC SD	FA_FTIR Mean	FA_FTIR SD	$R^2_{CG-FTIR}$
Butyric acid	C4:0	2.68	0.37	2.67	0.34	0.79
Caproic acid	C6:0	1.76	0.36	1.76	0.34	0.87
Caprylic acid	C8:0	1.61	0.45	1.60	0.43	0.89
Capric acid	C10:0	5.55	1.73	5.53	1.67	0.91
Lauric acid	C12:0	3.50	0.99	3.49	0.94	0.87
Myristic acid	C14:0	10.85	1.52	10.83	1.39	0.79
Palmitic acid	C16:0	25.97	2.95	25.97	2.58	0.69
Stearic acid	C18:0	10.24	2.49	10.25	2.20	0.72
Vaccenic acid (VA)	C18:1t11	2.06	1.04	2.05	0.92	0.75
Oleic acid	C18:1c9	17.14	3.58	17.20	3.34	0.85
Linoleic acid	C18:2ω6	2.09	0.50	2.09	0.40	0.51
α-Linolenic acid	C18:3ω3	0.89	0.50	0.89	0.43	0.68
Conjugated linoleic acid	CLAc9,t11	1.03	0.47	1.03	0.41	0.72
Saturated fatty acids	SFA	67.72	3.88	67.67	3.60	0.82
Monounsaturated	MUFA	25.83	3.58	25.88	3.29	0.81
Polyunsaturated	PUFA	6.44	1.45	6.44	1.32	0.79
PUFA ω6:PUFA ω3	ω6:ω3	2.47	1.15	2.48	1.01	0.70
Trans fatty acids - VA	TFAnoVA	4.56	1.52	4.55	1.35	0.77
de novo synthesized FA[1]	Denovo[1]	23.56	4.62	23.74	4.30	0.90

[1] Denovo = C6:0, C8:0, C10:0, C11:0, C12:0, iso-C13:0, C14:0 that are de novo synthesized in the mammary gland.

Genetic Parameters of Milk Fatty Acid profile

Heritability estimates showed relevant variations across different FA, phenotyping methods (GC vs FTIR), and models (Table 4). Overall, low to moderate values were

obtained, apart from C4:0 and C16:0. Largest heritabilities were observed for the C4:0 FA_FTIR in the GBLUP (0.56), and for the C16:0 FA_GC in the ABLUP (0.46) (Table 4), respectively. A similar pattern was detected for intra-flock heritabilities (Table 5), that exhibited larger values compared to h^2, especially for FA characterized by a larger flock-test date variance (Table 6) (e.g. C18:0, C18:1t11, C18:1c9, C18:2ω6, C18:3ω3, CLAc9,t11 and ω6:ω3). Lowest estimates (nearly zero) were obtained for SFA and MUFA in the ABLUP, and for C18:2ω6 in all the three prediction models for FA_FTIR.

The considered phenotype, FA_GC or FA_FTIR, affected the h^2 results, even though no defined patterns were observed. For example, FA_GC estimates were markedly larger than FA_FTIR for C16:0 in all models (Table 4). On the contrary, FA_GC estimates were smaller for C4:0, especially for the two genomic models. It should be also noticed that the h^2 estimated with ABLUP were close to zero for SFA and MUFA using FA_FTIR phenotypes. In order to highlight recurrent pattern in the additive genetic component, σ_a^2 for FA_GC was regressed onto σ_a^2 for FA_FTIR (Figure 1) for the three models used.

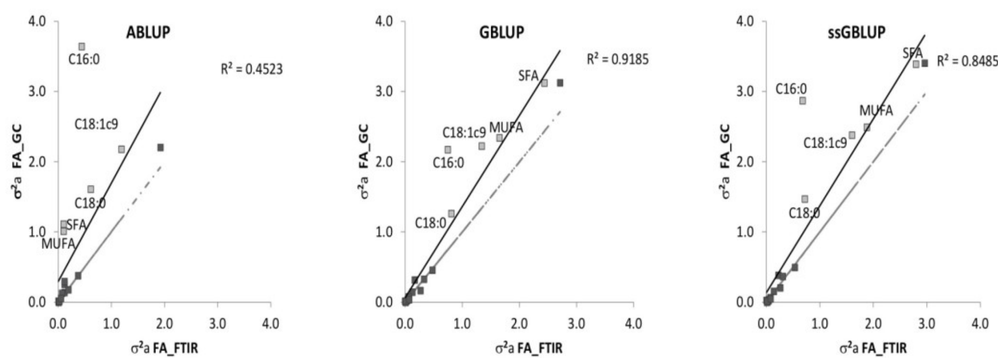


Figure 1. Regressions of additive genetic variance estimated using fatty acids measured through gas chromatography (FA_FC) and fatty acids predicted by Fourier Transform Infrared Spectra (FA_FTIR) within each investigated method: pedigree relationship matrix (ABLUP), genomic relationship matrix (GBLUP), blended genomic-pedigree matrix (ssGBLUP). Dashed line represents the equivalent line (y=x).

Additive genetic variances estimated using FA_GC and FA_FTIR were from moderately to strongly correlated depending on (co)variance matrix used.

The h^2 and h^2_{IF} estimated with ABLUP were generally lower than those obtained with the two genomic approaches, both for FA_GC and FA_FTIR (Tables 4 and 5). Exceptions were the C16:0 and C18:0, that showed an opposite behavior. In particular, largest differences were found for C4:0 and C16:0 as individual FA, and for SFA and MUFA as groups, respectively. GBLUP and ssGBLUP estimates were very similar (Tables 4 and 5).

Table 4. Heritability (h^2) for milk fatty acid composition measured by gas chromatography (FA_GC) or predicted by Fourier Transform Infrared Spectra (FA_FTIR) using pedigree relationship matrix (ABLUP), genomic relationship matrix (GBLUP), blended genomic-pedigree matrix (ssGBLUP), respectively. SE of heritability were reported in brackets.

	Ablup		Gblup		ssGblup	
Trait	FA_GC	FA_FTIR	FA_GC	FA_FTIR	FA_GC	FA_FTIR
C4:0	0.22 (.10)	0.27 (.11)	0.36 (.09)	0.56 (.10)	0.34 (.09)	0.49 (.10)
C6:0	0.04 (.06)	0.12 (.07)	0.16 (.06)	0.23 (.06)	0.17 (.06)	0.25 (.06)
C8:0	0.10 (.06)	0.12 (.06)	0.16 (.06)	0.20 (.06)	0.17 (.06)	0.22 (.06)
C10:0	0.13 (.06)	0.14 (.06)	0.16 (.07)	0.18 (.06)	0.17 (.06)	0.19 (.06)
C12:0	0.15 (.07)	0.15 (.07)	0.16 (.07)	0.16 (.06)	0.17 (.06)	0.17 (.06)
C14:0	0.12 (.09)	0.07 (.08)	0.15 (.08)	0.10 (.07)	0.19 (.08)	0.12 (.07)
C16:0	0.46 (.11)	0.07 (.07)	0.26 (.08)	0.12 (.07)	0.35 (.09)	0.11 (.07)
C18:0	0.29 (.10)	0.14 (.08)	0.23 (.08)	0.19 (.07)	0.26 (.08)	0.16 (.07)
C18:1t11	0.14 (.06)	0.09 (.05)	0.09 (.05)	0.08 (.00)	0.07 (.05)	0.09 (.04)
C18:1c9	0.17 (.07)	0.10 (.06)	0.17 (.06)	0.12 (.07)	0.18 (.06)	0.14 (.05)
C18:2ω6	0.07 (.06)	0.00 (.00)	0.08 (.06)	0.00 (.00)	0.12 (.06)	0.00 (.00)
C18:3ω3	0.03 (.02)	0.03 (.04)	0.01 (.01)	0.07 (.04)	0.02 (.02)	0.08 (.04)
CLAc9,t11	0.12 (.06)	0.13 (.06)	0.10 (.06)	0.09 (.05)	0.08 (.06)	0.10 (.05)
SFA[1]	0.07 (.09)	0.01 (.08)	0.20 (.08)	0.18 (.08)	0.22 (.08)	0.20 (.08)
MUFA[2]	0.08 (.07)	0.01 (.07)	0.18 (.07)	0.15 (.07)	0.19 (.07)	0.17 (.07)
PUFA[3]	0.09 (.05)	0.11 (.07)	0.08 (.05)	0.15 (.06)	0.10 (.05)	0.14 (.06)
ω6:ω3[4]	0.05 (.02)	0.05 (.03)	0.04 (.02)	0.08 (.03)	0.04 (.02)	0.08 (.03)
TFAnoVA[5]	0.14 (.07)	0.06 (.06)	0.15 (.06)	0.18 (.06)	0.16 (.06)	0.17 (.06)
Denovo[6]	0.11 (.07)	0.11 (.07)	0.15 (.06)	0.15 (.06)	0.16 (.06)	0.16 (.06)

[1]Sum of the individual saturated fatty acids.
[2]Sum of the individual monounsaturated fatty acids.
[3]Sum of the individual polyunsaturated fatty acids; odd- and branched-chain fatty acids.
[4]Ratio between the sum of individual PUFA ω6 fatty acids and the sum of individual PUFA ω3 fatty acids.
[5]Trans Fatty Acid (TFA) without Vaccenic acid (VA).
[6]Sum of C6:0, C8:0, C10:0, C11:0, C12:0, iso-C13:0, C14:0 that are *de novo* synthesized in the mammary gland.

Table 5. Intra-Flock heritability (h^2_{IF}) for milk fatty acid composition measured by gas chromatography (FA_GC) or predicted by Fourier Transform Infrared Spectra (FA_FTIR) using pedigree relationship matrix (ABLUP), genomic relationship matrix (GBLUP), blended genomic-pedigree matrix (ssGBLUP), respectively. SE of h^2_{IF} were reported in brackets.

	Ablup		Gblup		ssGblup	
Trait	FA_GC	FA_FTIR	FA_GC	FA_FTIR	FA_GC	FA_FTIR
C4:0	0.28 (.12)	0.34 (.13)	0.45 (.11)	0.68 (.11)	0.42 (.11)	0.59 (.11)
C6:0	0.09 (.14)	0.29 (.15)	0.38 (.13)	0.55 (.12)	0.40 (.12)	0.58 (.11)
C8:0	0.25 (.15)	0.30 (.15)	0.41 (.13)	0.52 (.12)	0.43 (.12)	0.55 (.12)
C10:0	0.31 (.14)	0.34 (.15)	0.38 (.13)	0.45 (.12)	0.41 (.12)	0.48 (.12)
C12:0	0.29 (.14)	0.32 (.14)	0.33 (.12)	0.35 (.12)	0.33 (.12)	0.36 (.12)
C14:0	0.19 (.14)	0.11 (.13)	0.23 (.13)	0.16 (.12)	0.28 (.12)	0.20 (.12)
C16:0	0.76 (.15)	0.13 (.13)	0.47 (.13)	0.23 (.12)	0.59 (.12)	0.20 (.12)
C18:0	0.50 (.15)	0.24 (.14)	0.40 (.14)	0.33 (.13)	0.44 (.13)	0.29 (.12)
C18:1t11	0.38 (.14)	0.31 (.15)	0.24 (.12)	0.27 (.14)	0.19 (.12)	0.30 (.13)
C18:1c9	0.44 (.16)	0.30 (.15)	0.45 (.13)	0.34 (.12)	0.47 (.12)	0.39 (.12)
C18:2ω6	0.17 (.14)	0.00 (.00)	0.18 (.14)	0.00 (.00)	0.28 (.13)	0.00 (.00)
C18:3ω3	0.22 (.13)	0.10 (.13)	0.06 (.09)	0.23 (.13)	0.13 (.10)	0.27 (.13)
CLAc9,t11	0.28 (.14)	0.35 (.15)	0.24 (.13)	0.24 (.14)	0.19 (.13)	0.27 (.13)
SFA[1]	0.12 (.14)	0.01 (.13)	0.33 (.13)	0.29 (.13)	0.35 (.12)	0.33 (.12)
MUFA[2]	0.16 (.15)	0.01 (.13)	0.36 (.13)	0.29 (.12)	0.38 (.10)	0.33 (.12)
PUFA[3]	0.26 (.15)	0.26 (.15)	0.25 (.13)	0.38 (.14)	0.30 (.13)	0.35 (.14)
ω6:ω3[4]	0.42 (.16)	0.23 (.14)	0.30 (.13)	0.37 (.13)	0.30 (.12)	0.36 (.13)
TFAnoVA[5]	0.30 (.16)	0.16 (.15)	0.33 (.13)	0.44 (.14)	0.35 (.13)	0.40 (.14)
Denovo[6]	0.23 (.14)	0.23 (.14)	0.32 (.13)	0.32 (.13)	0.35 (.12)	0.35 (.12)

[1]Sum of the individual saturated fatty acids.
[2]Sum of the individual monounsaturated fatty acids.
[3]Sum of the individual polyunsaturated fatty acids; odd- and branched-chain fatty acids.
[4]Ratio between the sum of individual PUFA ω6 fatty acids and the sum of individual PUFA ω3 fatty acids.
[5]Trans Fatty Acid (TFA) without Vaccenic acid (VA).
[6]Sum of C6:0, C8:0, C10:0, C11:0, C12:0, iso-C13:0, C14:0 that are *de novo* synthesized in the mammary gland.

Differences among h^2 estimates were mainly due to changes in the additive genetic components as shown in Appendix (Table S2). In particular, for most of the FA analyzed no differences in σ_a^2 were observed with genomic methods. In our study, largest values of R^2 of the regression between σ_a^2 FA_GC and σ_a^2 FA_FTIR were observed using genomic models (0.84 and 0.91) in comparison to the traditional pedigree models (0.45, Figure 1). Finally, σ_a^2 estimates of C16:0, C18:0, C18:1c9, SFA and MUFA were always higher for FA_GC than FA_FTIR.

The FTD contribution to total phenotypic variance was moderate to large. It was on average >0.5 across all different prediction models and phenotypes (Table 6), ranging from 0.17 to 0.88. The variance components for FTD were nearly the same in the three different models, while differences (up to 15%) were highlighted between FA_GC and FA_FTIR (e.g. C4:0, C14:0, C18:1t11, C18:2ω6, C18:3ω3, CLA, PUFA, ω3:ω6 and TFAnoVA).

Accuracy of EBV and GEBV predictions

Accuracies of breeding values were low to moderate, ranging from 0.05 to 0.84, and from 0.02 to 0.45 in the oldest and youngest cohort, respectively (Table 7).

Table 6. Proportion of phenotypic variance[1] explained by FTD (r^2_{FTD}) estimated in the three approaches

	Ablup		Gblup		ssGblup	
Trait	FA_GC	FA_FTIR	FA_GC	FA_FTIR	FA_GC	FA_FTIR
C4:0	0.22	0.20	0.22	0.18	0.20	0.17
C6:0	0.59	0.58	0.59	0.59	0.58	0.58
C8:0	0.61	0.62	0.62	0.62	0.61	0.61
C10:0	0.59	0.60	0.60	0.61	0.59	0.60
C12:0	0.50	0.55	0.51	0.55	0.50	0.55
C14:0	0.35	0.41	0.36	0.41	0.35	0.41
C16:0	0.40	0.47	0.44	0.48	0.41	0.47
C18:0	0.42	0.43	0.43	0.44	0.42	0.43
C18:1t11	0.63	0.71	0.64	0.71	0.64	0.71
C18:1c9	0.63	0.67	0.62	0.66	0.62	0.66
C18:2ω6	0.59	0.47	0.58	0.47	0.58	0.47
C18:3ω3	0.86	0.72	0.86	0.71	0.86	0.71
CLAc9,t11	0.58	0.64	0.59	0.64	0.58	0.64
SFA[2]	0.40	0.40	0.40	0.39	0.39	0.39
MUFA[3]	0.52	0.50	0.51	0.49	0.51	0.49
PUFA[4]	0.68	0.60	0.68	0.60	0.67	0.59
ω6:ω3[5]	0.88	0.79	0.88	0.79	0.88	0.78
TFAnoVA[6]	0.56	0.61	0.56	0.60	0.55	0.60
Denovo[7]	0.54	0.54	0.55	0.55	0.54	0.54
Mean±sd	0.55±0.16	0.55±0.14	0.56±0.16	0.55±0.14	0.55±0.16	0.55±0.14

[1]SE between 0.02 and 0.06 for FA_GC and ranging from 0.04 to 0.04 for FA_FTIR.
[2]Sum of the individual saturated fatty acids.
[3]Sum of the individual monounsaturated fatty acids.
[4]Sum of the individual polyunsaturated fatty acids; odd- and branched-chain fatty acids;
[5]Ratio between the sum of individual PUFA ω6 fatty acids and the sum of individual PUFA ω3 fatty acids.
[6]Trans Fatty Acid (TFA) without Vaccenic acid (VA).
[7]Sum of C6:0, C8:0, C10:0, C11:0, C12:0, iso-C13:0, C14:0 that are *de novo* synthesized in the mammary gland

Table 7. EBV and GEBV accuracy of prediction for milk fatty acids obtained with gas chromatography (FA_GC) or predicted by Fourier Transform Infrared spectra (FA_FTIR) using the three relationship matrices: pedigree (**A,** Ablup), genomic (**G**, Gblup) or pedigree and SNP blended using a single-step genomic approach (**H**, ssGblup)

	Oldest animals[1]						Youngest aninals[2]					
	FA_GC			FA_FTIR			FA_GC			FA_FTIR		
Trait	Ablup	Gblup	ssGblup	Ablup	Gblup	ssGblup	Ablup	Gblup	ssGblup	Ablup	Gblup	ssGblup
C4:0	0.52	0.54	0.56	0.57	0.59	0.60	0.19	0.28	0.35	0.21	0.31	0.37
C6:0	0.29	0.32	0.36	0.52	0.54	0.55	0.10	0.18	0.27	0.18	0.28	0.34
C8:0	0.48	0.50	0.52	0.53	0.55	0.56	0.17	0.26	0.33	0.18	0.28	0.34
C10:0	0.54	0.56	0.57	0.56	0.58	0.59	0.19	0.29	0.35	0.20	0.30	0.35
C12:0	0.52	0.54	0.56	0.55	0.56	0.58	0.18	0.28	0.34	0.19	0.29	0.35
C14:0	0.43	0.45	0.48	0.32	0.35	0.39	0.15	0.24	0.32	0.11	0.20	0.28
C16:0	0.83	0.84	0.83	0.35	0.38	0.41	0.29	0.41	0.45	0.12	0.21	0.29
C18:0	0.68	0.69	0.70	0.48	0.50	0.52	0.24	0.35	0.40	0.17	0.26	0.33
C18:1t11	0.59	0.60	0.61	0.54	0.56	0.57	0.20	0.31	0.36	0.19	0.29	0.34
C18:1c9	0.63	0.65	0.65	0.53	0.55	0.56	0.22	0.32	0.38	0.18	0.28	0.34
C18:2ω6	0.39	0.42	0.45	0.05	0.09	0.21	0.14	0.23	0.30	0.02	0.10	0.23

C18:3ω3	0.45	0.47	0.50	0.30	0.33	0.37	0.16	0.25	0.32	0.10	0.19	0.28
CLAc9,t11	0.51	0.53	0.55	0.57	0.58	0.60	0.18	0.28	0.34	0.20	0.30	0.35
SFA[3]	0.33	0.36	0.40	0.09	0.12	0.23	0.12	0.20	0.29	0.03	0.11	0.23
MUFA[4]	0.38	0.41	0.44	0.11	0.14	0.24	0.13	0.22	0.30	0.04	0.11	0.24
PUFA[5]	0.49	0.52	0.53	0.49	0.51	0.53	0.17	0.27	0.33	0.17	0.27	0.33
ω6:ω3[6]	0.61	0.63	0.64	0.46	0.48	0.50	0.21	0.32	0.37	0.16	0.25	0.32
TFAnoVA[7]	0.53	0.55	0.56	0.38	0.41	0.44	0.18	0.28	0.34	0.13	0.22	0.30
Denovo[8]	0.46	0.48	0.50	0.49	0.51	0.53	0.16	0.25	0.32	0.17	0.27	0.33
Mean	0.51	0.53	0.55	0.42	0.44	0.47	0.18	0.27	0.34	0.14	0.24	0.31
SD	0.13	0.12	0.11	0.17	0.16	0.13	0.04	0.05	0.04	0.06	0.06	0.04

[1]Cohort of sheep born before December 2012 with SNP genotypes and own milk FA records available.
[2]Cohort of sheep born after November 2012 with SNP genotypes available and own milk FA records masked to mimic a candidate set of younger sheep.
[3]Sum of the individual saturated fatty acids.
[4]Sum of the individual monounsaturated fatty acids.
[5]Sum of the individual polyunsaturated fatty acids; odd- and branched-chain fatty acids.
[6]Ratio between the sum of individual PUFA ω6 fatty acids and the sum of individual PUFA ω3 fatty acids.
[7]Trans Fatty Acid (TFA) without Vaccenic acid (VA).
[8]Sum of C6:0, C8:0, C10:0, C11:0, C12:0, iso-C13:0, C14:0 that are *de novo* synthesized in the mammary gland.

The palmitic acid (C16:0) showed the largest accuracy for FA_GC across the different prediction models, both for oldest (0.84) and youngest animals (0.45). The largest GEBV accuracy for FA_FTIR was observed for the butyric acid (C4:0). The linoleic acid (C18:2ω6) showed the lowest accuracy in most of the scenarios considered. Accuracies of FA groups reflected their composition, with saturated FA showing the lowest and PUFA and TFAnoVA the highest accuracies, respectively.

The cohort of animals with own phenotypes exhibited larger prediction accuracies compared to young animals without phenotype (overall average difference +0.24) in all scenarios (Table 7). The largest difference (+0.30) was observed for the stearic acid (C18:0), whereas the smallest for the saturated FA group (+0.09).

Differences were also observed between the phenotype (FA_GC vs FA_FTIR) for all the three models and for the two cohorts of animals (Table 7), even though without a defined pattern. The major difference between FA_GC and FA_FTIR were observed in the older cohort (from -0.23 up to 0.48 for C6:0 and C16:0, respectively). Accuracies differed mainly in the ABLUP approach for both young and older cohorts. The difference between FA_GC and FA_FTIR tended to reduce in genomic methods applied to young animals (Table 7). Regardless of the statistical model used, the largest difference between FA_GC and FA_FTIR was observed for the C16:0 (on average difference of 0.45 and 0.18 for old and young animals, respectively). Relevant differences (at least >15%) between FA_CG and FA_FTIR were observed also for C18:0, C18:2ω6, SFA and MUFA both in older and younger animals.

As far as the three models are concerned, genomic prediction accuracies were constantly higher than in ABLUP (Table 7). In particular, differences between ABLUP

and genomic methods were larger in young animals. In this cohort, positive changes up to +0.12 (+0.17) and +0.10 (+0.21) were observed in the comparison GBLUP-ABLUP (ssGBLUP-ABLUP) for FA_GC and FA_FTIR, respectively. Among the two genomic approaches, the ssGBLUP accuracies were always larger than GBLUP ones both in young and old animal cohorts.

Discussion

Fatty acid composition is a key feature in defining sheep milk nutritional quality. Its genetic improvement is an appealing option for enhancing market value of dairy sheep products. However, breeding for milk FA composition in sheep is hampered by difficulties in phenotyping and in implementing appropriate selection strategies. Use of equations for predicting FA from milk FTIR spectra is widely recognized as a cost-effective solution for obtaining FA profiles in milk of different ruminant species (Ferrand-Calmels *et al*., 2014). At the same time, early experiences of genomic selection on meat, wool (Daetwyler *et al*., 2012) and dairy sheep (e.g Duchemin *et al*., 2012; Legarra *et al*., 2014; Baloche *et al*., 2014) have reported an increase of breeding value accuracy and selection response compared to the traditional pedigree-based method.

Results of the present study showed an effect of both investigated phenotypes (i.e. FA_GC or FA_FTIR) and of the information used to structure the genetic covariance among animals (pedigree, genomic, or both) on genetic parameter estimates and breeding value prediction accuracies.

Genetic Parameters of Milk Fatty Acid profile

Heritability estimates based on pedigree models were consistent with a previous work carried out on a similar data set (Correddu *et al*., 2018), whereas genomic based h2 resulted higher and lower than pedigree based for saturated (<C14) and unsaturated FA, respectively. A large variation among different FA was observed, regardless the considered approach or the phenotype used, in agreement with previous studies (Sanchez *et al*., 2010; Boichard *et al*., 2014). Differences among FA are mainly related to their metabolic pathway. Some FA are synthetized de novo in the mammary gland, others are mostly related to the animal diet, and others came from of body reserve mobilization. Thus, larger heritability is expected for FA whose milk concentration is under enzymatic control (i.e. de novo FA) compared to FA that are related to the animal diet (Arnould and Soyeurt, 2009). The higher value of heritability observed for Denovo FA compared to those coming from diet or body fat reserve (e.g.: C18 FA) seemed to confirm the stronger genetic regulation for the former group of FA (e.g. Bastin *et al*., 2011; Narayana *et al*., 2017). Morever, lowest h2 values were highlighted for C18:2ω6 and C18:3ω3 (Tables 4 and 5), regardless the model used. It is well known that these two FA are strongly dependent on their concentration in animals' diet (e.g. Fleming *et al*., 2016; Pegolo *et al*. 2017).

Differences between h^2 estimated using FA_GC and FA_FTIR were in most of cases low to moderate. FA_FTIR produced larger h^2 estimates for short chain FA (Figures 1), whereas an opposite trend can be observed for medium and long-chain FA. A similar pattern was also observed in cattle using GC (Stoop *et al*., 2008; Duchemin *et al*., 2013). The largest differences were found for FA (e.g.C16:0 and C4:0) that

exhibited lowest FTIR prediction accuracies. In dairy cattle, larger heritabilities for FA_GC compared to FA_FTIR have been reported (Rutten *et al*., 2010; Bonfatti *et al*., 2017). In particular, Bonfatti *et al* (2017) pointed out that the differences were due to a reduction of the σ_a^2 in FA_FTIR (-0.52%) compared to FA_GC. In the present work, the use of FA_FTIR phenotypes resulted in most of cases (short chain FAs) in smaller estimates for all the three variance components (Table S2).

Apart from the values obtained for palmitic and stearic acids, pedigree based h^2 were in most of cases lower than those obtained using genomic information. In particular, most of FA showed an increase of σ_a^2 and a reduction of σ_e^2 (especially for FA_FTIR) when moving from traditional pedigree to genomic methods, respectively (Table S2). Veerkamp *et al*. (2011) working on a dairy cattle sample of comparable size, found larger heritabilities for milk yield, dry matter intake and body weight, when **A** instead of **G** was used. This result, due to a reduction of σ_a^2 when genomic information was used, was explained with the different structure of the two relationship matrices, especially as far as the base population is considered.

The higher heritability observed in the present work for genomic models can be ascribed to a series of reasons. The first are the considered traits. Milk FA content is characterized by a relevant sensitivity to environmental conditions. This peculiarity is enhanced in the typical farming system of the Sarda sheep, where natural pastures represent the main feeding source (Carta *et al*., 2009; Nudda *et al*., 2014). Moreover, it should be remembered that only one record per animal was available. This condition, that undoubtedly reduces the reliability of the measure, is rather frequent in studies on FA genetic parameter estimation using FA_GC also in cattle (*e.g.* Stoop *et al*., 2008;

Mele *et al.*, 2009; Pegolo *et al.*, 2016). On the other hand, the recording of a single measure per animal is more representative of the practical situation of a breeding scheme where innovative phenotypes are considered among the selection goals. A second reason is represented by the structure of the considered dairy sheep population, quite different from usual dairy cattle populations of genomic studies. It consisted of only females, sired by 445 rams (2.07±1.7 with a maximum of 15 daughter per ram). Such a structure can be considered representative of the Sarda breed, in which natural mating is the main reproductive technique (Carta *et al.*, 2009). A third reason can be found in the genetic structure of dairy sheep populations. Contrarily to what observed in the present study, larger heritabilities were found when **A** was fitted in comparison with G on dairy cattle (Veerkamp *et al.*, 2010; Haile-Mariam *et al.*, 2013; Loberg *et al.*, 2015). The authors explained these results with the imperfect linkage disequilibrium (LD) existing between SNP and causative mutations that makes **G** unable for capturing all the genetic variance of the trait in comparison with **A**. Such a limitation of **G** is likely to be more pronounced in sheep populations that, in comparison to cattle, are characterized by a lower LD at relatively short distance (Kijas *et al.*, 2014). However, the reliability of pedigrees in sheep is often questionable due to the uncorrected parentage assignment or the high number of unknown parents. Thus, the use of genomic relationship matrices could allow to estimate more accurately relationship among animals because the realized fraction of allele shared between individual is directly computed (Hayes and Goddard, 2008; Legarra *et al.*, 2014), with subsequent large heritability estimates.

Accuracy of EBV and GEBV predictions

In our study breeding value accuracies for FA milk profile were low to moderate. Considering the sample size, the genetic architecture of milk FA composition, and the number of records per ewe our results are in accordance to genomic selection theory (Goddard and Hayes, 2009). Animals with their own phenotypes exhibited larger accuracies compared to young animals. However, the addition of genotype information to the breeding value prediction resulted in an improvement of accuracy, also in latter group. Other studies in sheep underlined the higher accuracy of genomic methods compared to the pedigree-based approach for milk and meat production traits (Daetwyler *et al*., 2012; Legarra *et al*., 2014; Baloche *et al*., 2014). Moreover, GS studies carried out in beef cattle on muscle FA composition reported for some of FA investigated also in this study a similar pattern of GEBV accuracy (Chang *et al*., 2015; Chiaia *et al*., 2017; Zhu *et al*., 2017).

The similar magnitude of GEBV accuracy for FA_FTIR and FA_GC is an interesting for a possible implementation breeding program for milk FA composition in dairy sheep, due to the considerable reduction of phenotyping cost. The predictive ability of FTIR spectra (R2GC-FTIR, see Table 3) might have affected the accuracy of genomic predictions: a moderate correlation between R^2 GC-FTIR and (G)EBV accuracy were observed (0.46 and 0.45 in ssGBLUP for old and young cohort, respectively).

Regarding the prediction model, the slightly higher accuracies found using ssGBLUP could be ascribable to the blended (co)variance structure that can takes benefits from the inclusion of all relatives of non-genotyped and genotypes ewes with recorded traits (Aguilar *et al*., 2010; Legarra *et al*., 2014). Finally, when the selection intensity is not

so high (as in Sarda sheep), the use of genomic selection with genotyped females may help to improve milk composition traits even of un-phenotyped animals (young cohort) as already suggested in a simulation study by Gorjanc *et al*. (2015). However, the complex genetic correlation pattern that exist among the different FA should be carefully considered (Carta *et al*., 2008; Sanchez *et al*., 2010) when implementing a coherent selection goal aimed at improving the milk FA composition.

Conclusions

The Fourier Transform Infrared spectrography is commonly used in dairy industry for milk composition recordings, as well as genomic selection is an effective tool to rank the best candidates for breeding purpose. The results presented in the current investigation confirmed that in dairy sheep FTIR predicted FA are heritable traits, exhibiting from low to moderate heritabilities. These figures are comparable to those estimated from more expensive and time-consuming GC measured phenotypes. Moreover, breeding value accuracies obtained with genomic selection methods were always higher than those estimated with traditional pedigree-based approach, and ssGBLUP outperformed the GBLUP method. Results of the present study suggest that the combination of FTIR predictions and genomic selection technology could represent an interesting option for the genetic improvement of milk FA composition in dairy sheep. The inclusion of fat composition in breeding programs could have some interesting practical implications due the connection between milk fat profile and human health. Aim of the selection for FA profile would be the reduction of the amount

of saturated fatty acid and the simultaneous increase of PUFA, omega-3 and CLA because of their recognized healthy effects for humans.

Acknowledgements

This research was funded by Regional Government (Grant no. CRP 61608 “Il latte Ovino della Sardegna” and by G) and Italian Research Project “GenHome”. The authors would like to acknowledge the Provincial Breeders Associations (AIPA) of Cagliari, Nuoro, Sassari, and Oristano (Italy); the laboratory of Sardinian Breeders Association (ARA, Oristano, Italy) for providing milk spectra; the Italian Associations of Animal Breeders (AIA). Authors are grateful to Daniela Lourenco (University of Georgia, Athens, GA, USA) for her useful suggestions on implementing genomic models.

References

Aguilar I., Misztal I., Johnson D. L., Legarra A., Tsuruta S., & Lawlor T. J (2010) Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science, 93:743-752.

