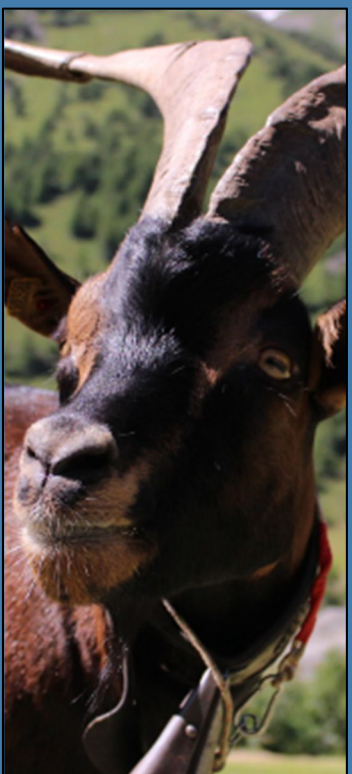
Università degli Studi di Milano

Analysis of genomic data in ruminant species for parentage, product tracing and population structure studies



Scuola di dottorato in Scienze Veterinarie e dell'allevamento

Doctoral School of Veterinary and Animal Sciences

XXX Cycle

PhD Candidate: Andrea Talenti

R10830

Tutor: Prof.ssa Paola Crepaldi

Academic Year 2016/2017

Acknowledgements

Trovo sempre molto difficile scrivere dei ringraziamenti, soprattutto quando sono sentiti e quando le persone da ringraziare sono tante.

I primi ringraziamenti vanno senza dubbio alla prof.ssa Paola Crepaldi, mio tutor che con pazienza e fiducia mi ha sostenuto fin dai tempi della tesi di laurea. Questa esperienza di crescita, sia professionale che personale, durata tre anni è stata possibile grazie alla sua guida.

Un grande grazie va ai miei compagni di ufficio Stefano e Beatrice, prima colleghi e poi amici. Con loro ho condiviso momenti di tristezza, gioia, sconforto e allegria. Sono stati una pietra miliare nella mia vita, e grazie alla loro presenza mi è stato possibile affrontare questo viaggio con il sorriso sulle labbra. Grazie a Letizia, che mi ha insegnato molto.

Ringrazio i professori Giulio Pagnacco e Gustavo Gandini per avermi sempre sostenuto nel percorso lavorativo. Grazie anche al prof. Michele Polli, che mi ha avvicinato al mondo cinofilo. Grazie anche a Marco, Licia, Elia, Elisa e Mario che, guidati dal professor Paolo Ajmone-Marsan, dai quali ho imparato molto e con cui spero di continuare a collaborare in futuro. Ringrazio anche i colleghi d'oltreoceano, ed in particolare prof. Max F. Rothschild e dott.ssa Elaine A. Ostrander per avermi accolto nei loro gruppi. Un ringraziamento speciale a Ezequiel Luis Nicolazzi, maestro nella via della bioinformatica e che mi ha insegnato una passione anomala come quella dell'analisi dei dati. Grazie a Francesca Bertolini, Muhammed Walugembe e Ahmed Elbeltagy, che mi hanno accolto in Iowa, aiutato e fatto sempre sentire il benvenuto. Grazie anche a Dayna, Andrew, Heidi, Jocelyn, Felina, Jaemin, Falina, Brian, Alex and Stephanie per l'accoglienza mostratami durante il periodo nel laboratorio della dott.ssa Elaine A. Ostrander. Un grande grazie va a Sonia Barnes, il cui aiuto come insegnante di inglese prima e reviewer della tesi dopo, è stato di cruciale importanza nell'avvio e nella chiusura di questa esperienza.

Un grazie ai miei colleghi del XXX ciclo del dottorato in scienze veterinarie e dell'allevamento, ed in particolare a Giada, Beppe e Marco. Le pause caffè con voi sono state una costante fonte di allegria durante questi tre anni.

Un grazie enorme a Marika, mia amica, compagna e sostegno in questo e in molti altri percorsi, e il cui supporto è difficile da esprimere a parole. Semplicemente grazie di esserci, sempre.

Infine, un grazie enorme a tutta la mia famiglia: mia madre e mio padre, che mi hanno sempre consigliato e sostenuto, il mio fratellino Luca, la cui dedizione è sempre stata per me fonte di ispirazione. I miei cugini Viviana, Irene, Francesco e Filippo, più fratelli che cugini, la cui allegria è stata spesso la mia fonte di forza. E poi i miei nonni Emilio e Giovanna, mio esempio di vita, i miei zii/genitori Elisa e Claudio con la loro costante disponibilità a sostenermi, i miei prozii Luisa e Lucio con il loro affetto… I nonni che non ci sono più, Nina e Bruno… E poi i parenti "acquisiti" che mi hanno sostenuto per tutta la vita Barbara, Dedo, Francesca e Giancarlo… I suoceri Silvana e Mario, i cognati Debora e Ale, Il piccolo Niki… Gli amici di una vita Elisa, Alberto e Marco, e poi Fabio, la cui mancanza rimane una ferita aperta…

A tutti voi, un ultimo e grande Grazie.

Index

FIGURES

threshold; modified from Talenti et al. (accepted on Journal of Heredity).

TABLES

Section 1 - Abstract

1.1 English version

The availability of genomic and high performance computing technologies gave access to new approaches and methods to exploit the high number of genetic information provided, even in livestock species. These new tools can be used to unveil the relationships between different populations or individuals, to understand the genetic background of breeds, to define the genetic architecture of phenotypes or pathologies and to detect selection signatures that shaped the cattle, sheep and goat reared today.

In this context, these technologies may provide breeders and breeders' associations with new tools that may be used in livestock management. In fact, the estimation of genomic parentage among individuals can help in defining the least related animals in case of identical values of additive parentage and detect errors in pedigree declarations as well. Moreover, since DNA is immutable within animal life but very polymorphic between individuals, these tools may be of help in the definition of the breed or the individual of origin, improving food safety.

In this thesis, I faced the complexity of genomic data in ruminant species, focusing in particular on goats. I reported three studies with distinct aims: i) study the signatures of selection in 369 animals of 16 Italian populations, and then focusing on the Valdostana goat breed which presents a breed-specific broad signature of selection in chromosome 7; ii) define a new method to develop panels of SNPs for parentage assessment that could be applied on species with unreliable genomic assembly; and iii) study the selection signatures in 929 animals of 41 Pakistani and Italian goat populations associated with a specific pigmentation pattern, the roan, detecting a candidate gene important in understanding the coat color genetics in goats and which could be also used as a marker for traceability of breeds that carry this peculiar phenotypic pattern.

1.2 Italian version

La disponibilità di nuove tecnologie nell'ambito della genomica e dell'informatica ha consentito di applicare nuovi metodi capaci di esprimere il pieno potenziale insito nell'elevato numero di informazioni genetiche prodotto da queste nuove tecnologie anche negli animali da reddito. Questi strumenti possono infatti essere applicati per definire le relazioni tra le popolazioni e tra gli individui, per evidenziare la composizione genetica delle razze, per comprendere l'architettura genetica di fenotipi e patologie, nonché per identificare le firme di selezione lasciate sul genoma durante la formazione delle moderne razze bovine, ovine e caprine.

In questo contesto, la genomica potrebbe fornire agli allevatori e alle associazioni di razza strumenti innovativi per la gestione delle specie da reddito. Ad esempio, la stima delle parentele genomiche potrebbe aiutare a definire, a parità di parentela additiva, gli animali meno imparentati tra loro negli schemi di accoppiamento nonché consentire l'individuazione degli errore nella registrazione dei pedigree. Inoltre, l'invariabilità del DNA nell'arco di vita di un individuo, unita alla sua variabilità tra individui, lo rende anche uno strumento ideale per l'assegnazione di razza, dei singoli animali e dei prodotti, migliorando così la sicurezza alimentare per i consumatori.

In questa tesi, abbiamo affrontato la complessità dei dati genomici nei ruminanti, ed in particolare della specie caprina. In questa tesi, ho riportato tre studi con tre obiettivi distinti: i) studiare le firme di selezione in 369 animali di 16 popolazioni italiane, focalizzandoci poi sulla razza caprina Valdostana, la quale presenta una ampia firma di selezione razza specifica sul cromosoma 7; ii) definire un nuovo metodo per sviluppare pannelli di SNP per l'assegnazione della parentela applicabile a specie con una bassa attendibilità della mappa genomica; e iii) studiare le firme di selezione nel genoma di 929 animali di 41 diverse popolazioni caprine italiane e pakistane associate ad un particolare *pattern* di pigmentazione, il roano, identificando un gene candidato importante per comprendere la genetica del colore del

mantello in capra e, se confermato, diventare un potenziale *marker* per la tracciabilità delle razze che presentano questo particolare fenotipo.

Section 2 - Introduction

2.1 The Genomic Revolution

The increasing affordability of high-throughput technologies such as whole genome sequencing, exome sequencing and SNP array genotyping allowed the diffusion of these tools also in species other than human (Kadarmideen, 2014). In the past decades, the genome of more than 800 animal species, including domestic and wild mammals, has been sequenced (NCBI https://www.ncbi.nlm.nih.gov/genome/browse/). The whole genome of dog (Lindblad-Toh et al., 2005), cat (Pontius et al.), cattle (Eck et al., 2009), chicken (Wallis et al., 2004), goat (Bickhart et al.), sheep (Jiang et al., 2014), pig (Swine Genome Seqencing Consortium, 2012), camel (The Bactrian Camels Genome Sequencing and Analysis Consortium, 2012) and horse (Wade et al., 2009) have been completed and publicly released. This led to strong competition on the development of the technologies needed for this purpose, and allowed the implementation of tools that became cheaper, faster and more reliable year by year. While the first two tools are species independent but still expansive (Sims et al., 2014), the third is much cheaper and saw a great fragmentation of the market, with several companies producing different solutions for different species.

Even if next generation sequencing provided genotyping for a great number of individuals in the different species, the availability of alternative genomic techniques, such as SNP arrays, varies greatly among them, with some having more options compared to others (Table 2.1). An example of this fragmentation can be seen in the three main ruminant species: cattle, sheep and goat. In cattle, there is a wide range of choices, from low (3,000 to 20,000 markers) to high (more than 600,000 markers) density SNP arrays, produced from different vendors. Customers can choose from SNP arrays with 3k, 7k, 33k, 54k and 700k SNP arrays, that provide different resolutions and prices to this kind of analysis (Nicolazzi et al., 2015). When working with small ruminants, such as sheep and goats, the situation

greatly changes. In sheep, researchers can choose from a couple of panels currently available: the mid-density 50k (Kijas et al., 2009) and the high-density 600k (Anderson, 2014; Kijas et al., 2014) SNP arrays. In goats, end users have only one choice: the mid-density SNP array developed by the International Goat Genome Consortium (Tosser-Klopp et al., 2014; IGGC). As a consequence of the differences in marker density, the feasibility and impact of specific researches can be hampered (e.g. copy number variant detection; Mason-Suares et al., 2013). Therefore, a well-defined experiment, with a proper choice of the tool, and a representative sample for the objective is pivotal when working with genomic tools to achieve specific goals.

Chip Density	Cow		Pig		Horse		Sheep		Goat		Chicken	
	Name	SNP n	Name	SNP n	Name	SNP n	Name	SNP n	Name	SNP n	Name	SNP n
Low-Density	3k	2,900										
	LDv1	6,909										
	LDv2	6,912										
	GGPLD v1	8,610	GGPLD v1	10,241								
	GGPLD v2	19,721										
	GGPLD v3	26,151										
Mid-density	SNP50 v1	54,001	SNP60 v1	62,163	SNP50 v1	54,602	SNP50 v1	54,241	SNP50 v1	53,347		
	SNP50 v2	54,609	SNP60 v2	61,565								
	GGPHD	76,879	SNP80	68,528	SNP70 v1	65,157						
High-Density	AxiomBos1	648,875					HD	606,006			Axom Chicken	580,961
	HD	777,962										

Table 2.1 - SNP genotyping array available for the different livestock species, from the three major produced (Illumina in green, Affymetrix in blue and Axiom in yellow)

Even though technological improvement in the genomic field has led to a significant reduction in the costs of analysis, important issues still exist.

A major problem related to genomics is tied to the data analysis and management. This problem arises from the high amount of data produced from genomic technologies that may reach several terabytes of size when working with high-coverage sequencing. In addition, advanced informatic skills are mandatory to the processing of these data, impossible to handle with standard methods and software. In addition to skilled bioinformaticians, computational power is another pre-requisite to handle the big dataset produced by these technologies. In fact, genotyping and sequencing can produce huge amount of data which make them difficult to handle. As a consequence of this need, the community released software aimed at specific goals first, such as programs for sequence alignment (Flicek and Birney, 2010) or for variant calling (Sandmann et al., 2017), and then moved to user-friendly integrated platforms such as Galaxy (Afgan et al., 2016), that allow also non-bioinformaticians to perform such analyses.

Even after obtaining the necessary skills and infrastructure required, the overall budget required for large-scale projects could still be unaffordable for single or small laboratories. In fact, the price per sample may vary from a few to several thousands of dollars, depending on the technology needed. And for some projects, several hundreds, or even thousands, of individuals are needed to achieve the goal. A number of research groups started to cooperate to overcome this limitation, joining their efforts and creating national and international projects or consortia. These groups' aims are to produce, collect and analyze the vast amount of genomic data in the different species.

This way, consortia for the production of de novo sequencing, resequencing and genotyping of species such as cattle, sheep and goat, collected thousands of samples for each of these species, making it

available to all contributors. Using this data, researchers performed a broad variety of analyses, that allow for a better understanding of the biodiversity, relationships and genetic background of breeds. But it also allowed researchers to unveil the role of genes in phenotypes of interest, in the adaptation to the environment and to quantitative trait. Its usage in management of populations allowed the production of highly reliable genomic indexes, relationship matrices and assessment of pedigree registrations. All of this has been possible after the release of the new genomic tools made available in the past decades.

Italy adopted genomic technologies later compared to other countries by promoting different national projects. The aim of my thesis is to analyze the genotypic data produced by these projects for the three main ruminant species using the different SNP arrays available. I will focus in particular on the goat species, trying to keep the applicative aspect of the analysis performed.

2.2 Genomic on livestock species

2.2.1 Cattle

Cattle (*Bos Taurus*) is mankind's largest domesticated animal. Its domestication probably started after that of the domestic sheep and goat, smaller in size and easier to manage (Conolly et al., 2012). Its extensive impact on human history is due to it being one of the earliest forms of capital, a source of meat, milk, hides and plowing fields for thousands of years (Felius et al., 2014).

Among ruminants, cattle genome was sequenced first in 2009 using an Hereford bull (Elsik et al., 2009; Liu et al., 2009). From that date onward, several projects provided new releases of the genome sequence, improving the map and releasing two alternative references: the BTau 5.0.1 (Cattle

Genome Sequencing International Consortium, 2015) and the UMD 3.1.1 (Zimin et al., 2009). After the production of the draft sequence, several projects aimed to identify the highest number of variants with the cattle genome. The first project with this aim was the Bovine HapMap project that re-sequenced several individuals of different breeds and defined the relationships between them using more than 40,000 SNPs (Bovine HapMap Consortium et al., 2009). These projects highlighted a common problem of these tools, which is the sampling bias of the SNPs. In fact, it is common that markers selected within a small number of populations are likely show anomalous patterns in genetically distant breeds. To overcome this limitation, the 1,000-bull genome project was proposed with the aim of re-sequencing a high number of bulls of different breeds or populations with the aim of improving the cattle variants database (1000 Bulls Genome Consortium, 2013). This project initially re-sequenced 236 bulls and 2 cows, and identified 28 million SNPs and InDels in the cattle genome (Daetwyler et al., 2014). In the latest update, this project has undergone several additions and can account for more than 1,100 bulls of 26 different breeds (Sanchez et al., 2017). The re-sequencing projects allowed the production of the SNP chip panels, available at low (3,000-7,000 SNP), mid (50,000-54,000 SNPs) and high (~600,000-800,000 SNPs) density (Nicolazzi et al., 2015) for this species. After the technology became available, several national and international projects produced huge amounts of data for cattle species. One example of the international effort is the Bovine Genome Project (Bovine Genome Project 2009). In Italy, the SELMOL and INNOVAGEN projects (Capomaccio et al., 2015; Marras et al., 2015) were funded by the Ministry of Agriculture Food and Forestry Policies (MiPAAF), and a third project, the BOVITA, was recently presented at a congress (Mastrangelo et al., 2017). These two in particular allowed the production of a huge dataset of several hundreds or thousands of individuals with mid- and/or high-density SNP array of several breeds and populations.