Arnould V. R., & Soyeurt H. (2009). Genetic variability of milk fatty acids. Journal of Applied Genetics, 50:29-39.

Astruc J. M., Barillet F., Carta A., Fioretti M., Gootwine E., Kompan D., Romberg F. J., & Ugarte E. (2008). Report of the working group on milk recording of sheep. Pages 275–282 in Proc. 36th Biennial Session of the International Committee for Animal Recording (ICAR), Niagara Falls, NY. ICAR technical series no. 13. Identification, breeding, production, health and recording on farm animals. J. D. Slatter, ed. ICAR, Rome, Italy

Baloche G., Legarra A., Sallé G., Larroque H., Astruc J. M., Robert-Granié C., & Barillet F. (2014) Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. Journal of Dairy Science, 97:1107-1116.

Banni S., Heys S. D., & Wahle K. W. J. (2003) Conjugated linoleic acids as anticancer nutrients: studies in vivo and cellular mechanisms. Pages 267–282 in Advances in conjugated linoleic acid research, W. W. Christie, J-L. Sébédio, and R. Adlof, ed. AOCS press. Illinois.

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

Bastin C., Soyeurt H. & Gengler N. (2013) Genetic parameters of milk production traits and fatty acid contents in milk for Holstein cows in parity 1–3. Journal of Animal Breeding and Genetics, 130:118-127

Bhattacharya A., Banu J., Rahman M., Causey J., & Fernandes G. (2006) Biological effects of conjugated linoleic acids in health and disease. The Journal of Nutritional Biochemistry, 17:789-810.

Boichard D., Govignon-Gion A., Larroque H., Maroteau C., Palhière I., Tosser-Klopp G., Rupp R., Sanchez M. P., & Brochard M. (2014) Déterminisme génétique de la composition en acides gras et protéines du lait des ruminants, et potentialités de sélection. Inra Prod. Animal, 27:283-298.

Boichard D., & Brochard M. (2012). New phenotypes for new breeding goals in dairy cattle. Animal, 6:544-550.

Bonfatti V., Vicario D., Lugo A., & Carnier P. (2017) Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle. Journal of Dairy Science, 100:5526-5540.

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

Buitenhuis B., Janss L. L., Poulsen N. A., Larsen L. B., Larsen M. K., & Sørensen P. (2014) Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. BMC genomics, 15:1112.

Cabiddu A., Decandia M., Addis M., Piredda G., Pirisi A., & Molle G (2005) Managing Mediterranean pastures in order to enhance the level of beneficial fatty acids in sheep milk. Small Ruminant Research, 59:169-180.

Caredda M., Addis M., Ibba I., Leardi R., Scintu M. F., Piredda G., & Sanna G. (2016) Prediction of fatty acid content in sheep milk by mid-infrared spectrometry with a selection of wavelengths by genetic algorithms. Food Science and Technology Leb, 65:503-510.

Carta A., Casu S., Usai M. G., Addis M., Fiori M., Fraghì A., Miari S., Mura L., Piredda G., Schibler L., Sechi T., Elsen J. M., & Barillet F. (2008). Investigating the genetic component of fatty acid content in sheep milk. Small Rumin. Res. 79:22-28.

Carta A., Casu S., & Salaris S. (2009). Invited review: Current state of genetic improvement in dairy sheep. Journal of Dairy Science, 92:5814–5833.

Cecchinato A., De Marchi M., Gallo L., Bittante G., & Carnier P. (2009) Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. Journal of Dairy Science, 92:5304-5313.

Chen L., Ekine-Dzivenu C., Vinsky M., Basarab J., Aalhus J., Dugan M. E., Fitzsimmons C., Stothard P., & Li C. (2015) Genome-wide association and genomic prediction of breeding values for fatty acid composition in subcutaneous adipose and longissimus lumborum muscle of beef cattle. BMC genetics, 16:135.

Correddu F., Cellesi M., Serdino J., Manca M. G., Contu M., Dimauro C., Ibba I., & Macciotta N. P. P. (2018). Genetic parameters of milk fatty acid profile in

sheep: comparison between gas chromatographic measurements and Fourier-Transform Infrared Spectroscopy predictions. Animal, 1:8.

Crisà A., Marchitelli C., Pariset L., Contarini G., Signorelli F., Napolitano F., Catillo G., Valentini A., and Moioli B. (2010) Exploring polymorphisms and effects of candidate genes on milk fat quality in dairy sheep. Journal of Dairy Science, 93:3834-3845.

Daetwyler H. D., Swan A. A., van der Werf J. H., & Hayes B. J. (2012) Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. Genetic Selection Evolution, 44:33.

De Marchi M., Penasa M., Cecchinato A., Mele M., Secchiari P. & Bittante G. (2011) Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal, 5:1653-1658.

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

Duchemin S. I., Colombani C., Legarra A., Baloche G., Larroque H., Astruc J. M., Barillet F., Robert-Granié C., & Manfredi E. (2012) Genomic selection in the French Lacaune dairy sheep breed. Journal of Dary Science, 95:2723–2733.

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

Ferrand-Calmels M., Palhière I., Brochard M., Leray O., Astruc J. M., Aurel M. R., Barbey S., Bouvier F., Brunschwig P., Caillat H., Douguet M.., Faucon-Lahalle F., Gelé M., Thomas G., Trommenschlager J. M., & Douguet M. (2014) Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. Journal of Dairy Science, 97:17-35.

Fleming A. (2016) Phenotypic and Genetic Variation of Milk Fat Components Incorporating Mid-Infrared Technology. Doctoral Thesis. University of Guelph, Ontario, Canada.

Gorjanc G., Bijma P., & Hickey J. M. (2015) Reliability of pedigree-based and genomic evaluations in selected populations. Genetics Selection Evolution, 47:65.

Haile-Mariam M., Nieuwhof G. J., Beard K. T., Konstatinov K. V., & Hayes B. J. (2013) Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. Journal of Animal Breeding and Genetics, 130:20-31.

Hayes B. J., Visscher P. M., & Goddard M. E. (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research, 91:47-60.

Kijas J. W., Porto-Neto L., Dominik S., Reverter A., Bunch R., McCulloch R., Hayes B. J., Brauning R., McEwan J., & the International Sheep Genomics Consortium (2014) Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. Anim. Genet. 45:754-757.

Legarra A., Baloche G., Barillet F., Astruc J. M., Soulas C., Aguerre X., Arrese F., Mintigi L., Lasarte M., Maeztu F., Beltrán de Heredia I., & Ugarte E. (2014) Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. Journal of Dairy Science, 97: 3200-3212.

Loberg A., Dürr J. W., Fikse W. F., Jorjani H., & Crooks L. (2015) Estimates of genetic variance and variance of predicted genetic merits using pedigree or genomic relationship matrices in six Brown Swiss cattle populations for different traits. Journal of Animal Breeding and Genetics, 132:376-385.

Macciotta N. P. P., Mele M., Cappio-Borlino A., & Secchiari P. (2005) Issues and perspectives in dairy sheep breeding. Italian Journal of Animal Science, 4:5-23

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

Mele M., Contarini G., Cercaci L., Serra A., Buccioni A., Povolo M., Conte G., Funaro A., Banni S., Lercker G., & Secchiari P. (2011) Enrichment of Pecorino cheese with conjugated linoleic acid by feeding dairy ewes with extruded linseed: effect on fatty acid and triglycerides composition and on oxidative stability. International Dairy Journal, 21:365-372.

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

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

Misztal I., Tsuruta S., Lourenco D. L., Aguilar I., Legarra A., & Vitezica Z. (2015) Manual for BLUPF90 family of programs. Retrieved from http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf

Moioli B., Contarini G., Pariset L., Marchitelli C., Crisà A., Catillo G., & Napolitano F. (2012) Genetic variation of C18:1 and C18:2 isomers in sheep milk fat. Small Ruminant Research, 103:187-193.

Narayana S. G., Schenkel F. S., Fleming A., Koeck A., Malchiodi F., Jamrozik J., Johnston J., Sargolzai M., & Miglior F. (2017) Genetic analysis of groups of mid-infrared predicted fatty acids in milk. Journal of Dairy Science, 100:4731-4744.

Nudda A., Battacone G., Boaventura Neto O., Cannas A., Francesconi A. H. D., Atzori A. S., & Pulina G. (2014) Feeding strategies to design the fatty acid profile of sheep milk and cheese. Revista da Sociedade Brasileira de Zootecnia, 43:445-456.

Pegolo S., Cecchinato A., Casellas J., Conte G., Mele M., Schiavon S., & Bittante G. (2016) Genetic and environmental relationships of detailed milk fatty acids profile determined by gas chromatography in Brown Swiss cows. Journal of Dairy Science, 99:1315-1330.

Pegolo S., Stocco G., Mele M., Schiavon S., Bittante G., & Cecchinato A. (2017) Factors affecting variations in the detailed fatty acid profile of Mediterranean

buffalo milk determined by 2-dimensional gas chromatography. Journal of Dairy Science, 100:2564-2576.

Pulina G., Milán M. J., Lavín M. P., Theodoridis A., Morin E., Capote J., Thomas D. L., Francesconi A. H. D., & Caja G. (2018) Invited review: Current production trends, farm structures, and economics of the dairy sheep and goat sectors. Journal of Dairy Science. In press.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., Bender D., Maller J., Sklar P., de Bakker P. I. W., Daly M. J., & Sham P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81:559-575.

Rupp R., Mucha S., Larroque H., McEwan J., & Conington J. (2016) Genomic application in sheep and goat breeding. Animal Frontiers, 6:39-44.

Rutten M. J. M., Bovenhuis H., Hettinga K. A., van Valenberg H. J. F., & van Arendonk J. A. M. (2009) Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. Journal of Dairy Science, 92:6202-6209

Rutten M. J. M., Bovenhuis H., & Van Arendonk J. A. M. (2010) The effect of the number of observations used for Fourier transform infrared model calibration for bovine milk fat composition on the estimated genetic parameters of the predicted data. Journal of Dairy Science, 93:4872-4882.

Sánchez J. P., San Primitivo F., Barbosa E., Varona L., & De La Fuente L. F. (2010) Genetic determination of fatty acid composition in Spanish Churra sheep milk. Journal of Dairy Science, 93:330-339.

Stoop W. M., van Arendonk J. A. M., Heck J. M. L., van Valenberg H. J. F., & Bovenhuis H. (2008) Genetic Parameters for Major Milk Fatty Acids and Milk Production Traits of Dutch Holstein-Friesians. Journal of Dairy Science, 91:385-394.

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

Tsuruta S., Misztal I., & Lawlor T. J. (2013) Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. Journal of Dairy Science, 96:3332-3335.

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

VanRaden P. M. (2008) Efficient methods to compute genomic predictions. Journal of Dairy Science, 91:4414-4423.

Veerkamp R. F., Mulder H. A., Thompson R., & Calus M. P. L. (2011) Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. Journal of Dairy Science, 94:4189-4197.

Zhu B., Niu H., Zhang W., Wang Z., Liang Y., Guan L., Guo P., Chen Y., Zhang L., Guo Y., Ni H., Gao X., Xu L., & Li J. (2017) Genome wide association study

and genomic prediction for fatty acid composition in Chinese Simmental beef cattle using high density SNP array. BMC genomics, 18:464.

Appendix Chapter 3

Genomic selection of milk fatty acid composition in Sarda dairy sheep: effect of different phenotypes and relationship matrices on heritability and breeding values accuracy

Table S1. Single FA used to define groups of FA analyzed.

Table S2. Variance components estimation (animal, flock test date and residual) for measured and predicted fatty acids across the three methods

Table S1. Single FA used to define groups of FA analyzed.

Group of FA	Single fatty acid
SFA: sum of individual saturated fatty acids	C4:0, C6:0, C0, C8:0, C9:0, C10:0, C11:0, C12:0, isoC13:0, anteisoC13:0, isoC14:0, C14:0, isoC15:0, anteisoC15:0, C15:0, isoC16:0, C16:0, isoC17:0, anteisoC17:0, C17:0, isoC18:0, C18:0, C19:0, C20:0, C22:0, C23:0, C24:0, C25:0, C26:0
MUFA: sum of individual monounsaturated fatty acids	C10:1, C14:1c9, C15:1, C16:1t4, C16:1t5, C16:1t6+t7, C16:1t9, C16:1t10, C16:1t11+t12, C16:1c7, C16:1c9, C16:1c10, C16:1c11, C17:1c6+c7, C17:1c8, C17:1c9, C18:1t4, C18:1t5, C18:1t6+t8, C18:1t9, C18:1t10, C18:1t11, C18:1t12, C18:1t13+t14, C18:1c9, C18:1t15+c10, C18:1c11, C18:1c12, C18:1c13, C18:1t16+c14, C18:1c15, C18:1c16, C20:1c5, C20:1c9, C20:1c11, C20:1c15, C22:1ω9, C24:1c15
PUFA: sum of individual polyunsaturated fatty acids	C18:2t10t14, C18:2t11t15, C18:2t9t12, C18:2c9t13, C18:2t8c13, C18:2c9t12, C18:2t9c12, C18:2t11c15, C18:2ω6, C18:2t12c15, C18:2c12c15, CLAc9t11, CLAt9c11, CLAt10c12, CLAt11c13, CLAt12t14, CLAt11t13, CLAt9t11, C20:2ω9, C20:2ω6, C22:2ω6, C18:3ω6, C18:3ω3, C20:3ω9, C20:3ω6, C20:3, C20:3ω3, C22:3ω6, C18:4ω3, C20:4ω6, C20:4ω3, C22:4ω6, C20:5ω3, C22:5ω3, C22:6ω3
TFA-VA	sum of individual trans FA excluding C18:1t11 (Vaccenic acid)
PUFA ω6 : PUFA ω3	ratio between the sum of individual PUFA ω6 and the sum of all individual PUFA ω3
Denovo de novo synthesized in the mammary gland.	C6:0, C8:0, C10:0, C11:0, C12:0, iso-C13:0, C14:0

Table S2. Variance components estimation (animal, flock test date and residual) for measured and predicted fatty acids across the three methods

	ABLUP						GBLUP						ssGBLUP					
	FA_GC			FA_FTIR			FA_GC			FA_FTIR			FA_GC			FA_FTIR		
	σ_a^2	σ_f^2	σ_e^2	σ_a^2	σ_f^2	σ_e^2	σ_a^2	σ_f^2	σ_e^2	σ_a^2	σ_f^2	σ_e^2	σ_a^2	σ_f^2	σ_e^2	σ_a^2	σ_f^2	σ_e^2
C4:0	0.02	0.02	0.06	0.02	0.02	0.05	0.04	0.02	0.04	0.05	0.02	0.02	0.04	0.02	0.05	0.04	0.02	0.03
C6:0	0.00	0.07	0.05	0.01	0.06	0.03	0.02	0.07	0.03	0.02	0.06	0.02	0.02	0.07	0.03	0.03	0.06	0.02
C8:0	0.02	0.12	0.06	0.02	0.11	0.05	0.03	0.12	0.04	0.04	0.11	0.03	0.03	0.12	0.04	0.04	0.11	0.03
C10:0	0.38	1.70	0.81	0.37	1.62	0.70	0.46	1.75	0.73	0.48	1.65	0.59	0.50	1.74	0.71	0.53	1.65	0.57
C12:0	0.14	0.46	0.33	0.12	0.45	0.25	0.15	0.48	0.31	0.13	0.46	0.24	0.16	0.47	0.31	0.14	0.46	0.24
C14:0	0.26	0.73	1.07	0.12	0.73	0.93	0.32	0.74	1.02	0.17	0.74	0.88	0.39	0.72	0.97	0.22	0.73	0.84
C16:0	3.64	3.19	1.10	0.44	2.90	2.79	2.17	3.68	2.44	0.75	2.96	2.50	2.87	3.42	1.98	0.68	2.93	2.61
C18:0	1.61	2.28	1.56	0.61	1.82	1.83	1.26	2.41	1.89	0.81	1.90	1.64	1.47	2.32	1.79	0.72	1.86	1.77
C18:1t11	0.13	0.59	0.21	0.07	0.56	0.16	0.08	0.60	0.26	0.06	0.58	0.17	0.07	0.60	0.28	0.07	0.57	0.16
C18:1c9	2.18	8.22	2.73	1.19	7.80	2.72	2.22	8.10	2.69	1.34	7.72	2.58	2.38	8.22	2.67	1.60	7.81	2.42
C18:2ω6	0.02	0.13	0.08	0.00	0.07	0.08	0.02	0.13	0.07	0.00	0.07	0.08	0.03	0.13	0.07	0.00	0.07	0.08
C18:3ω3	0.01	0.21	0.03	0.01	0.13	0.05	0.00	0.21	0.03	0.01	0.13	0.04	0.00	0.21	0.03	0.01	0.13	0.04
CLAc9t11	0.02	0.12	0.06	0.02	0.10	0.04	0.02	0.12	0.06	0.01	0.11	0.04	0.02	0.12	0.07	0.02	0.10	0.04
SFA[1]	1.11	6.17	8.00	0.10	5.31	7.97	3.12	6.16	6.14	2.44	5.36	5.77	3.39	6.12	6.08	2.80	5.40	5.59
MUFA[2]	1.01	6.68	5.26	0.10	5.46	5.39	2.34	6.63	4.03	1.65	5.46	3.93	2.49	6.65	4.03	1.88	5.52	3.82
PUFA[3]	0.18	1.41	0.49	0.19	1.06	0.52	0.17	1.44	0.50	0.27	1.08	0.44	0.21	1.42	0.47	0.26	1.07	0.47
ω6:ω3[4]	0.06	1.11	0.09	0.05	0.72	0.15	0.04	1.12	0.10	0.07	0.73	0.12	0.05	1.11	0.10	0.08	0.72	0.13
TFAnoVA[5]	0.30	1.25	0.69	0.12	1.13	0.60	0.33	1.26	0.67	0.33	1.12	0.41	0.37	1.24	0.66	0.31	1.12	0.45
Denovo[6]	2.21	11.18	7.29	1.92	9.68	6.31	3.13	11.47	6.43	2.71	9.94	5.57	3.41	11.38	6.34	2.96	9.86	5.50

[1]Sum of the individual saturated fatty acids

[2]Sum of the individual monounsaturated fatty acids.

[3]Sum of the individual polyunsaturated fatty acids; odd- and branched-chain fatty acids.

[4]Ratio between the sum of individual PUFA ω6 fatty acids and the sum of individual PUFA ω3 fatty acids.

[5]Trans Fatty Acid (TFA) without Vaccenic acid (VA).

[6]Sum of C6:0, C8:0, C10:0, C11:0, C12:0, iso-C13:0, C14:0 that are *de novo* synthesized in the mammary gland.

CHAPTER 4

Genome-wide variability and selection signatures in Italian island cattle breeds

This is the published version in **Animal Genetics** of:

Cesarani A., Sorbolini S., Criscione, A., Bordonaro S., Pulina G., Battacone G., Marletta D., Gaspa G., and Macciotta N. P.P. (2018) Genome-wide variability and selection signatures in Italian island cattle breeds. Anim Genet, 49: 371-383. doi:10.1111/age.12697

Abstract

In the present study, a sample of 88 animals belonging to four local (Modicana, Sarda, Sardo-Bruna and Sardo-Modicana) and one cosmopolitan (Italian Brown Swiss) cattle breeds were genotyped with a medium density SNP beadchip and compared in order to investigate their genetic diversity and the existence of selection signatures. A total of 43,012 SNPs scattered across all twenty-nine autosomal chromosomes were retained after the data quality control. Basic population statistics, Wright Fixation Index (F_{ST}) and Runs of Homozygosity (ROH) analyses confirmed that Italian Brown genome was mainly shaped by selection, as underlined by the low values of heterozygosity and minor allele frequency. As expected, local cattle exhibited a large within breed genetic heterogeneity. The F_{ST} comparison with the largest number of significant SNPs was Sardo-Bruna vs Sardo-Modicana, whereas the smallest was observed for Italian Brown Swiss vs Sardo-Modicana, respectively. Modicana exhibited the largest number of detected ROH, whereas the smallest was observed for Sardo-Modicana. Signatures of selection were detected in genomic regions that harbor genes involved in milk production traits for the Italian Brown Swiss and fitness traits for local breeds. According to the results of Multi-Dimensional scaling and admixture analysis the Sardo-Bruna is more similar to the Sarda rather than to the Italian Brown Swiss. Moreover, the Sardo-Modicana is genetically closer to the Modicana rather than to the Sarda breed. Results of the present work confirm the usefulness of Single Nucleotide Polymorphisms in deciphering the genetic architecture of livestock breeds.

Keywords

Indigenous breeds, selection signatures, inbreeding, admixture, biodiversity

Introduction

The bovine domestication occurred presumably about 8-10 thousand years ago in southwest Asia (Zeder 2017). This process led to the zebuine and taurine breeds (Loftus *et al*. 1994; Upadhyay *et al.* 2016) derived both from the extinct wild aurochs (*Bos primigenius*) that spread in Europe and Africa in successive waves of migration. With domestication, cattle acquired a large variety of distinctive traits compared to their wild ancestors: for example, they became smaller in size and developed the capacity to adapt to various environments. During the Neolithic revolution, cattle accompanied human migrations and crosses between individuals of different ethnic groups generated a gene flow that changed the genetic makeup of their populations (Ajmone-Marsan *et al.* 2010).

The continuously increasing demand for work, milk and meat has enhanced between population differences over the centuries. In particular, changes in the farming systems, intense implementation of artificial selection, crossbreeding, and widespread use of artificial insemination that occurred in the last decades resulted in a huge genetic improvement of few highly specialized cattle breeds. However, as a consequence the within breed genetic variability has been seriously constrained in these populations (Brotherstone and Goddard, 2005). Biodiversity has been drastically endangered, a relevant reduction in the number of farmed cattle breeds has been observed leading to the extinction of many local breeds. Indigenous populations, better suited to extensive farming but not very productive, have been often abandoned in favor of highly productive breeds (Scherf, 2000; Medugorac *et al.* 2009).

Concerns about climate changes, ethical issues, and evolution of consumer needs, including ecosystem services and landscape protection, are bringing towards sustainable livestock farming systems. Such an evolving situation seems to offer new opportunities to indigenous breeds, because of their strong linkage to the production area, large genetic variability, and great fitness. Local breeds, are now considered as important reservoirs of resilience and biodiversity (Giovambattista *et al*. 2001). Their genomes represent an ideal model for studying and understanding the evolutionary history of livestock species, essential goal for evolutionary biology and population genetics. Moreover, local breeds represent a source of income in marginal areas (Ruto *et al*. 2008) and a chance to answer to the environmental changes (Medugorac *et al.* 2009). Their typical products support a sustainable development of the rural environment and respond to new consumer demands for healthy foods.

In Italy there is a particular attention for biodiversity, due to the high number of native animal and plant populations distributed throughout the whole country (Maiorano *et al*. 2007). Seventeen indigenous cattle breeds have been officially recognized by the Italian Ministry of Agriculture. Of particular interest is the situation of four cattle breeds farmed in extensive traditional systems in the two main Italian Islands, Sicily and Sardinia. The Sarda (SAR) breed is present in the Island of Sardinia since about 3,000 years BC. It originates from west Mediterranean cattle populations (mainly from the Iberic peninsula) with influences from North African and Middle East breeds (Della Maria 1936; Brandano *et al*. 1983a). At the end of the XIX century, crossbreeding with Brown Swiss (BSW) bulls imported from Switzerland and Modicana (MOD) bulls imported from Sicily were carried out in order to improve the

aptitude of SAR to draught, milk and meat production respectively. These crosses have led to the current Sardo-Bruna (SB) and Sardo-Modicana (SM) breeds, respectively. The three Sardinian breeds have been officially recognized in 1985 with the establishment of the Herd book. The current population size, based on the number of animals recorded in the Herd book, is 25,315 and 923 herds for the SAR, 2,822 and 150 herds for the SM, and 33,662 and 1,426 herds for the SB respectively (www.aia.it). The Modicana herdbook was established since 1952. Currently there are 5,209 animals recorded in the herd book, farmed in 235 herds (www.aia.it). An early genetic characterization of these breeds was carried out using morphologic measurements (Brandano *et al.*1983b), milk and blood protein polymorphisms (Brandano *et al.* 1983c). Recently SM and MOD were compared in a study on coat color genetic determinism using the *Melanocortin 1 receptor* gene (Guastella *et al.* 2011) and the distribution of Runs of Homozygosity (ROH) was studied in MOD by Mastrangelo *et al*. (2016).

The SAR, MOD, and BSW can be considered as founder breeds and SB and SM are the derived ones. In this work, a comparison between the five breeds is carried out using a medium density (50K) SNP panel in order to investigate the genetic diversity and in particular to assess the extent of diversity between pure-breeds and derived crosses. Moreover, gene discovery was performed in the genomic regions that exhibited difference between breeds.

Materials and methods

Animals and genotypic data

A total of 88 animals of five different breeds were genotyped in outsourcing with the Illumina BovineSNP50 beadchip: 22 BSW, 27 MOD, 19 SAR, 10 SB, and 12 SM, respectively. Genomic DNA was obtained from blood samples for SB, MOD, SM, and from nasal swab for SAR, using the NucleoSpin DNA rapidLyse Kit (Macherey-Nagel) according to manufacturer's instructions. For BSW animals, genotype data were generated within the SELMOL research project using the Genomix kit (Talent, Trieste, Italy). Animals of local breeds were randomly sampled from different herds located in various areas of Sardinia and Sicily. Given the difficulty in gathering large samples in local breeds, criteria used in the present work to include animals in the analysis were absence of relatedness, distribution in the territory, morphological appearance and information based on farmer interviews.

Since BSW animals were genotyped using Illumina BovineSNP50 v1 BeadChip in contrast to the other genotypic data (Illumina BovineSNP50 v2), common markers were retained and remapped on the UMD 3.1 release of the Bovine genome assembly. Only autosomal SNPs were considered. Quality control was performed with Plink 1.9 (Purcell *et al*. 2007). Animals with a call rate > 95% were retained. SNP selection was based on call rate (>97.5%), minor allele frequency (MAF>0.05), and significant deviation for Hardy -Weimberg equilibrium (P<0.00001). After quality control, 43,012 common SNPs between the two Beadchip versions were retained. Missing genotypes were imputed using Beagle 4 (Browning and Browning, 2016).

Heterozygosity, Minor allele frequency and Linkage Disequilibrium

Heterozygote count (HET) and the minor allele frequency (MAF) were calculated for each SNP separately by breed using Plink 1.9. Linkage disequilibrium (LD) between markers was calculated within 1000 kb distance (McKay *et al*. 2007) using Haploview (Barrett *et al*. 2005).

Multi-dimensional scaling and admixture analysis

The Multi-Dimensional scaling plot (MDS) and admixture analysis were performed using the Zanardi pipeline (Marras *et al.*, 2016) and "*ggplot2*" R package (Wickham, 2009). In MDS analysis, a principal component (PC) analysis is performed on the genomic correlation matrix G and PC scores are calculated for each individual. In order to confirm the animal classification in five different breeds, the K parameter of admixture was fixed at 5.

Wright Fixation Index and LOWESS

Ten pair-wise comparisons were performed using the Wright fixation index (FST) calculated using the equation proposed by Nei (1977):

$$FST = (HT - HS) / HT$$

where HT is the observed total heterozygosity and HS is the observed heterozygosity in each population, respectively. For the FST calculation, an in-house Python script was used. In order to simplify the graphic interpretation of raw FST data, a Locally Weighted Scatterplot Smoothing (LOWESS) procedure was used (Pintus *et al*. 2014). The LOWESS is a local smoothing regression in which the space of the independent

variable (in this case the progressive order of adjacent SNPs along the chromosome) is fragmented into different intervals for which separate regressions are fitted. The method is aimed at removing noise from raw data and at improving graphical representation. A smoothing parameter corresponding to an interval of 20 SNPs for each local regression was used.

A common problem when interpreting genetic difference metrics is the lack of proper statistical tests. Some authors have proposed to fix a threshold based on the Fst distribution (Kijas *et al*., 2012; Pintus *et al*., 2014). Although the distribution of raw Fst values tends to be skewed, LOWESS smoothed values could be considered approximately normally distributed. Thus, the significance threshold in the present work was set to three standard deviation from the mean. Such a stringent threshold was adopted considering the limited sample size.

Runs of homozygosity

Runs of Homozygosity (ROH) were detected using the Zanardi pipeline. Some constraints were fixed in order to limit the number of spurious ROH segments (Marras *et al*., 2015): the minimum length of ROH was set at 1 Mb, homozygous segments of minimum fifteen SNPs were considered and neither heterozygous or missing genotypes were allowed. The following ROH statistics were calculated by animal and by breed: number of ROH, the average ROH length (in Mb) and the sum of all ROH segments by animal (SROH, in Mb). ROH were grouped into five classes of length ($1 < Mb \leq 2$, $2 < Mb \leq 4$, $4 < Mb \leq 8$, $8 < Mb \leq 16$ and $Mb > 16$).

The ROH-based inbreeding coefficient (F_{ROH}) for each animal was calculated as

$$F_{ROH} = \frac{\sum S_{ROH(>8Mb)}}{Lgen}$$

where Lgen is the total length of genome. The minimum length of ROH to be included in the calculation was fixed to 8 Mb based on previous reports in cattle (Marras *et al*., 2015). Moreover, the ROH count per SNP (SNPROH), i.e. the number of animals having a given SNP included in a ROH (Nothnagel *et al*., 2010) was calculated. A threshold of 50% was fixed to consider a SNP_{ROH} value as significant.

Gene discovery

Gene discovery was performed in regions flagged by F_{ST} values exceeding the control chart upper limit. Intervals spanning 0.25 Mb upstream and downstream the significant marker were considered. Moreover, regions identified by ROH distribution were studied. In particular, markers having $SNP_{ROH} > 50\%$ within a breed were considered as significant and the region spanning 0.25 Mb upstream and downstream surrounding them was investigated. Annotated genes were retrieved from UCSC Genome Browser Gateway (http://genome.ucsc.edu./) and National Centre for Biotechnology Information (*NCBI*) (www.ncbi.nlm.nih.gov) databases.

Results

HET and MAF showed a little variation between the five considered breeds (Table 1).

Table 1. Mean value of heterozygosity (HET) and Minor allele frequency (MAF) in the five breeds.

	HET		MAF	
	Mean	s.d.	Mean	s.d.
BSW	0.318	0.011	0.232	0.010
MOD	0.348	0.008	0.249	0.006
SAR	0.335	0.011	0.252	0.005
SB	0.343	0.012	0.251	0.007
SM	0.347	0.013	0.251	0.006

BSW = Italian Brown Swiss; MOD = Modicana; SAR =Sarda; SB =Sardo Bruna; SM = Sardo Modicana.

BSW showed smallest values of both HET and MAF, whereas MOD and SAR exhibited the largest values for these parameters, respectively.

A clear distinction between the breeds could be observed along the first axis (PC1) of the MDS plot (Figure 1).

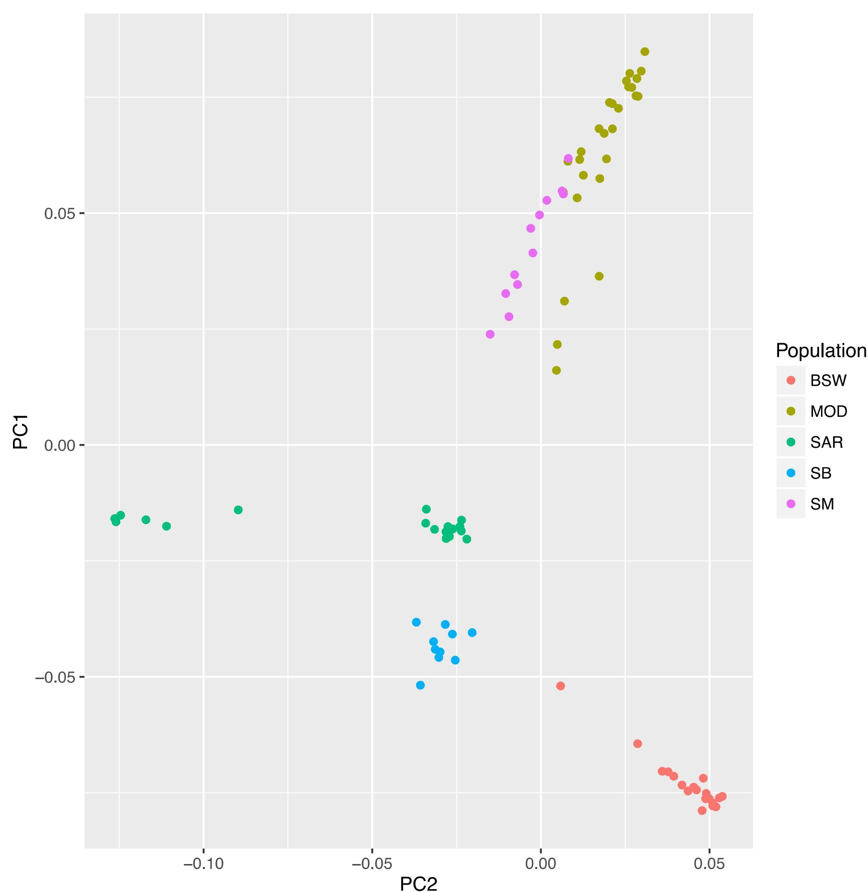


Figure 1 Multi-Dimensional Scaling plot of the five investigated breeds: Italian Brown Swiss (BSW), Modicana (MOD), Sarda (SAR), Sardo-Bruna (SB) and Sardo-Modicana (SM).

In particular the PC1, that explains about 5.4% of the total variance, depicts a geographic cline: starting from the bottom of the graph there are individuals from BSW (origin from the Switzerland, North of Italy), then SAR and SB (centre of Italy), and at the top SM and MOD (native of Sicily, Southern Italy). Furthermore, it could be seen that along this dimension, SM breed is more similar to MOD than SAR. The second axis (PC2), explaining about 3% of the total variance, highlights a separation within the SAR breed. The PC2 seemed to be able to discriminate animals according

to the percentage of SAR genetic contribution: an increase in PC2 scores indicates the passage from SAR purebred to crosses, and then to MOD and BSW breeds. Population structure analyzed by admixture (Figure 2) revealed a clear definition of BSW animals (95% assigned to a single cluster, the one of red color), and less precise for MOD and SAR (90% and 93% assigned to two different clusters, respectively).

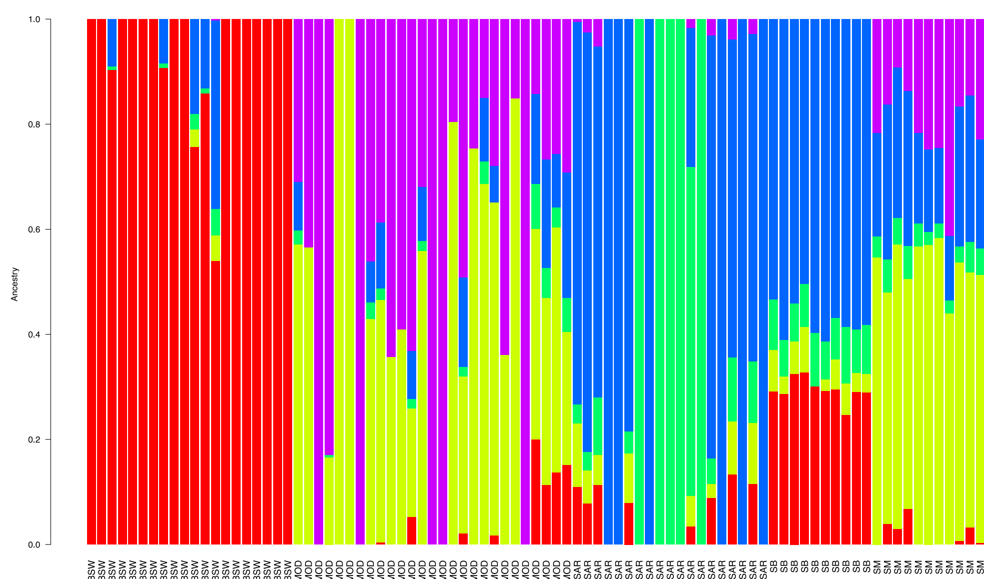


Figure 2 Genetic structure and admixture plot obtained through coefficients of individual membership to clusters (K=5) assumed to be present in the sample of investigated breeds. Red columns = cluster 1; Light green columns = cluster 2; Blue columns = cluster 3; Green columns = cluster 4; Purple columns = cluster 5.

Finally, also the derived breeds were grouped into two distinct clusters (70% of both SB and SM cattle). The LD pattern (Figure 3) shows the lowest value for MOD, the highest for BSW and SB, respectively.

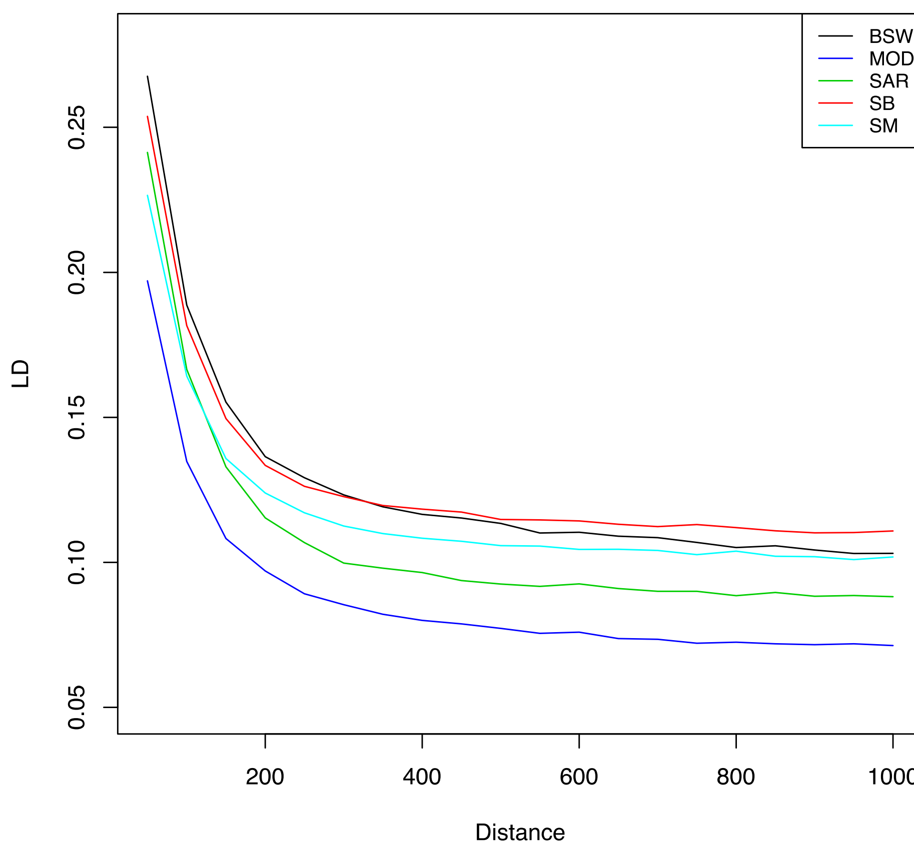


Figure 3 Average LD (r^2) between markers within an interval of 1000 kb in the five Italian cattle breeds: Italian Brown Swiss (BSW), Modicana (MOD), Sarda (SAR), Sardo-Bruna (SB) and Sardo-Modicana (SM).

The FST comparison with the largest number of significant SNPs was SB vs SM, whereas the smallest was observed for BSW vs SM (Table S1). Figure 4 reports Manhattan plots of FST predicted by LOWESS for the comparisons between pure breeds and crosses.

It can be observed that the highest FST values between BSW and SB were found for BTA6 (Figure 4a), with the top significant markers (Table S2) located between 38.20

and 38.83 Mb. In this region map some known genes controlling milk production traits (ABCG2, PKD2, SPP1, LAP3), and body size (NCAPG and LCORL) in cattle.

BTA8 and BTA13 showed the highest FST peaks in the SAR vs SB comparison (Figure 4b) with seven and three significant markers respectively (Table S2). In the region highlighted on BTA8 is located the microRNA2471 (MIR2471), whereas in the highlighted segment of BTA13 is annotated the Eukaryotic translation initiation factor 6 (EIF6) gene. SAR and SM were different mainly on BTAs 7, 14, and 21 (Figure 4c and Table S2). An interesting gene retrieved from the database was the Ubiquitin Protein Ligase E3A (UBE3A) that maps in the region between 2.1 and 2.3 Mb of BTA21.

As far as the comparison between SM and MOD is concerned (Figure 4d), the highest values of FST have been found on BTAs 5, 16 and 20 (Table S2). On BTA20 the region from 70.9 to 71.7 Mb presents a QTL associated with milk somatic cell score. Moreover, this segment contains several annotated genes, among which of interest is the Solute Carrier Family 9 Member A3 (SLC9A3). Finally, for the SM vs SB comparison the highest values of FST have been detected on chromosomes 7 and 24 (Figure 4e and Table S2). On BTA7, five significant markers define a region (47.2-47.3 Mbp) were the Transcription Factor 7 (T-Cell Specific, HMG-Box) (TCF7) gene maps.

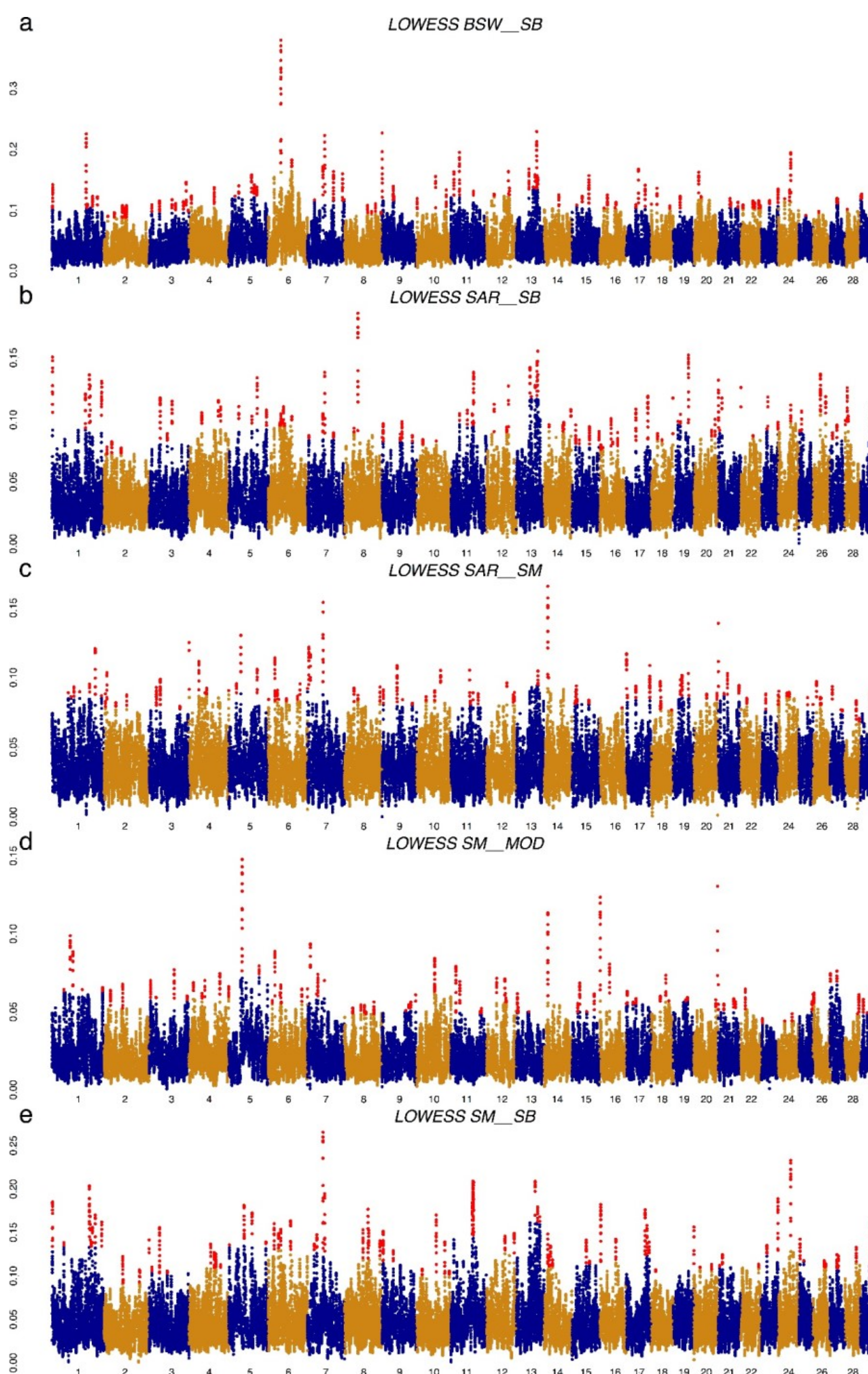


Figure 4 Manhattan plot of F_{ST} values predicted by the LOWESS. **a**) Comparison between Italian Brown and Sardo-Bruna. **b**) Comparison between Sarda and Sardo-Bruna. **c**) Comparison between Sarda and Sardo-Modicana. **d**) Comparison between Sardo-Modicana and Modicana. **e**) Comparison between Sardo-Modicana and Sardo-Bruna. Red color dots indicate significant F_{ST} values (i.e. greater than 3 standard deviations from the mean).

The total number of detected ROH (Table 2) exhibited a large variation between breeds, with MOD and SM having the largest and the smallest value, respectively. The BSW had the largest average ROH length, even if together with a huge variability as evidenced by the value of the standard deviation (Table 2).

Table 2. Statistics of ROH size and frequency in the five investigated cattle breeds.

	BSW	MOD	SAR	SB	SM
Average length (Mb)	3.9 ± 5.0	2.3 ± 1.8	2.9 ± 2.4	2.6 ± 2.3	2.4 ± 2.0
Average SNP/ROH	67.2 ±85.8	40.2 ±30.3	49.1 ±40.8	44.7 ±39.1	41.2 ±33.6
Number of ROH					
1-2 Mb	780	1270	834	423	447
2-4 Mb	404	571	420	220	195
4-8 Mb	231	242	251	87	83
8-16 Mb	138	34	74	21	13
>16 Mb	56	2	2	4	2
Total	1609	2119	1581	755	740

BSW = Italian Brown Swiss; MOD =Modicana; SAR =Sarda; SB =Sardo Bruna; SM = Sardo Modicana.

This breed had also the highest average number of SNP per ROH (Table 2). On the contrary MOD showed the smallest values of both statistics. As expected, most represented ROH classes in all breeds were those of length <4Mb (relative frequency ranging from 0.736 in BSW to 0.868 in MOD and SM, respectively). The largest number of ROH in the class of highest length (>16 Mb) was observed in BSW, and it was markedly larger than in all the other considered breeds (Table 2).

ROH count per SNP showed some interesting peaks along the genome. The highest peak was observed on BTA6 for BSW at approximately 30-40Mb (Figure 5a).

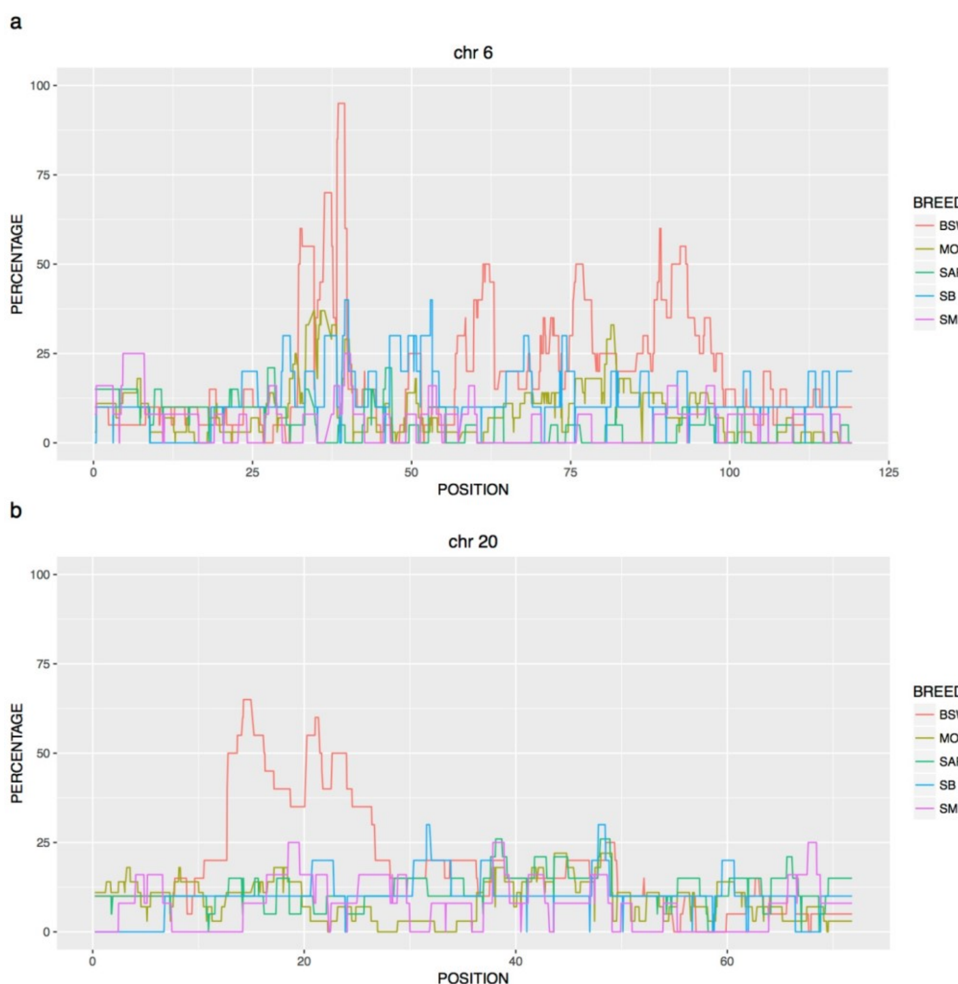


Figure 5. Occurrence of SNP counted in a ROH measured by the percentage of animals belonging to the five investigated breeds for which a particular SNP falls into a ROH versus the position along the chromosome. **a**) Comparison of BTA6. **b**) Comparison of BTA20.

In this region map several known genes as ABCG2, SPP1, LCORL, NCAPG. BSW exhibited another signal between 10-30 Mb on BTA20 (Figure 5b). Moreover, BTA1, BTA10 and BTA11 showed interesting signals of SNPs in homozygosity for over 50% of the animals. In particular, BSW showed a SNPROH peak on BTA1 (Figure 6a) between 103.5 and 105.5 Mb. On the same chromosome, a peak was detected for MOD at 139.0 Mb. On BTA10 an interesting homozygous region was observed in the SAR

breed between 72.2 and 72.8 Mb (Figure 6b). Among the genes that map in this region the Dehydrogenase/Reductase 7 (DHRS7) can be mentioned. Finally, the SB showed a relevant value of SNPROH on BTA11 (Figure 6c) between 65.0 and 67.0 Mb where the Ewing Tumor Associated Antigen 1 (ETAA1) was annotated.

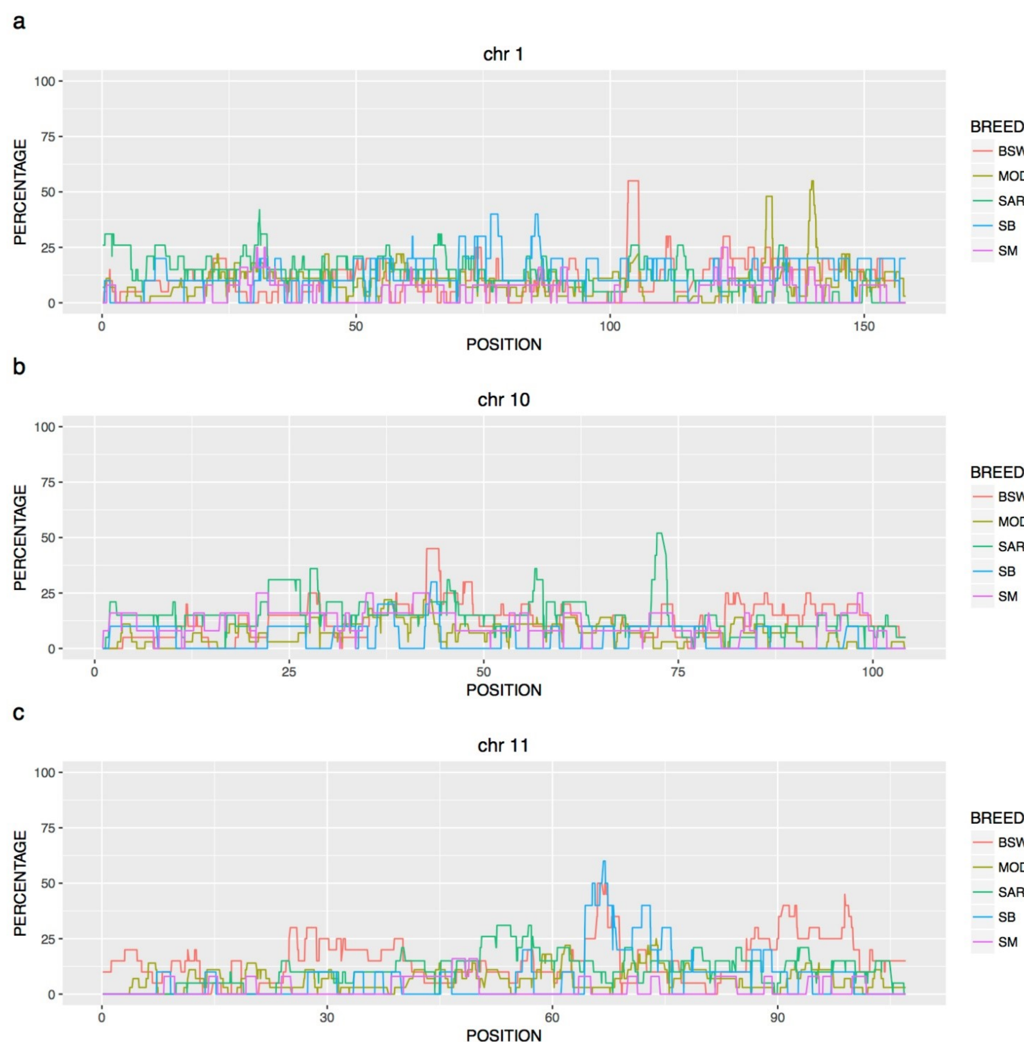


Figure 6 Occurrence of SNP counted in a ROH measured by the percentage of animals belonging to the five investigated breeds for which a particular SNP falls into a ROH versus the position along the chromosome. **a**) Comparison of BTA1. **b**) Comparison of BTA10. **c**) Comparison of BTA11.

BSW exhibited also the largest average FROH (Table 3) whereas the smallest value was observed by SM.

Table 3. ROH-based inbreeding coefficient (F_{ROH}) calculated using ROH>8Mb.

	F_{ROH}			
	Mean	s.d.	Max	Min
BSW	0.127	0.043	0.210	0.043
MOD	0.073	0.056	0.290	0.031
SAR	0.095	0.086	0.360	0.015
SB	0.080	0.078	0.282	0.019
SM	0.060	0.058	0.227	0.023

BSW = Italian Brown Swiss; MOD =Modicana; SAR =Sarda; SB =Sardo Bruna; SM = Sardo Modicana.

Discussion

The practice of crossbreeding has represented a major cause of gene flow across cattle populations, providing a relevant contribution to the shaping of worldwide current breeds. The history of the Sarda breed and its crosses with Modicana and Brown Swiss represents a typical example. Results of the present study confirm the genetic relationships between the considered breeds. The admixture analysis (Figure 2) clearly detected the five different genetic groups, highlighting the genetic background of the crossbred derived population in comparison of the original purebreds. Furthermore, the analyses of the genome features with different approaches gave useful insights on effects of selection and environmental adaptation on the cattle genome.

A first indication was provided by basic population statistics. The lower genetic variability exhibited by the BSW in comparison with the other two pure-breeds, SAR and MOD, was expected due to the intense artificial selection this breed has been subjected to in the last decades (www.anarb.it). A low allelic diversity for BSW cattle in comparison with other cattle breeds has been already reported (Schmid *et al*., 1999; Melka and Schenkel 2012).

The genome feature analysis carried out using the MDS decomposition, and the ROH detection highlighted an interesting structure of the considered sample of animals. The North-South geographical gradient highlighted by the first axis of the MDS is in agreement with several studies where a dimension reduction method is applied to molecular data on populations from different geographical origin (Price 2006; Chessa *et al*., 2009; Ciani *et al*., 2014). Also, the variation of the ROH statistics and of the inbreeding coefficient FROH exhibited the same cline. In particular the average ROH length, the average number of SNP per ROH, and the FROH showed an increase moving from South to North. This gradient was also confirmed by the LD analysis (Figure 3). Purfield *et al*. (2012) found a higher number of ROH in cattle breeds of British Isles compared to other European breeds and ascribed such a diversity to the closed population histories of these cattle. Results obtained in the present study can be probably due to a low effective population size of BSW and to the population history of the SAR, MOD, and their crosses. A geographical South-North gradient in ROH feature distribution has been observed also in human populations (Nothnagel *et al*., 2010), and it has been explained with the most pronounced genetic isolation of Northern populations compared to Mediterraneans. The second axis of the MDS

analysis highlights two clusters in the sample of Sarda cattle (Figure 1). Previous studies on this population highlighted a large morphological heterogeneity (Brandano *et al*., 1984). Moreover, in the traditional extensive cattle farming system of Sardinia it is not very common to exchange bulls between herds, resulting in a high average relatedness of individuals within farm and a low degree of kinship among farms.

Different degree of genetic relationships between original and derived breeds have been observed. The similarity between SM and MOD was quite expected (Figure 1). Although the first importation of MOD bulls from Sicily started at the end of the nineteenth century in the Montiferru area (Center-North Sardinia), it probably occurred again in more recent times and therefore the genetic component of Modicana purebred is still preserved into current SM. On the other hand, the separation between SB and the two founder breeds, i.e. BSW and SAR (Figure 1), seems to indicate an absence of recent genetic exchange.

The genetic history of the breeds is also depicted by other structural elements of their genome, as their linkage disequilibrium (Figure 3) and the extent of regions of autozygosity (Figure 5 and 6). The intensive genetic selection of BSW in comparison with the other investigated breeds resulted in the highest level of LD and in the largest values of all ROH statistics. These results agree with previous reports on this breed (Ferenčaković *et al*., 2013; Marras *et al*., 2015). A previous study on MOD breed reported a smaller value of F_{ROH} (Mastrangelo *et al*., 2016) but using different ROH settings (i.e. minimum number of SNP in a ROH equal to 40, minimum ROH length 4Mb, two missing SNP allowed in a ROH etc.). An interesting result is the distribution across individuals of specific ROHs, i.e. a segment that starts and ends exactly in the

same position. The largest ROH frequency was about 0.06 (Table 4) and it can be seen that in general local breeds tend to share ROH whereas the autozygous segment detected on BTA6 can be found only within the BSW breed. In particular, the latter ROH flagged a region where several known genes affecting milk traits are located.

Table 4. Most frequent ROH detected in the five breeds

BTA	Start	End	Length (Mb)	Frequency[1]	Breed
1	73924347	75505402	1.58	5	SB, MOD, SAR
29	23762023	25780595	2.02	5	SB, SM, MOD, SAR
6	32241952	34661866	2.41	5	BSW
9	27516531	28538817	1.02	5	SB, SM, SAR
9	821062	2677236	1.86	5	SM, MOD, SAR, BSW

BSW = Italian Brown Swiss; MOD =Modicana; SAR =Sarda; SB =Sardo Bruna; SM = Sardo Modicana.
[1]Number of individuals that possess the specific ROH across breeds

These results confirm the role of ROH as indicators not only of inbreeding but also of signatures of selection (Marras *et al*. 2014; Kim *et al*. 2015).

Signatures of selection were highlighted in the present study. Some of them flagged genome regions already detected in many studies on cattle. An example is represented by the markers exhibiting the largest FST values in the BSW vs SB comparison, all located in the region of BTA6 spanning between 36-39Mb that harbors some known genes controlling milk production traits (*ABCG2*, *PKD2*, *SPP1*, *LAP3*) (Olsen *et al*., 2005; Cohen-Zinder *et al*., 2005) and body size (*NCAPG* and *LCORL*) (Takasuga

2016) (Table S2). This region was also flagged by a significant value of ROH count per SNP in BSW.

Other two well-known selection signatures were detected in BSW on BTA6 (Figure 5a) by SNPROH significant values (>50%). The first was located at around 70 Mb, where the *V-Kit Hardy-Zuckerman 4 Feline Sarcoma Viral Oncogene Homolog* (*KIT*) locus maps. This gene is involved in mammalian coat color determinism (Fontanesi *et al*., 2010; Stella *et al*., 2010). The second signature of selection, at around 85 Mb, identified the caseins cluster (Blott *et al*., 2003). Another interesting peak value of SNPROH was found on BTA20 (14-25 Mb) (Figure 5b), in a region where a large QTL associated with milk protein percentage was reported (Ashwell *et al*., 2004). Among the several genes that map in this region, of interest is the *Importin 11* (*IPO11*) locus. This gene has been found to be associated with the displacement of the abomasum in German Holstein cattle breed (Mömke *et al*., 2013).

Interestingly, the FST pairwise comparison between the SAR and the SB did not detect SNPs located in genomic regions known to contain genes associated with milk production traits. These results, together with the pattern highlighted by the MDS, confirm that current SB is closer to SAR than to BSW, probably due to backcrossing.

Of interest are the signatures of selection found in the comparisons between local breeds. Some of them include interesting genes that were found to be associated with fitness traits. In the comparison between the SAR and its derived SB, the seven highly significant SNPs found on BTA8 between 40.4 and 40.6 Mbp (Figure 4b) identified a region where maps the *microRNA2471* (*MIR2471*). In animals, microRNAs are molecules involved in diverse biological processes such as development, cell

differentiation, proliferation and metabolism. They are among major post-transcriptional regulators of gene expression through promoting mRNA degradation or translational repression (Glazov *et al*., 2009; Guo *et al*., 2010; Meunier *et al*., 2013). Recently they have been found to be essential for the regulation of the immune response (Xiao and Rajewsky, 2009). The highest peak of F_{ST} comparison between the SAR and the other derived breed, the SM, was detected on BTA14 (Figure 4c and Table S2) in a region where maps the *gasdermin C* (*GSDMC*) locus. This gene was associated to UV-protective eye pigmentation in Fleckvieh cattle (Pusch *et al*., 2012). Another peak was located on BTA21, between 2.1 and 2.3 Mb, where the *Ubiquitin Protein Ligase E3A* (*UBE3A*) gene is annotated. This locus has been associated with the calving ease (Pausch *et al*., 2011; Meszaros *et al*., 2016) in cattle. This trait represents very often a distinguishing feature in indigenous breed that are mainly reared in extensive and semi-extensive systems (Boggio *et al*., 1988).

Other genes detected in the local breeds are related to milk production traits and fatty acid metabolism. Among genetic differences found between SM and MOD, of interest is the region located on BTA20, from 70.9 to 71.7 Mb. Among the annotated genes, is worth of mention the *Solute Carrier Family 9 Member A3* (*SLC9A3*), involved in the rumen sodium transport (Rabbani *et al*. 2011). A high Na^{2+} tissue concentration improves milk production in warm/humid conditions (Granzin and Gaughan, 2002). Moreover, a QTL associated with milk somatic cell score was reported in this region (Durán Aguilar *et al*., 2016). The comparison between the two derived breeds, i.e., SB vs SM. found a selection signature defined by five significant markers (47.2-47.3 Mbp) on BTA7, where maps the *Transcription Factor 7 (T-Cell Specific, HMG-Box)* (*TCF7*)

gene. Recently, this locus was associated with milk production in Chinese Holstein (Mao *et al*., 2015).

An interesting candidate gene highlighted by SNP_{ROH} in the SAR breed on BTA10 is the *Dehydrogenase/Reductase 7* (*DHRS7*) locus. It catalyzes the oxidation/reduction of a wide range of substrates, including retinoid and steroids (Haeseleer and Palczewski, 2000) and it has high expression levels in adipocytes and skeletal muscles (Wu *et al*., 2009). In addition, this gene is responsible for the final step in the cholesterol production (Porter, 2000). This gene was already associated in Nellore cattle with the intramuscular fat deposition and composition (Cesar *et al*., 2014). Finally, another signature of selection that included a gene involved in fatty acid metabolism was found in the SAR vs SB comparison (three significant markers on BTA13 between 65.1 and 65.2 Mb) (Figure 4b). This region harbors the *Eukaryotic translation initiation factor 6* (*EIF6*) locus. This gene controls fatty acid synthesis and glycolysis in tissues responsive to insulin such as adipose and muscular.