The availability of datasets with a high number of markers, individuals and populations allowed research teams to perform studies that helped in unveiling the genome of this species. A recent study based on whole genome sequencing highlighted the relationships of modern cattle with their wild ancestor, the Auroch, identifying it as a consistent outgroup in the phylogeny tree (Figure 2.1; Park et al., 2015) and confirming the previous studies based on mitochondrial (Edwards et al., 2007; Achilli et al., 2008) and Y-DNA (Bollongino et al., 2008). These studies also detected the populations mostly related with it and several genes involved in neurobiology, growth, metabolism and immunobiology as the most relevant in the domestication process (Park et al., 2015).

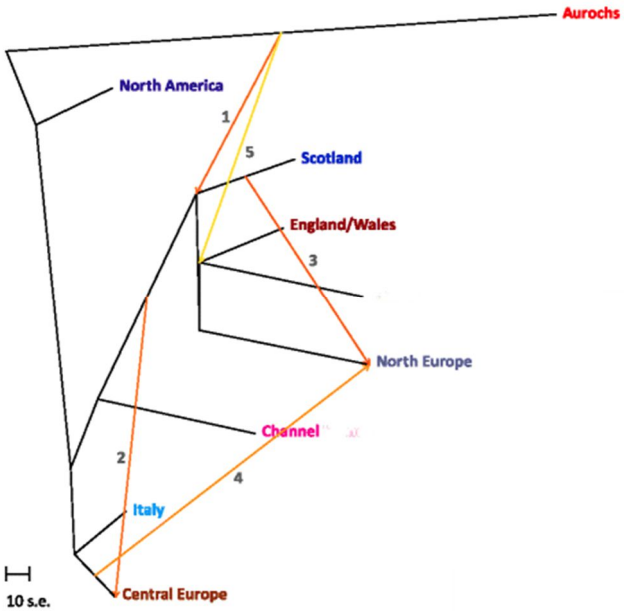


Figure 2.1 - Phylogeny tree of worldwide cattle from different European and their relationships with the Aurochs defined by Whole genome sequencing (modified from Park et al., 2015)

Other studies focused on the definition of the relationships between populations and breeds using diverse datasets. An analysis that aimed to define the genetic structure of several cattle populations was performed in 2009 by the Bovine HapMap consortium using 40 thousand SNPs for 500 animals from 19 diverse worldwide populations (Bovine HapMap Consortium et al., 2009). Another recent study, performed by Decker et al. in 2014, defined the ancestry, divergence and admixture between 1,543 animals belonging to 134 cattle populations from all over the world using 43 thousands markers (Decker et al., 2014). At a local level, some studies have been performed to outline the genetic diversity in specific groups of populations, such as for French (Gautier et al., 2010), New World (Mctavish et al.) and recently also for Italian (Mastrangelo et al., 2017) cattle breeds.

The genomic tools were also used to define the inbreeding levels within different cosmopolitan and local breeds, such as the cosmopolitan Holstein-Friesian (Kim et al., 2013; Zhang et al., 2015) and Jersey (Zhang et al., 2015) and the local Chillingham cattle (Williams et al., 2016).

Genomic tools allowed a finer mapping of QTLs and detection of genes associated with productive and reproductive traits (Sharma et al., 2015; Streit et al., 2013; Olsen et al., 2011; Bolormaa et al., 2011). Using SNP chip data imputed to sequence (i.e. the missing information was statistically defined based on other sequences available), a group of researchers identified 22 different genes that explain the majority of the variation in the milk composition in three different cattle breeds (Sanchez et al., 2017). This list includes both well-known genes (i.e. *DGAT1* and caseins genes) and new candidate as well, and it identified unique, or in other cases, multiple candidate variants. In other cases, the genomic association studies allowed the detection of genes associated with specific pathologies. Using genomics, genes associated predisposition to pathology such as

Paratuberculosis (Purdie et al., 2011), bovine Tuberculosis (*SLC6A6*; Finlay et al., 2012) and abomasum displacement (*SLITRK*; Biffani et al., 2014).

A similar approach to association studies that has been used on this data is the identification of the signatures of selection in this species' genome. These signs of natural and artificial selection are present in the genome, and could give hints on the genetic basis of adaptation to environment and morphology of the breeds (Stella et al., 2010; Taye et al., 2017b). Using these approaches, geneticists identified genes related to coat color pattern (Edea et al., 2017), adaptation to high temperatures (Taye et al., 2017a), polledness and coat color (Stella et al., 2010), resistance to parasites such as trypanosome (Noyes et al.). Randhawa et al. (2016) recently proposed a comprehensive meta-assembly (i.e. a collection) of the selection signatures defined in the cattle species.

Finally, the most important application of genomics in cattle breeding was by the adoption of genomic selection (Meuwissen et al., 2001; VanRaden et al., 2009; Hayes et al., 2009b). Genomic selection is performed through the usage of SNP arrays by estimating the effect for each marker in the panel. Once the effect has been estimated, using different statistical models including Bayesian approaches, it can be used to estimate the breeding values for newly genotyped individuals (Dekkers, 2012). Moreover, the improvement in the estimation of breeding values was achieved by the implementation of genomic relationships matrices (i.e. the relationships occurring between every pair of individuals based on their genotypes) to replace additive relationships matrix (i.e. the relationships between individuals based on their pedigree). These methods have the advantage of providing an index for newly, young genotyped bulls which has a higher reliability in comparison with standard estimation (Hayes et al., 2009b).

2.2.2 Sheep

Sheep have been domesticated for around 9,000 years BC in the middle-east (Demirci et al., 2013). Findings suggest the possibility of multiple domestication events, with South Turkey, the Near-East Zagros and the Indus Basin as the most probable (Pedrosa et al., 2005). Initially reared for meat, this species was later specialized for other productive purposes, such as wool and milk, around 4000 to 5000 years ago (Liu et al., 2016). Thanks to their small size compared to cattle, to their docility, to the extensive breeding and to the broad variety of production, which includes wool, sheep quickly spread from the domestication centers to the Mediterranean basin (Zeder, 2008), India (Singh et al., 2013) and Africa (Muigai and Hanotte, 2013).

This quick spread all over the world led to the creation of several local varieties, or landraces, adapted to the different environmental conditions and with different phenotypes reflected human preferences. These are still the vast majority of all populations distributed worldwide, and this is especially true for marginal areas of the world. However, in the Victorian era, in Europe these locally adapted populations became the raw material for the creation of the first purebred breeds. One of the main criteria behind the standardization of the breeds was the specialization for a productive purpose. For sheep in particular, the most important production is wool. This production was important even in the past, with the standardization of the Merinos breed even before the Victorian era. The great economical interest led to the creation in some countries of ad-hoc breeding schemes with the additions of genomic tools similarly to what is done for cattle, even at low density (Raoul et al., 2017).

The first reference sequence for this species' genome was released in 2011 (Kijas et al., 2012), shortly anticipated by the release of the mid (54K SNPs), and later the high (600K SNPs) density SNP chips by the sheep

HapMap project (International Sheep Genomics Consortium, 2009), shortly followed by projects such as the International Sheep Genomics Consortium (ISCG, 2009). At the European level, one of the most recent international projects is the iSAGE project, which aims to assess and develop solutions for future challenges for the ovine and caprine sector, such as climate changes, food security, efficiency and deprivation in marginal areas (Arsenos, 2016). At a national level, there is the Italian Project for Sheep Biodiversity which genotyped 496 animals from 19 different breeds, and highlighted the concordance of geographical and genetic distances, defining both known and unknown introgression between breeds (Ciani et al., 2014).

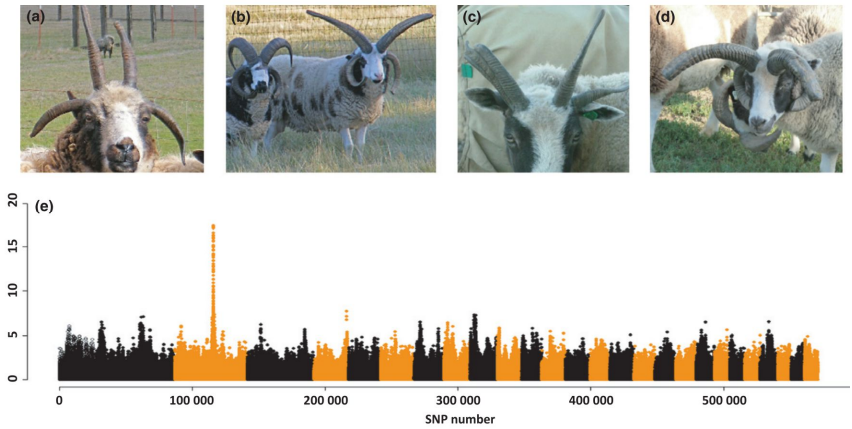


Figure 2.2 - Example of genome-wide analysis to identify the genes involved in four-horned phenotype in sheep (modified from Kijas et al., 2016)

This project is particularly active, and at a congress recently presented preliminary results on the selection signatures discovery in the different Italian sheep populations (Sorbolini et al., 2017). Several studies using microsatellites, mitochondrial DNA (Demirci et al., 2013) and then SNP array genotyping, produced several biodiversity and phylogenetic

studies both at international (Kijas et al., 2012) and national (Ciani et al., 2014) levels.

Similarly to cattle, the high standardization achieved by human-driven selection allowed the researcher to scan the genome of this species using different approaches. The signatures of selection studies, for example, unveiled the genetic basis of several morphological traits, such as horn phenotypes: a region on chromosome 2 containing four *HOXD* genes has been associated with four-horned sheep (Kijas et al., 2016) and the gene *RXFP2* has been associated with polledness (Wiedemar and Drögemüller, 2015). Another recent study detected several genes associated with coat color and production traits (Fariello et al., 2014).

Even though this species has a high economic interest due to wool production, the single individual value remains low if compared with cattle. In this context, the development of affordable genomic tools remains a major goal for this species, and in particular the development of tools for parentage assessment. A study of Heaton et al. (2014) identified a small panel for parentage assessment and breed traceability in worldwide sheep breeds (Heaton et al., 2014). Thanks to the increasing affordability of these instruments, the authors estimate a decrease between 20 and 60 % for a single parentage assessment with respect to the current microsatellite test (Heaton et al., 2014).

2.2.3 Goat

The goat species was domesticated in the same period as that of its close relative, the sheep, from its wild counterpart, the Bezoar, around 11,000 years ago (Zeder et al., 2006). Archaeological studies, further supported by mitochondrial DNA evidence, suppose the presence of two domestication sites in the Zagros and in Turkish mountains (Zeder and

Hesse, 2000; Zeder, 2008). However, mitochondrial DNA haplotypes shows that animals from Turkey are the ancestors of most of the modern goats (Colli et al., 2015). After their domestication, the species started to spread from their original site around 7,000 years ago (Vigne, 2011) to the rest of the world. In the following centuries, this species became a particularly important food resource in the marginal rural areas of the world thanks to its adaptability to different environmental and nutrition conditions (Nicoloso et al., 2015).

Goat genome was sequenced later compared to other major livestock species, in 2013 (Dong et al., 2013) and shortly followed by the release of a mid-density SNP chip, with more than 53K SNPs (Tosser-Klopp et al., 2014), and recently from the release of a more accurate genome sequence (Bickhart et al., 2017). The Italian Goat Consortium (Nicoloso et al., 2015) genotyped 350 animals accounting to 14 different breeds to perform population genetic studies. In 2015, this data conveyed in the ADAPTmap dataset, an international unfunded initiative which is collecting goat genotyping data. This initiative will put its focus on the adaptation of this species to different environments all over the world (ADAPTmap, 2014). This initiative will be soon followed by VarGoat, a project supported by France Genomique that will re-sequence a high number of goats, belonging to different breeds from all over the world, and aims to be the first step towards the 1,000 goat genome project (VarGoats, 2017). This new initiative will start from the results of the previous ADAPTmap initiative to properly choose the animals to sequence. Other international projects involve studies on this three species are the International Goat Genome Consortium (IGGC) and, at a European level, the NEXTGEN and 3SR projects were funded with the aim of studying the genetic variability of ruminant species (NEXTGEN, 2009; 3SR, 2014).

Despite the availability of genomic technologies for this species, genome variability studies on goat species started later with respect to

other species. In the past few years several studies have been published on biodiversity, population structure, parentage assessment, genomic improvement and selection signatures selection that expand what was already achieved using fewer markers.

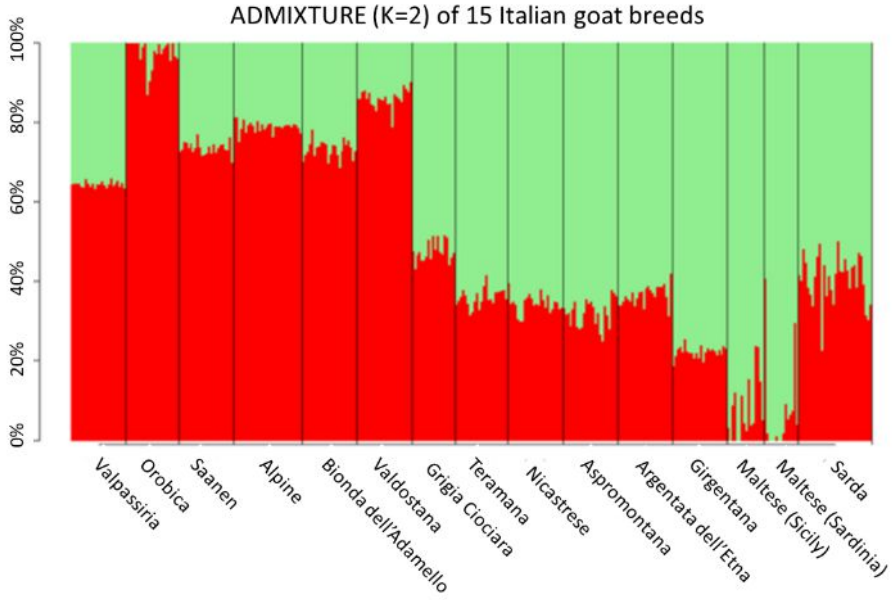


Figure 2.3 - Genetic background of Italian goat breeds, which highlights the difference between northern (red) and southern (green) breeds (modified from Nicoloso et al., 2015)

For example, the Italian goat consortium characterized the genetic diversity of population, which is due to a combined effect of drift, presence/absence of gene flow and consequence of traditional management practices (Nicoloso et al., 2015). The same approach has been applied to Spanish (Manunza et al., 2016) and Swiss (Burren et al., 2016) goat breeds. Other studies focused on defining the population structure of specific breeds, such as the cosmopolitan Angora (Lashmar et al., 2016, 2015). A recent study tried to unveil the genes involved in milk, meat and fiber production in several highly specialized breeds, such as Alpine and

Toggenburg for milk, Boer for meat and Cashmere for fiber (Brito et al., 2017). Finally, a French study tried to set up a the genomic selection in Alpine breeds (Mucha et al., 2015). Another study from Kim et al. (2016) studied the adaptation to hot, arid environment in goat and sheep adapted to Egyptian environment (Kim et al., 2015).

In this thesis, I will introduce several studies, performed within national and international projects and using different datasets, that aim to better understand the architecture of goat breeds genome. First, I contributed to Nicoloso et al. (2015) work aiming to unveil the Italian goat biodiversity. We then updated the same dataset, showing its extended version, that now count more than 1,000 animals of more than 30 populations. I developed a method to define small SNP panels for parentage assessment and tried this method on three different goat breeds. I also applied the same method to the Italian populations provided by the Italian Goat Consortium that, at that time, counted 369 animals of 16 populations and finally to the ADAPTmap dataset, counting more than 4,000 animals of more than 100 worldwide populations. In a subsequent work, we studied the selection signatures in the Italian goat populations and highlighted a strong signature of selection in chromosome 7 of Valdostana, an Italian goat breed. Similarly, I studied the signatures of selection in 929 animals of 41 breeds and identified a signal associated in two Pakistani goat breeds to the roan coat color pattern that can be, if confirmed, a candidate marker for breed and product traceability.

2.3 Application of genomic

As seen before, the availability of genomic technologies allowed researchers to deeply understand the process that brought the domesticated animals to the breeds in the different species. However, these findings vary greatly from species to species and could be affected by

several factors, such as the availability of the wild ancestor, the presence of intermediate evolutionary steps, the presence of ecotypes and of standardized breeds.

An example of this is the dog, which is probably the best model to study the process from wild animals to domestic breeds. In fact, its ancestral wild population, the gray wolf, still exists even if a recent study suggested that the actual ancestor is an extinct wolf subspecies (Thalmann et al., 2013). In addition, the dog can count for an intermediate evolutionary step between domestic and wild, the Village dogs, diffused all over the world and that can help in defining the origin of domestic dogs (Shannon et al., 2015). These, together with the local ecotypes, the breeds that are undergoing official recognition and the officially recognized ones allow researchers to perform deep phylogenetic studies to understand the dog evolutionary history (Ostrander et al., 2017). This long history produced changes in the dog genome, that led to changes in the architecture of the genome of this species. These can be summarized in three distinct points: i) fixation of novel traits and phenotypes in every breed, that lead to an increase in variability; ii) fixation of QTLs with large effects increasing the selective pressure on them; iii) reduction in the natural and sexual selection on the remaining genome, allowing stronger variation in phenotypes that otherwise would be limited by them (Boyko, 2011). These variations in the genome of this species make it particularly interesting also to define the genetic basis of pathologies, simple mendelian and even quantitative traits (Parker et al., 2010; Boyko et al., 2010; Boyko, 2011). Moreover, the divergence led by the selective pressure relaxation allow the detection of novel, breed specific markers that could be used to perform individual assignment to its breed, a process that could be associated with the traceability in economically relevant species. All these advances could be partially applied also to ruminant species, with some differences among the three considered.

If we consider cattle, we do not have the wild ancestor of this species anymore, things that make a direct comparison difficult, although some studies using archaeological remains are emerging (Park et al., 2015). However, the strong standardization undergone by several cattle breeds in the past two centuries allow the performing of both studies to detect the genetic basis of different traits and to detect markers to perform breed traceability.

For sheep and goats, we still have the supposed wild ancestors, but these species undergone a lower extend of standardization with several populations that are still ecotypes, locally adapted to the environmental conditions.

Starting from these considerations, genomic approaches and techniques need to be thoroughly evaluated to properly infer phylogenies and define the relationships between individuals with a very high reliability. In this chapter we will discuss several possible solutions that could be used, underlining the criteria behind their choice.

2.3.1 Population structure

Genomic revolution allowed geneticists to increase the resolution of population genetic studies. Using genomic it is possible to deepen genetic structure and/or evolutionary history of a species, highlighting how populations diverged and mixed in their history. In addition to that, it is possible to detect the so-known selection signatures, trace left on the genome from natural or artificial selection forces. The latter in particular, produced changes in domesticated animals' morphology, physiology and behavior depending on the needs of humans. As a consequence of the process, different breeds for the different domesticated species have been selected.

Even if in some part of the world breeders kept the pedigree of their animals, such as for Arab horses, this was more part of oral traditions, and not a supervised practice. The first officially recognized breeds were standardized and registered in the 18th century, when Robert Bakewell began his selection processes on his own animals (Stanley, 1995). Thanks to his approach, that relies on choosing the contributors to the next generation based on their phenotype, he gave birth to the Longhorn cattle breed, to the Leicester sheep breed and to the Shire horse. After him, several breeders started using the same approach in selecting their animals, giving birth to several new breeds (Hall and Clutton-Brock, 1989). Later in the 19th century several breeders joined and formed associations with the aim of defining the standard for the breeds, which is the set of characteristics that animals belonging to the breed must have (Lush, 1943). These associations started to manage the individuals of each breed by the adoption of studbook and herdbook. The first herdbook was the General Studbook, created in 1791 to record Thoroughbred horses (Van Vleck et al., 1987). These registers where all animals of a specific breeds are recorded, together with their parental information, allowing a better management of matings.

Nowadays, these registers could be closed or open to new registrations depending on the dimension and history of each single breed, that may have complex histories and include several admixing events (exchanges of genetic material) with living wild ancestors or different breeds with particular traits. These can be either crossing events, exchanges of reproducers, or even splitting of previously existing populations in different subgroups. These events are well known in some cases thanks to historical recordings, but often the breeding practices applied in some species make the reconstruction tough.

In this situation, the genetic markers have proven to be particularly effective (Goldstein and Pollock, 1997; Leaché et al., 2015). Several

measures have been proposed to evaluate the genetic differences occurring between individuals, populations or species (Lawler, 2017). These measures are designed to increase linearly on time with a low variance level, improving the performances of phylogenetic reconstructions (Goldstein and Pollock, 1997). Genetic distances from genotyping data could be estimated in different ways, using both statistical visualization techniques and *ad-hoc* calculations.

The first category includes a series of heavy-calculation approaches, that could be performed thanks to the availability of high-computing machines. Among statistical approaches, the most diffused to analyze genomic data is the Principal Component Analysis (PCA). This is a multivariate technique that describes several variables in newly created orthogonal variables called principal components (Abdi and Williams, 2010). This approach allows the identification of population structure by relative genetic distances (Galinsky et al., 2016), identifying outlier animals and providing a simple representation of groups.

This kind of analysis is particularly useful in the preprocessing of data by detecting substructures or misclassified individuals that need to be evaluated. These approaches prove to be useful in defining the hidden structure in animal populations, allowing different degrees of precision, from species to breed and to farm. Considering the relatively low computational time, the low amount of resources needed and the availability of several types of software that allow to perform it easily, geneticists generally use this analysis routinely as can be seen in the works reported in this thesis.

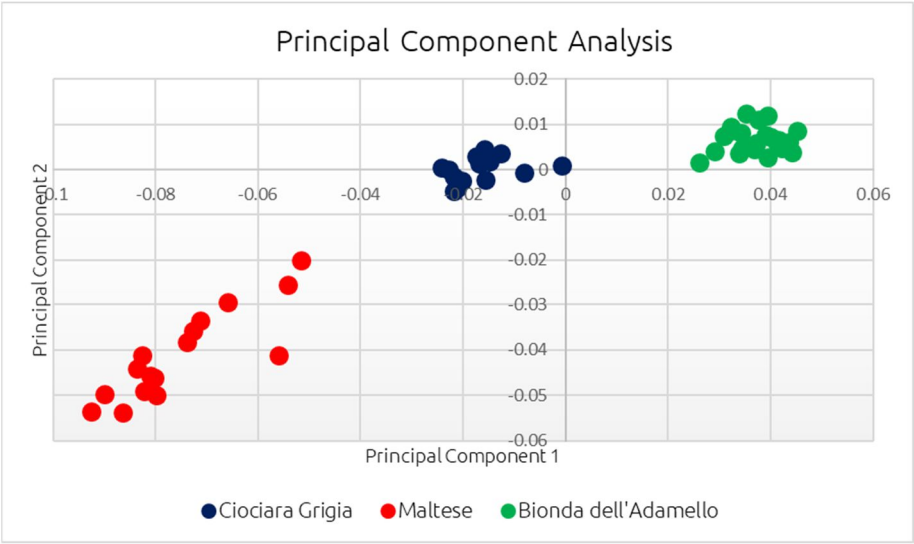


Figure 2.4 – Scatter plot of a PCA output of three different caprine breeds; each of the two axis is a different component, decreasingly representative of the variability of the sample; each point in the figure represent a different goat and is colored by its classification (i.e. breed, population, family, etc.). The Euclidean distance between each point is representative of the genetic distance separating them.

The availability of genetic markers made it possible to apply several distance measures, previously developed for other markers, that are based on different genetic models that may explain the events that shaped the genome of the different populations (Lawler, 2017). Different distance measures have been proposed in different times and aim to answer to different biological questions, using different kinds of markers and focusing on individuals or populations distances. The application of such measures on the very high number of markers provided by SNP arrays allows us estimate these distances with a high reliability. One of the most commonly used measures of genetic distance between populations is the Fst, proposed by Sewall Wright in the 50' (Bhatia et al., 2013), and that have now several variants such as the one proposed by Weir and Cockerham in 1984 (Reynolds et al., 1983). In its original design, the Wright Fst index was:

$$Fst = \frac{H_T - H_S}{H_T}$$

Where H_T is the expected heterozygosity in the global population and H_S is the weighted average of expected heterozygosity across populations. The easiness and speed of calculation, and the high number of variants and implementations, make it a popular measure to define the genetic differences between populations.

Other popular distances proposed at several times are Cavalli-Sforza's Chord distance (D_{CH}) (Cavalli-Sforza and Edwards, 1967), Nei's standard (D) (Nei, 1972) and D_A distances (Saitou and Nei, 1987) and Goldstein's distance for microsatellites (Goldstein et al., 1995).

Genetic distance	Model	Year	Type	Marker
Nei D	Mutation and Drift	1972	Population	All
Nei Da	Mutation and Drift	1978	Population	All
Cavalli-Sforza's Chord	Genetic drift, no mutation	1967	Population	All
Goldstein	Stepwise-mutation	1995	Population	MS
Reynolds Θ	Genetic drift, no mutation	1983	Population	All
Hamming	Mutation and drift	1950	Individual	SNP

Table 2.2 - Examples of genetic distances for microsatellites (MS) and SNP markers

These distance values consider different evolutionary models, and therefore will provide different measurements of the divergence between populations. For example, if the aim is to study a divergence happened in a distant past a model that takes into account both mutation and drift is preferred, whereas in case of a divergence happened in a relatively short time a model that accounts mainly for the drift should be the choice. In any case, they could be used as a simple heat map to find the greatest and smallest difference, or alternatively can be represented graphically building phylogenetic trees. These are representations of the divergences of species or populations from an initial common ancestor. A phylogenetic tree is a group of branches connected by nodes: a branch is a genetic lineage

conserved through time, nodes are splitting points of populations leading to a new lineage (Yang and Rannala, 2012). A group of branches with a common ancestor is called a clade (Yang and Rannala, 2012). The length of the branches and their position in the tree reflect the genetic distances between the groups in the sample analyzed (Lawler, 2017). Generally, researchers should try to build phylogenetic trees using different distance measures, and in case of consistency of the different representations, the relationships among populations should be considered more reliable.

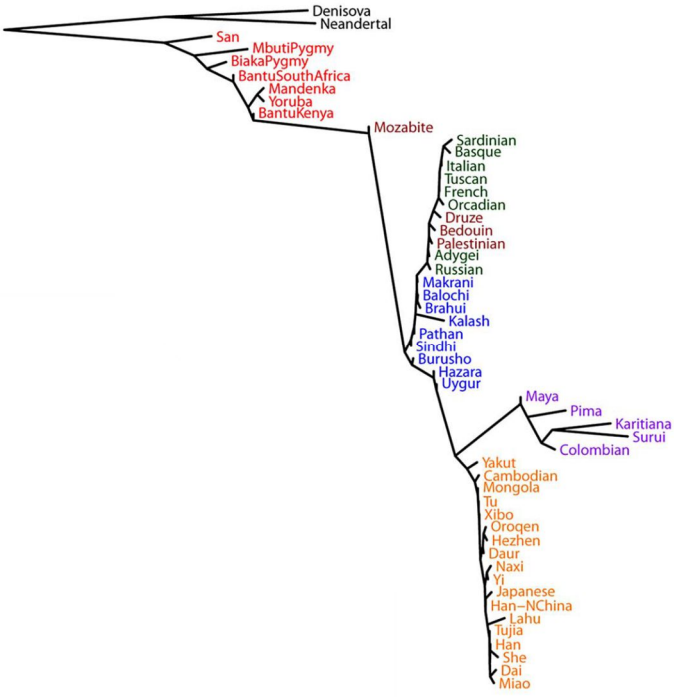


Figure 2.5 - Phylogenetic tree of human populations (picture modified from Pickrell et al., 2012)

Phylogenetic trees provide important insight in the relationships occurring among populations. However, these estimates are based on groups of individuals, that should be not homogeneous due to recent or

past admixing events. These events could be detected only with a thorough analysis of a high number of genetic markers and using *ad-hoc* models that create phylogenies considering admixing events.

An example of these tools is TreeMix: this program uses allelic frequencies to detect a user defined number of admixing events, which populations are involved, the direction and the strength of the admix (Pickrell et al., 2012). Despite that, this program still considers groups of individuals, and does not account for the similarities between individuals. This can be achieved using individual-based genetic distances, allowing a precise definition of clades of individuals that could consist of genetically similar animals. These clades allow the detection of the "core" population, composed of individuals that are more genetically similar to each other, and allow the detection of more distant animals. An relevant example of this is described in dog by Parker et al. (2017), who used identity by state between pairs of individuals (i.e. the proportion of alleles two individuals share) as a distance measure to define the relationships between more than 1,346 animals of 161 breed.

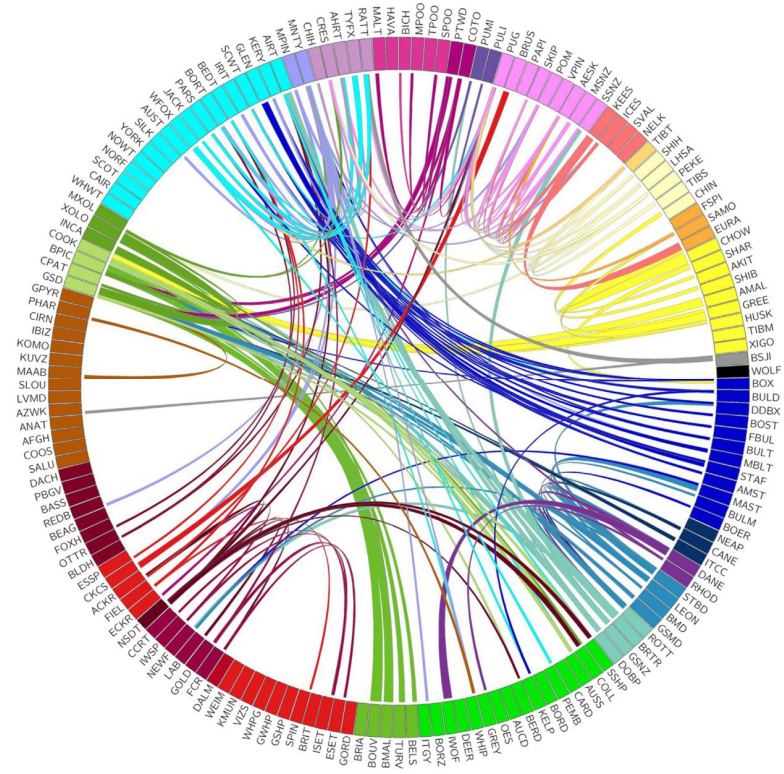


Figure 2.6 - Haplotype sharing between dog populations proposed by Parker et al. (2017)

In the same work, Parker et al. (2017) proposed an alternative solution to define the relationships and the admixing events between populations. This approach uses the Beagle software V4.1 (Browning and Browning, 2013, 2008, 2016) to phase the haplotypes (i.e. assign the paternal and maternal alleles on a probabilistic base) (Browning and Browning, 2007) and define the regions identical by descent (IBD) between pairs of individuals (Browning and Browning, 2013). Using these data, together with historical records, the researchers estimated the actual dates of admixing or splitting events between pairs of populations.

Another recent approach is the one used from the GLOBETROTTER software. The researchers developed the program to work with the companion software Chromopainter (Falush et al., 2003) and estimated the admixing events occurring between human populations in a time span of 4000 years (Hellenthal et al., 2014).

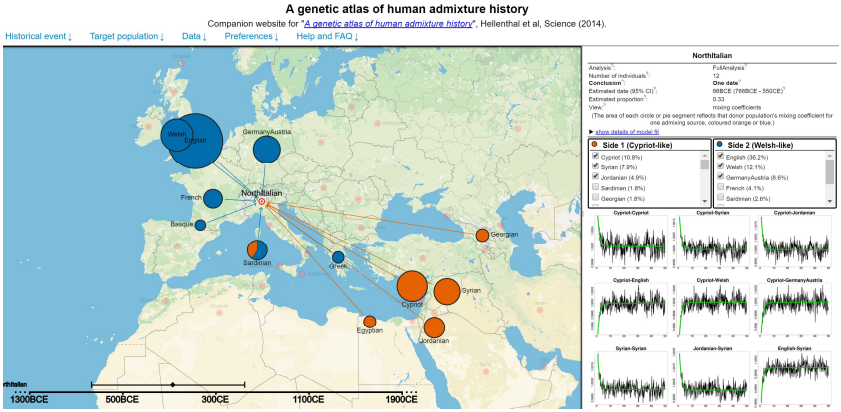


Figure 2.7 - An example of admixing events for an Italian population using GLOBETROTTER by Hellenthal et al. (2014, www.paintmychromosomes.com)

In the current PhD study, the understanding of the relationships between populations was one of the major goals of this thesis. In this context I learned and implemented the above mentioned techniques, and recently used them to define the relationships that occur among Italian breeds of two different species, dog and goats, presented the results at the 22nd congress of Italian association of animal production sciences and the paper showing these findings is in preparation.

Phylogenetic analysis gave new hints in modeling the evolutionary events that shaped modern breeds and landraces. These findings are helpful not only to unveil the relationships occurring between populations, but also to improve the detection of genomic regions of interest by selection signatures identification.