Conclusion

Results of the present work confirm the usefulness of genome structural features in deciphering the genetic architecture of livestock breeds. The different approaches used to explore medium density SNP genotypes gave a comprehensive picture of genetic relationships between the three original and the two derived breeds, reflecting their recent genetic history. As expected, a larger heterogeneity was highlighted for the local breeds. Signatures of selection located in genomic regions harboring candidate genes for milk production traits have been detected in the comparisons

involving the specialized BSW breed, whereas for local breeds the flagged genes involved in fitness and fatty acid metabolism. The study confirmed the importance of these populations as reservoir of biodiversity and as models for studying the genetic basis of adaptability.

Acknowledgments

This Research was funded by the Banco di Sardegna Foundation and by the Regional Government of Sardinia (project grant 07/G1-20, POR-FSE 2007-13). The authors wish to thank Dr. Salvatore Mele of Agenzia FORESTAS for his contribution to the animal sampling.

References

Ajmone-Marsan P., Garcia J. F., & Lenstra J.A. (2010) On the origin of cattle: How aurochs became cattle and colonized the world. Evolutionary Anthropology: Issues, News, and Reviews 19, 148-157.

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

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

Blott S., Kim J. J., Moisio S., Schmidt-Küntzel A., Cornet A., Berzi P., Cambisano N., Ford C., Grisart B., Johnson D., Karim L., Simon P., Snell R., Spelman R., Wong J., Vikki J., Georges M., Farnir F., & Coppieters W. (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163, 253–266.

Boggio F., Pracchi R., & Asole A. (1998) Atlante economico della Sardegna 1. Jaca Book, Edizioni Universitarie Jaca, Italia.

Borg I., & Groenen P. (2003) Modern multidimensional scaling: theory and applications. Journal of Educational Measurement 40, 277-280.

Brandano P., Asara P., Pulina G., Bolla P., & Crimella C. (1983a). The Sardinian cattle. 1. Morphological and biological characters. Annals of the Faculty of Agriculture of the University of Sassari 30, 161-177.

Brandano P., Asara P., Pulina G., Bolla P., & Crimella C. (1983b) The Sardo-Modicana cattle Breed. 1. Annals of the Faculty of Agriculture of the University of Sassari 30, 197-214.

Brandano P., Pulina G., & Asara P. (1983c) The indigenous cattle of Sardinia. Breeds and herds characteristics. Annals of the Faculty of Agriculture of the University of Sassari 30, 1-23.

Brotherstone S., & Goddard M. (2005). Artificial selection and maintenance of genetic variance in the global dairy cow population. Philos Trans R Soc Lond B Biol Sci. 360: 1479–1488.

Browning B. L., & Browning S. R. (2016) Genotype imputation with millions of reference samples. The American Journal of Human Genetics 98, 116-126.

Cesar A. S., Regitano L. C., Mourão G. B., Tullio R. R., Lanna D. P., Nassu R. T., Mudado M. A., Oliveira P. S. N., do Nascimento M. L., Chaves A. S., Alencar M. M., Sonstegard T. S., Garrick D.J., Reecy J. M., & Coutinho L. L. (2014) Genome-wide association study for intramuscular fat deposition and composition in Nellore cattle. BMC genetics 15, 39.

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

Ciani E., Crepaldi P., Nicoloso L. *et al*. (2014) Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. Animal Genetics, 45, 356-366.

Cohen-Zinder M., Seroussi E., Larkin D. M., Loor J. J., der Wind A. E., Lee J. H., Drackley J. K., Band M. R., Hernandez A. G., Shani M., Lewin H. A., Weller J.

I., & Ron M. (2005) Identification of a missense mutation in the bovine *ABCG2* gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Research 15, 936-944.

Della Maria G. (1936) Ancient stories of Sardinian bovine breed. Rivista di Zootecnia 13, 47-57.

Durán Aguilar M., Román Ponce S. I., Ruiz López F. J., González Padilla E., Vásquez Peláez C. G., Bagnato A., & Strillacci M. G. (2016) Genome-wide association study for milk somatic cell score in Holstein cattle using copy number variation as markers. Journal of Animal Breeding and Genetics 134, 49-59.

Ferenčaković M., Hamzić E., Gredler B., Solberg T. R., Klemetsdal G., Curik I., & Sölkner J. (2013) Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. Journal of Animal Breeding and Genetics 130, 286–293.

Fontanesi L., Scotti E., & Russo V. (2010) Analysis of SNPs in the KIT Gene of Cattle with Different Coat Colour Patterns and Perspectives to Use These Markers for Breed Traceability and Authentication of Beef and Dairy Products. Italian Journal of Animal Science 9, e42.

Giovambattista G., Ripoli M. V., Peral-Garcia P., & Bouzat J. L. (2001) Indigenous domestic breeds as reservoirs of genetic diversity: the Argentinean Creole cattle. Animal Genetics 32, 240–247.

Glazov E. A., Kongsuwan K., Assavalapsakul W., Horwood P. F., Mitter N., & Mahony, T. J. (2009) Repertoire of Bovine miRNA and miRNA-Like Small Regulatory RNAs Expressed upon Viral Infection. Plos One 4, e6349.

Granzin B. C., & Gaughan J. B. (2002) The effect of sodium chloride supplementation on the milk production of grazing Holstein Friesian cows during summer and autumn in a humid sub-tropical environment. Animal Feed Science and Technology 96, 147-160.

Guastella A. M., Sorbolini S., Zuccaro A., Pintus E., Bordonaro S., Marletta D., & Macciotta N. P. P. (2011) Melanocortin 1 receptor (MC1R) gene polymorphisms in three Italian cattle breeds. Animal Production Science 51, 1039–1043.

Guo L., & Lu Z. (2010) Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data. Computational Biology and Chemistry 34, 165-171.

Haeseleer F., & Palczewski K. (2000) Short-chain dehydrogenases/reductases in retina. Methods in enzymology 316, 372-383.

Kijas J. W., Lenstra J. A., Hayes B., Boitard S., Porto Neto L. R., San Cristobal M., Servin B., McCulloch R., Whan V., Gietzen K., Paiva S., Barendse W., Ciani E., Raadsma H., McEwan L., Dalrymple B., & International Sheep Genomics Consortium (2012) Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PloS Biology* 10, e 10001258.

Loftus R. T., MacHugh D. E., Bradley D. G., Sharp P. M., & Cunningham P. (1994) Evidence for two independent domestications of cattle. Proceedings of the National Academy of Sciences 91, 2757–2761.

Maiorano L., Falcucci A., Garton E.O., & Boitani L (2007) Contribution of the Natura 2000 network to biodiversity conservation in Italy. Conservation Biology 21, 1433-1444.

Mao Y., Zhu X., Xin S., Zhang M., Wang X., Cheng D., Zhang H., Konig S., Yang Z., & Yang L. (2015) Polymorphisms in the promoter of interleukin-12β2 and interleukin-23 receptor genes influence milk production traits in Chinese Holstein cows. Livestock Science 178, 1–8.

Marras G., Gaspa G., Sorbolini S., Dimauro C., Ajmone-Marsan P., Valentini A., Williams J. L., & Macciotta N. P. P. (2015) Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Animal Genetics 46, 110–121.

Marras G., Rossoni A., Schwarzenbacher H., Biffani S., Biscarini F., & Nicolazzi E. L. (2016) Zanardi: an open-source pipeline for multiple-species genomic analysis of SNP array data. Animal Genetics 48, 121-128.

Mastrangelo S., Di Gerlando R., Tolone M., Tortorici L., Sardina M. T., & Portolano B. (2014) Genome wide linkage disequilibrium and genetic structure in Sicilian dairy sheep breeds. BMC Genetics 15, 108.

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

McKay S. D., Schnabel R. D., Murdoch B. M., Matukumalli L. K., Aerts J., Coppieters W., Crews D., Dias Neto E., Gill C. A., Gao C., Mannen H., Stothard P., Wang

Z., Van Tassell C. P., Williams J. L., Taylor J. F., & Moore S. S. (2007) Whole genome linkage disequilibrium maps in cattle. BMC Genetics 8, 74.

Medugorac I., Medugorac A., Russ I., Veit-Kensch C. E., Taberlet P., Luntz, B., Mix H. M., & Förster M. (2009) Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. Molecular Ecology 18, 3394-3410.

Melka M. G., & Schenkel F. S. (2012) Analysis of genetic diversity in Brown Swiss, Jersey and Holstein populations using genome-wide single nucleotide polymorphism markers. BMC Research Notes 5, 161.

Mészáros G., Taferner R., & Sölkner J. (2016) Pleiotropic and epistatic interactions between stillbirth and calving ease in cattle. Acta Agriculture Slovenica 5, 56.

Meunier J., Lemoine F., Soumillon M., Liechti A., Weier M., Guschanski K., Hu H., Khaitovich P., & Kaessmann, H. (2013) Birth and expression evolution of mammalian microRNA genes. Genome Research 23, 34-45.

Mömke S., Sickinger M., Lichtner P., Doll K., Rehage J., & Distl O. (2013) Genome-wide association analysis identifies loci for left-sided displacement of the abomasum in German Holstein cattle. Journal of Dairy Science 96, 3959–3964.

Nei M. (1977) F-statistics and analysis of gene diversity in subdivided populations. Annals of Human Genetics 41, 225-233.

Nothnagel M., Lu T. T., Kayser M., & Krawczak M. (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Human Molecular. Genetics 19, 2927–2935.

Olsen H. G., Lien S., Gautier M., Nilsen H., Roseth A., Berg P. R., Sundaasen K. K., Svendsen M., & Meuwissen T. H. E. (2005) Mapping of a Milk Production Quantitative Trait Locus to a 420-kb Region on Bovine Chromosome 6. Genetics 169, 275–283.

Pausch H., Wang X., Jung S., Krogmeier D., Edel C., *et al*. (2012) Identification of QTL for UV-Protective Eye Area Pigmentation in Cattle by Progeny Phenotyping and Genome-Wide Association Analysis. PLoS ONE 7: e36346.

Pausch H., Flisikowski K., Jung S., Emmerling R., Edel C., Götz K. U.. & Fries R. (2011) Genome-Wide Association Study Identifies Two Major Loci Affecting Calving Ease and Growth-Related Traits in Cattle. Genetics 187, 289–297.

Pintus E., Sorbolini S., Albera A., Gaspa G., Dimauro C., Steri R., Marras G.. & Macciotta N. P. P. (2014) Use of locally weighted scatterplot smoothing (LOWESS) regression to study selection signatures in Piedmontese and Italian Brown cattle breeds. Animal Genetics 45, 1–11.

Porter F. D. (2000) RSH/Smith–Lemli–Opitz syndrome: a multiple congenital anomaly/mental retardation syndrome due to an inborn error of cholesterol biosynthesis. Molecular Genetics and Metabolism 71, 163-174.

Price A. L., Patterson N. J., Plenge R. M., Weimblatt M. E., Shadick N. A.. & Reich D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38, 904-909.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., Bender D., Maller J., Sklar P., de Bakker P. I. W., Daly M. J., & Sham P. C. (2007) PLINK: a

toolset for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81, 559–575.

Purfield D. C., Berry D. P., McParland S., & Bradley D. G. (2012) Runs of homozygosity and population history in cattle. BMC Genetics 13, 70.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ruto E., Garrod G., & Scarpa R. (2008) Valuing animal genetic resources: a choice modeling application to indigenous cattle in Kenya. Agricultural Economics 38, 89–98.

Scherf B. D. (2000) World watch list for domestic animal diversity (No. Ed. 3) Food and Agriculture Organization (FAO).

Schmid B. M., Saitbekova N., Gaillard C., & Dolf G. (1999) Genetic diversity in Swiss cattle breeds. Journal of Animal Breeding and Genetics 116, 1-8.

Stella A., Ajmone-Marsan P., Lazzari B., & Boettcher P. (2010) Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production. Genetics 185, 1451.

Takasuga A. (2016) PLAG1 and NCAPG-LCORL in livestock. Animal Science Journal 87, 159–167.

Upadhyay M. R., Chen W., Lenstra J. A., Goderie C. R. J., MacHugh D. E., Park S. D. E., Magee D. A., Matassino D., Ciani F., Megens H. J., van Arendonk J. A. M., Groenen M. A. M., & European Cattle Genetic Diversity Consortium, Crooijmans R. P. M. A. (2016) Genetic origin, admixture and population history

of aurochs (*Bos primigenius*) and primitive European cattle. Heredity 118, 169-176.

Wickham H. (2009) Gplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Xiao C., & Rajewsky K. (2009) MicroRNA control in the immune system: basic principles. Cell 136, 26-36.

Zeder M. (2017) Domestication and early agriculture in the Mediterranean basin: Origins, diffusion, and impact. Proceedings of the National Academy of Sciences 105, 11592-11604.

Appendix Chapter 4

Genome-wide variability and selection signatures in Italian island cattle breeds

Table S1. Significant markers per chromosome detected in the FST comparisons.

Table S2. Top significant markers of each FST pairwise comparison.

Table S1. Significant markers per chromosome detected in the FST comparisons.

Chromosome	BSW *vs* SB	SAR *vs* SB	SAR *vs* SM	SM *vs* MOD	SM *vs* SB
1	40	46	21	26	60
2	47	15	22	20	26
3	40	26	29	31	23
4	17	22	16	31	28
5	27	24	23	20	21
6	27	27	35	30	34
7	40	27	29	18	36
8	20	15	16	28	32
9	25	31	23	17	23
10	21	5	14	24	29
11	24	27	19	25	52
12	11	13	11	22	18
13	38	32	8	14	16
14	6	16	13	14	26
15	20	23	19	22	14
16	15	24	1	20	23
17	17	22	33	17	29
18	8	9	14	13	2
19	8	19	22	7	6
20	9	13	10	8	10
21	13	16	22	22	9
22	24	3	5	9	3
23	14	11	10	5	18
24	20	6	10	6	11
25	3	10	3	6	9
26	2	23	13	16	13
27	13	9	11	12	14
28	1	17	14	12	6
29	12	14	8	8	11
Total	562	545	474	503	602

BSW = Italian Brown Swiss; MOD =Modicana; SAR =Sarda; SB =Sardo Bruna; SM = Sardo Modicana.

Table S2. Top significant markers of each FST pairwise comparison.

Marker	BTA	Position	Value	Breeds
BTA-100891-no-rs	6	38,689,886	0.379	BSW-SB
ARS-BFGL-NGS-112812	6	38,627,070	0.370	BSW -SB
Hapmap26308-BTC-057761	6	38,576,012	0.363	BSW -SB
Hapmap43470-BTA-114677	6	38,746,212	0.360	BSW -SB
Hapmap30134-BTC-034283	6	38,464,203	0.346	BSW -SB
Hapmap27083-BTC-041166	6	38,825,835	0.334	BSW -SB
Hapmap26555-BTC-033429	6	38,366,100	0.332	BSW -SB
Hapmap26258-BTC-033509	6	38,343,712	0.316	BSW -SB
Hapmap26259-BTC-033526	6	38,321,808	0.300	BSW -SB
Hapmap29922-BTC-033565	6	38,286,952	0.274	BSW -SB
ARS-BFGL-NGS-5595	7	47,349,962	0.265	SB-SM
ARS-BFGL-NGS-73679	7	47,313,107	0.260	SB-SM
BTA-78954-no-rs	7	47,384,327	0.256	SB-SM
ARS-BFGL-NGS-20141	7	47,274,866	0.254	SB-SM
ARS-BFGL-NGS-12557	7	47,252,135	0.235	SB-SM
Hapmap59568-rs29024289	24	37,844,515	0.233	SB-SM
ARS-BFGL-NGS-49210	24	37,819,368	0.230	SB-SM
BTB-00888233	24	37,763,861	0.222	SB-SM
BTB-00887818	24	37,938,500	0.217	SB-SM
BTB-00887858	24	37,958,693	0.214	SB-SM
Hapmap33220-BTA-149236	8	40,585,048	0.186	SAR-SB
Hapmap33243-BTA-158375	8	40,610,167	0.182	SAR-SB
Hapmap59547-rs29026130	8	40,538,141	0.181	SAR-SB
Hapmap36155-SCAFFOLD226597_1989	8	40,470,452	0.174	SAR-SB
ARS-BFGL-NGS-91112	8	40,681,934	0.170	SAR-SB
ARS-BFGL-NGS-8821	8	40,422,559	0.170	SAR-SB
Hapmap27456-BTA-158380	8	40,389,762	0.166	SAR-SB
Hapmap23621-BTC-064696	14	11,776,673	0.165	SAR-SM
Hapmap27348-BTC-064649	14	11,816,360	0.157	SAR-SM
ARS-BFGL-NGS-100338	13	65,164,792	0.155	SAR-SB
ARS-BFGL-NGS-790	13	65,201,054	0.155	SAR-SB
ARS-BFGL-NGS-89409	13	65,236,809	0.155	SAR-SB
BTB-01846474	7	47,717,578	0.154	SAR-SM
Hapmap36733-SCAFFOLD230838_1182	14	11,736,525	0.152	SAR-SM
Hapmap24455-BTC-064551	14	11,848,870	0.150	SAR-SM
Hapmap48064-BTA-73407	5	40,554,369	0.148	SM-MOD
BTB-00226032	5	40,553,720	0.148	SM-MOD

Hapmap50586-BTA-118480	7	47,617,783	0.147	SAR-SM
ARS-BFGL-NGS-17803	14	11,881,261	0.143	SAR-SM
UA-IFASA-8207	5	40,580,237	0.143	SM-MOD
Hapmap42591-BTA-57149	14	11,711,053	0.143	SAR-SM
ARS-BFGL-NGS-65543	5	40,512,359	0.139	SM-MOD
ARS-BFGL-NGS-53975	21	2,151,256	0.139	SAR-SM
Hapmap53246-rs29026986	5	40,605,553	0.138	SM-MOD
ARS-BFGL-NGS-112232	14	11,909,609	0.133	SAR-SM
BTB-01700124	5	40,638,405	0.132	SM-MOD
ARS-BFGL-NGS-118166	20	71,793,734	0.130	SM-MOD
BTB-01700108	5	40,664,008	0.127	SM-MOD
ARS-BFGL-NGS-96258	16	1,328,125	0.123	SM-MOD
Hapmap35625-SCAFFOLD312099_6800	16	1,296,954	0.120	SM-MOD

BSW = Italian Brown Swiss; MOD =Modicana; SAR =Sarda; SB =Sardo Bruna; SM = Sardo Modicana.

CHAPTER 5

The use of different statistical approaches highlights consensus signatures of selection harboring diversity among sheep breeds

Abstract

In a population, demographic events such as bottlenecks, genetic drifts, migrations, inbreeding and artificial selection alter allele frequencies and combinations. On this genetic variation, natural selection and local adaptation act by ensuring the transmission of favorable genotypes and leading to an increase in physical fitness.

In domestic animals, genetic diversity has often been tackled looking only signs of artificial selection (for example comparing groups with different productive capacity) ignoring the process of adaptive evolution because difficult to investigate. Sheep breeds, due to their pronounced ability to tolerate different climates and environments, to adapt to different production systems, and owing to different levels of selection pressures, represent a good animal model to investigate signatures of selection and adaptation. In this study, genomic regions subject to selection were investigated by using three statistical approaches were used. Fixation index for population differentiation, the canonical discriminant analysis and runs of homozygosity methodologies were applied to twenty sheep breeds representative of the Italian ovine population.

Obtained results revealed the presence of selection signals harboring known genes (*RXFP2, MC1R, PRL, ABCG2, EDAR, PLAG1, NPR2).* New genes involved in morphological traits such as body size (*RALY*), skull development (*GAB1*, *NPR3*), and milk, meat or fiber production (*SGCA, SGCD*, *MFG-E8, EGFR*) were highlighted. Moreover, selective sweeps in *loci* related to sensory perception (*TAS* family, *ORs* family, *TRPM8*), circadian rhythm (*CLOCK, PER3*), and diseases resistance (*DEFB134, IL6*, *CDH26, PRP/PRND*) were found.

Selection signatures highlighted in this work suggest in sheep the action of artificial selection but also denote strong adaptation to the environment.

Keywords

Ovine, selection signatures, SNP polymorphisms, productive traits,

adaptation to environment

Introduction

Animal domestication occurred 8,000-12,000 years ago (Vigne, 2011). It can be considered the first attempt of genetic improvement performed by man. Initially, this practice was unconscious and based on phenotypic traits such as coat color, size or behavioral features like docility. Afterwards, other more conscious and profitable criteria such as flexible diet, fast growth rate and adaptability to captivity governed the choice of individuals for breeding. Therefore, such an empirical practice evolved into a methodical approach called artificial selection aimed at pursuing a definite standard of breed. The main effect of selection, either natural or artificial, is the over-time change of frequency and type of variants between or within a population. At genome level, this effect results in the appearance of areas (selective sweeps) in which signatures left by selection are recognizable with adequate statistical metrics (Oleksyk *et al.*, 2010).

In farm animals, the main purpose of selection is to increase the number of individuals who exhibit a desired feature, causing however, a permanent effect in the population genetic variability. On the other hand, selection may lead to an increase of phenotypic diversity between and within domestic breeds. An emblematic example is the morphological variability in body size, coat color, tail or legs length, floppy ears, wavy or curly hairs shown by domestic animals compared to their wild ascendants (Andersson, 2001; Groeneveld *et al.*, 2010).

Natural and artificial selection have therefore played a pivotal role in the evolution of domestic species. Over the centuries, for many species of domestic animals such as dog (Akey *et al.*, 2010), cattle (Brotherstone and Goddard, 2010; Flori *et al.*, 2009),

horse (Petersen *et al*., 2013; Metzger *et al*., 2015) and chicken (Rubin *et al*., 2010) artificial selection and breeding have accelerated evolutionary processes resulting in highly productive strains. For other species, such as sheep and goat, adaptation to the environment and natural selection were more effective than artificial selection in shaping breeds (Kim *et al*., 2014).

Decoding the nature of genetic differentiation, understanding the mechanisms involved in it and identifying at molecular level genes/nucleotides underlying quantitative and adaptive traits are main goals of evolutionary and population genetics (Lee *et al*., 2014). However, it is difficult to assess whether a specific feature is the result of natural or artificial selection.

The current availability of advanced genomic technologies, statistical approaches and calculation tools allows to manipulate huge amounts of molecular data, thus enabling the investigation of selective loci, the survey of mechanisms underlying genetic variability and to highlight levels of differentiation within/between breeds. However, the choice of the best metric to use for selective sweep mapping is challenging when genomic data are considered. Statistical tests based on different approaches can highlight different selective events. For example, metrics based on site frequency spectrum detect skew in the allele frequency distribution and are best suited to capture sweeps close to the fixation, whereas those based on the extent of LD are able to highlight incomplete and ongoing sweeps (Biswas and Akey, 2006). Recently, studies on population and evolutionary genomics have addressed this difficulty with the application of multiple approaches simultaneously (Ma *et al*., 2015).

Ovine were among the first animals to live with mankind and several studies date the primordial events of domestication of wild mouflon in the ancient Mesopotamia (Hiendleder *et al*., 2002; Chessa *et al*., 2009; Pariset *et al*., 2011). Sheep represented a source of food and commodities, assuring availability of fresh meat and milk and of secondary products such as skin, horns and wool useful for handicraft creations. Due to their reduced size, to the ability to tolerate different climates and environments and to adapt to different production systems, sheep have colonized many terrestrial habitats (Kijas *et al*., 2009). Currently, sheep are one of the most widespread ruminant species in the world with about 200 different pure breeds and 400 composite ones (Rasali *et al*., 2005) and represent a large part of livestock productions for many countries.

The Italian sheep stock, with about 8 million heads divided into more than 50 different breeds with different productive capacity and morphological traits (www.assonapa.it), is characterized by a relevant diversity. High levels of past gene flows and admixture, strong and continuous north-to-south geographic cline, long history of traditional farming in accordance with local environmental conditions, natural mating and moderate artificial selection pressure have contributed to the today breeds make-up (Italian Institute of Statistics, 2010; Ciani *et al*., 2013). For all these reasons, Italian ovine breeds can represent the ideal subjects for population diversity studies.

In Italy, 95% of the sheep are farmed in the Mediterranean south-central and insular areas. Sheep farming is diversified according to different environmental conditions. A sedentary system (with herds of 500-1000 animals) is the most widespread in the center of the peninsula. The transhumance system, with flocks of 1000-3000 heads, is instead a widespread form of farming in southern and insular Italy. Contrariwise, in

alpine areas, sheep graze in the mountains in summer and use fodder during the winter season (Malorgio *et al*., 1995; Pardini and Nori, 2011).

The aim of this study was to investigate signatures of selection in sheep breeds representative of the Italian ovine stock by using three different metrics used to detect signatures of selection in livestock: the classical fixation index (F_{ST}) method proposed by Nei (1977) (Kim *et al*., 2014; Liu *et al*., 2016), a multivariate approach using Canonical Discriminant Analysis (CDA) (Sorbolini *et al*., 2016; Dimauro *et al*., 2015) and a haplotype-frequency based approach by Runs of Homozygosity (ROHs) (Marras *et al*., 2015; Iacolina *et al*., 2016). These three methods were chosen because able to investigate the genetic diversity at different levels: differences between groups (F_{ST}), genetic diversity both intra and inter groups (CDA), and diversity within group (ROH).

Materials and methods

Breeds and animals

A total of 20 Italian sheep breeds were chosen for the analysis according to their geographic distribution, morphology and production purpose (Table S1). Breeds have been identified in order to have a complete representation of the Italian ovine population and of the different geographical areas of Italy. A total of 496 animals were sampled from different flocks to avoid as much as possible closely related individuals. In this study animals belonging to different breeds were grouped into four types of production purposes: milk (140), meat (117), dual (191) and wool (48) (Table S1).

Genotyping

Genotypes were provided by the Italian Sheep Consortium (BiOvIta). Animals were genotyped using the OvineSNP50 BeadChip (Illumina, Inc.) (http://www.illumina.com). Data editing was performed using PLINK software. Markers were retained if SNP call rate and SNP minor allele frequency (MAF) were ≥ 99% and if mapped on autosomes. A threshold of ≥ 90% for animal call rate was considered. Missing data were imputed by Beagle 4 software (Browning and Browning, 2016). At the end of editing procedure, a total of 46.827 SNPs was retained for the analysis.

Population stratification and admixture

Genetic diversity, genetic structure and admixture analyses of the breeds and individuals considered in this study were recently investigated in a genome-wide survey (Ciani *et al*., 2013) and therefore, these issues have not been further dealt.

Fixation Index (F_{ST})

Pairwise comparisons between production aptitudes (meat *vs* wool, wool *vs* milk, dual *vs* wool, meat *vs* milk, dual *vs* milk and meat *vs* dual) were considered for the calculation of the fixation index (F_{ST}). Using a homemade script elaborated for Python 2.7 (Van Rossum, 1995) F_{ST} was calculated at each *locus* according to the formula proposed by Nei (1977):

$$F_{ST} = (H_T - H_S) / H_T$$

where HT and HS were the weighted expected heterozygosity in the total population and in the two sub-population, respectively.

To determine significant SNP that had higher F_{ST} values than expected under a neutral model of selection, a classical outlier approach was considered. Raw F_{ST} values were smoothed using a locally weighted scatterplot smoothing (LOWESS) regression, combined with a control chart approach using the PROC LOWESS procedure of SAS software (SAS/STAT® software version 9.2, SAS Institute, Inc., Cary) as proposed by (Pintus *et al*., 2014). A SNP was considered significant when the corresponding F_{ST} value exceeded the threshold of 3 standard deviations from the mean.

Canonical Discriminant Analysis (CDA)

The canonical discriminant analysis (CDA) performs a dimension reduction of a multivariate system by extracting the canonical variables (CVA), that are linear combinations of the original variables. The CVA are extracted in order to maximize the variation between predefined groups (Krzanowsky, 2003). In the present study, CDA was carried out on the SNP data matrix, separately for each chromosome, and the groups were the production aptitudes. In order to have full rank SNP matrices for each chromosome (i.e. the number of SNP should not exceed the number of animals) a preliminary screening was performed using a stepwise discriminant analysis (SDA) (Dimauro *et al*., 2015). A total of 8,950 SNP was selected and further used for CDA.

SNPs for which the correlation with the canonical variable was in the 99th percentile for a specific chromosome were considered to be significantly associated with that

variable (Sorbolini *et al*., 2016). Analyses were carried out using proc STEPDISC and CANDISC of SAS (SAS/STAT® software version 9.2, SAS Institute, Inc., Cary).

Runs of Homozygosity Analysis (ROH)

Runs of homozygosity (ROHs) were detected using the Zanardi pipeline. Parameters defining a ROH were fixed as following: i) a ROH should contain no less than 15 contiguous homozygous SNPs, ii) no missing or heterozygous SNP were allowed, and iii) the minimum ROH length was set at 1Mb. According to Kirin *et al*. (2010) the ROHs segments were grouped into five classes of length (< 2, 2-4, 4-8, 8-16, >16 Mb). A SNP was considered significantly associated to a homozygous region if it was included in a ROH in more than 20% of animals of the same productive purpose group.

Genes Annotation, QTLs and Functional Analysis

Annotated genes within the genomic regions that contained the SNPs flagged in the F_{ST}, CDA and ROH approaches were retrieved from NCBI (http://www.ensembl.org/index.html) using the *Ovis aries* OAR v3.1 release. Intervals of 500 kb (250 kb upstream and 250 downstream of the significant SNP) were considered. Genecards (http://www.genecards.org/), Animal QTL (http://www.animalgenome.org), Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Szklarczyk *et al*., 2017) interaction network (http://string-db.org/cgi/network) databases were used for the putative candidate gene functional associations analysis and QTL comparisons. To assess the functional annotation of the putative candidate genes, a gene ontology analysis (GO) using STRING v10.5

database for human orthologues was performed. To calculate and correct enrichment p-values STRING uses the Hypergeometric test (Rivals *et al*., 2007) and the method of (Benjamini and Hochberg, 1995), respectively. GO enrichment terms were considered statistically significant if the p-value was ≤ 0.05. The GO terms were categorized in three major groups: Biological Process, Molecular Function and Cellular Component.

Finally, the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.kegg.jp/kegg) database was considered for pathway analyses.

Results

Detection of significant SNPs with F_{ST}, CDA and ROH approaches

Population diversity measured by F_{ST} flagged 3,809 of SNPs as significant (Table S2). Several SNPs were shared between-different-pairwise comparisons. The number of SNPs in common differed according to the considered pairwise comparison and ranged from a minimum of 46 to a maximum of 245 (Table 1).

Table 1. Number of common SNPs among F_{ST} pairwise comparisons.

	Meat *vs* Wool	Dual *vs* Wool	Milk *vs* Wool	Meat *vs* Milk	Dual *vs* Milk	Dual *vs* Meat
Meat *vs* Wool	641	**245**	160	119	57	98
Dual *vs* Wool		634	223	63	47	**46**
Milk *vs* Wool			617	142	72	57
Meat *vs* Milk				656	128	213
Dual *vs* Milk					622	75
Dual *vs* Meat						639

The largest number of significant SNPs per chromosome were observed for meat *vs* milk and meat *vs* wool F_{ST} comparisons on OAR 1 (81 SNPs) and on OAR2 (77 SNPs), respectively (Table S2). The OAR1 showed also 68 significant SNPs in dual *vs* wool comparison. In milk *vs* dual and meat *vs* dual comparisons, OAR2 showed the largest number of significant SNPs: 102 and 92, respectively (Table S2). This autosome exhibited also the largest number of significant SNPs (66), in wool *vs* milk comparison.

The CDA based approach detected 2,655 significant SNPs across the whole genome. Some of them (260) were found in common among different CVAs. OAR 18 was the chromosome with the largest number of significant markers (n=141) equally divided into the three CVAs (n=47). Chromosomes with the lowest number of significant SNPs were OARs 24 and 26 with 16 SNPs for each CVA, respectively. Plots of animals on the canonical space for the 26 ovine autosomes are reported in Figure S1. The ROH count per SNP (SNP_{ROH}), i.e. the number of animals in which a particular SNP is included in a ROH, is reported in Table 2.

Table 2. Count of SNP in a ROH for each productive aptitude group.

Productive aptitude	Mean	SD[1]	Max	Min
Meat	69.13	104.08	1,902	16
Dual	92.10	128.4	1,581	16
Wool	95.29	146.39	1,276	16
Milk	78.31	118.45	2,296	16

[1] Standard Deviation

Wool and meat breeds exhibited the largest and the smallest average SNP_{ROH}, respectively. A similar pattern can be observed for the average of ROH length (Table 3).

Table 3. Length of ROH for each productive aptitude group.

Productive aptitude	Mean	SD[1]	Max	Min
Meat	3,611,678	5,328,567	96,110,653	1,000,069
Dual	4,815,977	6,572,184	80,721,921	1,000,116
Wool	4,990,344	7,469,866	65,859,169	1,003,910
Milk	4,089,332	6,070,363	112,756,965	1,000,069

[1] Standard Deviation

As far as the ROH distribution among length classes is concerned (Table 4), dual-purpose breeds showed the largest number of homozygous segments in the classes of greater amplitude, whereas wool breeds showed the smallest value.

Table 4. Number of ROH for each length class for the productive aptitude groups.

	Meat		Dual		Wool		Milk	
	N°	Freq,	N°	Freq,	N°	Freq,	N°	Freq,
< 2 Mb	2379	0.56	3287	0.47	758	0.50	3229	0.50
2 – 4 Mb	946	0.22	1413	0.20	287	0.20	1517	0.24
4 – 8 Mb	482	0.11	1105	0.16	200	0.13	874	0.14
8 – 16 Mb	267	0.07	707	0.10	141	0.10	463	0.07
> 16 Mb	145	0.03	433	0.06	104	0.07	270	0.04

All the four productive aptitude groups showed about 50% of homozygous segments in the smallest class of length (<2Mb).

A total of 21 significant SNPs were detected by all the three different approaches (F_{ST}, CDA and ROH). Five SNPs were located on OAR 2 and OAR 19, five on OAR 6, two on OAR 3 and 9 and one on OAR 13 respectively (Table 5).

Table 5. Common SNPs and putative candidate genes shared by the different statistical approaches.

OAR	Selection Signature position	SNPs	Candidate genes
2	52,322,304–53,411,840	3	RECK, FAM221B, NPR2
	84,525,056–84,705,280	2	SNORA2, TLE1, PHF24, BNC2, SAXO1, SLC24A2, ADAMTSL1
3	153,976,832–154,173,440	2	MBSRP3
6	36,814,848–37,986,304	5	ABCG2, PKD, SPP1, MEPE, IBSP, MED28, FAM184B, LCORL, NCAPG
9	77,332,480–77,627,392	2	VPS13B, SNORA70, STK3, KCNS2, NIPAL2
13	62,607,216–62,857,216	1	RALY, NECAB3, PXMP4, ZNF341, CHMP4B, EIF2F2, SNORA73
19	515,968-1,289,728	6	DBNL, EGFR, LANCL2, PGAM2, UBE2D1, VOPP1

Detection of signatures of selection

F_{ST} comparisons pointed out a total of 510 selection signatures genome-wide (meat *vs* wool = 84; meat *vs* milk = 85; dual *vs* wool = 84; wool *vs* milk= 78; meat *vs* dual = 83 and dual *vs* milk = 96, respectively). Selection signatures detected with the F_{ST} approach were characterized by a large number of significant markers (most of them included more than 15 significant adjacent SNPs).

The CDA approach highlighted a large number of selection signatures (n = 1401) mostly characterized by only one or few consecutive SNPs. Only 18 signatures of selection detected across the whole genome contained at least five adjacent SNPs. The largest number of selection signatures was detected on OAR 4 (n = 109) and OAR 5 (n = 106), whereas the smallest number was on OAR 25 (n = 24).

The ROH approach detected signatures of selection on nine chromosomes (OARs 1, 2, 3, 6, 9, 10, 13, 19 and 22) (Figure S2). Twenty-two selection signatures were

detected at genome-wide level. For example, on OAR 13 two regions in homozygosity were found, one at 50 Mb for milk breeds and another at 65 Mb for wool breeds, respectively (Figure S2).

Selection signatures shared among procedures

Some of the signatures of selection highlighted in the present study were shared by two or by all the three detection approaches (Figures 1 A-F). In particular, signatures of selection detected by all the three approaches were located on OARs 2, 3, 6, 9, 13 and 19 (Table 5). For example, F_{ST} exhibited a selection signature in meat *vs* milk (Figure 1B), and meat *vs* dual (Figure 1E) comparisons, on OAR6 at about 37-39 Mb. In figure 2, an enlargement of OAR 6 was reported. The first CVA extracted on OAR 6 highlighted a clear separation between production aptitudes, with milk and wool breeds located at the two extremes, whereas CVA2 separated milk and meat from dual and wool (Figure 3). Among the SNPs that showed largest correlations, three were located between 37.8 and 38.9 Mb (CVA1) and one at 37.6 Mb (CVA2), respectively. Using ROH approach significant peaks were obtained for meat, milk and dual groups (Figure 4). In this region the *ATP Binding Cassette Subfamily G Member 2* (*ABCG2*), *Polycystin 2, Transient Receptor Potential Cation Channel* (*PKD2*), *Secreted Phosphoprotein 1 (SPP1), Leucine Aminopeptidase 3* (*LAP3*), *Ligand Dependent Nuclear Receptor Corepressor Like* (*LCOR*L) and *Non-SMC Condensin I Complex Subunit G* (*NCAPG*) genes were annotated. Analyzing animal QTL database two large QTLs (one for milk traits and one for meat traits) were reported in this region.

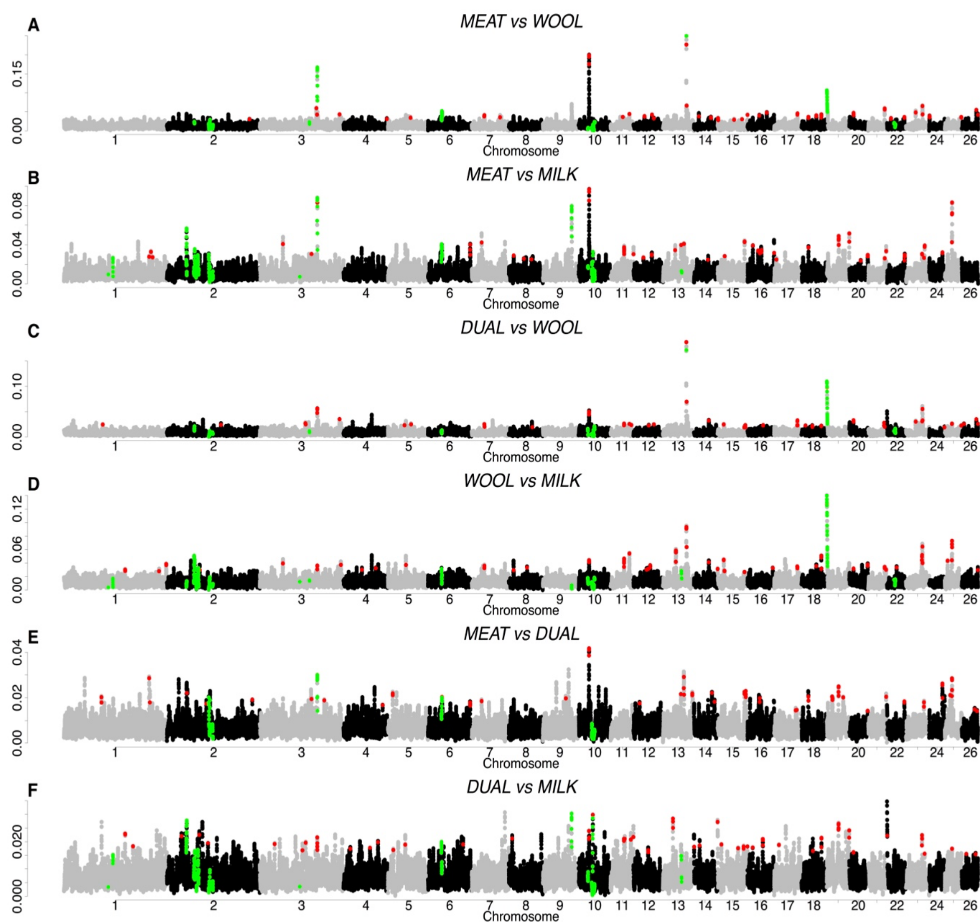


Figure 1. Manhattan plots of FST pairwise comparisons. Y axis showed the smoothed FST values. High values represented the significant SNPs for this technique. Black and grey dots represent FST values of different chromosomes. Red dots highlighted the significant SNPs found using CDA. Green dots evidenced the significant SNPs detected using ROH approach. The occurrence of red and/or green dots in FST peaks indicates that significant SNPs are shared among approaches. (A) Meat vs Wool (B) Meat vs Milk (C) Dual vs Wool (D) Wool vs Milk (E) Meat vs Dual (F) Dual vs Milk.

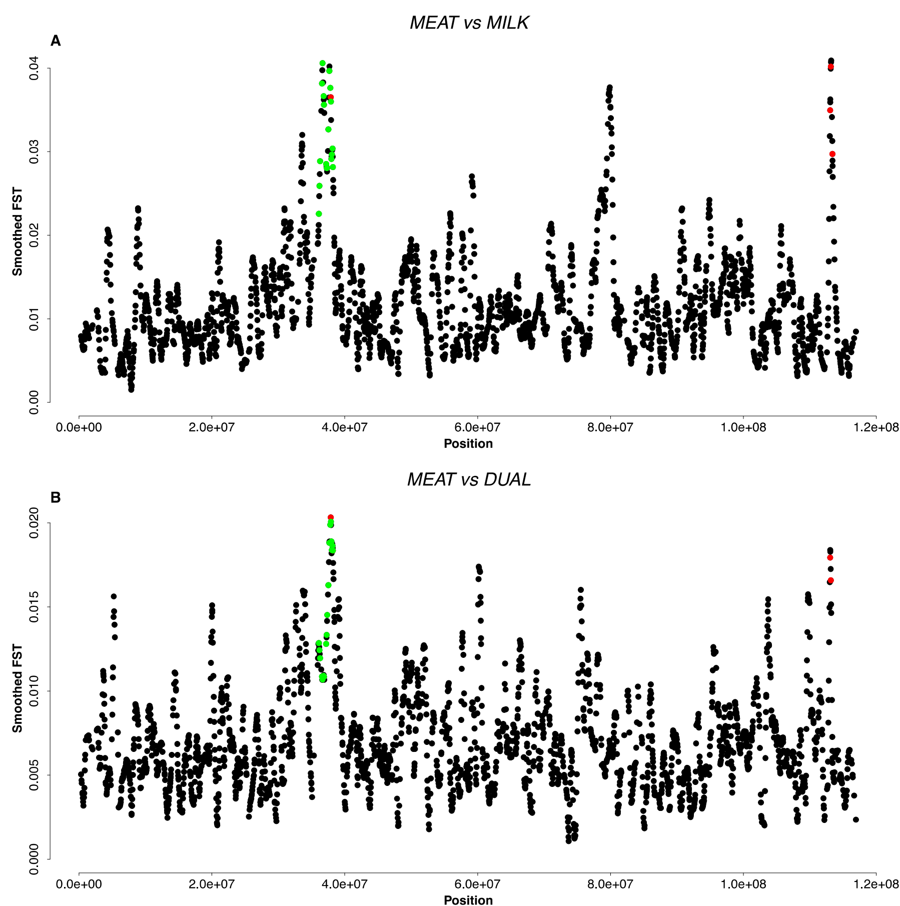


Figure 2. Manhattan plots enlargement of FST pairwise comparisons for OAR6. Y axis showed the smoothed FST values. High values represented the significant SNPs for this technique. Red dots highlighted the significant SNPs found using CDA. Green dots evidenced the significant SNPs detected using ROH approach. When colored dots were present in FST peaks significant SNPS were shared among approaches. (A) Meat vs Milk (B) Meat vs Dual.

Among the other signatures of selection shared by all the three procedures, three were worthy of note: i) one on OAR 2 between 84.5-84.7 Mb where is annotated the *Basonuclin 2 (BCN2)* locus; ii) one on OAR 13 at 62.8Mb containing the *RALY*

Heterogeneous Nuclear Ribonucleoprotein (*RALY*) gene; iii) and the last on OAR 19 harboring the *Epidermal growth factor receptor* (*EGFR*).

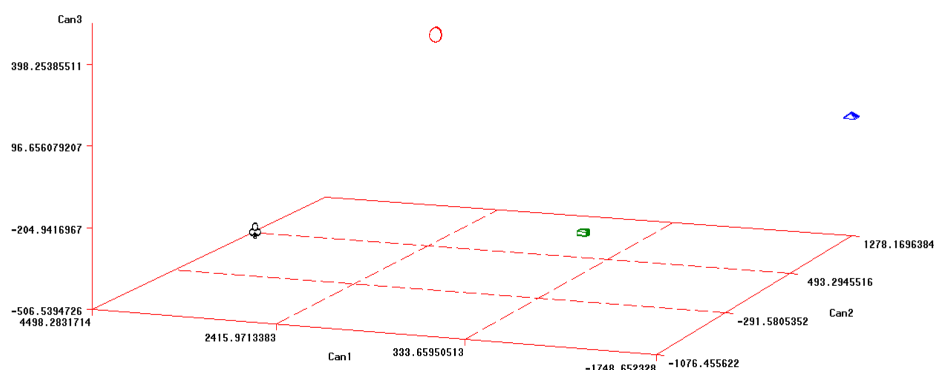


Figure 3. Plot of the Canonical Discriminant Analysis of OAR6. Iindividual scores of the first, second and third canonical values (Can1, Can2, Can3) extracted from OAR 6 in the four productive aptitude groups. Red circle = Meat. Black flower = Wool. Green square = Dual. Blue triangle = Milk.

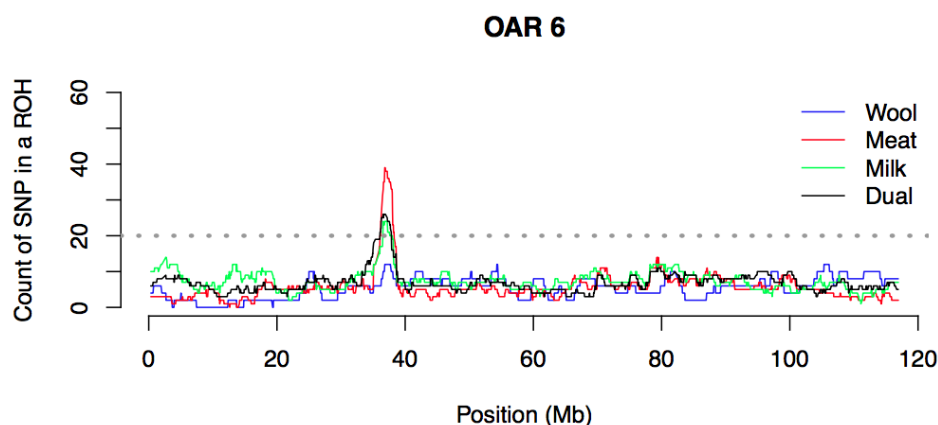


Figure 4. Plot of Runs of Homozygosity in OAR6. Occurrence of SNP counted in a ROH measured by the percentage of animals belonging to the four investigated productive aptitude groups for which a particular SNP falls into a ROH versus the position along the OAR 6.

Some genomic regions detected with at least two metrics have also been found in this study. For example, a genomic region spanning from 46 to 47 Mb on OAR 22 was detected by both F_{ST} (Figures 1A, 1B, 1E) and CDA (Figure S1) approaches, respectively. In this region, three non-protein coding RNA genes were annotated.

Finally, many regions of genomic diversity were detected by only one of the three statistical approaches. For example, selection signatures were found exclusively with ROH in meat on OAR 2 (122.0-123.4 Mb) or in milk breeds on OAR 1 (118.5-118.6 Mb and 131.0-131.5 Mb). For what concerns wool breeds significant unique signals were highlighted on OAR 3 at 130.0-133.4 Mb and the closest genes annotated were several members of the Keratin gene family (Figures S2).

Candidate gene discovery, functional annotation and QTLs

About 400 candidate genes potentially under selection were identified. Several loci were already identified as candidate for important traits in sheep (Table 6).

Table 6. List of known candidate genes detected in this study using multi statistical approach.

OAR	Gene symbol	References
1	*CSTA, ADIPOQ, TRPM8*	Liu et al., 2015; An et al., 2015; Farriello et al., 2014
2	***NPR2,*** *ACSL3,* ***BNC2***	Kijas et al., 2012; Liu et al 2012; Farriello et al., 2014
3	*EDAR, FSHR, TRHDE, HMGA2, PTHLH,* ***MSRB3****, LEMD3*	Zhao et al, 2016; Kijas et al., 2012; Farriello et al., 2014; Wei et al.,2015; Wang et al., 2014; Manunza et al., 2016
4	*LEP, CALCR, GHRHR, SLC13A1*	Zhou et al., 2009
6	*ALB,* ***LCORL****,* ***NCAPG****,* ***ABCG2****,* ***PKD2****,* ***SPP1****, FGF5, BMPR1B*	Blunt 1975; Kijas et al., 2012; Farriello et al., 2014; Wei et al., 2015
7	*TSHR*	Kijas et al., 2012
9	*PLAG1*	de Simoni Gouveia et al., 2017
10	*RXFP2*	Kijas et al., 2012; Farriello et al., 2014
11	*FASN*	Zhu et al., 2016; Suarez-Vega et al., 2017
12	*TGFB2*	Zhang et al., 2017
13	*PRP/PRND, BMP2, ASIP*	Kijas et al., 2012; Farriello et al., 2014
14	*FTO, MC1R*	Kijas et al., 2012; Farriello et al., 2014; Wang et al., 2015
16	*GHR, PRLR*	Wang et al., 2015; Periasamy et al., 2014
19	*MITF*	Kijas et al., 2012
20	*PRL*	Kijas et al., 2012; Liu et al.,2016
21	*TYR, IGF2*	Deng et al., 2008; Lan et al., 2013
22	*SCD, PRLHR*	García-Fernández et al., 2009
26	*ACSL1*	Bolormaa et al., 2016; Suárez-Vega et al., 2017

Gene symbol in bold= highlighted by three methods

The multiple approach used in the present study identified genomic regions harboring genes involved in milk production and mammary gland biology (*BTN1A1*, *CSNK2A2* and *MFGE8*) others in the muscle and bone development (*PPIE, GYG1, HES1, SPP2, OBSCN, TLN2, TAB2, SGCD, SGCA* and *CHAD*). Several candidate genes affecting wool traits were identified such as *FGF18, LSS, CLND 16, PADI 2, PADI 3, BMP7, EGFR, WNT10B* and *PROP1*. In this survey, a number of putative candidate genes

controlling food intake/energy balance and lipid/fatty acid metabolism were found across the sheep genome. In figure 5 the fatty acid biosynthesis scheme for *Homo sapiens* (http://www.wikipathways.org) was reported. Several orthologous genes involved in the human pathway were highlighted in this study (*FASN, ACSL1, ACSL3, SCD, ACLY, ECH1* and *ECHDC3*).

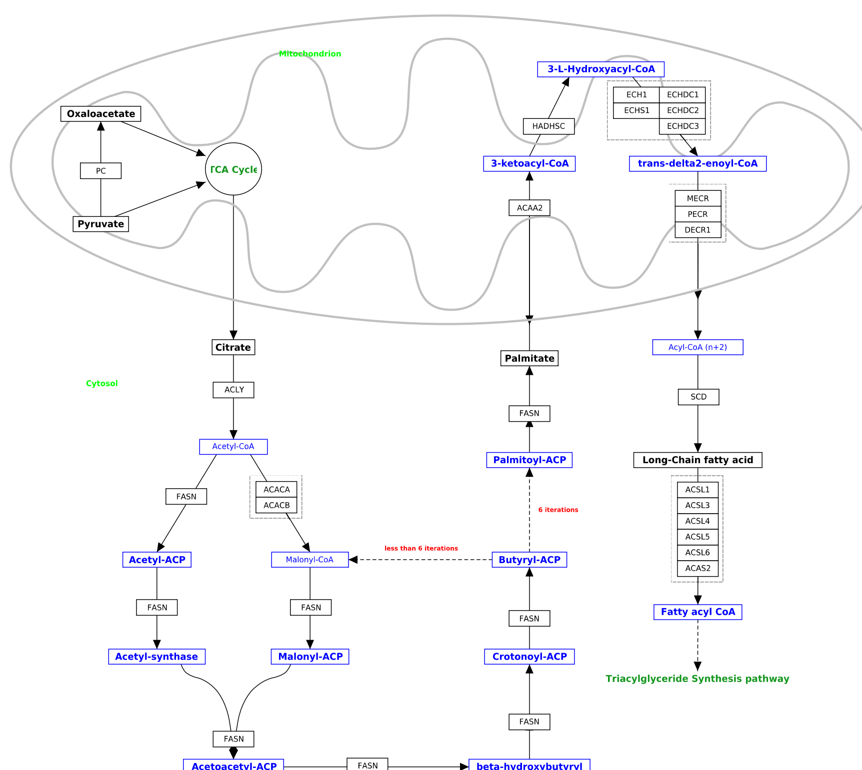


Figure 5. Fatty acid biosynthesis scheme for Homo sapiens extracted from http://www.wikipathways.org. FASN, PC, ACSL1, ACSL3, SCD, ACLY, ECH1 and ECHDC3 were putative candidate genes also in this study.

In addition to the genes involved in productive traits, many significant SNPs were found near loci that control adaptation to environment. Many genes controlling

sensory function such as hearing (*OTGL, OTOR, OTOG, OTOA*), taste (*TAS2R10, KCNK12, GNAT3, TAS1R1*), smell (*ORs family, NTPN, G1 and GFY*) and sight (*OPA, LENEP, CRYZ, EYA4*), regulation of body temperature (*TRMP8*) and circadian cycle (*CLOCK, TIPIN, PER3, DBP, FBXL3*) were detected. In figure 6, the protein-protein interaction network derived from STRING database for the orthologues genes detected in this study was depicted. These genes are involved in the control of circadian rhythm in human.

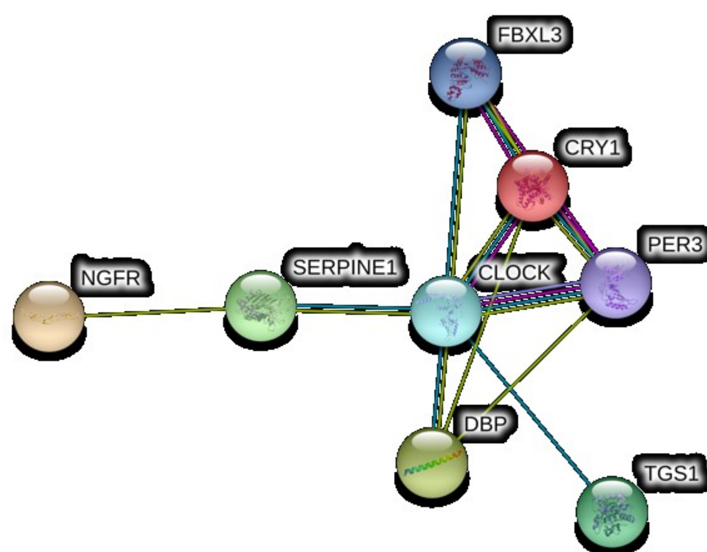


Figure 6. The protein-protein interaction network derived from STRING v10 database for the orthologues genes involved in the control of circadian rhythm in human. Purple line = experimentally determined interactions. Turquoise line = interactions from curated databases. Lightgreen line = textmining interactions. Violet line = protein homology interactions. Filled nodes = known 3D structure protein.

Several loci associated with craniofacial and dental development were identified (*DLX3, GAB1, TUFT1, ODAM*). Analysis of GO term enrichment was performed on the sheep genes set identified. A total of 513 GO terms were found for biological process, 46 for molecular function, 36 for cellular component and finally 21 pathways significantly enriched for Kyoto Encyclopedia Genes and Genomes (KEGG) pathways. Finally, in tables from S3 to S6 were reported the lists of known QTLs associated with productive traits extracted from sheep QTLdb database release 30 and also found in this study.

Discussion

The study of genetic variability and the research of selection signatures is generally carried out by comparing breeds farmed for different production purposes and subjected to different selective pressures (Gutierrez-Gil *et al*., 2007; McRae *et al*., 2014; Zhao *et al*., 2015). On the other hand, it is also possible to investigate the genetic variability using samples composed of individuals belonging to different breeds and comparing them among different productive purposes (Fariello *et al*., 2014; Randhawa *et al*., 2014; Wei *et al*., 2015). The choice of the most suitable statistical metrics to capture variability, however, represents a challenge for this type of surveys. Recently, to solve this puzzle, researchers have opted for the simultaneous use of different tests on the same dataset (Baye, 2011; Grossman *et al*., 2013; Yang *et al*. 2016; Brito *et al*., 2017).

In the present study, the application of a multiple statistical approach resulted in the detection of a large number of selection signatures and candidate genes.

Notwithstanding with the different mathematical logic and sensitivity of the three methodologies used, obtained results showed a good degree of concordance (Figure 1A-F). Moreover, the three methods produced consistent results with existing literature in sheep (Zhang *et al*., 2013; Wang *et al*., 2015; Manunza *et al*., 2016; de Simoni Gouveia *et al*., 2017).

The detection of genes involved in the biology of the mammary gland and lactation, in the development of bone and muscle tissue, and in fiber production, was quite expected, being the data set composed by meat, dairy, dual-purpose and wool breeds. Moreover, in addition to genes widely known such as *ABCG2, SCD, PRL, NPR2, FTO, GHR, EDAR, HR* (Table 6), many interesting new putative candidates for important traits were found.

In this study, the *milk fat globule epidermal growth factor (EGF*)-*factor VIII* (*MFG-E8*) and the *butyrophilin 1A1* (*BTN1A1*) genes were detected. These loci are involved in the formation of milk fat globules (MFGs) (Jeong *et al*., 2013). MFGs are composed up to 98% of triglycerides packaged and secreted into milk as plasma membrane trilayer-coated structures (McManaman, 2009). Milk fat globules features depend on several factors such as breed, lactation phase, parity and feeding (Martini *et al*., 2004). MFG-E8 and BTN1A1 are plasma membrane proteins on the surface of MFGs (Yasueda *et al*., 2015) and playing a critical role in regulating the dynamics of MFGs after weaning or during the suckling (Jeong *et al*., 2013; Nakatani *et al*., 2012). Moreover, numerous studies conducted on dairy animals demonstrate the relationship between MFGs and milk quality (Wiking *et al*., 2004; Couvreur *et al*., 2007; Ménard *et al*., 2010). Milk lipids contained in the globules represent one of the most important

factors affecting dairy productions. Polymorphisms at these genes are already reported associated with quantitative milk traits in dairy goats (Qu *et al*., 2011). Because most of the ovine milk produced is transformed into cheese, polymorphisms of *MFG-E8* and *BTN1A1* genes are of particular interest also in sheep. Among loci influencing muscle biology, the *alpha-sarcoglycan precursor* (*SGCA*), the *delta-sarcoglycan isoform 1* (*SGCD*) and *calpain 3* (*CAPN3*) were highlighted. In the striated muscular cells, the *SGCA* and *SGCD* genes code for two members of transmembrane glycoproteins forming the sarcoglycan complex. This complex, binding the actin cytoskeleton to the extracellular matrix, is essential for muscle integrity and functionality (Tarakci and Berger, 2016). Tenderness is one of the most important traits in meat quality. The transformation of muscle into meat has been associated with *postmortem* proteolysis. During the tenderization process the main proteins of myocytes cytoskeleton such as actin, myosin and filamin are degraded by calpain/calpastatin system (Lana and Zolla, 2016). The calpain 3, is one member of the calpain family and is considered to play a major role in meat tenderization (Ilian *et al*., 2004; Lian *et al*., 2013). Moreover, *in vitro* studies have shown that the calpain 3 degrades filamin regulating the interactions with sarcoglycan and affects the functionality of muscle cells (Guyon *et al*., 2003). In addition, *CAPN3* was recently indicated as putative candidate gene in Chinese and Mongolian fat tailed sheep breeds (Wang *et al*., 2015).

Wool is a crimped and elastic textile fiber obtained by sheep. It is produced by small cells located in the epidermis and called follicles. The quality of wool is determined by several factors such as fiber diameter, length and strength, color and crimp (Wang

et al., 2014). Ancestral sheep's fleece was composed of outer fibers, long and coarse with mechanical function and innermost fine fibers with a thermoregulatory function. Domestication has led to profound changes in sheep's coat fixing in populations those mutations that led to a fleece composed of more homogeneous fibers. For fiber industry, a prominent resource is represented by wool production. In sheep with the aim of improve the quality and quantity of fine wool numerous studies have focused on the search for molecules that control this trait. Several possible candidate genes such as *Wnt Family Member 10B* (*WTN10B*), *Paired-Like Homeobox 1* (*PITX1*), *Transforming Growth Factor Beta 2* (*TGFB2*), *Bone morphogenetic protein 7* (*BMP7*) and *Fibroblast growth factor 18* (*FGF18*) expressed in skin and fibers were suggested (Liu *et al*., 2016; Zeng *et al*., 2011; Liu *et al*., 2015; Lv *et al*., 2016). Results reported here are in agreement with the literature detecting all the above-mentioned loci. An interesting locus affecting the physiology of follicles associated with wool traits and not yet reported in sheep is the *epidermal growth factor receptor* (*EGFR*). In humans and mice, interacting with its specific ligands (EGF, BTC, TGF-α, AEREG), *EGFR* performs a pivotal role in several aspects of cutaneous biology such as normal hair follicle morphogenesis, cycling and regulation of proliferation/differentiation of follicular keratinocytes (Murillas *et al*., 1995; Mak and Chan, 2003; Schneider *et al*., 2008a; Schneider *et al*., 2008b; Namba *et al*., 2013).

The production of milk, meat and wool are processes involving the biosynthesis and processing of many types of fatty acids. The fat content is a very important parameter in the production of milk, meat and wool. For example, the marbling fat (the intramuscular fat deposits that are found between myofibrils) determines the degree of

leanness of the meat (Wood *et al*., 2008; Dodson *et al*., 2010) and in milk, fatty acid quality and quantity influences cheese fat content as well as rheological and sensorial qualities (Soyeurt *et al*., 2006; Morand-Fehr *et al*., 2007; Nudda *et al*., 2014). Also in wool production, waxes such as lanolin are essential for fiber quality (Jiang *et al*., 2014). In ruminants, volatile fatty acids fermented in the rumen are mostly absorbed through ruminal epithelium (Dieho *et al*., 2016) and blood-transported to different tissues and organs such as liver, adipose tissue, muscle and mammary gland. In the eukaryotic cells, fatty acids biosynthesis occurs in two biosynthetic processes completely independent, one in the cytosol and one in the mitochondrion (Hiltunen *et al*., 2009). In their final cellular compartment, the fatty acids adsorbed from diet are subjected to further metabolic modifications to produce lipids. In this study, several selection signatures containing candidate genes (*ACLY*, *FASN* and *PC*) involved in the pathway responsible of fatty acid metabolism were identified (Figure 5). Key enzymes of the cytosol are the *ATP Citrate Lyase (ACLY)* and Fatty *Acid Synthase* (*FASN*). *ACLY* has a central role in *de novo* lipid synthesis and is responsible for the synthesis of cytosolic acetyl-CoA and oxaloacetate that are important in several biosynthetic pathways, including lipogenesis and cholesterolgenesis. *FASN* catalyze the synthesis of palmitate from acetyl-CoA and malonyl-CoA, into long-chain saturated fatty acids and is associated with obesity. In ruminants about 90% of fatty acid synthesis occur in the adipose tissue (Bauman, 1976). In brown and white adipocytes, P*yruvate Carboxylase (PC)* is located exclusively in the mitochondrial matrix and is involved in the lipogenesis (Jitrapakdee *et al*., 2006). Due to the several roles exerted by adipose tissue (e.g. glucose homeostasis, thermal insulation and control of energy balance)

understanding the molecular bases of lipogenesis is of utmost importance in mammals and more in particular, in ruminants, because adipose tissue physiology can have significant economic repercussions on the production system influencing meat, milk and wool quality traits (Laliotis *et al*., 2010).

Animals use senses to interact with environment. Through the sensory system, they process the information (stimuli) coming from inside or outside the body. Therefore, the sensory system is one of the most involved in the mechanisms of animal adaptation to environment. In general, sheep possess a highly developed sensory system capable of perceiving even minimal stimuli. Since they are mostly farmed with extensive systems, sight, smell, taste and hearing ensure animal survival. In this study, several adaptive loci involved in the physiology of the sensory perception were highlighted. For example, a high number of candidate loci belonging to the olfactory receptor (OR) family gene or the development and functioning of ears and eyes were detected. Sheep have an excellent sense of smell, taste and sight to recognize predators, to locate lambs or help rams identify ewes, to find water and to detect the differences in pasture plants (Baldwin *et al*., 1977; Blissett *et al*., 1990; Piggins and Phillipds, 1996). In mammals, also craniofacial development and dental conformation are considered adaptive traits (Tsuboi *et al*., 2014; Parés Casanova and Bravi, 2014; de Moura Bubadué *et al*., 2016). In this study, several genes involved in the dental and skull development were detected. A possible explanation to this variability could be that sheep breeds analyzed in this study belonging to different geographical areas and present differences in morphological and ecological adaptation to their habitats.