These traces on the genome may change depending on the intensity of selection, on the time span the events happened and on the type of selection needed. Considering the different events that may shape the genome, different specific approaches that may detect selection signatures have been developed.

Initially, the detection of regions associated with different traits was performed using microsatellite markers located in different genomic regions (Khatkar et al., 2004). However, the availability of SNP arrays, with thousands of markers, allowed mapping and associations of QTLs and genes to traits of interest thanks to the higher resolution (Tiwari et al., 2016).

In addition to classical QTL detection approaches, the improvement in genomic and informatics technologies allowed the application of several statistics to these data. An example of classical statistics applied to genomic data is the linear discriminant analysis proposed by Fisher in 1930 (Fisher, 1936). This method is a multivariate analysis that, using a high number of variables for groups of individuals, maximizes the difference between groups and detects the parameters that mostly distinguish them (Fisher, 1936). Even if this approach has several limitations (i.e. the needs to remove collinear variables and to keep the number of variables below the number of observation) some studies using this method have been published (Dimauro et al., 2013).

Statistical approaches applied to genetic data are actively progressing, with the development of newer techniques and methods. A field that is becoming increasingly important even in biological sciences is the machine learning, a series of algorithms that allows us to efficiently perform clustering, classification, feature selection and more. A very popular example of this family of algorithms is the machine learning, that has been used, in combination with other approaches, to perform the

selection of informative SNP (Bertolini et al., 2015; Chen and Ishwaran, 2013; Menze et al., 2009).

Similarly to what has been described for genetic distance, classical statistics and algorithm for SNP selection have been accompanied by a broad variety of genetic methods that came on the scene thanks to the high resolution of genomic technologies and to the increasing computational power. These approaches try to detect different selection events at different times, and are based on different genetic bases. In fact, the detection of selection signatures can be performed by approaches based on allelic frequencies, Linkage Disequilibrium (LD), homozygosity regions detection or haplotype structures, and can lead for example to the identification of key genes for domestication [31].

Approach	Type of selection signature	Method	Reference
Frequency-based methods	Detection of increasing in rare allelic variants' frequencies otherwise expected to be rare in the global population	Ewens-Watterson Test	Ewens 1972
		Tajima's D	Tajima 1989
		Fay & Wu's H	Fay and Wu 2000
		Runs of Homozygosity (ROH)	Peripolli et al. 2016
Linkage Disequilibrium-based methods	Detecting extended regions at a high prevalence in a population	Long-range Haplotype (LHR) test	Sabeti et al. 2002
		Long-range Haplotype similarity test	Hanchard et al. 2006
		Integrated Haplotype score (iHS)	Voight et al. 2006
		Cross-population extended haplotype homozygosity (XP-EHH)	Sabeti et al. 2007
		Linkage disequilibrium decay (LDD)	Wang et al. 2006
		Identity by Descent (IBD) analysis	Cai et al. 2011
Population differentiation-based methods	Detecting markers that underwent to differential selection in two different populations, which likely changed the frequencies of different alleles	Lewontin-Krakauer Test (LKT)	Bonhomme et al. 2010
		Locus specific Branch length (LSBL)	Shriver et al. 2004
		HapFLK	Fariello et al. 2014
		Fst	Karlsson et al. 2007
Composite methods	Detecting regions by combining scores from different methods or, alternatively, from multiple sites in a genomic region	Composite likelihood ratio (CLR)	Kim and Stephan 2002
		Cross-population Composite likelihood ratio (XP-CLR)	Chen et al. 2010
		DH Test	Zeng et al. 2006
		Composite of multiple signals (CMS)	Grossman et al. 2010

Table 2.3 - Examples of methods to detect signatures of selection

Statistics to detect selection signatures can be split into two main categories: to detect macroevolution phenomena the first, and to detect the microevolution phenomena the second.

Macroevolution is an event of separation between species, rather than within, whereas microevolution involves the split of populations within a species. Some methods that belong to this group are the ω (or Ka/Ks) index, the McDonald-Kreitman (MKT) test, Hudson-Kreitman-Aguadé (HKA) MKT and the accelerated regions identification (Vitti et al., 2013). This kind of event is not an aim of this thesis, and therefore will not be further discussed.

On the other hand, microevolutionary methods can be further classified in 4 distinct categories depending on the theory that lies behind their development (Table 2.3): frequency-based method, LD-based method, population differentiation-based methods or composite methods (Vitti et al., 2013). Frequency-based methods rely on the detection of increasing in rare allelic variants' frequencies otherwise expected to be rare in the global population. Probably the most famous representative of this category is Tajima's D, that is used to test the excess or absence of rare alleles in a population (Tajima, 1989). The Linkage Disequilibrium-based methods rely on detecting extended regions at a high prevalence in a population. In fact, the prevalence of an extended haplotype (i.e. the combination of alleles in a strand of DNA) in a population is likely due to the selective pressure on a genomic region that consequently underwent an increase in frequency (Szpiech and Hernandez, 2014; Gautier et al., 2017). Methods such as the integrated Haplotype Score (iHS; Voight et al., 2006) and the cross-populations extended haplotype homozygosity (XP-EHH; Sabeti et al., 2007) became very popular once phasing algorithms such as Beagle (Browning and Browning, 2016), capable of reliably defining the paternal and maternal allele, became available. The third category, Population differentiation-based methods, relies on detecting markers that

underwent to differential selection in two different populations, which likely increased the frequencies of two different alleles up to the fixation. Probably the most popular approach belonging to this category is the Fst, previously described as a general measure of the differences between two groups of subjects but that allows the detection of the genomic regions that underwent to differential selection when applied to single or to contiguous sets of markers (Karlsson et al., 2007). Finally, the last category, the Composite methods, could rely on combining scores from different methods or, alternatively, from multiple sites in a genomic region such as in cross-population composite likelihood ratio (XP-CLR; Chen et al., 2010) that look for selection signature that distinguish two different populations by modeling the allele frequency spectrum in two different populations.

The discrimination among methods to detect signatures of selection is much more complicated than this, and must consider both the type of selection and its strength. All the above mentioned approaches can then be further distinguished in methods to detect positive, negative and balancing selections that respectively fixate, eradicate and modulate allelic variants in a population. As can be easily guessed, positive and negative leave similar reduction in the frequencies of specific alleles, and it is hard to distinguish between them (Nielsen, 2005). Finally, methods to detect positive selection could be further distinguished in methods to detect strong events, that fixated long chromosomal regions in a relatively short time (also known as hard sweep), or a mild positive events that increased the frequencies of some allelic combination in a region without erasing the variability (soft sweep) (de Simoni Gouveia et al., 2014). To complicate the choice even more, several methods belonging to different categories often overlap one another, and could be used as a way to exclude false positive signals keeping only consensus (e.g. both XP-EHH and XP-CLR identify recent, strong selection). In my studies, I used a combined method of up to four different methods, belonging to three different categories. In particular, in my studies I combined the scores provided by the population

differentiation method Fst, the two haplotype-based methods Rsb (Tang et al., 2007) and XP-EHH (Sabeti et al., 2007) and the frequency-based method ROH both belonging to single population to detect signatures of selection in goat populations of interest.

Another very popular, yet criticized, approach used to identify genes and regions associated with traits of interest is the Genome-wide association study (GWAS). Using this technique, several genes associated with diseases, production and morphology have been identified in several species (Reverter and Fortes, 2013). Several types of software performing GWAS have been developed, including PLINK (Chang et al., 2015), GCTA64 (Yang et al., 2011), GEMMA (Zhou and Stephens, 2014, 2012; Zhou et al., 2013) and the GenABEL R package (Aulchenko, 2009; Package, 2013). Among the limitations, the need for a very high number of individuals with an extremely precise registration of the phenotypes, make it an approach not suited to all studies (Korte and Farlow, 2013). Even though GWAS was one of the most important techniques in the past years, it presents several criticisms that are important to consider prior to the experiment. Among them, probably the most important is the impossibility to identify variants with a small effect in complex traits and as well as the fact that the highest significant variant could be a spurious association (Visscher et al., 2012; Korte and Farlow, 2013).

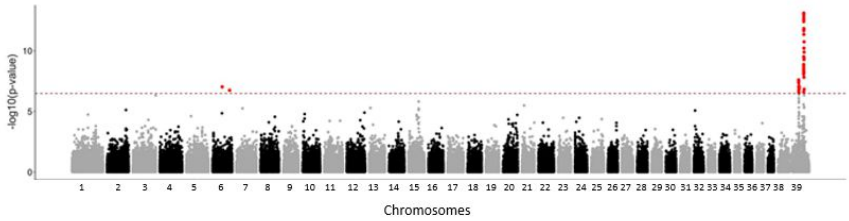


Figure 2.8 - An example of Manhattan plot resulting from a Genome-wide population study on dog size; every dot represents a SNP and the height is its significance represented as –log$_{10}$(Pvalue). The line is the threshold used to determine the strongest signals, and red dots are the SNPs significantly associated with the trait of interest (modifed from Plassais et al., 2017).

Given these drawbacks of the GWAS approaches, in my studies I decided to use a different method, the Bayesian association study implemented in GenSel (Fernando and Garrick, 2009). This approach does not considers single SNP, but entire chromosomal regions to perform the association, and provide a proportion of variance that a region explains. In this way it is possible to limit the number of false positives. In one of my studies, I used this approach in combination with the other four methods described to detect signatures of selection and chromosomal regions of interest in the different caprine populations.

All the methods previously described prove to be useful to defining the genes mostly involved in the development of populations, and even though these methods could lose effectiveness applied to define genes involved in complex traits (Kemper et al., 2014), they still help in understanding the genetic basis of productive, reproductive and morphological traits and diseases. These would likely help breeders' associations to perform a better genomic selection if already performed (e.g. in cattle), or to efficiently start the programs in these species and breeds where it is not applied yet. In fact, together with relationship management, the knowledge of positively and negatively associated

markers at every trait is important for effective genomic selection programs.

2.3.2 Parentage analysis

The reliable knowledge of relationships between individuals is an important tool in genetic resources management by allowing mating schemes planning (Jones et al., 2010). Having a clear knowledge of relationships occurring between individuals allows the estimation of the average amount of genetic material shared by a couple of individuals, also called the additive parentage. In this way, it is possible to properly choose the best unrelated individuals that allow an improvement in the performances of the next generation of animals. However, this is true only if pedigree registrations are correct, and otherwise propagating errors by negatively affecting the next generation. In particular in species such as goat, relationship recording can be hampered by commonly used practices such as seasonal pasturing, the use of mating groups and the collection of kids in collective nurseries.

To overcome these limitations, several tests have been developed to assess the trueness of the declared pedigree. The first parentage assessment methods were performed on biochemical markers that have a mendelian inheritance such as cattle blood group (Glowatzki-Mullis et al., 1995). Later, these approaches have been replaced by DNA-based markers. Microsatellites (MS) are currently the golden standard to assess parentage (Schnabel et al., 2000). They work using either the exclusion principle used also for the biochemical markers, which exclude a direct relationship if two individuals do not share any alleles, or exploit the allelic variability of the markers using classic and Bayesian statistics that returns a probability instead of a direct exclusion of parentage (Jones et al., 2010).

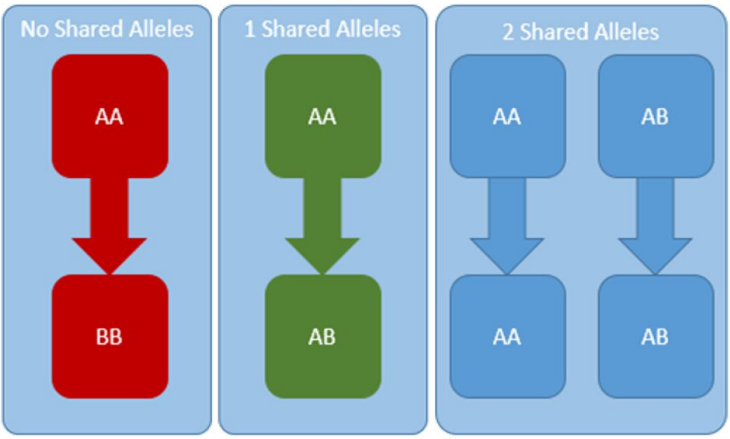


Figure 2.9 - Example of allelic comparison at one locus between individuals two individuals, showing the concordance, partial concordance or discordance of allelic state (also known as identity by state, IBS).

Despite their high variability, MS markers show several limitations: difficulties in standardization of allelic size among laboratories (Groeneveld et al., 2010), rely on human work (Vignal et al., 2002), low number of markers tested and, in species such as goat, they still have a high cost compared to the value of the animal. All these pitfalls led researchers to look for alternative ways to efficiently define the relationships between individuals.

The most promising alternative proposed to MS are SNP markers, that have proven to be successfully applied for parentage assignment. When compared to MS, despite their bi-allelic nature they can be analyzed in a very high number and can be cheaper, easier to analyze (Krawczak, 1999), to automatize (Kruglyak, 1997), and could be used to perform both direct exclusion and probability methods. Since SNP can be only biallelic, and thus less informative compared with MS, a higher number of markers is required to provide as much informativeness as a MS panel (Phillips et al., 2012). However, previous studies highlighted how 2-3 SNP markers provide

the same amount of informativeness of one MS marker (Fernández et al., 2013).

The adoption of SNP to perform relationship analysis still allows the direct exclusion of relatedness using discordant homozygotes count, similarly to what has been done with MS (Anderson, 2005). In addition, methods exploiting the high number of data to give a proportion of relatedness between pairs of individuals have been developed, such the algorithm in GCTA64 (Yang et al., 2011). This software estimates parentage between pairs of individuals weighting the number of shared alleles with their frequencies in the whole population (Yang et al., 2011). These values allow the control of pedigree information and the rebuilding of all relationships among individuals.

Availability of such methods to estimate the relationships occurring between individuals raises the problem of comparability with pedigree-based estimates. In fact, the estimates using dense SNP data and pedigree need to be comparable in terms of values or pattern to be correctly implemented in breeding schemes. A previous study showed that different molecular parentage from MS markers highly correlate with pedigree based methods (Toro et al., 2002). Recently, the same authors suggested that SNP array with a density of 500 SNP/Morgan should be enough for molecular to surpass genealogical co-ancestry (Toro et al., 2014).

Even though high density SNP arrays provide precise estimates of genealogical co-ancestry, in some species the cost required to perform the analysis is still unaffordable for breeders to perform it routinely. To overcome this problem, several studies have started the development of small SNP-based panels for parentage assignment in several species, livestock included.

In cattle, a panel of 100 + 100 SNPs for parentage assessment has already been released by the International Society for Animal Breeding

(ISAG) in 2012 (ISAG). However, the broad application of genomic selection in highly productive cosmopolitan cattle breeds is performed using a medium-density SNP arrays that provides more than 20K markers. This panel also allows the validation of relationships, and consequently limited the diffusion of this tool.

The situation changes dramatically in small ruminant species, where the low animal value may limit the diffusion of genomic tools. In this situation, the availability of cheap tools may help breeders in overcoming the limitations related to the management practices used.

For sheep species, a research group has recently published a 163 SNPs international panel for parentage assessment (Heaton et al., 2014), and an Italian group of researchers identified two panels of 110 and 108 SNPs for geographical and breed assignment (Dimauro et al., 2015).

In this thesis, I contributed in the development and comprehension of genomic tools for parentage. We published a study in which we developed a method to identify small SNP panels even in case of unreliable SNP positioning. Using this approach, I identified on three different Italian goat breeds two panels of 114 and 130 SNPs, respectively. In addition, I presented at an international congress a comparison of genomic with additive parentage estimates using an Italian small dog population genotyped with more than 130K SNPs as a model. In this preliminary study, I evidenced the high correlations between pedigree- and SNP-based parentage estimates.

These evidences suggests the possibility of implementation of these methods in genomic selection. Selecting a proper panel, with a well-defined reference population and SNP set, is a prerequisite to adequately implement this tool in breeding schemes. In addition to SNPs for panel assessment, the panel should also be extended including markers for

disease diagnosis, for important genes (i.e. caseins) and even for breed traceability and assessment.

2.3.3 Product tracing

The ability to reliably identify animals or animal products, from the farm to the retailer, is known as traceability (McKean, 2001). This aspect came to light in the past few years due to the increasing interest of consumers in food quality (Dalvit et al., 2007), safety (Goffaux et al., 2005; Barcos, 2001) and animal welfare (Opara and Mazaud, 2001). These concerns are also related to social changes in food habits (Cozzi and Ragno, 2003), losses of meat organoleptic properties (Kerry et al., 2001), attention to ecological and environmental matters (Opara and Mazaud, 2001) and the increase of industrialized farming (Ajmone-Marsan et al., 2004) amplified the need for reliable methods to trace back food products to their source. This need led to the definition of a European Union rule that makes label-based product tracing mandatory in all member states (Commission of the European Coomunities, 2000).