Genetic variation in the sheep DNA sequence was also observed in genes that control the circadian rhythm. In mammals, it is known that seasonal variations and circadian cycle regulate behavior and many biological functions (Lincoln *et al*., 2002). Moreover, living organisms have evolved endogenous long-term timing devices, which allow them to anticipate forthcoming seasonal variations in the environment, for example in food availability or climatic conditions (Dardente *et al*., 2014). Reproduction, lactation and food intake are the most relevant processes modulated by circadian rhythm (Boden and Kennaway, 2006; Challet, 2013; Wang *et al*., 2015). In the present study selection signals containing genes related to seasonal variations in the photoperiod were highlighted. In Figure 5 is reported a protein-protein interaction network among genes involved in the human circadian rhythm. In mammals, mutations on *Clock Circadian Regulator* (*CLOCK*), *Period Circadian Clock* 3 (*PER3*) and *Cryptochrome Circadian Clock 1* (*CRY1*) genes can alter the circadian period and rhythmicity (Reppert and Weaver, 2001). Polymorphisms on genes controlling molecular mechanisms of circadian clocks can be useful to understand the responses of an organism to environmental stimuli and use them to increase animal productions.

Conclusions

The search for signatures of selection carried out on various sheep breeds farmed in Italy confirmed the usefulness of a multiple statistical approach, providing interesting insights on the genetic basis of their differentiation. The three statistical approaches used, although different in the metrics used to make inference, provided a

good concordance of results, highlighting genomic regions where the most relevant differences among breeds were located.

Some selection signatures detected in this work were, as expected, located in genomic regions that harbor well known genes involved in the morphological features of animals or in their productive aptitudes. However, several interesting novel candidate genes related to physiological functions that contribute to the mechanism of environmental adaptation have been detected. These results confirm the important role of environment in the evolution of the sheep genome and the suitability of this species as model for studying the genetic basis of resilience.

References

Akey J. M., Ruhe A. L., Akey D. T., Wong A. K., Connelly C. F., Madeoy J., *et al*. (2010) Tracking footprints of artificial selection in the dog genome. Proceedings of the National Academy of Sciences of the United States of America 107, 1160-1165.

Andersson L. (2001) Genetic dissection of phenotypic diversity in farm animals. Nature Reviews Genetics 2, 130-138.

Baldwin B. A., McLaughlin C. L., & Baile C. A. (1977) The effect of ablation of the olfactory bulbs on feeding behaviour in sheep. Applied Animal Ethology 3, 151-161.

Bauman D. E. (1976) Intermediary metabolism of adipose tissue. Federation Proceedings 35, 2308-2313.

Baye T. M. (2011) Inter-chromosomal variation in the pattern of human population genetic structure. Human Genomics 5, 220-2240.

Benjamini Y., & Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B 57, 289–300.

Biswas S., & Akey J. M. (2006) Genomic insights into positive selection. Trends in Genetics 22, 437-446.

Blissett M. J., Boland K. P., & Cottrell D. F. (1990) Discrimination between odours of fresh oestrous and non-oestrous ewe urine by rams. Applied Animal Behaviour Science 25, 51–59.

Boden M. J., & Kennaway D. J. (2006) Circadian rhythms and reproduction. Reproduction 132, 379-392.

Brito L. F., McEwan J. C., Miller S. P., Pickering N. K., Bain W. E., Dodds K. G., *et al*. (2017) Genetic diversity of a New Zealand multi-breed sheep population and composite breeds' history revealed by a high-density SNP chip. BMC Genetics 18, 25.

Brotherstone S., & Goddard M. E. (2010) Artificial selection and maintenance of genetic variance in the global dairy cow population. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 36, 1479-1488.

Browning B. L., & Browning S. R. (2016) Genotype imputation with millions of reference samples. American Journal of Human Genetics 98, 116-126.

Challet E. (2013) Circadian clocks, food intake, and metabolism. Progress in molecular biology and translational science 119, 119-35.

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

Ciani E, Crepaldi P, Nicoloso L, Lasagna E, Sarti FM, Moioli B, *et al*. (2013) Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. Animal Genetics 45, 256-266.

Couvreur S., Hurtaud C., Marnet P. G. , Faverdin P. , Peyraud J. L. (2007) Composition of milk fat from cows selected for milk fat globule size and offered either fresh pasture or a corn silage-based diet. Journal of Dairy Science 90, 392-403.

Dardente H., Hazlerigg D. G., & Ebling F. J. (2014) Thyroid hormone and seasonal rhythmicity. Frontiers in endocrinology 2014, 5.

de Moura Bubadué J., Cáceres N., dos Santos Carvalho R., & Meloro C. (2016) Ecogeographical variation in skull shape of South-American canids: abiotic or biotic processes? Evolutionary biology 43, 145-159.

de Simoni Gouveia J. J., Paiva S. R., McManus C. M., Caetano A. R., Kijas J. W., Facó O., *et al*. (2017) Genome-wide search for signatures of selection in three major Brazilian locally adapted sheep breeds. Livestock Science 197, 36-45.

Dieho K., Dijkstra J., Schonewille J. T., & Bannink A. (2016) Changes in ruminal volatile fatty acid production and absorption rate during the dry period and early lactation as affected by rate of increase of concentrate allowance. Journal of Dairy Science 99, 5370-5384.

Dimauro C., Nicoloso L., Cellesi M., Macciotta N. P. P., Ciani E., Moioli B., *et al*. (2015) Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. Small Ruminant Research 128, 27-33.

Dodson M. V., Hausman G. J., Guan L., Du M., Rasmussen T. P., Poulos S. P., *et al*. (2010) Lipid metabolism, adipocyte depot physiology and utilization of meat animals as experimental models for metabolic research. International Journal of Biological Sciences 6, 691.

Fariello M. I., Servin B., Tosser-Klopp G., Rupp R., Moreno C., San Cristobal M., *et al*. (2014) Selection signatures in worldwide sheep populations. PLoS One 9, e103813.

Flori L., Fritz S., Jaffrézic F., Boussaha M., Gut I., Heath S., *et al*. (2009) The Genome Response to Artificial Selection: A Case Study in Dairy Cattle. PLoS One 4, e6595.

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, 6-31.

Grossman S. R., Andersen K. G., Shlyakhter I., Tabrizi S., Winnicki S., Yen A., *et al*. (2013) Identifying recent adaptations in large-scale genomic data. Cell 152, 703-713.

Gutiérrez-Gil B., Wiener P., & Williams J. L. (2007) Genetic effects on coat colour in cattle: dilution of eumelanin and phaeomelanin pigments in an F2-Backcross Charolais× Holstein population. BMC genetics 8, 56.

Guyon J. R., Kudryashova E., Potts A., Dalkilic I., Brosius M. A., Thompson T. G., *et al*. (2003) Calpain 3 cleaves filamin C and regulates its ability to interact with gamma- and delta-sarcoglycans. Muscle Nerve 28, 472-83

Hiendleder S., Kaupe B., Wassmuth R., & Janke A. (2002) Molecular analysis of wild and domestic sheep questions current nomenclature and provides evidence for domestication from two different subspecies. Proceedings of the Royal Society of London. Series B, Biological Sciences 269, 893–904

Hiltunen J. K., Schonauer M. S., Autio K. J., Mittelmeier T. M., Kastaniotis A. J., & Dieckmann C. L. (2009) Mitochondrial fatty acid synthesis type II: more than just fatty acids. Journal of Biological Chemistry 284, 9011-9015.

Iacolina L., Storen A. V., Pertoldi C., Tokarska M., Noorgard L., Munoz J., *et al*. (2016) Novel Graphical Analyses of Runs of Homozygosity among Species and Livestock Breeds. International Journal of Genomics 2016, 2152847.

Ilian M. A., Bekhit A. E. D., & Bickerstaffe R. (2004) The relationship between meat tenderization, myofibril fragmentation and autolysis of calpain 3 during post-mortem aging. Meat Science. 66, 387-397.

Italian Institute of Statistics (2010) 6th General Census of Agriculture.

Jeong J., Lisinski I., Kadegowda A. K. J., Shin H., Wooding F. B. P., Daniels B. R., *et al*. (2013) A test of current models for the mechanism of milk-lipid droplet secretion. Traffic 14, 974-986.

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, 1168–1173.

Jitrapakdee S., Vidal-Puig A., & Wallace J. C. (2006) Anaplerotic roles of pyruvate carboxylase in mammalian tissues. Cellular and Molecular Life Sciences 63, 843-854.

Kijas J. W., Townley D., Dalrymple B. P., Heaton M. P., Maddox J. F., McGrath A., *et al*. (2009) A Genome Wide Survey of SNP Variation Reveals the Genetic Structure of Sheep Breeds. PLoS One 4, e4668.

Kim Y. H. B., Warner R. D., & Rosenvold K. (2014) Influence of high pre-rigor temperature and fast pH fall on muscle proteins and meat quality: a review. Animal Production Science 54, 375-395.

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

Krzanowsky W. J. (2003) Principles of multivariate analysis. New York: Oxford University Press Inc.

Laliotis G. P., Bizelis I., & Rogdakis E. (2010) Comparative approach of the de novo fatty acid synthesis (Lipogenesis) between ruminant and non ruminant mammalian species: from biochemical level to the main regulatory lipogenic genes. Current genomics. 11, 168-183.

Lana A., & Zolla L. (2016) Proteolysis in meat tenderization from the point of view of each single protein: A proteomic perspective. Journal of Proteomics 147, 85-97

Lee Y. W., Gould B. A., & Stinchcombe J. R. (2014) identifying the genes underlying quantitative traits: a rationale for the QTN programme. AoB Plants 6, 1-14.

Lian T., Wang L., & Liu Y. (2013) A new insight into the role of calpains in post-mortem meat tenderization in domestic animals: a review. Asian-Australasian Journal of Animal Sciences 26, 443.

Lincoln G., Messager S., Andersson H., & Hazlerigg D. (2002) Temporal expression of seven clock genes in the suprachiasmatic nucleus and the pars tuberalis of the sheep: evidence for an internal coincidence timer. Proceedings of the National Academy of Sciences 99, 13890-13895.

Liu G., Liu R., Tang X., Cao J., Zhao S., & Yu M. (2015) Expression Profiling Reveals Genes Involved in the Regulation of Wool Follicle Bulb Regression and

Regeneration in Sheep. International journal of molecular sciences 16, 9152–9166.

Liu Z., Ji Z., Wang G., Chao T., Hou L., & Wang J. (2016) Genome-wide analysis reveals signatures of selection for important traits in domestic sheep from different ecoregions. BMC Genomics 17, 863.

Lv X. Y., Ni R., Sun W., Su R., Musa H. H., Yin J. F., *et al*. (2016) Candidate genes for the development of hair follicles in Hu sheep. Genetics and molecular research 15.

Ma Y., Ding X., Qanbari S., Weigend S., Zhang Q., & Simianer H. (2015) Properties of different selection signature statistics and a new strategy for combining them. Heredity 115, 426–436.

Mak K. K., & Chan S. Y. (2003) Epidermal growth factor as a biologic switch in hair growth cycle. Journal of Biological Chemistry 278, 26120-26126.

Malorgio G. (1995) Systems of production in the Italian beef and sheep industry: prospects for extensive systems. Paper presented at the Conference on Beef and Sheep production system, prospects for extensification. EU Working Group, Paris.

Manunza A., Cardoso T. F. , Noce A., Martínez A., Pons A., Bermejo L. A., *et al*. (2016) Population structure of eleven Spanish ovine breeds and detection of selective sweeps with BayeScan and hapFLK. Scientific Reports 6, 27296.

Marras G., Gaspa G., Sorbolini S., Dimauro C., Ajmone-Marsan P., Valentini A., *et al*. (2015) Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Animal Genetics 46; 110-121.

Martini M., Scolozzi C., Cecchi F., & Abramo F. (2004) Morphometric analysis of fat globules in ewe's milk and correlation with qualitative parameters. Italian Journal of Animal Science 3, 55-60.

McManaman J. L. (2009) Formation of milk lipids: a molecular perspective. Clinical lipidology 4, 391-401

McRae K. M., McEwan J. C., Dodds K. G., & Gemmell N. J. (2014) Signatures of selection in sheep bred for resistance or susceptibility to gastrointestinal nematodes. BMC genomics 15, 637.

Ménard O., Ahmad S., Rousseau F., Briard-Bion V., Gaucheron F., & Lopez C. (2010) Buffalo vs. cow milk fat globules: Size distribution, zeta-potential, compositions in total fatty acids and in polar lipids from the milk fat globule membrane. Food Chemistry 120, 544-551.

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

Morand-Fehr P., Fedele V., Decandia M., & Le Frileux Y. (2007) Influence of farming and feeding systems on composition and quality of goat and sheep milk. Small Ruminant Research 68, 20-34.

Murillas R., Larcher F., Conti C. J., Santos M., Ullrich A., & Jorcano J. L. (1995) Expression of a dominant negative mutant of epidermal growth factor receptor in the epidermis of transgenic mice elicits striking alterations in hair follicle development and skin structure. The EMBO journal 14, 5216-5223.

Nakatani H., Yasueda T., Oshima K., Okajima T., Nadano D., Flint D. J., *et al*. (2012) Post-weaning increases in the milk-fat globule EGF-factor VIII on fat globules in mouse milk and in the uptake of the fat globules by HC11 mammary epithelial cells. The Journal of Biochemistry 153, 31-41.

Namba D., Toki F., Barrandon Y., & Higashiyama S. (2013) Recent advances in the epidermal growth factor receptor/ligand system biology on skin homeostasis and keratinocyte stem cell regulation. Journal of Dermatological Science 72, 81-86.

Nei M. (1977) F-statistic and analysis of gene diversity in subdivided populations. Annals of human genetics 41, 225–233.

Nudda A., Battacone G., Boaventura Neto O., Cannas A., Francesconi A. H. D., Atzori A S., *et al*. (2014) Feeding strategies to design the fatty acid profile of sheep milk and cheese. Revista Brasileira de Zootecnia 43, 445–56.

Oleksyk T. K., Smith M. W., & O'Brien S. J. (2010) Genome-wide scans for footprints if natural selection. Philosophical Transactions of the Royal Society B Biological Science 365, 185-205.

Pardini A., & Nori M. (2011) Agro-silvo-pastoral systems in Italy: integration and diversification. Pastoralism: Research, Policy and Practice 1, 26-35

Parés Casanova P. M., & Bravi R. (2014) Directional and fluctuating asymmetries in domestic sheep skulls. Journal of Zoological and Bioscience Research 1-2, 11-17.

Pariset L., Mariotti M., Gargani M., Joost S., Negrini R., Perez T., *et al*. (2011) Genetic Diversity of Sheep Breeds from Albania, Greece, and Italy Assessed by

Mitochondrial DNA and Nuclear Polymorphisms (SNPs). Scientific World Journal 11, 1641–1659.

Petersen J. L., Mickelson J. R., Cothran E. G., Andersson L. S., Axelsson J., Bailey E., *et al*. (2013) Genetic Diversity in the Modern Horse Illustrated from Genome-Wide SNP Data. PLoS One 8, e54997.

Piggins D., & Phillips C. J. C. (1996) The eye of the domesticated sheep with implications for vision. Animal Science 62, 301–308.

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

Qu Y., Liu Y., Ma L., Sweeney S., Lan X., Chen Z., *et al*. (2011) Novel SNPs of butyrophilin (BTN1A1) and milk fat globule epidermal growth factor (EGF) 8 (MFG-E8) are associated with milk traits in dairy goat. Molecular biology reports 38, 371-377.

Randhawa I. A. S, Khatkar M. S., Thomson P. C., & Raadsma H. W. (2014) Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. BMC Genetics 15, 34.

Rasali D. P., Shrestha J. N. B., & Crow G. H. (2005) Development of composite sheep breeds in the world: a review. Canadian journal of animal science 86, 1-24.

Reppert S. M., & Weaver D. R. (2001) Molecular analysis of mammalian circadian rhythms. Annual Review of Physiology. 63, 647–76

Rivals I., Personnaz L., Taing L., & Potier M. C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23, 401–407.

Rubin C. G., Zody M. C., Eriksson J., Meadows J. R. S., Sherwood S., Webster M. T., *et al*. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464, 587-591.

Schneider M. R., Antsiferova M., Feldmeyer L., Dahlhoff M., Bugnon P., Hasse S., *et al*. (2008) Betacellulin regulates hair follicle development and hair cycle induction and enhances angiogenesis in wounded skin. Journal of Investigative Dermatology 128, 1256-1265.

Schneider M. R., Werner S., Paus S., & Wolf E. (2008) Beyond Wavy Hairs: The Epidermal Growth Factor Receptor and Its Ligands in Skin Biology and Pathology. The American journal of pathology 173, 14-24.

Sorbolini S., Gaspa G., Steri R., Dimauro C., Cellesi M., Stella A., *et al*. (2016) Use of canonical discriminant analysis to study signatures of selection in cattle. Genetics Selection Evolution 48, 58

Soyeurt H., Dardenne P., Gillon A., Croquet C., Vanderick S., Mayeres P., *et al*. (2006) Variation in fatty acid contents of milk and milk fat within and across breeds. Journal of Dairy Science 89, 4858-4865.

Szklarczyk D., Morris J. H., Cook H., Kuhn M., Wyder S., Simonovic M., Santos A., Doncheva N. T., Roth A., Bork P., Jensen L. J., & von Mering C. (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Research 45, 362-68

Tarakci H., & Berger J. (2016) The sarcoglycan complex in skeletal muscle. Frontiers in bioscience 21, 744-756.

Tsuboi M., Gonzalez-Voyer A., & Kolm N. (2014) Phenotypic integration of brain size and head morphology in Lake Tanganyika Cichlids. BMC evolutionary biology 14, 39.

Van Rossum G. (1995) Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam.

Vigne J. D. (2011) The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. C R Biol 334, 171-181.

Wang H., Zhang L., Cao J., Wu M., Ma X., Liu Z., *et al*. (2015) Genome-wide specific selection in three domestic sheep breeds. PloS One 10, e0128688.

Wang M., Jing Y., Hu L., Gao J., Ding L., & Zhang J. (2015) Recent advances on the circadian gene PER2 and metabolic rhythm of lactation of mammary gland. Animal Nutrition 1, 257-261.

Wang Z., Zhang H., Yang H., Wang S., Rong E., Pei W., *et al*. (2014) Genome-wide association study for wool production traits in a Chinese Merino sheep population. PLoS One 9, e107101.

Wei C., Wang H., Liu G., Wu M., Cao J., Liu Z., *et al*. (2015) Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. BMC Genomics 16, 194.

Wiking L., Stagsted J., Björck L., & Nielsen J. H. (2004) Milk fat globule size is affected by fat production in dairy cows. International Dairy Journal 14, 909-913.

Wood J. D., Enser M., Fisher A. V., Nute G. R., Sheard P. R., *et al*. (2008) Fat deposition, fatty acid composition and meat quality: A review. Meat science 78, 343-358.

Yang J., Li W. R., Lv F. H., He S. G., Tian S. L., Peng W. F., *et al*. (2016) Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments. Molecular Biology and Evolution 33, 2576–92.

Yasueda T., Oshima K., Nakatani H., Tabuchi K., Nadano D., & Matsuda T. (2015) A protective effect of milk fat globule EGF factor VIII (MFG-E8) on the spontaneous fusion of milk fat globules in breast milk. The Journal of Biochemistry 158, 25-35.

Zeng X. C., Chen H. Y., Jia B., Zhao Z. S., Hui W.Q., Wang Z.B., *et al*. (2011) Identification of SNPs within the sheep PROP1 gene and their effects on wool traits. Molecular biology reports 38, 2723–2728.

Zhang L., Mousel M. R., Wu X., Michal J. J., Zhou X., Ding B., *et al*. (2013) Genome-Wide Genetic Diversity and Differentially Selected Regions among Suffolk, Rambouillet, Columbia, Polypay, and Targhee Sheep. PLoS One 8, e65942.

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

Appendix Chapter 5

The use of different statistical approaches highlights consensus signatures of selection harboring diversity among sheep breeds

Table S1. Sheep breed Information: geographic origins, production purposes and morphological traits.

Table S2. Number of significant SNPs found using F_{ST} approach in each pairwise comparison.

Table S3. Known QTLs detected in this study for milk production traits extracted from sheep QTLdb database release 30 (http://animalgenome.org).

Table S4. Known QTLs detected in this study for meat production traits extracted from sheep QTLdb database release 30 (http://animalgenome.org).

Table S5. Known QTLs detected in this study for wool traits extracted from sheep QTLdb database release 30 (http://animalgenome.org).

Table S6. Known QTLs detected in this study for other phenotypic traits extracted from sheep QTLdb database release 30 (http://animalgenome.org).

Figure S1. Plot of the individual scores of the first, second and third canonical values (Can1, Can2, Can3) extracted from the 26 autosomes in the four productive aptitude groups.

Figure S2. Plot of Runs of Homozygosity across the 26 chromosomes.

Table S1. Sheep breed Information: geographic origins, production purposes and morphological traits.

Breed (n. animals)	Geographic origin	Production purposes and morphological traits
Alpagota (24)	Veneto	**Meat**. Uni coloured: white with dark spots on face and legs. Horned.
Altamurana (47)	Puglia	**Dual**. Uni coloured, white, occasionally dark spots on face. It is horned and has, small ears.
Appenninica (24)	Toscana, Umbria	**Meat**. Color: white. It is polled and has semi-lopped ears.
Bagnolese (23)	Campania	**Dual**. White coat with black spots small or large. The ears are long, wide and hanging with small black dots
Bergamasca (24)	Lombardia	**Meat**. Uni coloured: white. It is polled, and has long ears.
Biellese (22)	Piemonte	**Meat**. Uni colored: white. It is polled and has long ears.
Comisana (24)	Sicilia	**Milk**. Colour white and a reddish-brown face. The breed has long ears and is polled.
DelleLanghe (24)	Piemonte, Liguria	**Milk**. Uni colored: white. The breed has semi-lop ears
Fabrianese (23)	Marche	**Meat**. The breed is uni colored white and polled.
Gentile di Puglia (24)	Puglia	**Wool**. The breed is uni coloured: white, horned and with small ears
Istrian Pramenka (24)	Friuli Venezia Giulia	**Dual**. The breed is white, black, spotted, patchy. It is a short-eared and hornless sheep
Laticauda (24)	Campania	**Dual**. The breed is uni colored white, horned and with small ears
Leccese (25)	Puglia	**Milk**. Rosy skin and white coat, with black spots on the breastbone, black short-haired muzzle, plain black or spotted limbs. The breed is polled and with small ears.
Massese (24)	Toscana	**Milk**. Uni coloured: black, grey or brown with darker head. Animals are horned.
Pinzirita (24)	Sicilia	**Milk**. They have black or brown marks on the face and legs. The males are horned and the females are polled.
Sambucana (24)	Piemonte	**Meat**. The colour is yellow-white and both sexes are usually polled and with small ears.

Sardinian ancestral black (20)	Piemonte	**Milk**. The hair is black, sometimes with shades of gray lead. It is horned and with very small ears.
Sardinian white (24)	Sardegna	**Milk**. Uni coloured: white, horned and with small ears
Sopravissana (24)	Marche	**Wool**. Uni colored white, rams are horned and the ewes are polled. It has small ears.
Valle del Belice (24)	Sicilia	**Milk**. White coat; white head without horns in females while in males they may be present or not. Ears are small.

Dual = meat and milk

Table S2. Number of significant SNPs found using F_{ST} approach in each pairwise comparison.

OAR	Dual *vs* Wool	Wool *vs* Milk	Meat *vs* Wool	Meat *vs* Dual	Dual *vs* Milk	Meat *vs* Milk
1	**68**	42	39	57	63	**81**
2	64	**66**	**77**	**92**	**102**	53
3	61	49	29	51	64	25
4	26	36	31	28	36	31
5	38	23	20	26	29	35
6	17	24	30	29	25	49
7	34	38	47	24	18	37
8	26	18	24	22	24	28
9	18	22	33	39	29	20
10	17	15	19	27	31	23
11	18	22	17	4	18	13
12	17	23	23	16	11	20
13	9	25	8	29	15	36
14	19	11	18	15	17	8
15	24	23	26	22	23	19
16	8	15	30	23	20	43
17	22	33	16	10	20	13
18	25	18	26	15	16	16
19	20	22	22	26	22	18
20	19	16	17	7	4	15
21	13	11	10	7	**0**	13
22	10	10	15	15	12	15
23	21	25	26	10	11	18
24	9	**0**	15	14	1	9
25	14	18	7	21	1	14
26	17	12	16	10	10	4
Total	634	617	641	639	622	656

Table S3. Known QTLs detected in this study for milk production traits extracted from sheep QTLdb database release 30 (http://animalgenome.org).

QTL position (Mbp)	SNP name	Trait	Reference that had already reported the QTL
OAR6 (31-45)	*§*	MFA-C14:1	Bouwman *et al*., 2011
OAR9 (35.3-50.2)	*rs405612537**	MFY	Garcia-Gomez *et al*., 2013
OAR13 (23.4-23.5)	*§*	MFY	Garcia-Gomez *et al*., 2013
OAR14 (70.2)	*rs401264364**	MFY	Garcia-Gomez *et al*., 2012
OAR2 (54.7-55.5)	§	MPY	Garcia-Gomez *et al*., 2013
OAR19 (50.4-53.6)	§	MPY	Garcia-Gomez *et al*., 2013
OAR14 (33.8-33.9)	§	MY	Crisà *et al*., 2010
OAR19 (47.5-49.3)	§	MY	Garcia-Gomez *et al*., 2013
OAR13 (24.9-35.3)	§	MPP	Garcia-Gomez *et al*., 2013
OAR16 (30.6-30.7)	§	MF	Crisà *et al*., 2010
OAR16 (70.2)	§	PP	Garcia-Gomez *et al*., 2012
OAR16 (12.2-12.3)	§	PY	Garcia-Gomez *et al*., 2012
OAR19 (50.4-53.6)	§	PY	Garcia-Gomez *et al*., 2012

MFY=milk fat yield; MPY=milk protein yield; MY= milk yield; MPP= milk protein percentage; MF=milk fat percentage; PP= milk protein percentage; PY= milk protein yield; MFA-C14:1=milk fatty acid cis 9-C14:1 percentage
*= common SNP between literature and the present study;
§= common QTL position but different significant SNP between literature and the present study.

Table S4. Known QTLs detected in this study for meat production traits extracted from sheep QTLdb database release 30 (http://animalgenome.org).

QTL position (Mbp)	SNP name	Trait	Reference that had already reported the QTL
OAR3 (213.7)	§	ADG	Zhang at al., 2013
OAR6 (38-40)	§	ADG	Lindholm-Perry *et al.*, 2011
OAR8 (14.7)	§	ADG	Zhang at al., 2013
OAR13 (34.8)	§	ADG	Zhang at al., 2013
OAR14 (33.7)	§	ADG	Zhang at al., 2013
OAR16 (56.1-56.3)	§	ADG	Zhang at al., 2013
OAR18 (35.1-35.2)	§	ADG	Zhang at al., 2013
OAR20 (16.2)	§	ADG	Zhang at al., 2013
OAR22 (2.3)	*rs410274217**	ADG	Zhang *et al.*, 2013
OAR26 (8.4)	*rs406486309**	ADG	Zhang *et al.*, 2013
OAR6 (36.8)	*§*	DP	Matika *et al.*, 2016
OAR3 (120.9)	*rs427339855**	BW	Al-Mamun *et al.*, 2015
OAR6 (70.2)	*rs419653967**	BW	Al-Mamun *et al.*, 2015
OAR12 (70.2)	*§*	BW	Al-Mamun *et al.*, 2015
OAR13 (5.8)	*§*	BW	Gholizadeh *et al.*, 2015
OAR16 (42.8)	*§*	BW	Gholizadeh *et al.*, 2015
OAR19 (70.2)	*rs404704213**	BW	Al-Mamun *et al.*, 2015
OAR13 (75.4-81.4)	§	CC	Matika *et al.*, 2016
OAR19 (51.1-56.9)	*§*	MDEN	Matika *et al.*, 2016
OAR19 (0.3-6.2)	*§*	TBONE	Matika *et al.*, 2016

ADG=average daily gain; BW=body weight; CC= carcass composition; DP=dressing percentage; MDEN=muscle density; TBONE=total bone
*= common SNP between literature and the present study; §= common QTL position but different significant SNP between literature and the present study.

Table S5. Known QTLs detected in this study for wool traits extracted from sheep QTLdb database release 30 (http://animalgenome.org).

QTL position (Mbp)	SNP name	Trait	Reference that had already reported the QTL
OAR1 (208.7-232.6)	§	FLYD	Roldan *et al*., 2010
OAR1 (240.3-254.8)	§	FCURV	Roldan *et al*., 2010
OAR4 (52.8)	*rs421033324**	MFDIAM	Wang *et al*., 2014
OAR5 (36.0-41.0)	§	MFDIAM	Zeng *et al*., 2011
OAR6 (57.0-57.6)	§		Wang *et al*., 2014
OAR9 (29.8)	*rs424763565**	CRIMP	Wang *et al*., 2014
OAR11 (40.4–40.7)	*§*	FLYD	Roldan *et al*., 2010
OAR13 (9.8)	*§*	CVFD, FDSD	Wang *et al*., 2014
OAR13 (17.1)	*rs402243256**	MFDIAM	Wang *et al*., 2014
OAR23 (55.4)	*§*	CRIMP	Wang *et al*., 2014

FLYD= fleece yield; CVFD=coefficient variance of fiber diameter; FCURV=fibre curvature
MFDIAM=mean fiber diameter; CRIMP=wool crimp; FDSD=fiber diameter standard deviation
*= common SNP between literature and the present study;
§= common QTL position but different significant SNP between literature and the present study.

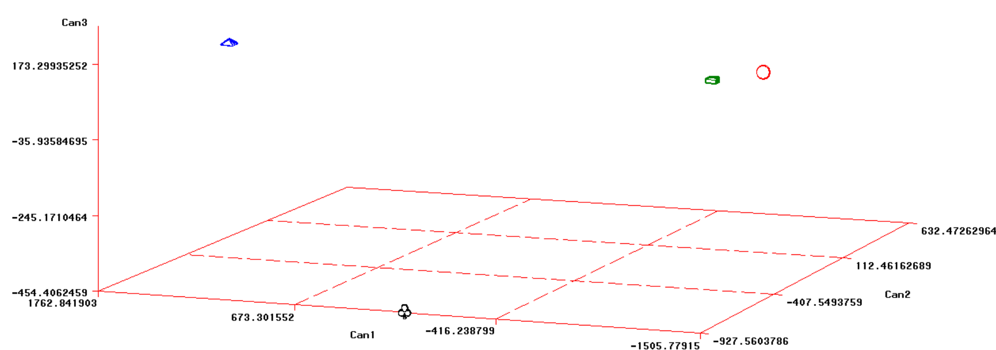
Table S6. Known QTLs detected in this study for other phenotypic traits extracted from sheep QTLdb database release 30 (http://animalgenome.org).

QTL position (Mbp)	SNP name	Trait	Reference that had already reported the QTL
OAR5 (32.6-32.7)	§	OR	Mullen and Hanrahan, 2014
OAR8 (68.5)	§	MVVS (host defence)	White *et al*., 2012
OAR13 (52.0)	§	MVVS (host defence)	White *et al*., 2012
OAR14 (14.2)	rs418443666*	COCO	Kijas *et al*., 2013

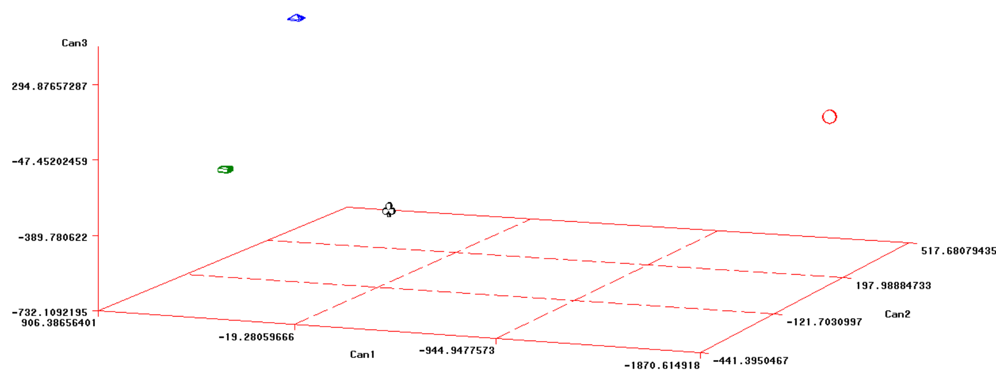
MVVS=maedi-visna virus susceptibility; COCO= coat color; OR=ovulation rate;
*= common SNP between literature and the present study;
§= common QTL position but different significant SNP between literature and the present study.