The strong interest behind traceability highlighted the need for a method capable of monitoring all the steps from the farm to the selling points. An effective solution should present several characteristics: it has to be cost-effective, easiness of use and interpretation, durable over time, fraud-proof and respectful towards both human and animal health (Dalvit et al., 2007). Today, two solutions are commonly applied to trace animal products: the conventional traceability and the geographical traceability.

The first solution traces animal products to the respective source using tags with a unique identifier for each individual. Even if this method is extremely cost-effective and simple, it is still prone to errors or fraud (Stanford et al., 2001).

The second kind of traceability does not aim to identify individuals but to associate a product to a specific geographical area. The European Union recognize two kind of protection: the protected designation of origin (PDO) and the protected geographical indication (PGI). These two kinds of protection are used to optimize local products, and are particularly important to promote less competitive economies optimizing local livestock systems (Dalvit et al., 2007).

Both approaches are economical and easy to use, but prone to errors or fraud and therefore a solution is needed. A possible solution is the adoption of genetic markers, that shows advantages such as high variability among individuals, ubiquity in all tissues, immutability during animal life span and stability during the processing and manipulation of the products.

Several kind of markers have been proposed to perform genetic traceability of individuals, breeds or species. These can be classified in two main categories of approaches: i) deterministic and ii) probabilistic approaches (Ajmone-Marsan et al., 2004). The deterministic approach aims to identify breed-specific allelic variants for genes and markers, without the need of any specific statistics. On the other hand, the probabilistic approach aims to identify allelic frequencies of a set of markers in several breeds, and assign an individual using maximum likelihood functions- (Paetkau et al., 1995), Bayesian- (Rannala and Mountain, 1997) or genetic distance-based (Cornuet et al., 1999) statistical approaches.

The first category accounts for genes related to coat color. In this case, such markers assign a product to its original breed by detecting allelic variants which are specific for one or for a group of breeds (Nicoloso et al., 2012). However, probably the most important family of markers in this category is the AFLP that seems to have a great discrimination power in comparison with other markers, such as MS, as stated by several authors (Óvilo et al., 2000; Negrini et al., 2007; De Marchi et al., 2006). Despite that great discrimination power, AFLP markers have several pitfalls that limit

their diffusion: methodology is expansive, complex and is hardly possible to implement it routinely.

The second category of analysis, which relies on statistical approaches instead of a direct exclusion uses different kind of markers. Probably the most used for this purpose are the microsatellites (Goffaux et al., 2005). The precision of this method is highly influenced by several factors, such as the variability of every marker, the number of markers considered and the genetic distances between populations: i.e. three MS markers could be enough to achieve 95% of precision as long as the populations are highly differentiated (Fst > 0.2) and the sample size is adequate (> 20; Bjørnstad and Røed, 2001). Several software systems that use MS markers to perform these analyses have been proposed, but the most famous is probably STRUCTURE (Falush et al., 2003) that together with CLUMPP for multiple run processing (Jakobsson and Rosenberg, 2007) and DISRUPT for graphical production (Rosenberg, 2003) allows detection of the genetic background of individuals.

Recently, dense SNP arrays have been proposed as a candidate technology to perform breed assignment (Dimauro et al., 2013; Heaton et al., 2014). Even if SNP markers show a lower informativeness per marker due to their biallelic nature, the high number of markers that can be genotyped overcome this limitation. Several software systems are available and use high density SNP data to infer population genetic structure, and can be classified in model-free and model-based methods (Wollstein and Lao, 2011). The first category includes methods that rely on multivariate statistics, such as PCA and MDS, previously described to also infer the population distances (Alexander et al., 2009).

Principle	Method	Name of package	Reference
These methods detect the population structure using classical and Bayesian statistical approaches, without considering any genetic model (**Model-Free**)	Principal component analysis	EIGENSOFTa	Price et al., 2006
	Principal components and Moran's/	adegenet	Jombart and Ahmed, 2011
	Multidimensional scaling	PLINK	Chang et al., 2015
	Principal coordinates	PCO-MC	Reeves and Richards, 2009
	Spectral graph theory	GemTools	Klei et al., 2011
	Spectral graph theory	SpectralGem	Lee et al., 2010
	Laplacian eigenfunction	LAPSTRUCT	Zhang et al., 2009
	Genetic algorithm coupled to AMOVA	GAGA	Lao et al., 2014
These methods use genetic models to identify the ancestry of individuals (**Model-Based**)	Log-likelihood HWE	ADMIXTURE	Alexander et al., 2009
	Log-likelihood HWE	FRAPPE	Tang et al., 2005
	Bayesian HWE	STRUCTURE	Falush et al., 2003
	Bayesian HWE	fastSTRUCTURE	Raj et al., 2014
	Nonnegative matrix factorization	sNMF	Frichot et al., 2014
	Bayesian	BAPS	Corander et al., 2004
	Chromopainting and Bayesian classifier	fineSTRUCTURE	Lawson et al., 2012
	Log-likelihood genotypic/haplotypic gradients	LOCO-LD	Baran et al., 2013
	Log-likelihood allelic gradients	SPA	Yang et al., 2012
	ADMIXTURE and linear regression	GPS	Elhaik et al., 2014
	Bayesian clustering with spatial information	TESS	Caye et al., 2016

Table 2.4 - Examples of software to infer population co-ancestry (Wollstein et al., 2015)

Instead, the latter includes software that uses different population statistic models to infer the co-ancestry of analyzed individuals. Some examples of this family are ADMIXTURE (Alexander et al., 2009), fastSTRUCTURE(Raj et al., 2014), FRAPPE (Tang et al., 2005), TESS (Caye et al., 2016), sNMF (Frichot et al., 2014) and fineSTRUCTURE (Lawson et al., 2012).

In my studies, I exploited dense SNP array to perform selection signature detection to define genes associated to coat color pattern specific to two Pakistani goat breeds, which are well-known as markers for breed traceability. The discovery of new allelic variants, and consequently of new molecular markers, may help in the optimization of local breeds and product certifying their origin and preventing fraud that may negatively affect the breeders economical conditions.

Section 3 - Project Articles

3.1 Population Structure Genetics

3.1.1 Aim

The knowledge of the genetic background of population could be of help in the breeding and management of breeds. This is even truer for small populations, where the genetic uniqueness should be preserved as an important biodiversity resource for future needs. In this context, selection signature discovery is particularly important since it allows us to detect breed-specific genomic regions, highlighting the genetic uniqueness of populations.

In this context, in my study I examined the selection signatures in 369 animals of 16 Italian goat populations. These analyses highlighted interesting signals on a 4 Mb region on chromosome 7 in the Valdostana goat, a small breed reared in Val d'Aosta, an Italian region in the Alps mountain ranges. This small population, counting around 600 animals, shows the typical alpine morphology, with the exception of the horns that are oversized in both males and females. This characteristic lets the breeders and researchers speculate about the introgression of the Alpine Ibex (*Capra ibex*).

It is also worth mentioning that this small population, bred for milk production, is also used in non-cruel fighting contests, the *Batailles the chevres*. This unique background make it a particularly interesting model to study to detect the population structure and the selection signatures in the genome of this breed.

3.1.2 The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep regions

Article published on Mammalian Genome (2017) 28:114–128 (DOI:10.1007/s00335-017-9678-7)

Authors and affiliations

Andrea Talenti[1*], Francesca Bertolini[2*], Giulio Pagnacco[1], Fabio Pilla[3], Paolo Ajmone-Marsan[4], Max F. Rothschild[2], Paola Crepaldi[1] and The Italian Goat Consortium

[1] Dipartimento di Medicina Veterinaria, Università degli Studi di Milano, Milan, Italy

[2] Department of Animal Science, Iowa State University, Ames, IA, USA

[3] Dipartimento Agricoltura, Ambiente e Alimenti, Università degli Studi del Molise, via Francesco De Sanctis s.n.c., 86100 Campobasso, Italy

[4] Istituto di Zootecnica, Università Cattolica del Sacro Cuore, via Emilia Parmense, 84, 29122 Piacenza, Italy

* Authors have contributed equally to the work

Abstract

The Valdostana goat is an alpine breed, raised only in the northern Italian region of the Aosta Valley. This breed's main purpose is to produce milk and meat, but is peculiar for its involvement in the "Batailles de Chèvres," a recent tradition of non-cruel fight tournaments. At both the genetic and genomic levels, only a very limited number of studies have been performed with this breed and there are no studies about the genomic signatures left by selection. In this work, 24 unrelated Valdostana animals were screened for runs of homozygosity to identify highly homozygous regions. Then, six different approaches (ROH comparison, Fst single SNPs and windows based, Bayesian, Rsb, and XP-EHH) were applied comparing the Valdostana dataset with 14 other Italian goat breeds to confirm regions that were different among the comparisons. A total of three regions of selection that were also unique among the Valdostana were identified and located on chromosomes 1, 7, and 12 and contained 144 genes. Enrichment analyses detected genes such as cytokines and lymphocyte/leukocyte proliferation genes involved in the regulation of the immune system. A genetic link between an aggressive challenge, cytokines, and immunity has been hypothesized in many studies both in humans and in other species. Possible hypotheses associated with the signals of selection detected could be therefore related to immune-related factors as well as with the peculiar battle competition, or other breed-specific traits, and provided insights for further investigation of these unique regions, for the understanding and safeguard of the Valdostana breed.

Introduction

Over the past several years, the increase of genomic technologies and molecular information has given researchers the chance of developing useful tools for genome-wide analyses in livestock. Since 2008, a series of single-nucleotide polymorphism (SNP) chips of medium and high density have been developed and assessed for the major livestock species (Nicolazzi et al., 2015). These tools have provided the opportunity to investigate the underlying structure of genomes for several purposes such as detection of selective sweeps, breed differentiation, genome- wide association studies (GWAS), and genomic selection in cattle, pigs, sheep, horses, and chickens (Meuwissen et al., 2013; Nicolazzi et al., 2015).

The selective sweep can be defined as a reduction or elimination of variation among the nucleotides in genomic regions adjacent to a mutation that become fixed from natural or artificial selective pressure. This selection tends to cause changes not only in the pattern of variation among selected loci, but also neutral loci linked to them via the well-known hitch-hiking effect. The effect due to selective pressure can affect different traits, from aesthetic to economical variants, and they could also be associated with deleterious phenotypes as well as behavioral traits. These regions of lower variability could be therefore seen as "genomic footprints" that allow identification of loci subjected to that selective pressure (de Simoni Gouveia et al., 2014). Several approaches have been used to detect these regions, such as run of homozygosity (ROH; Zhao et al., 2012; Fleming et al., 2016), fixation index analysis (Fst; Kijas et al., 2012; Porto-Neto et al., 2013), and haplotype- based analyses (e.g., de Simoni Gouveia et al., 2014). Other approaches, such as Bayesian methods, have also been successfully used on some occasions to detect selective sweeps as well (e.g., Druet et al., 2014).

Compared with the other major livestock species, the goat was one of the last for which medium-density SNP chips became available. In 2012,

through the international goat genome consortium, the first medium-density Goat 52 K SNP chip was designed and released (Tosser-Klopp et al., 2014). The first goat genome of a Yunnan black female goat was completely assembled and officially released about one year before in 2013 (Du et al., 2012; Tosser-Klopp et al., 2012; Dong et al., 2013). Since the Caprine 52 K SNP chip was recently developed, only a limited number of studies have been reported but they encompass a wide variety of aspects including (i) linkage disequilibrium, population distribution, and structure analyses in several goat breeds (Kijas et al., 2013; Nicoloso et al., 2015; Lashmar et al., 2015) (ii) implementation and development of marker-assisted breeding scheme strategies (Brito et al., 2015; Lashmar et al., 2015; Mucha et al., 2015); (iii) development of SNP chip-based caprine parentage tests (Talenti et al., 2016); and (iv) signatures of selection and GWAS analyses for phenotypic traits and adaptation (Becker et al., 2015; Kim et al., 2015; Reber et al., 2015). Italy is a country that can be considered an important reservoir of genetic resources for goat species in Europe. Nowadays, 36 breeds are officially recognized by the National Goat and Sheep Breeder Association and 14 of them are localized in the Alpine regions (ASSONAPA, http://www.assonapa.it). The Valdostana goat is an alpine breed, raised in the northern Italian region of the Aosta valley in the extreme north-west corner of the Alpine area, a natural border of Northern Italy. The Valdostana has been primarily used for the production of cheese (in 125 days of lactation, the production is approximately 249 Kg) and meat and for the production of traditional and seasoned products (e.g., the Mocetta). While this breed is from the alpine region, it differs from the other breeds of the same area primarily because of its larger size, and for the presence of well-developed horns in females (ASSONAPA, 2014). The Valdostana characteristics have been influenced by the natural selection of the mountain environment, but also by the selection of farmers for the maintenance of the recent traditional fighting tournaments that are organized in Valle d'Aosta. These non-cruel fights, called "*Batailles de*

Chèvres", are a recent event of fight tournaments that take place in the valley (Association Comité Régional Batailles des Chèvres, 2016). The current status of this population is 640 registered animals and this breed is considered at risk with a declining number of animals reared (Nicoloso et al., 2015). At both genetic and genomic levels, only a very limited number of studies have been performed on this breed (Colussi et al., 2008; Nicoloso et al., 2015) and there is no information about the genomic signatures left by selection.

The aim of this work was to identify unique selective sweep regions in the Valdostana goat genome resulting from man-made artificial selection and natural/environmental selection. These genomic regions could govern phenotypic traits of interest and may be linked to peculiar phenotypic characteristics of this breed. To accomplish this task, we used the medium-density Goat 52 K SNP chip to detect ROH and we compared the Valdostana genome with those of 14 other Italian breeds using ROH comparisons, Fst, haplotype-based, and Bayesian analyses.

Material and Methods

Goat sampling, genotyping, and multidimensional scaling analysis

Animals belonging to 15 different breeds were collected in Italy from different farms (approximately three from each farm) to collect animals as much unrelated as possible. For each animal, blood samples were collected following the European rules (Council of Europe, 1986) for animal care and DNA extraction was performed using a commercial kit (NucleoSpin Blood, Macherey–Nagel) according to the manufacturer's instructions. Then, DNA samples were genotyped using the CaprineSNP50 BeadChip (Illumina Inc., San Diego, CA; Tosser-Klopp et al., 2014). For further details, see Nicoloso et al., (2015). Goats (N = 369) and breeds (N =

15) included in this study are listed in Table 3.1. In addition to Valdostana (n = 24; 15 females and 9 males), a group of 14 other breeds (Argentata dell'Etna, Dell'Aspromonte, Ciociara Grigia, Girgentana, Maltese, Nicastrese, Sarda, Di Teramo, Bionda dell'Adamello, Camosciata delle Alpi, Nera di Verzasca, Orobica, Saanen, Valpassiria) was investigated in order to find the most unique and divergent genomic regions across the Valdostana genome. To further confirm the unrelatedness of the animals within the dataset, above all among the Valdostana goats, an in-house script was used for calculating the number of discordant homozygotes at each locus between all pairs of individuals in the dataset. A pair is defined related if the total number of discordant homozygotes is lower than 100 (< 0.5 %). Out of a total of 67,896 comparisons among individuals, only 32 pairs had a number of discordant homozygotes below the given threshold of 100 and were considered closely related, and none of them were individuals of the Valdostana breed (data not shown).

Breed name	N.
Valdostana	24
Argentata dell'Etna	24
Dell'Aspromonte	24
Bionda dell'Adamello	24
Camosciata delle Alpi	30
Ciociara Grigia	19
Girgentana	24
Maltese	31
Nicastrese	24
Nera di Verzasca	19
Orobica	23
Saanen	24
Sarda	32
Di Teramo	23
Valpassiria	24

Table 3.1 - Name of breeds and number of animals for each breed considered for the analyses. All animals except the Nera di Verzasca are already generally described in Nicoloso et al. 2015

SNPs with a call rate < 90 %, monomorphic SNPs, variants not mapped to the assembly or on the X chromosome were excluded from subsequent analyses using Plink v1.9 (Chang et al., 2015). Monomorphic SNPs can be considered fixed across all breeds, so they were not considered informative for the purpose of the analyses. After the SNP marker quality check, animals with an individual call rate < 0.95% as performed by Nicoloso et al. (2015) were removed from the dataset. The filtered dataset was then phased and imputed breed by breed for the missing genotypes using Beagle v3.3.2 (Browning and Browning, 2007, 2008; Browning, 2011). Multi-dimensional scaling (MDS) was calculated in two dimensions using the cluster algorithm of Plink v1.9 (Chang et al. 2015).

Runs of homozygosity in Valdostana goats and enrichment analyses of regions under selection

Analyses of high-homozygosity regions across the genome were conducted with the --homozyg command in Plink v1.9 (Chang et al. 2015), including in each window 20, 25, or 30 SNPs with the command --homozyg-snp, and allowing no heterozygotes (--het 0). The output files (.summary) contained for each SNP a raw value that indicated the number of animals and was normalized by dividing that number by the total number of animals included in the analysis, obtaining a locus homozygosity (H) range from 0 (0) to 1 (100%) as performed in Bertolini et al., (2016). Regions with H ≥ 0.62 at each SNP site, equivalent to the top 0.2% of the empirical distribution of all the SNPs, were considered as regions of higher homozygosity.

Annotation of all highly homozygous regions was obtained downloading the complete list of genes available for the Capra hircus genome CHIR_1.0 available in the CoGe (Comparative Genomics) database (Lyons and Freeling, 2008, https://genomevolution.org/coge/). Then the list of genes was screened at the desired positions using the BEDTools software (Quinlan and Hall, 2010). Enrichment analysis was performed using the web-based tool Enrichr (Chen et al., 2013; Kuleshov et al., 2016; http://amp.pharm.mssm.edu/Enrichr/), where "Wiki pathway" and "Gene Ontology biological processes" were investigated.

Valdostana vs other goat breeds

A total of six different analyses were performed comparing the Valdostana to the 14 other breeds considered separately (ROH comparison)

or comparing the Valdostana to the 14 other breeds as a whole (Fst, haplotype-based, and Bayesian analysis) in order to investigate whether the most homozygous regions detected in Valdostana could be considered as unique to the breed.

ROH comparison

For each of the remaining 14 breeds, homozygosity was determined as described above for the Valdostana and the results were separately H transformed. Summary statistics were calculated modifying the approach suggested by Akey et al., (2010) to compare the locus-specific divergence for each goat breed based on H scores:

$$SHD_i = \sum_{i \neq j} \frac{HD^{ij} - E(HD^{ij})}{sd(HD^{ij})}$$

where HD^{ij} is the difference of *H* between two breeds *i* and *j*, and $E(HD^{ij})$ and $sd(HD^{ij})$ are the expected value and standard deviation of HD between *i*th and *j*th breed. An SHD value > 6 was considered as the threshold which indicates the highest divergence at each locus, equivalent to approximately the top 0.2 % of the empirical distribution.

Fst analysis

Fst analysis between Valdostana compared to all the 14 other goat breeds of the dataset was performed for each SNP, using the formula reported by Karlsson et al., (2007). Then, a mean Fst value (mFst) was

calculated in 1 Mb slid- ing windows with 500 Kb overlapping using an in-house script. The window size was chosen to be consistent with the ROH, according to SNP density (20 SNP * 50,000 bp/ SNP = 1,000,000 Mb). Values >0.31 for the mFst and >0.56 for the single-SNP Fst represented approximately 0.2 and 0.05%, respectively, of the empirical distribution of all the values, and were the most divergent between the two groups and were therefore considered.

Bayesian analysis

A Bayesian approach called Bayes B implemented in GenSel software (Fernando and Garrick, 2009) was used to obtain the variance explained by SNPs in every genomic non-overlapping window of 1 Mb each, using categorical traits. Valdostana goats were treated as "case" and all the other breeds together were treated as "controls"; the comparison was performed between these two groups, with no fixed effects or covariates being added in the model. A prior probability (p_i) of 0.992 was used to fit 250–300 markers per iteration of the Markov chain in a mixture model for the estimation of individual SNP effects (Dekkers, 2012; Onteru et al., 2013), with VarG = 123.383, VarR = 1. Windows that explained more that 1% of the variance were considered.

Haplotype-based analysis

Two analyses, Rsb and XP-EHH, were performed. Rsb was defined as the standardized log-ratio of the integrated extended haplotype homozygosity (EHH) between pairs of populations (Tang et al., 2007), while Cross-population Extended Haplotype Homozygosity (XP-EHH) compares

the integrated EHH profiles between two populations at the same SNP (Sabeti et al., 2007). The Rsb statistic compares EHH for the same SNP in two different populations and can provide evidence of selection given the presence of high- frequency or fixed alleles in one population but not on the other (Tang et al., 2007). Similarly, the XP-EHH detected selective sweeps in which one allele had undergone strong directional selection in one population while remaining polymorphic in the population as a whole (Sabeti et al., 2007).

The *rehh* R package was used to compute Rsb values with default parameters (Gautier and Vitalis, 2012), whereas the selscan software was then used to compute XP-EHH (Szpiech and Hernandez, 2014). XP-EHH values were then normalized using the norm tool included in the selscan package. Ancestral allele information, which is important for this analysis, was identified starting from a dataset composed of eight Ibexes (data not shown) and seven Bezoars (produced by the NEXTGEN project, 2009) that were genotyped with the same GoatSNP50 BeadChip, in a manner similar to what has been previously performed in cattle (Matukumalli et al., 2009). These two caprine species are known to be geographically close (Alpine Ibex) or the closest ancestors of the modern goat (Bezoar, Colli et al., 2015). Values >8 and >4.5 that represented around 0.2% of the empirical distribution of all the normalized values for Rsb and XP-EHH, respectively, were considered as biologically relevant.

Results

The GoatSNP50 BeadChip contains 53,347 SNPs, and a total of 3,404 SNPs were mapped to the X chromosome or were unmapped, and 1,051 SNPs did not pass the quality-filtering step. All of these were excluded from further analyses. Therefore, the working dataset included 48,892 autosomal SNPs. All animals had a genotyping rate >0.95%. The MDS plot

shown in Figure 3.1 demonstrates a clear separation between breeds raised in the north and in the south of Italy, with the Valdostana (black dots) clearly belonging to the cluster of northern breeds, as already reported by Nicoloso et al. (2015), with some animals overlapping the Alpine and Nera di Verzasca breeds.

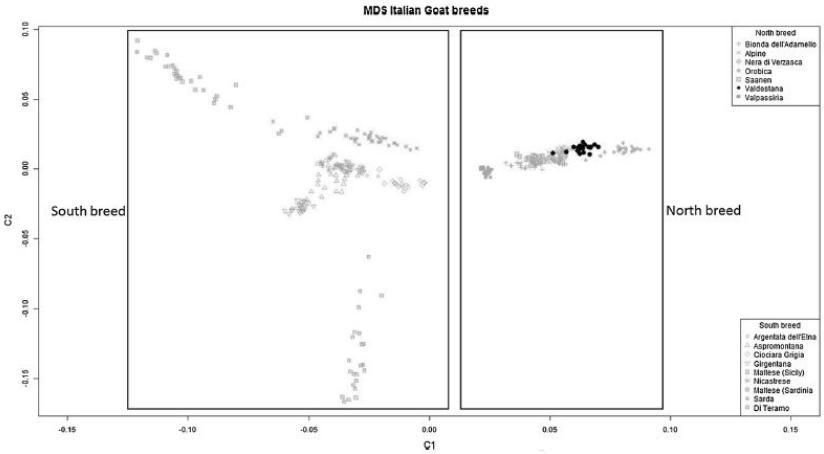


Figure 3.1 - Multidimensional scaling of Italian goat breeds and populations including Valdostana. The two clusters indicate breeds raised in the south and the north of Italy. Valdostana is black colored. Modified from Talenti et al. (2017).

Runs of homozygosity

For the runs of homozygosity, three SNP windows were considered. The window of 20 SNPs identified three peaks above the threshold (Figure 3.2), while using 25 and 30 SNPs showed a decay of one of the peaks (Figure S3.7 and S3.8). Therefore, the window with 20 SNPs was chosen for the following analysis. For the selected threshold, three regions with H ≥0.62 were detected (Figure 3.2).

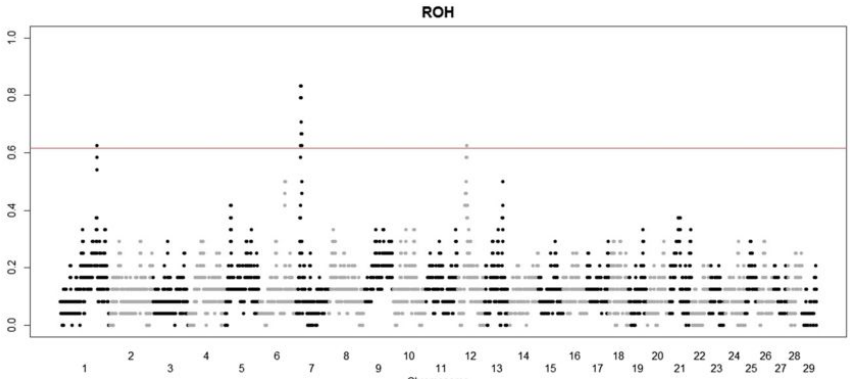


Figure 3.2 - Regions of homozygosity in the Valdostana dataset. The raw values obtained were normalized according to the number of animals used in the analysis. A threshold of H score = 0.62 was chosen to detect the regions with low heterozygosity (indicated with the red line)

One region was detected on chromosome 1 (from 112,414,563 to 113,060,421 bp), with the highest H value of 0.63 (Figure S3.9) and a length of 645 Kb. A second region located on chromosome 7 (from 15,057,327 and 19,670,982 bp) had the highest H value of 0.83 and was 4.6 Mb in length (Figure S3.10), and a third smaller region on chromosome 12 (from 28,544,783 to 28,664,628 bp) showed the highest H value of 0.63 (Figure S3.11) with 120 kb length. The list of the 129 annotated genes located in the three high-homozygosity regions is reported in Table 3.2. The region on chromosome 1 contained 4 genes and the second region on chromosome 7 had 116 genes. A total of 37 genes were included in the subregion on chromosome 7 within the region on the top of the peak, with all the SNPs having H = 0.83. These regions included the *MAP2K2*, *APBA3*, and *ATCAY* genes. The third region on chromosome 12 contained 1 annotated gene.

Chr	Start	End	Gene symbol	Gene name	Chr	Start	End	Gene symbol	Gene name
1	675002	123987664	TRNAG-UCC	transfer RNA glycine (anticodon UCC)	7	16863564	16866964	C7H19orf77	–
1	7624399	148400517	TRNAV-CAC	transfer RNA valine (anticodon CAC)	7	16878586	16930540	NFIC	Nucleus factor I/C (CCAAT-binding protein L54)
1	112400894	112492829	RAP2B	RAP2B, member of RAS oncogene family	7	17004688	17022568	CELF5	CUGBP, Elav-like family member 5
1	112845294	112849567	P2RY1	purinergic receptor P2Y, G-protein coupled, 1	7	17063113	17082730	NCLN	Nicalin
7	15053932	15058909	HSD11B1L	hydroxysteroid (11-beta) dehydrogenase 1-like	7	17081246	17083566	S1PR4	Sphingosine-1-phosphate receptor 4
7	15060184	15062136	C7H19orf70	–	7	17095410	17116109	GNA15	Guanine nucleotide-binding protein, alpha 15
7	15065044	15069283	LOC102169378	–	7	17126314	17137531	GNA11	Guanine nucleotide-binding protein, alpha 11
7	15069451	15075296	SAFB	scaffold attachment factor B1-like	7	17167056	17167370	LOC102137616	pseudogene
7	15077037	15126429	LOC102169953	–	7	17169117	17182692	AES	Amino-terminal enhancer of split
7	15126550	15154442	SAFB2	scaffold attachment factor B2	7	17191762	17218405	TLE2	Transducin-like enhancer of split 2
7	15277239	15278880	LOC102183866	–	7	17219310	17234875	TLE6	Transducin-like enhancer of Split 6
7	15420303	15542566	PTPRS	Protein tyrosine phosphatase, receptor Type, S	7	17247811	17267239	ZNF77	Zinc finger 77
7	15541392	15647804	KDM4B	Lysine (K)-specific Demethylase 4B	7	17273536	17300375	ZNF555	Zinc finger 555
7	15669170	15703813	UHRF1	Ubiquitin-like with PHD and ring finger domains 1	7	17316607	17326730	ZNF554	Zinc finger 554
7	15708356	15722205	ARRDC5	Arrestin domain-containing 5	7	17333511	17350262	THOP1	Thimet oligopeptidase 1
7	15739749	15761331	PLIN3	Perilipin 3	7	17357031	17364911	SGTA	Small glutamine-rich tetratricopeptide repeat-containing, alpha
7	15770780	15786827	TICAM1	Toll-like receptor adaptor molecule 1	7	17377647	17384058	SLC39A3	Solute carrier family 39 (zinc transporter), member 3
7	15802328	15807760	FEM1A	Fem-1 Homolog A	7	17394483	17395184	DIRAS1	DIRAS family, GTP-binding RAS-like 1
7	15879539	15915659	DPP9*	Dipeptidil-peptidase 9	7	17444465	17541139	GNG7	Guanine nucleotide-binding protein
7	15920677	15932753	C7H19orf10*		7	17561663	17563284	GADD45B	Growth arrest and DNA damage-inducible, Beta
7	15936585	15953655	TNFAIP8L1*	Tumor necrosis factor, alpha-induced protein 8-Like 1	7	17587771	17597983	LMNB2	Lamin B2

*Table 3.2 - List of genes included in the highly homozygous regions. Genes located in the region with the highest H value (H = 0.83) were indicated with the * symbol. Modifed from Talenti et al. (2017).*

Chr	Start	End	Gene symbol	Gene name	Chr	Start	End	Gene symbol	Gene name
7	16015285	16040063	SEMA6B*	Sema domain, transmembrane domain (TM), And cytoplasmic domain, (Semaphorin) 6B	7	17600811	17601233	TIMM13	Translocase of inner mitochondrial membrane 13 Homolog
7	16041644	16044715	LRG1*	Leucine-rich repeat	7	17608098	17641028	TMPRSS9	Transmembrane protease, serine 2
7	16046725	16057485	PLIN5*	Perilipin 5	7	17669940	17680121	SPPL2B	Signal peptide peptidase-like 2B
7	16061550	16071410	PLIN4*	Perilipin 4	7	17679041	17687505	LSM7	LSM7 homolog, U6 small nuclear RNA and mRNA degradation associated
7	16075055	16099421	HDGFRP2*	hepatoma-derived growth factor-related protein 2	7	17689919	17711175	LINGO3	Leucine-rich repeat and Ig domain-containing 3
7	16114145	16123589	UBXN6*	UBX domain protein 6	7	17719888	17725516	C7H19orf35	-
7	16125311	16148160	CHAF1A*	Chromatin assembly factor 1	7	17727496	17778465	OAZ1	Ornithine decarboxylase Antizyme 1
7	16166155	16180027	SH3GL1*	SH3 Domain GRB2-Like 1	7	17783151	17822107	DOT1L	DOT1-Like histone H3K79 methyltransferase
7	16180192	16189526	MPND*	MPN Domain-containing	7	17827659	17831621	PLEKHJ1	Pleckstrin Homology domain-containing J1
7	16195347	16205742	STAP2*	Signal-transducing adaptor family member 2	7	17831541	17840494	SF3A2	Splicing factor 3a, subunit 2, 66 kDa
7	16206199	16221966	FSD1*	Fibronectin type III and SPRY domain-containing 1	7	17841203	17842810	AMH	Anti-mullerian hormone
7	16222367	16227311	TMIGD2*	Transmembrane and immunoglobulin domain-containing 2	7	17844234	17848014	JSRP1	Junctional Sarcoplasmic reticulum protein 1
7	16228061	16236382	SHD*	Src Homolog 2	7	17853923	17886867	AP3D1	Adaptor-related protein complex 3, Delta 1 Subunit
7	16243123	16257235	CCDC94*	Coiled-Coil domain-containing 94	7	17859329	17892234	IZUMO4	IZUMO Family member 4
7	16259166	16265047	EBI3*	Epstein–Barr virus-induced 3	7	17903199	17911300	MOB3A	MOB kinase activator 3 A
7	16265767	16283544	ANKRD24*	Ankyrin repeat domain 24	7	17929169	17949970	MKNK2	MAP kinase-interacting Serine/Threonine Kinase 2
7	16296333	16305024	SIRT6*	Sirtuin 6	7	18017834	18025614	SEPT8	Septin 8
7	16305157	16316543	CREB3L3*	CAMP Responsive element-binding protein 3-like 3	7	18029881	18036074	CCNI2	Cyclin I family, member 2
7	16349231	16366609	MAP2K2*	Mitogen-activated protein kinase 2	7	18044228	18120787	KIF3A	Kinesin family member 3 A
7	16399643	16403577	ZBTB7A*	Zinc finger and BTB domain-containing 7 A	7	18128067	18135921	IL4	interleukin 4
7	16416362	16447151	PIAS4*	Protein inhibitor of activated STAT, 4	7	18152706	18155442	IL13	interleukin 13