Figure S1. Plot of the individual scores of the first, second and third canonical values (Can1, Can2, Can3) extracted from the 26 autosomes in the four productive aptitude groups.

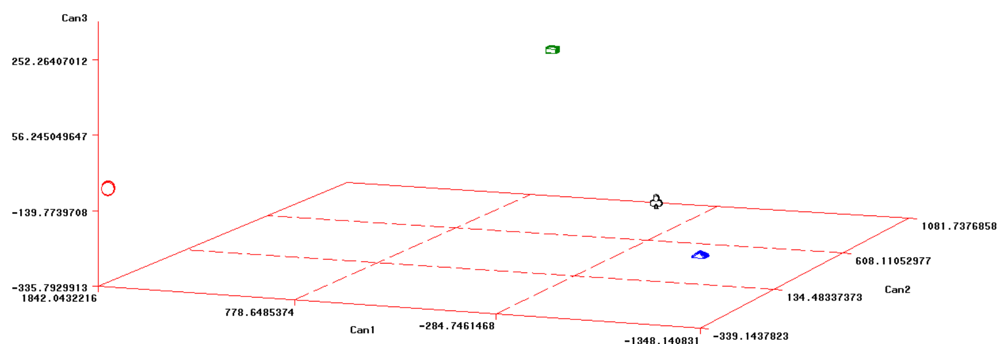
Chromosome 1



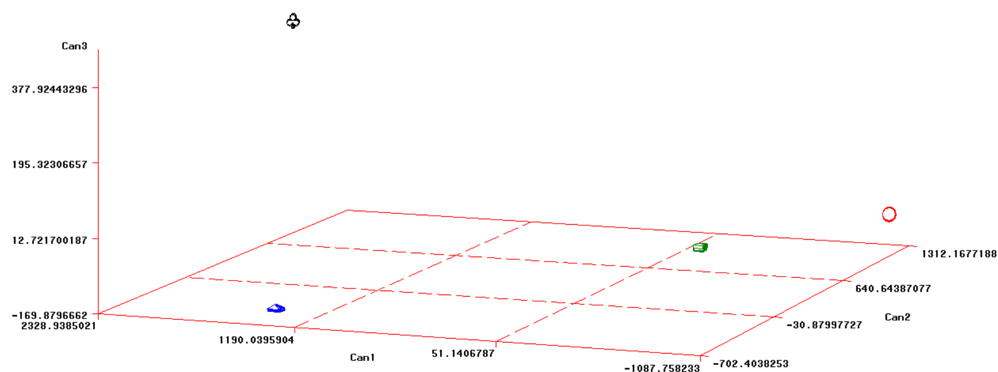
Chromosome 2



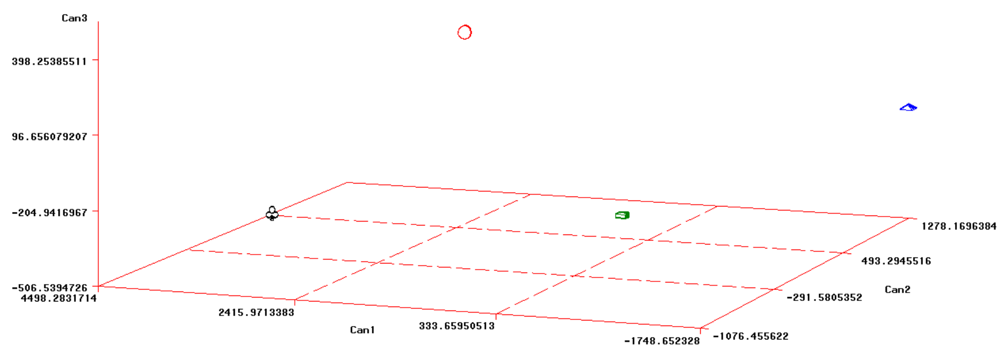
Chromosome 3



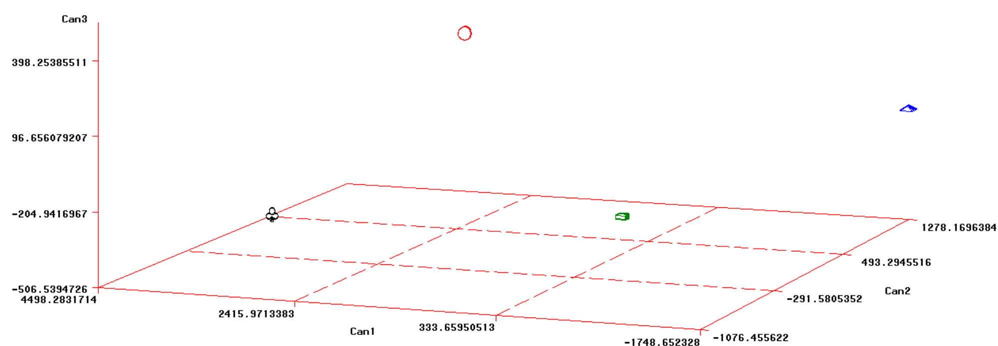
Chromosome 4



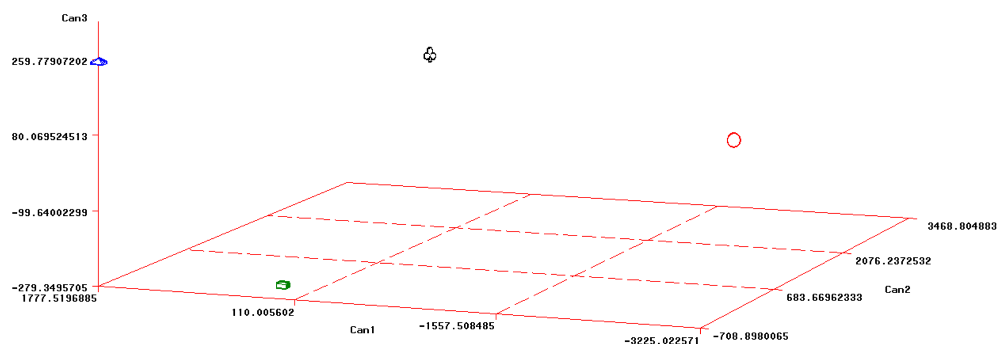
Chromosome 5



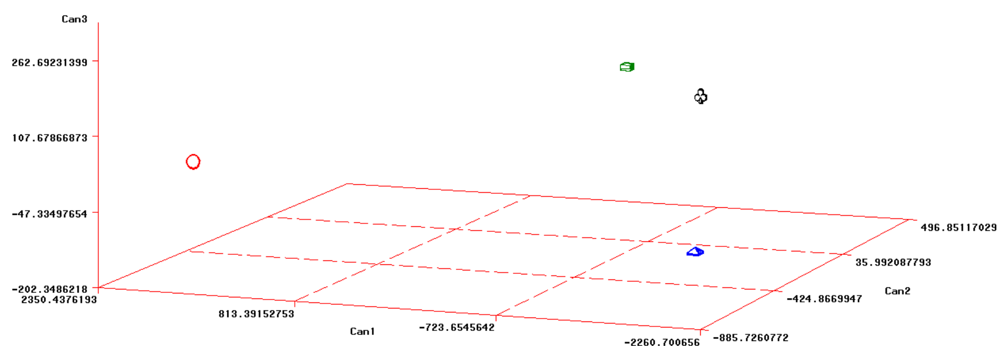
Chromosome 6



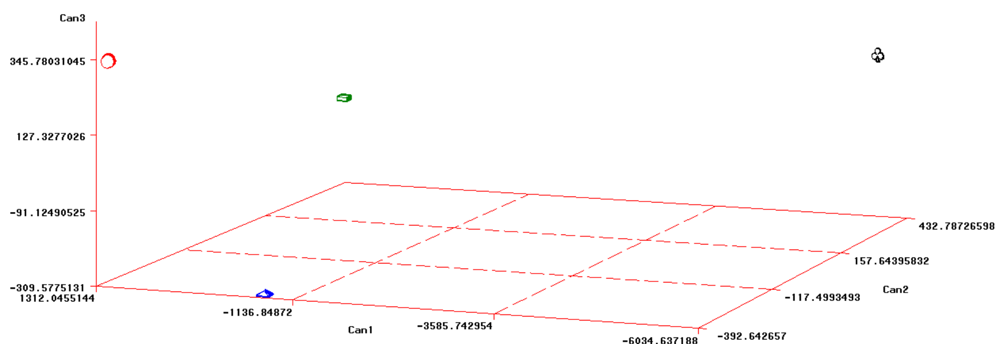
Chromosome 7



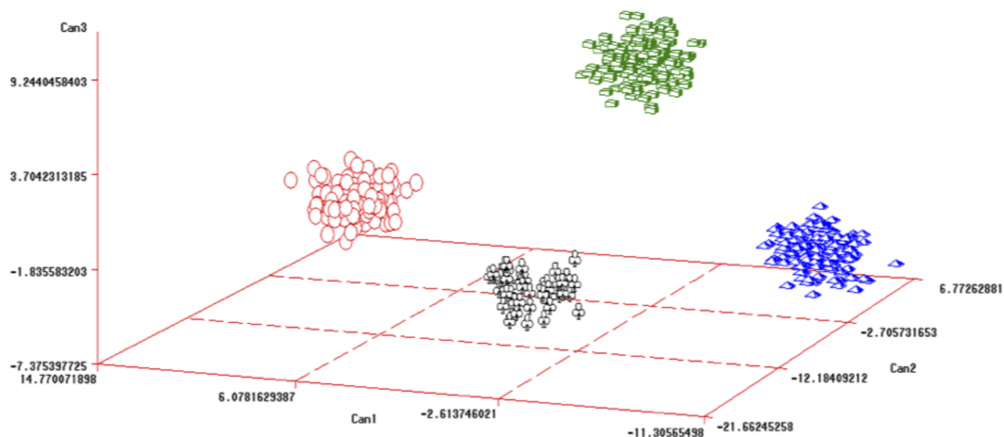
Chromosome 8



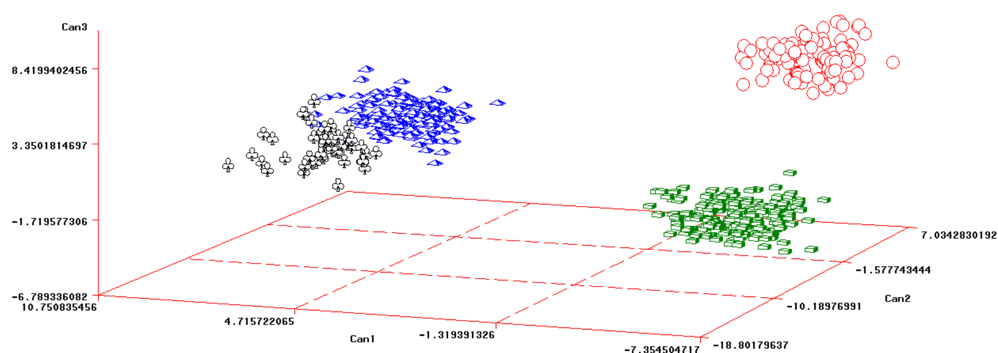
Chromosome 9



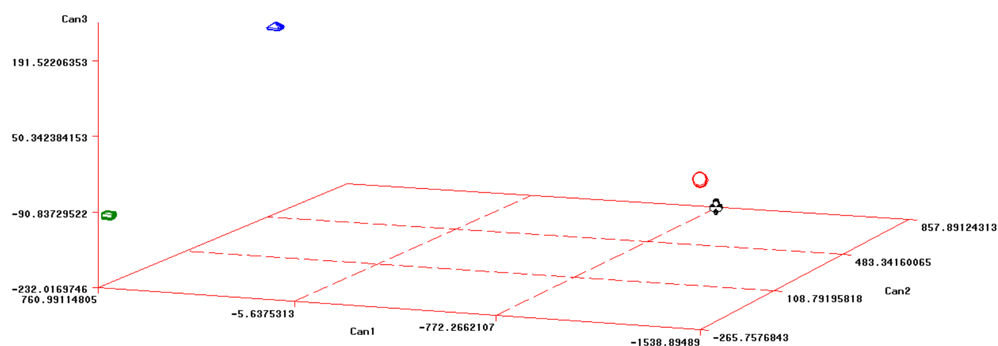
Chromosome 10



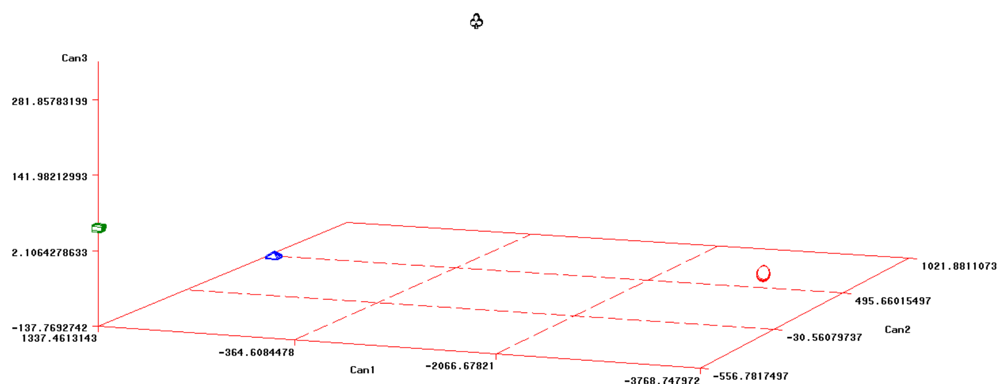
Chromosome 11



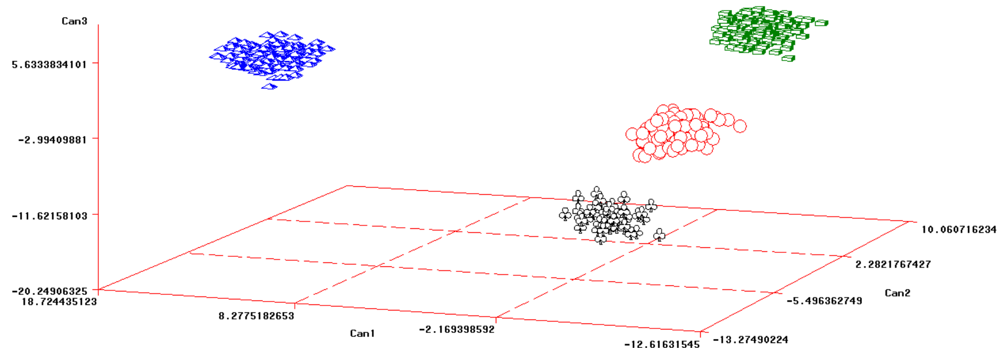
Chromosome 12



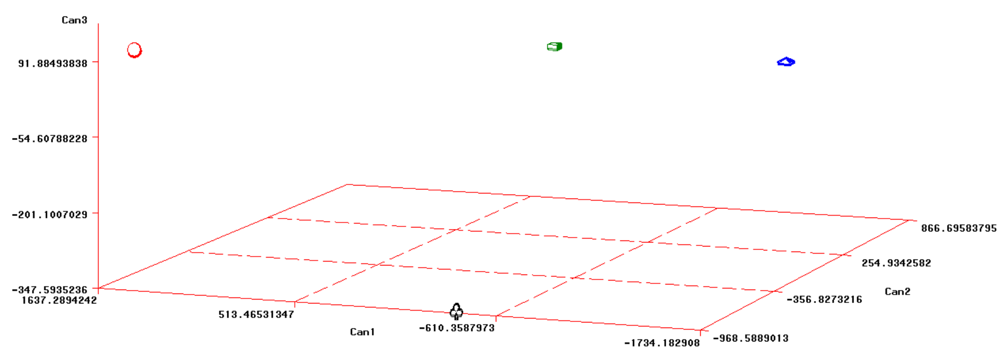
Chromosome 13



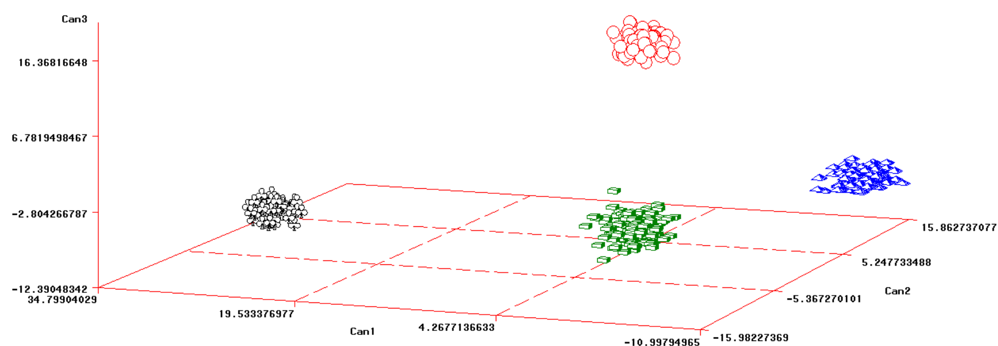
Chromosome 14



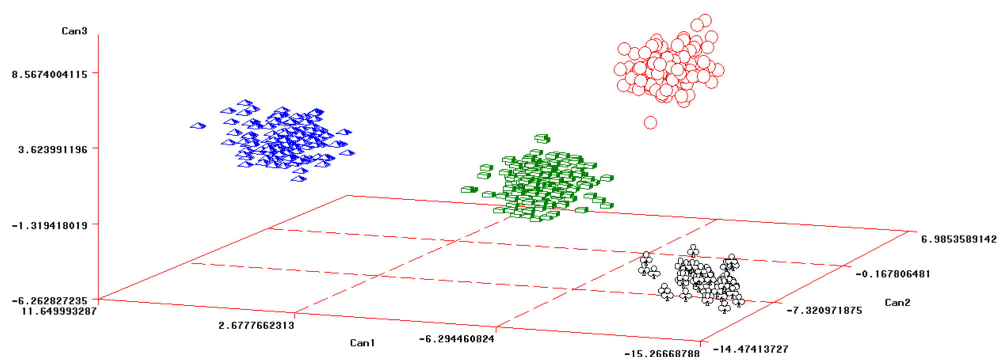
Chromosome 15



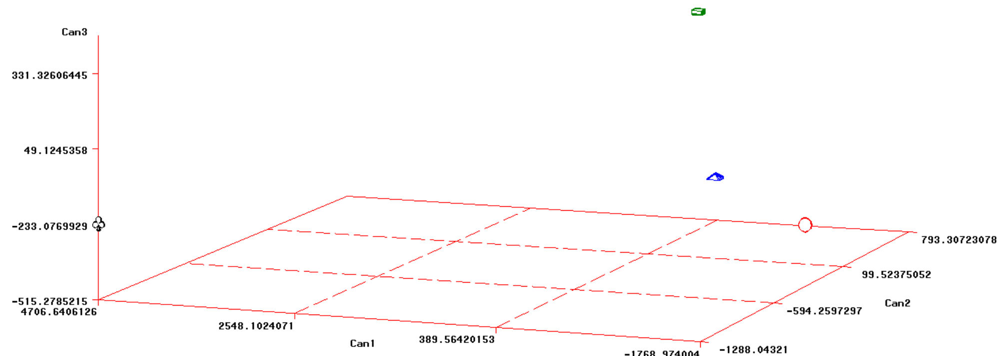
Chromosome 16



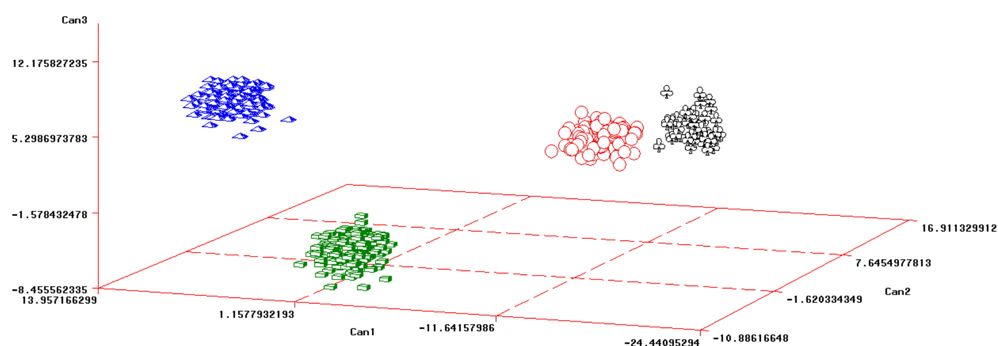
Chromosome 17



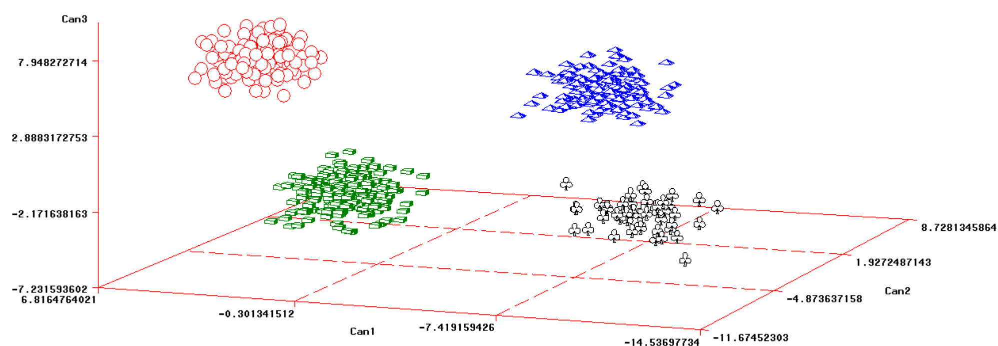
Chromosome 18



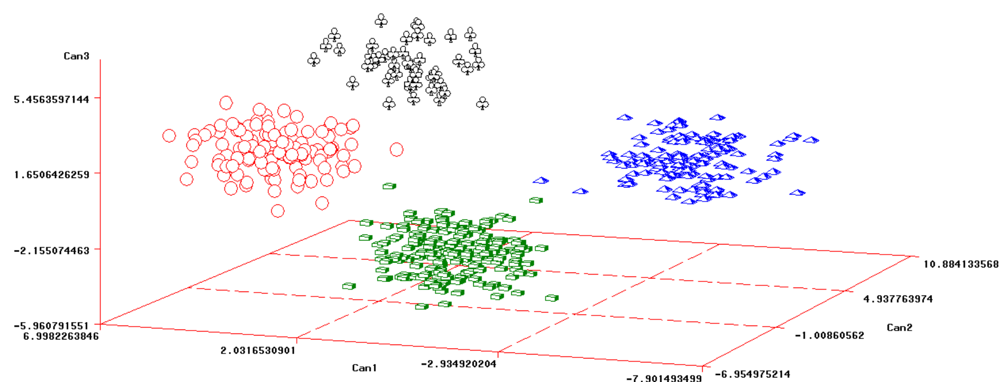
Chromosome 19



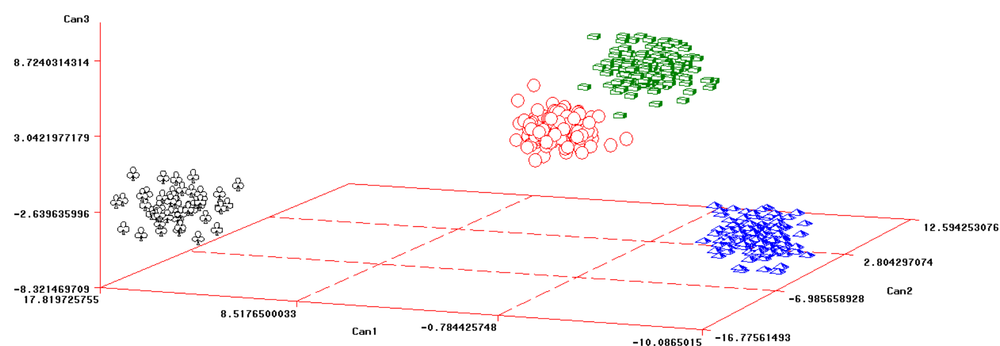
Chromosome 20



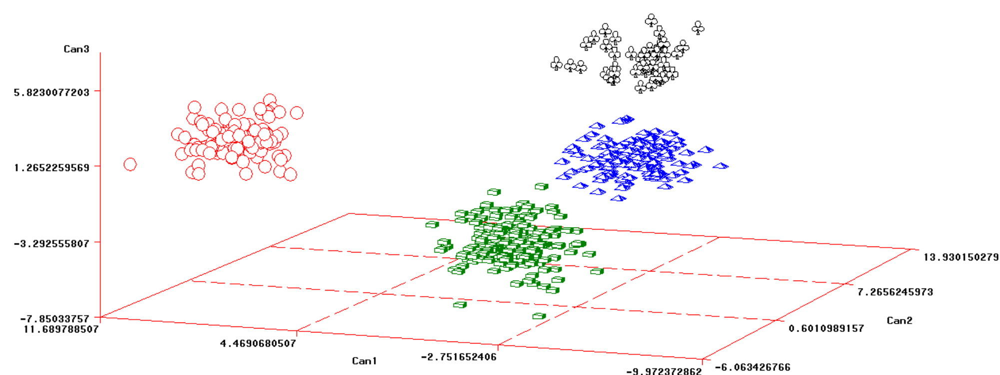
Chromosome 21



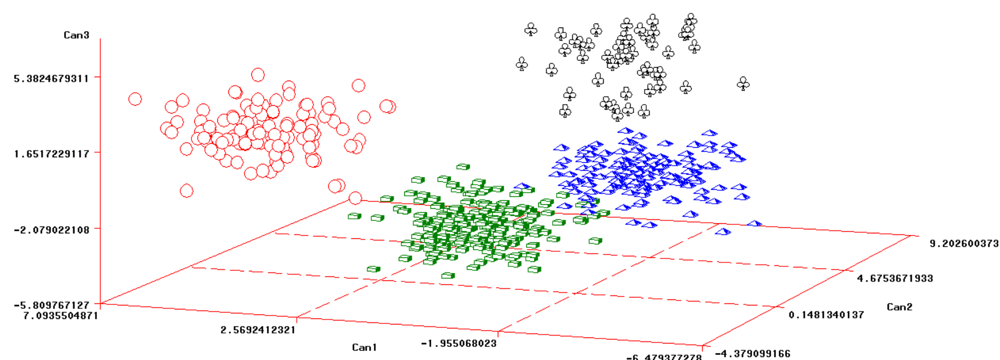
Chromosome 22



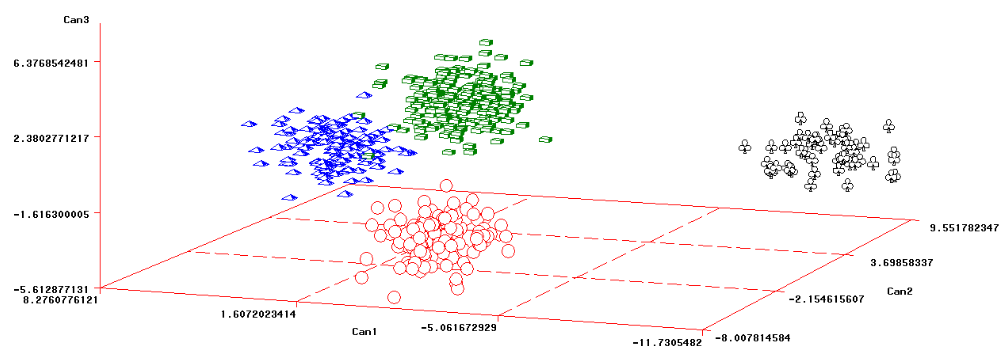
Chromosome 23



Chromosome 24



Chromosome 25



Chromosome 26

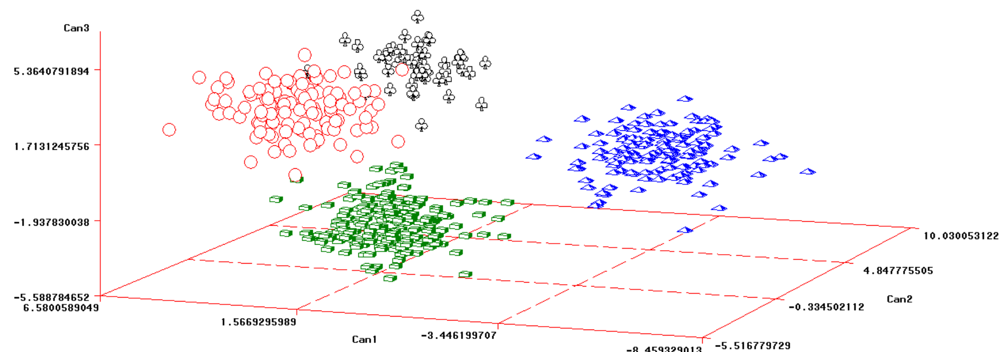
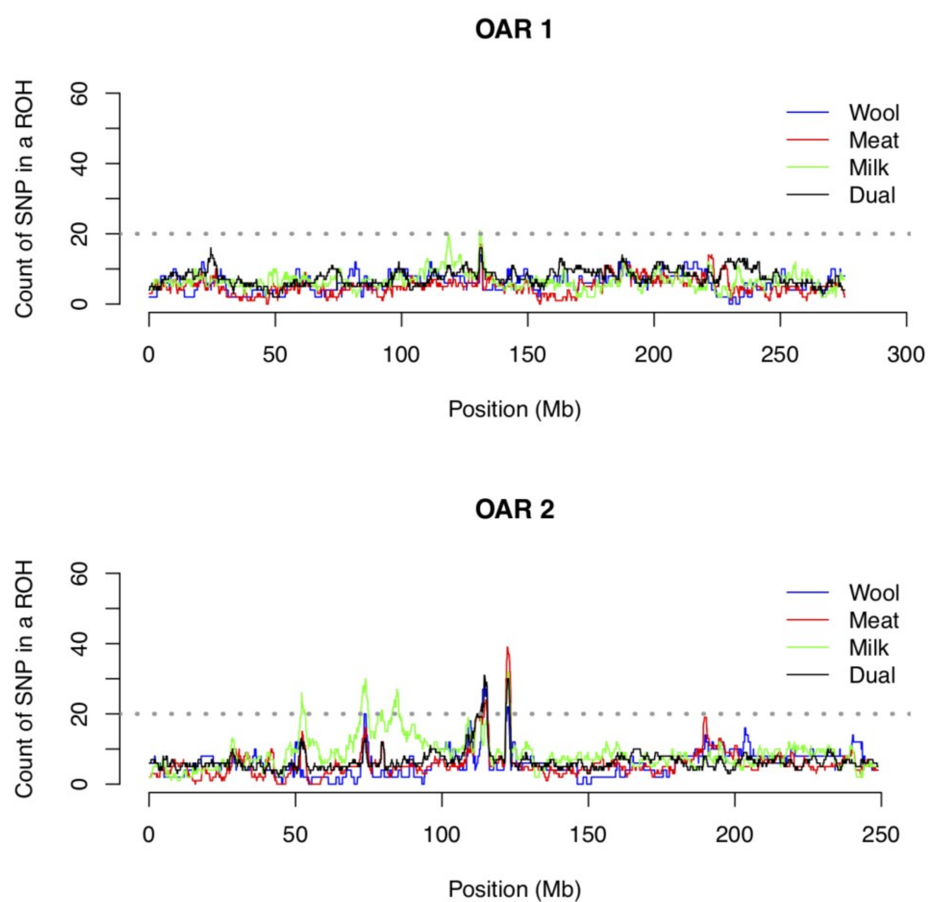
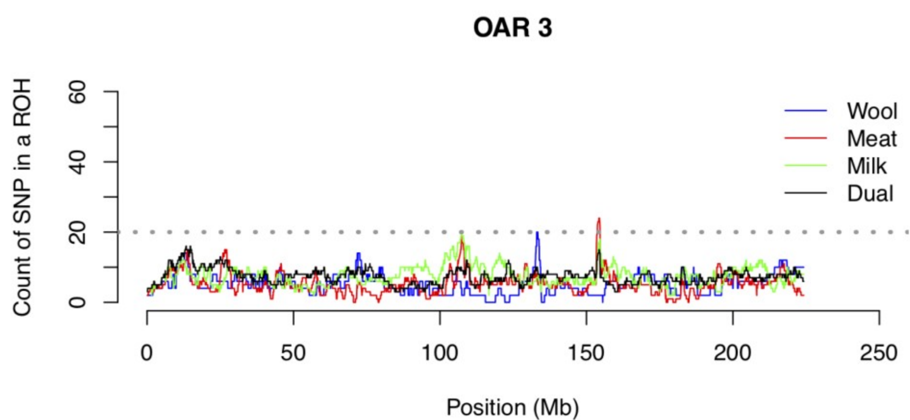
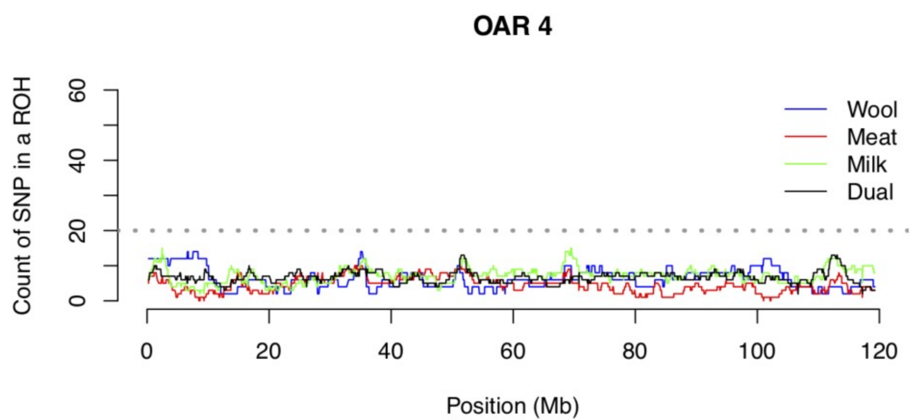
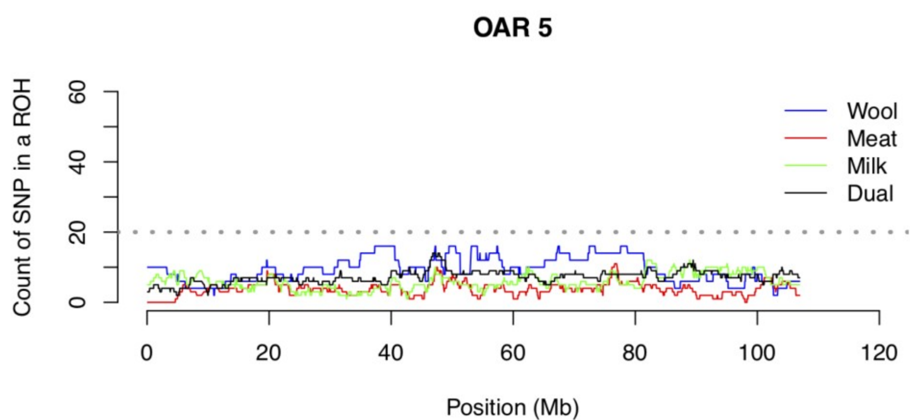


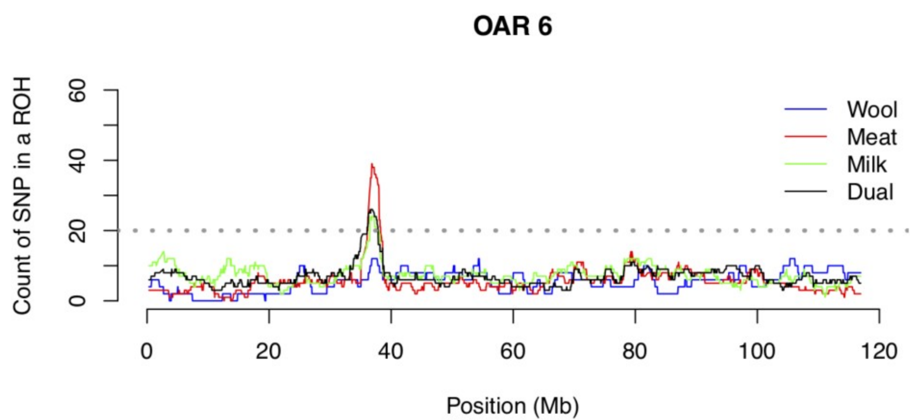
Figure S2. Plot of Runs of Homozygosity across the 26 chromosomes.