Chr	Start	End	Gene symbol	Gene name	Chr	Start	End	Gene symbol	Gene name
7	16462232	16471129	EEF2*	Eukaryotic translation elongation factor 2	7	18198088	18310565	RAD50	RAD50 homolog, double-strand break repair protein
7	16474787	16486621	DAPK3*	Death-associated protein kinase 3	7	18326235	18328495	IL5	interleukin 5
7	16499574	16,503,730	NMRK2*	Nicotinamide riboside kinase 2	7	18360575	18360980	LOC102188306	pseudogene
7	16515642	16550762	ATCAY*	Ataxia, cerebellar, Cayman type	7	18375,010	18382102	IRF1	interferon regulatory factor 1
7	16587868	16611947	ZFR2*	Zinc finger RNA-binding protein 2	7	18474665	18500421	SLC22A5	solute carrier family 22 (organic cation/carnitine transporter), member 5
7	16629035	16636141	MATK*	Megakaryocyte-associated tyrosine kinase	7	18524881	18567090	SLC22A4	Solute carrier family 22 (organic cation/zwitterion transporter), member 4
7	16640063	16641228	RAX2*	Retina and anterior neural fold homeobox	7	18584900	18605166	PDLIM4	PDZ and LIM domain 4
7	16644654	16646660	MRPL54*	Mitochondrial ribosomal protein L54	7	18623324	18655016	P4HA2	prolyl 4-hydroxylase, alpha polypeptide II
7	16648979	16656398	APBA3*	Amyloid beta (A4) precursor protein-binding, family A, member 3	7	18678180	18823311	LOC102182028	pseudogene
7	16656486	16686391	TJP3*	Tight junction protein 3	7	18847086	18849410	GM-CSF	–
7	16694357	16737658	PIP5K1C*	Phosphatidylinositol-4-Phosphate 5-Kinase, Type I, Gamma	7	18862567	18864359	IL3	interleukin 3
7	16745649	16757967	CACTIN*	Spliceosome C complex subunit	7	18922344	18985869	ACSL6	acyl-CoA synthetase long-chain family member 6
7	16762240	16773455	TBXA2R*	Thromboxane A2 receptor	7	18991713	19128723	MEIKIN	meiotic kinetochore factor
7	16776308	16779707	GIPC3	GIPC PDZ domain-containing family, member 3	7	19162923	19262640	FNIP1	folliculin-interacting protein 1
7	16782955	16788034	HMG20B	high-mobility group 20B	7	19268176	19485469	RAPGEF6	Rap guanine nucleotide exchange factor 6
7	16795853	16809504	MFSD12	Major facilitator superfamily domain-containing 12	7	19503530	19596261	CDC42SE2	CDC42 small effector 2
7	16809268	16815892	C7H19orf71	–	7	19670441	19693055	LYRM7	LYR motif-containing 7
7	16816243	16828319	FZR1	Fizzy/cell division cycle 20-related 1	12	28572340	28620141	UBL3	ubiquitin-like 3
7	16847991	16857637	DOHH	Deoxyhypusine hydroxylase/monooxygenase					

The enrichment analyses of the genes reported clusters of genes (adjusted P value <0.05) that are involved in activities related to the immune system such as regulation of immunoglobulin production, lymphocyte, T cells, mononuclear and leukocyte proliferation, as well as regulation of the JAK–STAT cascade (Table 3.3).

Biological process name	P-value	Adjusted P-value
regulation of lymphocyte proliferation (GO:0050670)	0.0001429	0.02775
regulation of mononuclear cell proliferation (GO:0032944)	0.0001474	0.02775
regulation of leukocyte proliferation (GO:0070663)	0.0001716	0.02775
positive regulation of lymphocyte proliferation (GO:0050671)	0.0002095	0.02775
positive regulation of mononuclear cell proliferation (GO:0032946)	0.0002178	0.02775
positive regulation of JAK-STAT cascade (GO:0046427)	0.0001554	0.02775
positive regulation of leukocyte proliferation (GO:0070665)	0.0002441	0.02775
regulation of JAK-STAT cascade (GO:0046425)	0.00003227	0.02775
regulation of alpha-beta T cell activation (GO:0046634)	0.0002385	0.02775
Positive regulation of immunoglobulin production (GO:0002639)	0.0003314	0.0339

Table 3.3 - Gene enrichment for genes inside the genomic window on chromosome 7

Comparison of the Valdostana breed with the other breeds

Six different approaches were tested to find regions across the genome that differentiated the Valdostana from the other goat breeds.

ROH comparisons, shown in Figure S3.12, identified three regions of highest divergence between the ROH of Valdostana and the ROH of the other breeds examined separately with the same parameters. The first region was located on chromosome 1 (from 112,301,140 to 113,060,421 bp), the second on chromosome 7 (from 15,057,327 to 19,670,982), and the last on chromosome 12 (from 27,763,600, to 28,664,628). These regions included the windows of high homozygosity detected analyzing the Valdostana separately, and is shown in detail in Figures S3.9-S3.11.

The results of the single-SNP Fst analysis are shown in Figure 3.3a and identified SNPs on 4 chromosomes: chromosome 1 (8 SNPs from 110,663,697 to 124,748,543 bp), chromosome 7 (12 SNPs from 15,992,536 to 19,504,658 bp), chromosome 9 (1 SNP 61,687,558 bp), and chromosome 12 (3 SNPs from 25,743,128 to 28,327,291 bp). These results were confirmed also performing the Fst analysis in 1 Mb partially overlapping windows and is shown in Figure 3.3b. The analysis identified nine windows that had values higher than the selected threshold of 0.31. This included two overlapping windows located on chromosome 1 (from 112 Mb to 113.5 Mb bp) and seven continuous and mainly overlapping windows located on chromosome 7 (from 15 Mb to 19.5 Mb). The window that included the markers identified with the single-SNP approach on chromosome 12 was right under/below the established threshold. Considering the two approaches, the windows detected with the Fst analyses were overlapping the three homozygous regions detected on chromosomes 1, 7, and 12 through runs of homozygosity.

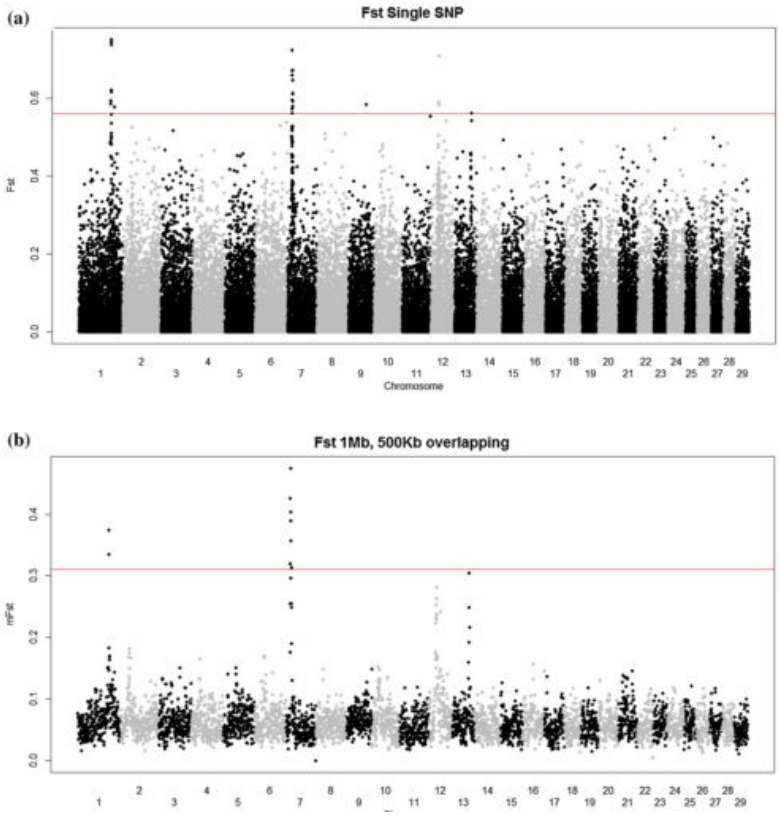


Figure 3.3 - Fst plot considering the single SNPs (a) and 1 Mb, 500 Kb overlapping window (b). On the Y-axis, mean Fst (mFst) values are plotted, while on the X-axis chromosomes are plotted. The red line across the plot indicated the fixed threshold of 0.56 for the single SNPs (a) and 0.32 for the mFst (b). Modified from Talenti et al. (2017)

GenSel analysis identified two 1 Mb windows that explained more than 1% of the variance. One window was located on chromosome 7 (from 16,043,582 to 16,974,423 bp) and explained 8.86% of the total variance. This window was included in the highly homozygous sub-region and in the Fst analysis. The second window was located on chromosome 13 (61,006,494 to 61,971,928 bp) that explained 1.58% of the variance (Figure 3.4).

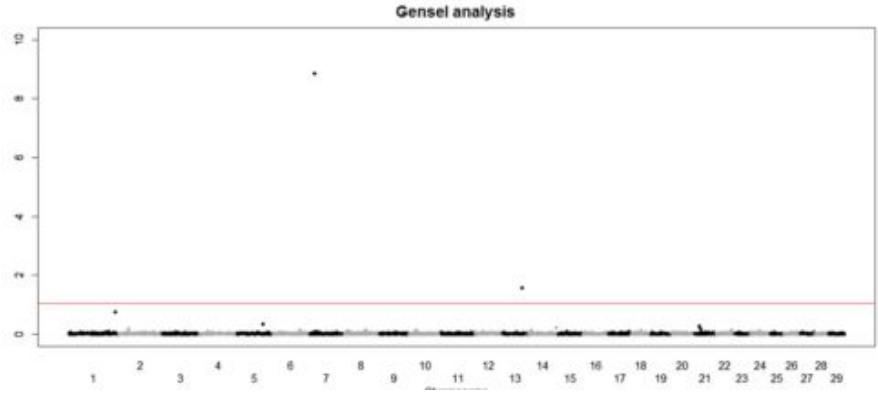


Figure 3.4 - 1 Mb non-overlapping window plot of Bayes B analysis. The percentage of the overall explained variance is plotted on the Y-axis and chromosomes are plotted on the X-axis. The red line indicates the threshold of 1% of explained variance. Modifed from Talenti et al. (2017)

The region on chromosome 7 was also confirmed by the Rsb analysis (Figure 3.5a) that identified 24 SNPs in the range of 15,221,110–20,065,201 bp above the threshold. A total of 13 SNPs were continuous from 15,221,110 to 15,948,105 bp. 1 SNP was located at position 17,028,582, and 8 and 2 SNPs were continuous in the ranges of 18,446,344–18,816,632 bp and 19,718,859–20,065,201 bp, respectively. Another non-continuous region was detected on chromosome 12 from 22,054,337 to 29,826,735 bp and contained 64 SNPs above the threshold. The XP-EHH analysis (Figure 3.5b) was concordant for the region on chromosome 7, with 54 non-continuous SNPs above the threshold that span from 14,464,313 to 20,737,623 bp and chromosome 12, with 10 SNPs from 24,467,948 to 28,489,734 bp. A continuous region was identified on chromosome 1, from 112,270,731 to 113,060,421 bp, which was therefore concordant with the ROH and Fst analyses. Two other SNPs on chromosome 13 (60,072,974 and 60,128,943 bp) were also above the threshold.

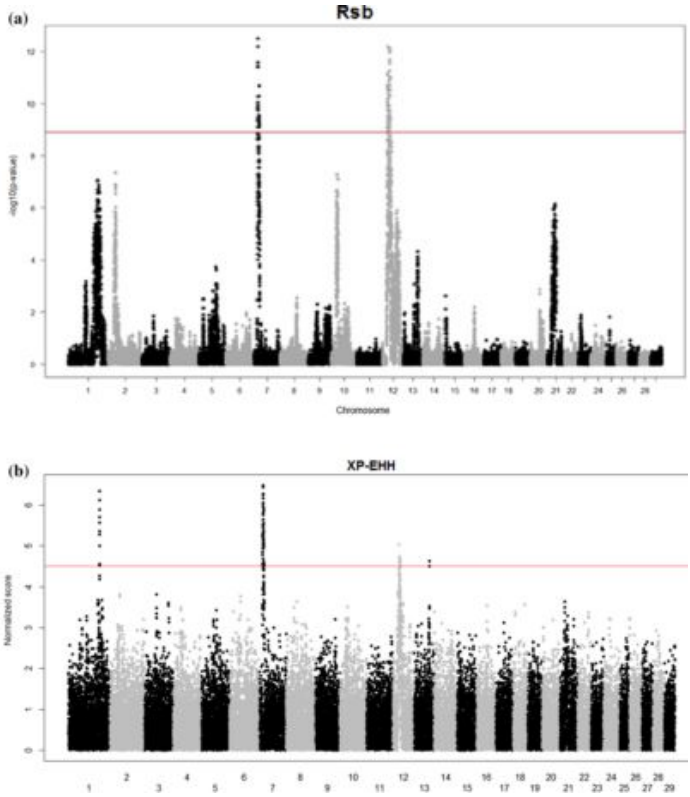


Figure 3.5 - Rsb (a) and XP-EHH (b) analyses. The normalized score for each SNP locus is plotted on the Y-axis and chromosomes are plotted on the X-axis. The red lines indicate the threshold values of 8 and 4.5 for Rsb and XP-EHH, respectively. Modifed from Talenti et al. (2017)

Discussion

Selective sweep analysis is a useful tool to investigate regions under selection in livestock, not only in animals under strong selection such as cattle, but also in those species that are reared for human consumption without a specific breeding scheme, such as goats (e.g., Andersson and Georges, 2004; Kim et al., 2015). Among the 36 officially recognized Italian breeds (http://www.assonapa. it), 21 are considered not to be at risk

(number of registered animals >1200 registered head), 11 are endangered (number of registered animals <1200 with a declining trend), and four are classified as in critical status (number of animals <100), as reported by (FAO, 2013). With 600 officially registered animals, the Valdostana could therefore be considered an endangered breed.

The multidimensional scaling (MDS) plot confirmed the division between breeds raised in the north and those in the south of Italy (Nicoloso et al., 2015). This division is probably due to several factors, such as the climate difference between the two Italian regions, where the north is colder and humid and the south is generally hotter and arid. Despite these conditions, some breeds can be raised in both parts, but this climatic difference facilitates the selection of more specific breeds in the different regions. As expected, the Valdostana breed fits in the northern cluster, with some animals that overlap the Alpine and Nera di Verzasca breeds. This fact is probably due to some gene flow that occurred between these three breeds, because they have always been reared free-range on pastures in the same regions. In the case of the Alpine breed, these two breeds also share the same coat color and pattern.

To consider a region as highly homozygous, a threshold of H >0.62 was chosen. This value was chosen also considering the presence of possible genotyping errors and the possibility that some of the Valdostana goats analyzed may have a few recent non-Valdostana ancestors. All these factors could reduce the number of animals that share a com- mon homozygous region. The runs of homozygosity analyses revealed the presence of a long region of about 4 Mb located on chromosome 7 and two other shorter regions (645 and 120 Kb) located on chromosomes 1 and 12, respectively.

The uniqueness of the region on chromosome 7 in the Valdostana breed was demonstrated by all five different analyses that compared the

Valdostana genome with a group of 14 non-Valdostana goat breeds sampled across Italy. Despite a slightly different number of regions detected, all the five statistical analyses were concordant in showing the region on chromosome 7 as the most divergent between Valdostana and the other breeds. The regions identified on chromosomes 1 and 12 were also found divergent in almost all the comparisons, except for the Bayesian analysis.

Three of the genes within the highest homozygous H score on chromosome 7 (H = 0.85) were the *MAP2K2* (Mitogen-Activated Protein Kinase Kinase 2) gene, the *APBA3* gene (Amyloid Beta (A4) Precursor Protein-Binding, Family A, Member 3), and the *ATCAY* gene (Ataxia, cerebellar, Cayman type). These genes could be directly or indirectly involved in modulating scrapie or Yersinia Pseudotuberculosis, two widespread diseases of sheep and goat (Tanahashi and Tabira, 1999; King and Turner, 2004; Nordström et al., 2005; Gossner and Hopkins, 2015). It has been observed that Valdostana goats have a difference in several alleles of *PRNP* (Prion Protein gene: the major gene involved in scrapie) compared to the other breeds of northern Italy even if this difference was not significant (Colussi et al., 2008). The uniqueness of the region in Valdostana may provide interesting insights for future studies directed in this direction.