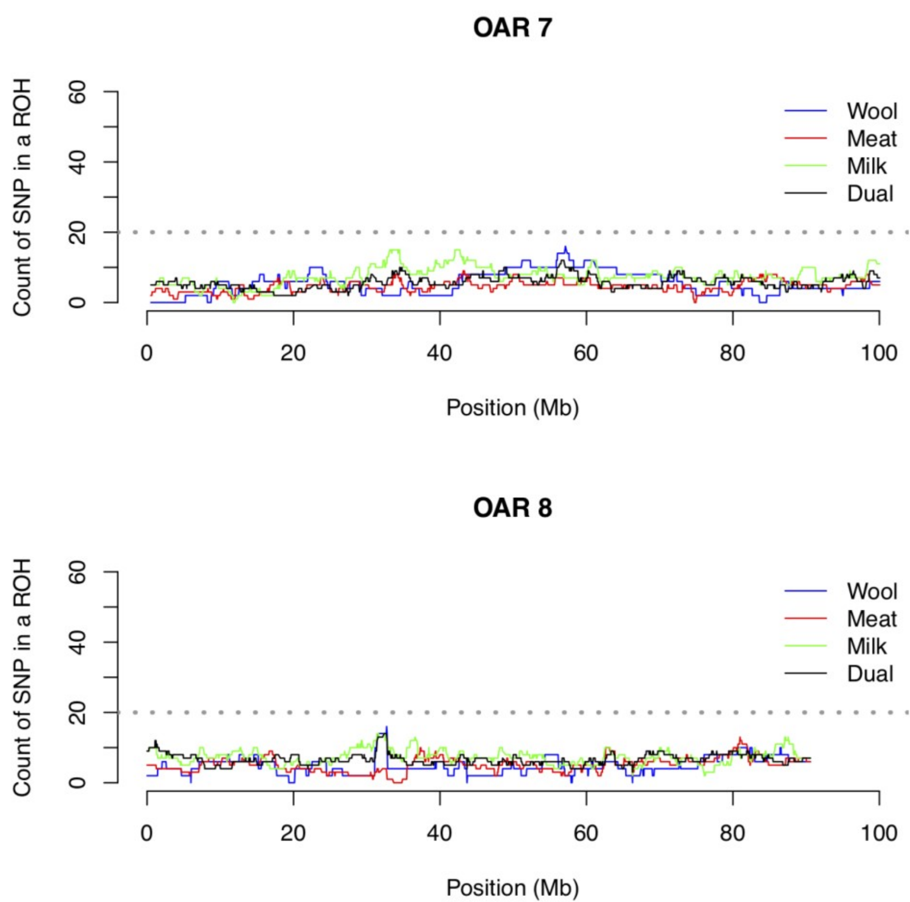


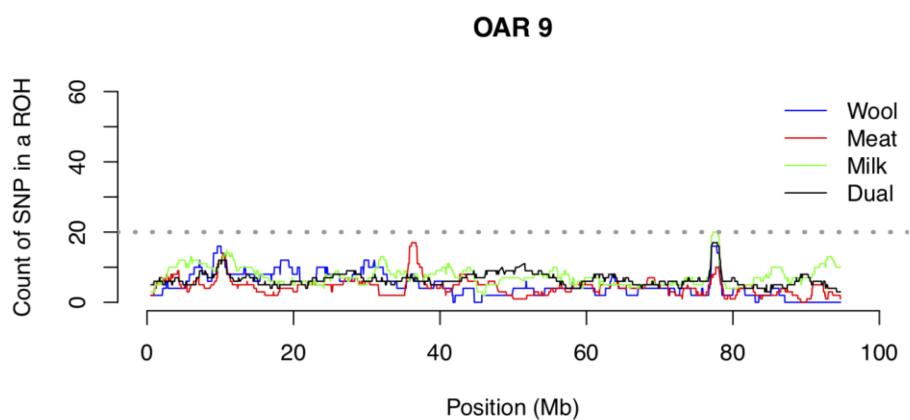

OAR 3
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
50
100
150
200
250
Position (Mb)

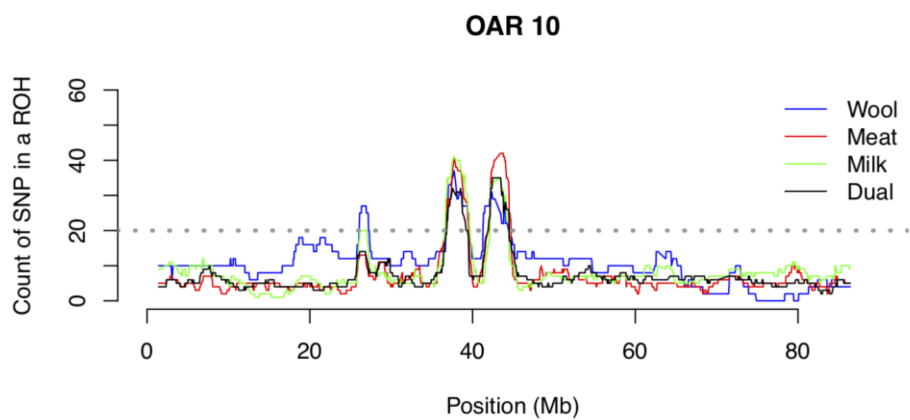

OAR 4
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
100
120
Position (Mb)

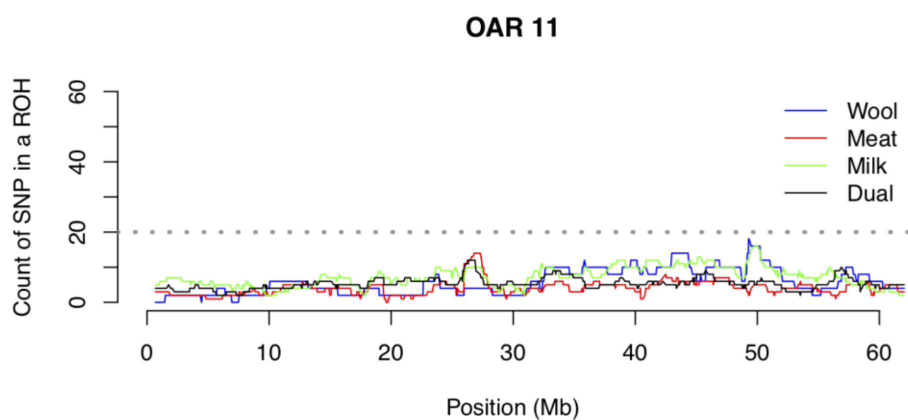

OAR 5
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
100
120
Position (Mb)

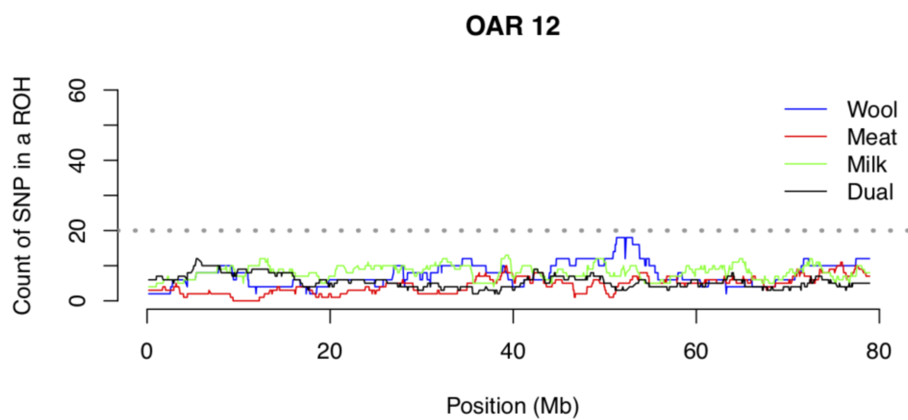

OAR 6
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
100
120
Position (Mb)

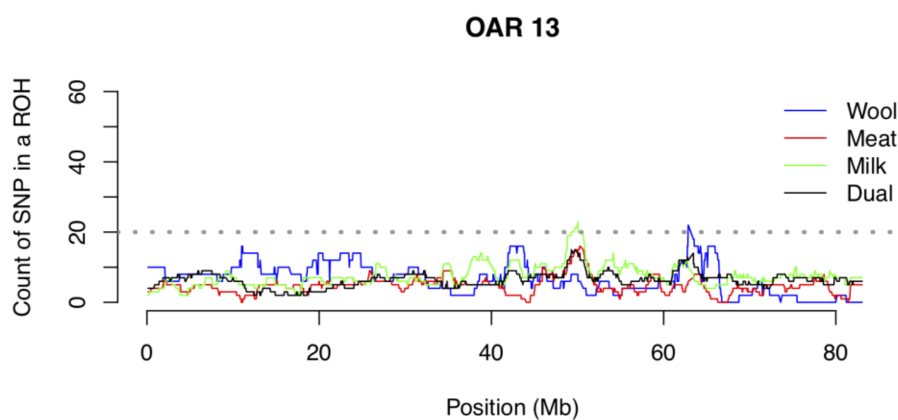

OAR 7
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
100
Position (Mb)
OAR 8
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
100
Position (Mb)

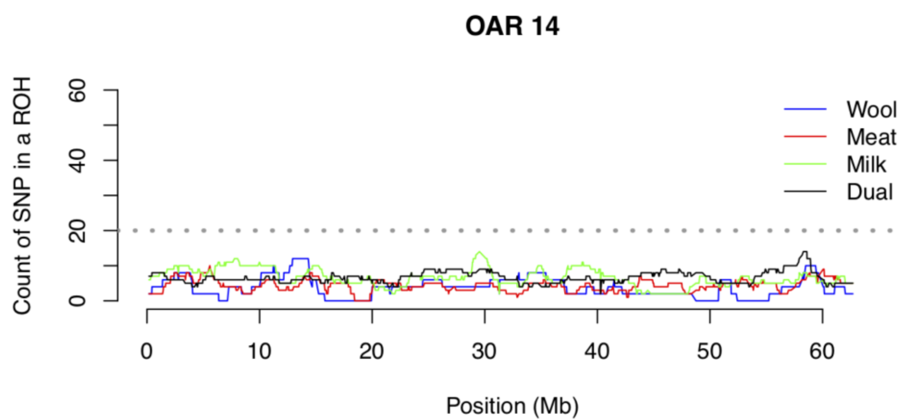

OAR 9
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
100
Position (Mb)

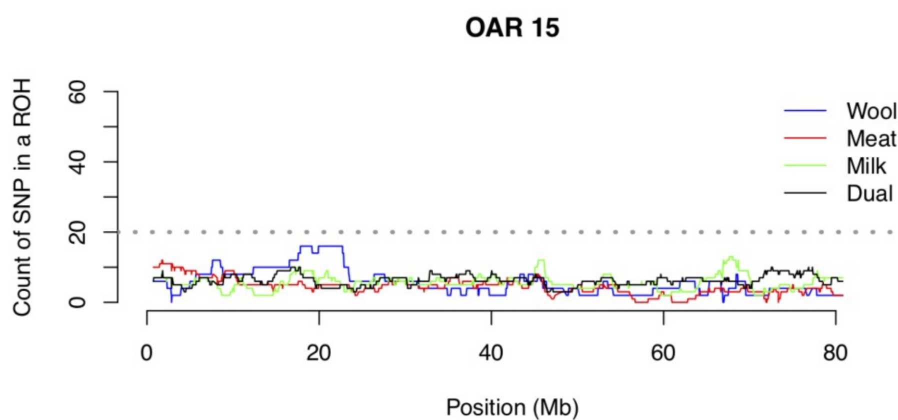

OAR 10
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
Position (Mb)

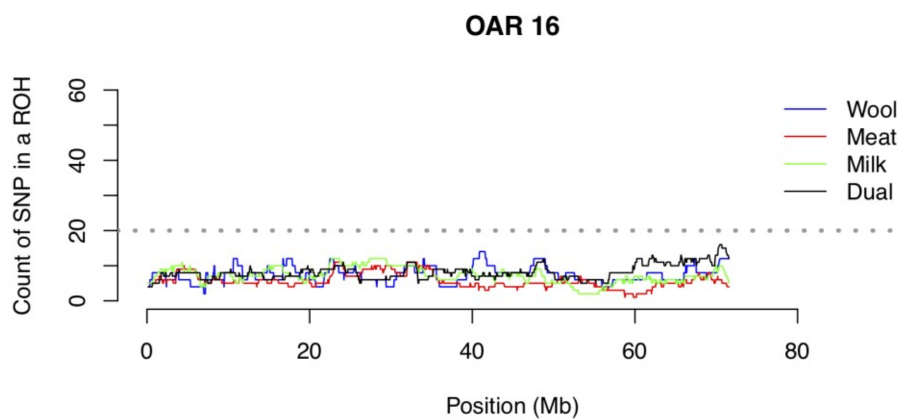

OAR 11
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
60
Position (Mb)

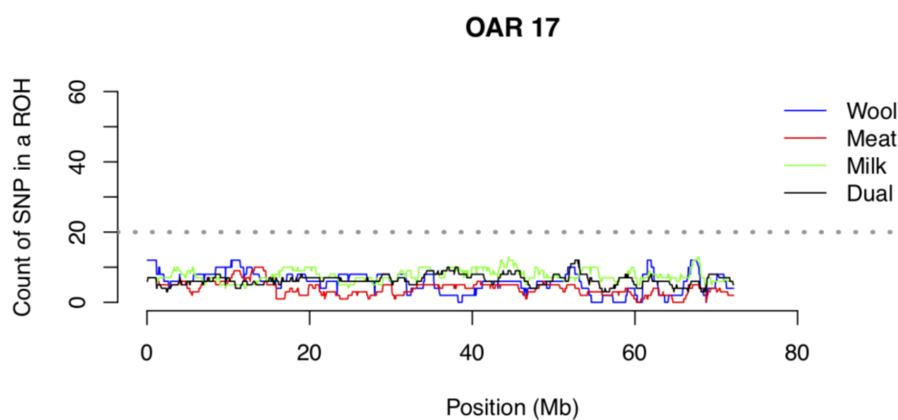

OAR 12
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
Position (Mb)

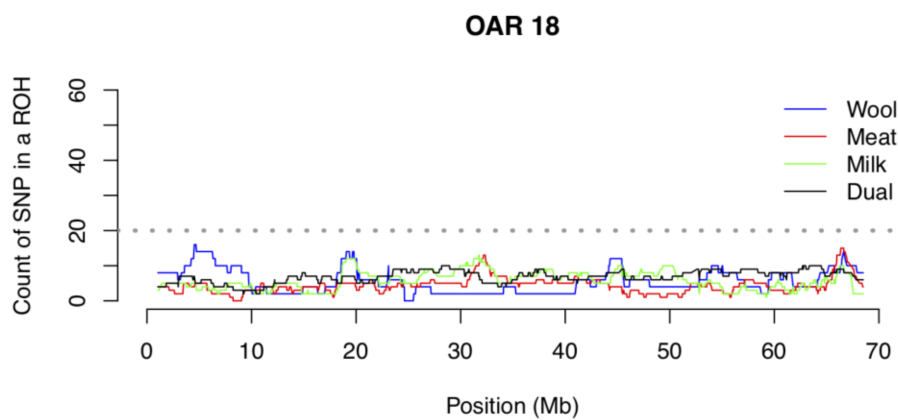

OAR 13
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
Position (Mb)

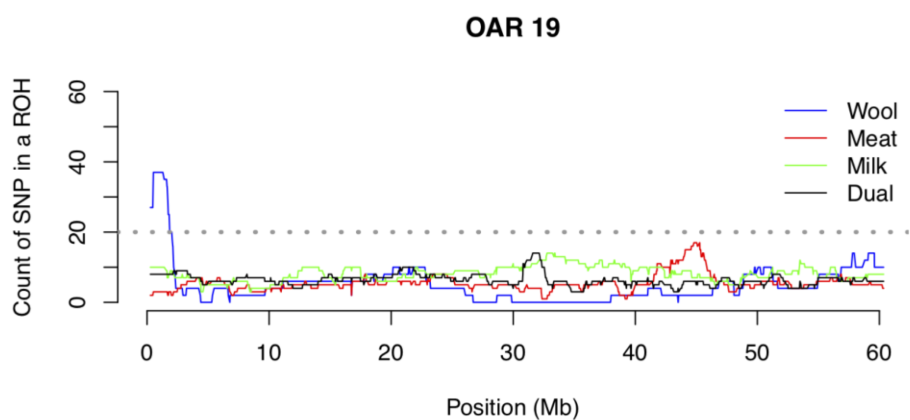

OAR 14
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
60
Position (Mb)

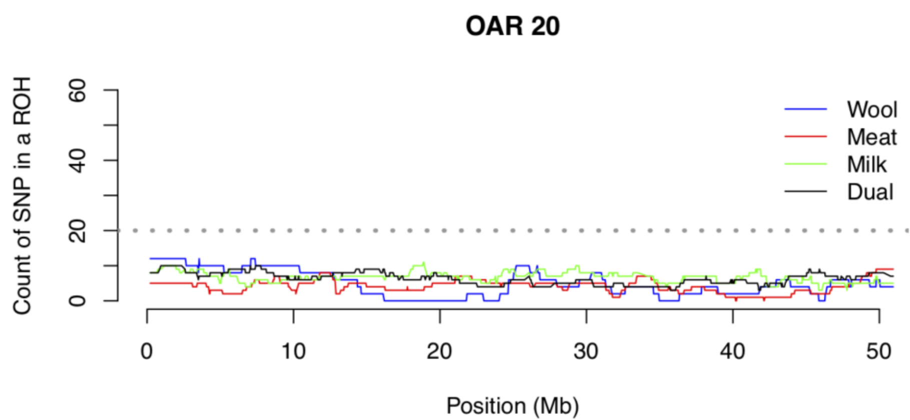

OAR 15
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
Position (Mb)

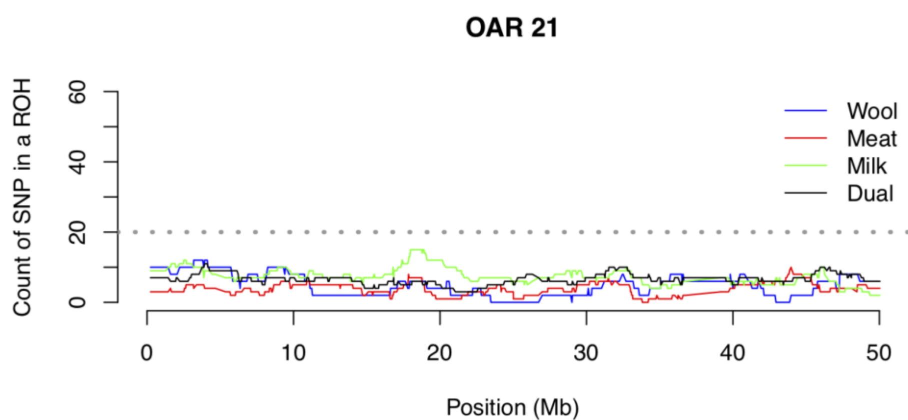

OAR 16
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
Position (Mb)

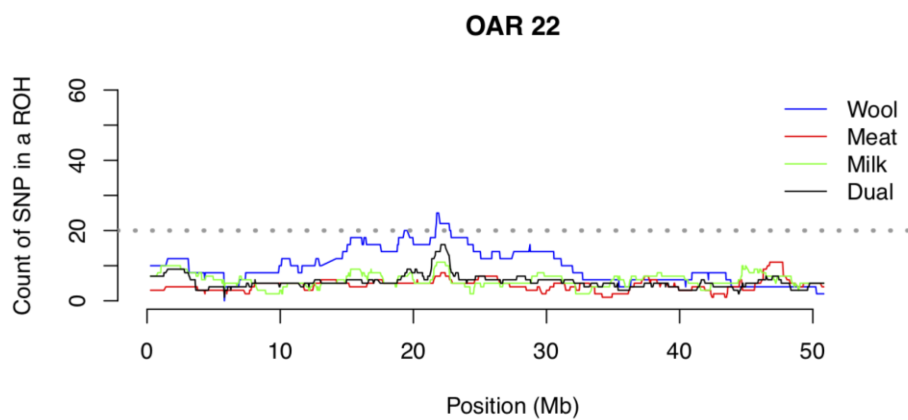

OAR 17
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
20
40
60
80
Position (Mb)

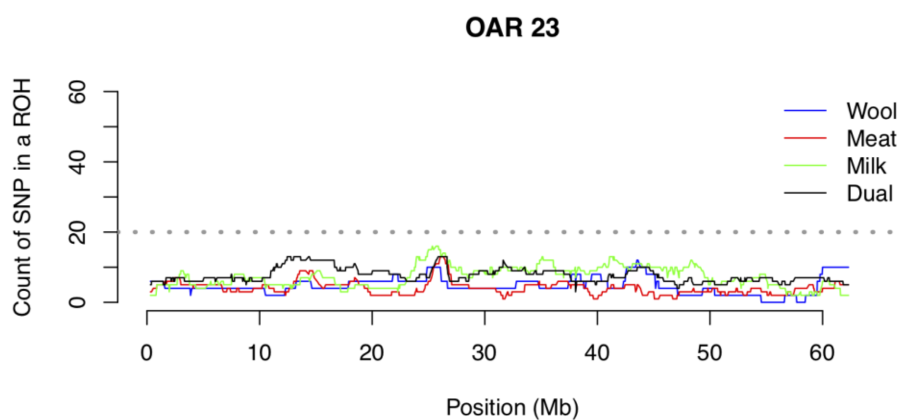

OAR 18
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
60
70
Position (Mb)

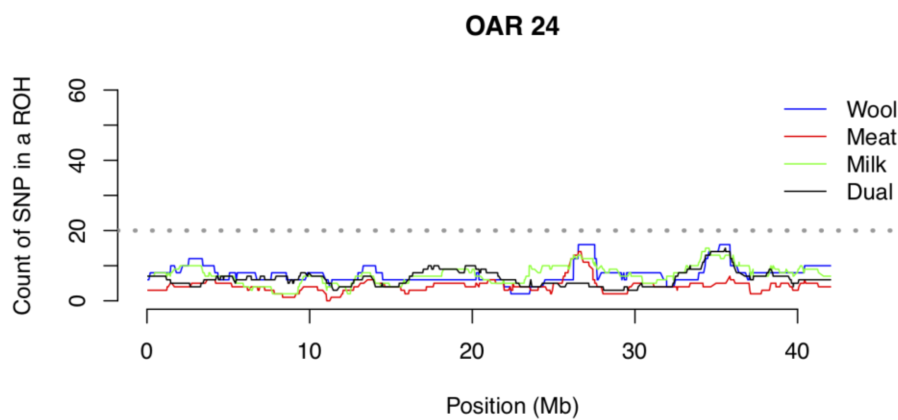

OAR 19
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
60
Position (Mb)

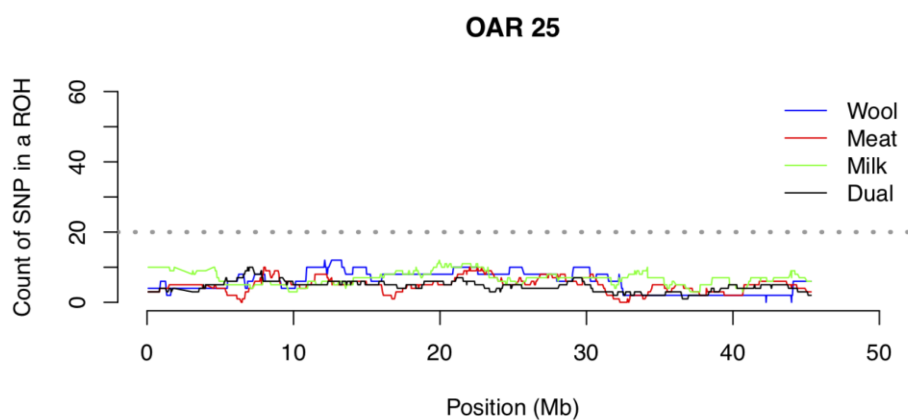

OAR 20
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
Position (Mb)

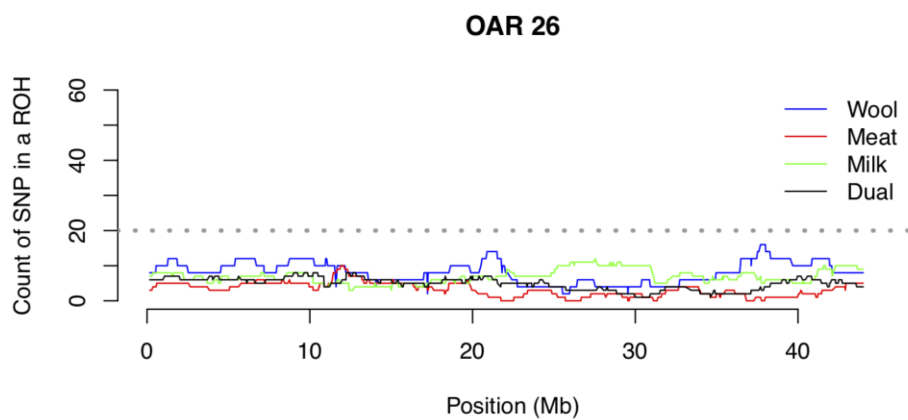

OAR 21
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
Position (Mb)


OAR 22
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
Position (Mb)


OAR 23
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
60
Position (Mb)


OAR 24
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
Position (Mb)


OAR 25
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
50
Position (Mb)


OAR 26
Count of SNP in a ROH
0
20
40
60
Wool
Meat
Milk
Dual
0
10
20
30
40
Position (Mb)

CHAPTER 6

General conclusion

This thesis debated the usefulness of use genomic information in animal science: thanks to its several applications, this kind of information has become fundamental for the modern livestock sector. More precisely, the attention was focused on role of the genetic (co)variance structure among animals. This structure – calculated using SNP markers – is largely implemented in different branches: animal breeding, biodiversity, products valorization *etc*.

In the first two experimental contributions applications of genetic (co)variance structure on animal breeding were evaluated. Genomic tools (e.g. *SNP beadchip*) are successfully implemented in almost all breeding programs for cosmopolitan breeds: the estimated breeding values predicted using also genomic information are more accurate than traditional pedigree-based. Since the breeding values depend also on relationship among animals, the use of genotypes allows to compute more precisely the exact amount of DNA shared by two animals and estimate more reliable genomic breeding values. Moreover, in order to be sure of relationship among animals, genomic tools are implemented in the parentage verification and discovery: with genotypes available for one animal and its parents, it can be accepted or excluded this parentage; in the case that one parent has a status of excluded, the possible parent can be discovered among the genotyped animals of the population.

The second chapter of this thesis demonstrated the importance of choosing animals that must be genotyped to construct the genetic structure for heritability estimation purposes. In order to have a pool of genotyped animals that can better represent the entire genetic structure of the population, the best strategy seemed to be the random genotyping of females; while the worst case was to select the best animals,

especially when a population is under selection. In the simulation, the numbers of breeding male and females were chosen according to a supposed breeding nucleus for sheep breeds. The number of one male for 50 females was decided according to the real ratio used in Sarda breed for natural matings.

In the third chapter, the genetic structure was analyzed with different prediction models (only pedigree, only genotypes or both pedigree and genotypes) to verify the possibility of including FA milk profile as breeding goal in dairy sheep. Genomic models (i.e. using genotypes) gave better GEBV accuracies for both animals with and without phenotypic information.

Another field that took huge advantages of genomic tools is the study and safeguard of animal biodiversity. Genotypes can be used to establish how modern breeds are born and to calculate genetic distance among these breeds. Moreover, genetic structure of animal populations can give evidences of selective pressure that livestock has withstood to increase its production. The fourth and fifth chapters of this thesis investigated these two aspects: genetic diversity and relationship between purebreds and their crossbreeds and selection signatures. Results of the 4th chapter confirmed the goodness of genomic tools to study distances among different breeds and to identify genomic regions that can distinguish or assemble breeds. In this case, the genomic analysis gave information also about the geographic localization of cattle breeds raised in Italy: results of these analysis showed not only differences among breeds, but also among the latitude on which animals are raised. In the last chapter of this thesis, *SNP beadchips* have been used to identify selection sweeps for different productive aptitudes (milk, meat, dual and wool). Once again, genetic structure offered

a deeper possibility to understand the complicated state of livestock production that with only phenotypic manifestations is not possible to decipher.