The enrichment analysis revealed that several of the genes within the region are linked to the development/regulation of several components of the immune system. It is interesting to underline that a genetic link between behavior and immunity systems has been hypothesized (Petitto et al., 1994). These authors showed that cytokines and T-cell proliferation were higher in mice bred for high aggression than in mice bred for low aggression. Since that initial research, the association between immune cell activity and various measures of aggressive behavior has been described in several studies and documented in humans, mice, and cats. The factors that have been found in these studies include pathways that mainly involved

inflammatory cytokines and T cells (reviewed by Zalcman and Siegel, 2006). Interleukins modulate neurotransmitters and neurocrine activity influencing the individual's behavioral response to potentially threatening environmental stimuli (Bhatt and Siegel, 2006).

Figure 3.6 - Two Valdostana goats during the "Batailles des Chevres." The image was provided by Cinzia Finotto Association Comité Régional Batailles des Chèvres

These findings may be linked with the peculiar activity, battle competition, for which the Valdostana has been employed. This characteristic non-cruel "*Batailles de Chèvres*" has a recent origin and is officially recognized, with the first competition having taken place in 1981. In addition, with the Valdostana cow traditional battle, *Bataille de Reine*, these bloodless competitions use the animal's natural behavior to fight (Figure 3.6). Each match ends when one of the two competitors recognizes the superiority of the other. This event represents an attraction for the tourists and an economic opportunity for the farmers that own the strongest animals. Even if directed selection for the traits related to this competition were not performed, a recent estimation of heritability of the "fighting ability" trait in Valdostana cattle showed that selection for battle performance would be successful (Sartori and Mantovani 2010). The large region on chromosome 7 is probably an event of recent selection, and maybe it can be partially explained by the new fighting activity of this breed of goat.

In conclusion, we found evidence of selective sweep regions on three different chromosomes in the Valdostana goat breed. These regions showed unique homozygosity patterns when compared to a wide representation of the Italian goat breeds. Interestingly, these regions contained genes involved in the immune system development/ regulation. Our findings suggest that this region could be linked with the very recent, non-cruel battle events that are uniquely involved with these breeds. Further analyses will need to be performed to investigate in detail the three regions that could also be related to other breed-specific traits. All these are insights for further investigations of these unique genomic regions, for the understanding and safeguard of the Valdostana breed.

Acknowledgements

We thank the goat breeders for providing the samples. Financial support was provided in part by State of Iowa funds, Iowa State University, and the Ensminger International Animal Agriculture fund. This work was partially funded by the Italian Min- istry of Agriculture (grant INNOVAGEN). This study makes use of data generated by the NextGen Consortium. The European Union's Seventh Framework Programme (FP7/2010–2014) provided funding for the project under grant agreement no. 244356 - "NextGen."

Supplementary material

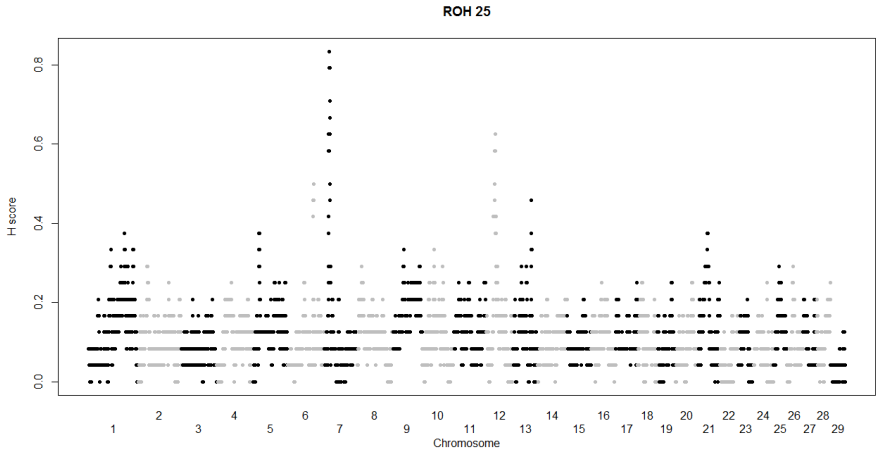


Figure S3.7 - Regions of homozygosity in the Valdostana dataset with the option winidow-snp 25. . The raw values obtained were normalized according to the number of the animals used in the analysis

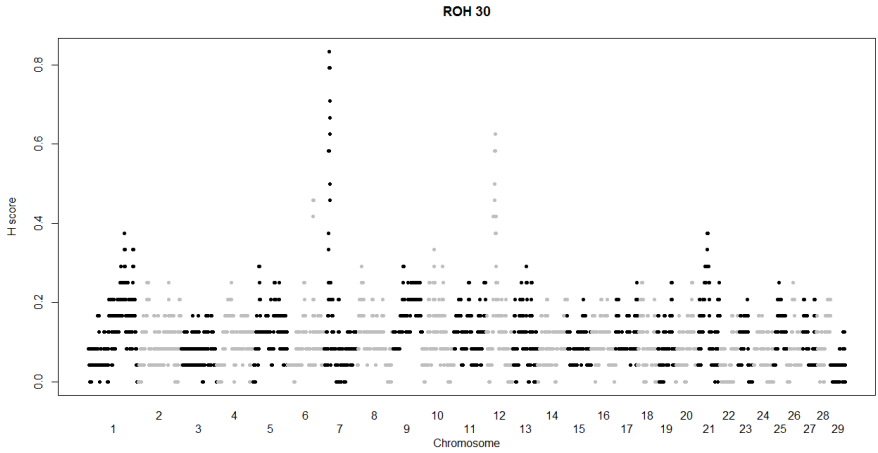


Figure S3.8 - Regions of homozygosity in the Valdostana dataset with the option winidow-snp 30. The raw values obtained were normalized according to the number of the animals used in the analysis

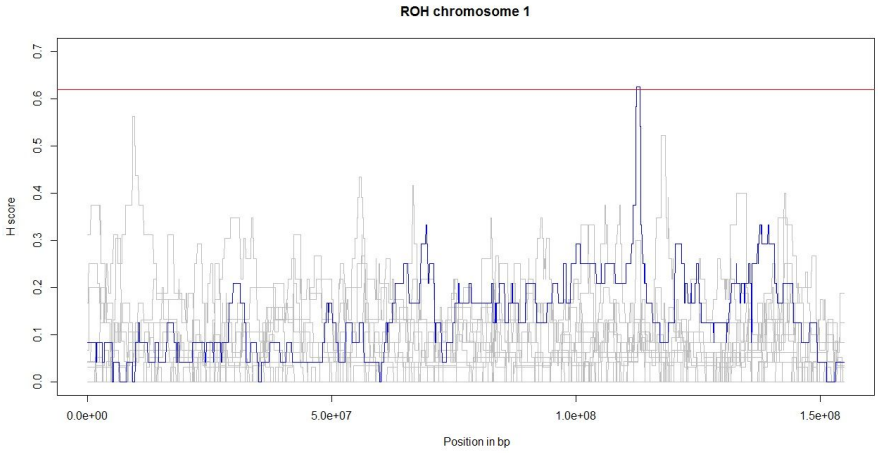


Figure S3.9 - ROH on chromosome 1. Details of runs of homozygosity on chromosome 1 (on the X axis is reported the chromosome length, on the Y axis the H score). The red line indicates H = 0.62 threshold. The Valdostana breed is represented by the blue line, while the other 14 breeds are represented by the grey lines.

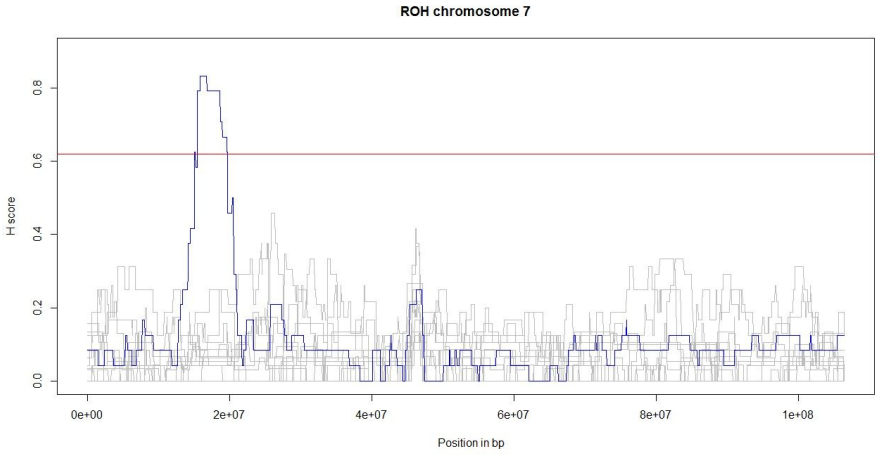


Figure S3.10 - ROH on chromosome 7. Details of runs of homozygosity on the chromosome 7 (on the X axis is reported the chromosome length, on the Y axis the H score). The red line indicates H = 0.62 threshold. The Valdostana breed is represented by the blue line, while the other 14 breeds are represented by the grey lines.

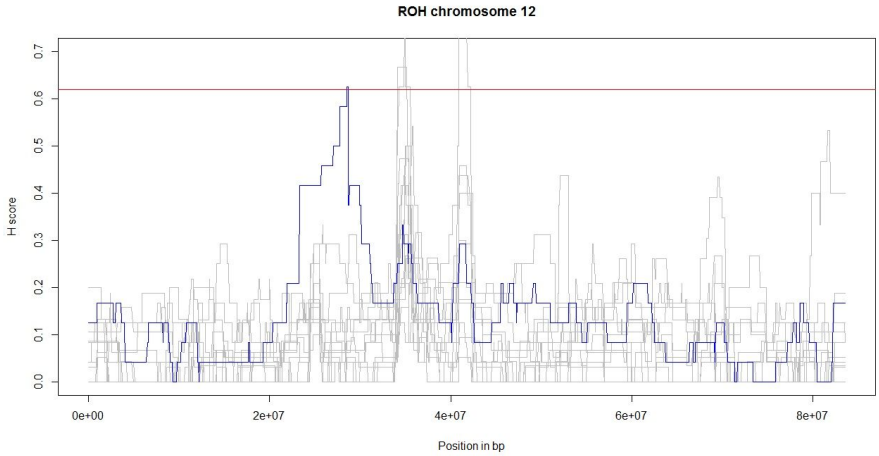


Figure S3.11 - ROH on chromosome 12. Details of runs of homozygosity on the chromosome 12 (on the X axis is reported the chromosome length, on the Y axis the H score). The red line indicates H = 0.62 threshold. The Valdostana breed is represented by the blue line, while the other 14 breeds are represented by the grey lines.

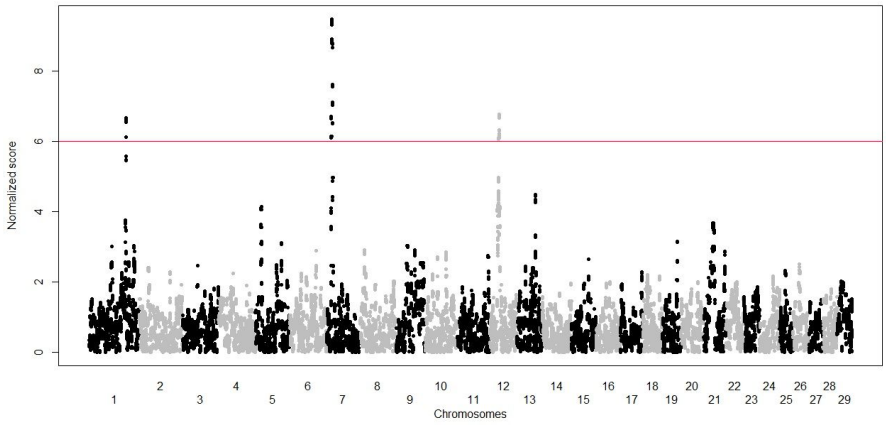


Figure S3.12 - ROH comparison between Valdostana and the other 14 breeds. Overall difference is related to the value. The higher the value, the bigger is the differences between Valdostana and other breeds. The red line indicates the value=6.

3.2 Parentage analysis

3.2.1 Aim

The knowledge of relationships between individuals is one of the most important tools in animal breeding and genetic improvement. Wrong recording of parentage can both negatively affect the planning of mating schemes, that will negatively affect the genetic improvement, the inbreeding levels and the frequency of genetic diseases in a population. In this context, the availability of cheap and reliable molecular tools for parentage assessment is extremely important.

Currently, the official tests for parentage assessment are performed by microsatellites, but the availability of dense SNP arrays is rapidly replacing them as standard. In this chapter, I developed a method to choose the most informative markers for parentage assessment and tested it on three Italian goat breeds. This method proved to be useful also in case of low quality of the genome assembly, overcoming the problems of position-based methods based on linkage disequilibrium.

The method described in this chapter was a first step in the detection of a panel for parentage assessment in the goat species. In fact, this approach was improved and tested first on a dataset of 15 Italian goat populations, and then on the ADAPTmap dataset, that count more than 4,000 animals of 130 populations from all over the world, to define a standard set of SNP to assess parentage in the caprine species. If properly developed, this findings could likely help in improving the mating scheme and genetic improvements of the goat species.

3.2.2 A method for single nucleotide polymorphism selection for parentage assessment in goats

Article published on Journal of Dairy Science (2016) 99(5):3646–3653 (DOI: 10.3168/jds.2015-10077)

Authors and Affiliations

Andrea Talenti*, Ezequiel L. Nicolazzi†, Stefania Chessa‡, Stefano Frattini*, Riccardo Moretti‡, Beatrice Coizet*, Letizia Nicoloso*, Licia Colli§, Giulio Pagnacco*, Alessandra Stella†‡, Paolo Ajmone-Marsan§, Grazyna Ptak#• and Paola Crepaldi*

*Dipartimento di Scienze Veterinarie e Sanità Pubblica, Università di Milano, via Celoria 10, Milano, 20133, Italy

†Fondazione Parco Tecnologico Padano (PTP), Via Einstein, Cascina Codazza, Lodi, 26900, Italy

‡Istituto di Biologia e Biotecnologia Agraria (IBBA-CNR), Consiglio Nazionale delle Ricerche, Via Einstein, Cascina Codazza, Lodi, 26900, Italy

§Istituto di Zootecnica, Università Cattolica del Sacro Cuore Via Emilia Parmense, Piacenza, 29122, Italy

#Dipartimento di Scienze Biomediche Comparate, Università di Teramo, Piazza Aldo Moro 45, Teramo, 64100, Italy

•Institute of Animal Production, ul. Sarego 2, Krakow 31-047, Poland

Abstract

Accurate pedigrees are essential to optimize genetic improvement and conservation of animal genetic resources. In goats, the use of mating groups and kidding management procedures hamper the identification of parentage. Small panels of single nucleotide polymorphisms (SNP) have been proposed in other species to substitute microsatellites for parentage assessment. Using data from the current GoatSNP50 chip, we developed a new 3-step procedure to identify a low-density SNP panel for highly accurate parentage assessment. Methodologies for SNP selection used in other species are less suitable in the goat because of uncertainties in the genome assembly. The procedure developed in this study is based on parent–offspring identification and on estimation of Mendelian errors, followed by canonical discriminant analysis identification and stepwise regression reduction. Starting from a reference sample of 109 Alpine goats with known pedigree relationships, we first identified a panel of 200 SNP that was further reduced to 2 final panels of 130 and 114 SNP with random coincidental match inclusion of $1.51 \times 10{-}57$ and $2.94 \times 10{-}34$, respectively. In our reference data set, all panels correctly identified all parent–offspring combinations, revealing a 40% pedigree error rate in the information provided by breeders. All reference trios were confirmed by official tests based on microsatellites. Panels were also tested on Saanen and Teramana breeds. Although the testing on a larger set of breeds in the reference population is still needed to validate these results, our findings suggest that our procedure could identify SNP panels for accurate parentage assessment in goats or in other species with unreliable marker positioning.

Introduction

Accurate assessment of relationships between individuals in a population is one of the main requirements for a successful genetic improvement program. In goats, the accurate registration of parentage is often hampered by some widespread management practices, such as the use of mating groups, summer pasturing, and collective nurseries for kids at early stages of life. These practices usually lead to high rates of pedigree registration errors, which may involve one or both parents. Consequently, genetic progress slows down.

The use of molecular markers to determine parentage has been extensively studied in livestock (Heaton et al., 2002; Werner et al., 2004; Fisher et al., 2009; Matukumalli et al., 2009; Hayes, 2011; Heaton et al., 2014). Currently, DNA-based parentage analysis is shifting from the use of microsatellite (**MS**) to SNP markers. A panel of SNP has been recently adopted by the International Committee of Animal Recording (http://www.icar.org/) and by the International Society for Animal Genetics (**ISAG**, http://www.isag.us/) for cattle parentage testing (http://www.isag.us/Docs/Cattle-SNP-ISAG-core-additional-panel-2013.xlsx).

Other comparison ring tests for parentage with a core panel of 100 SNP plus an additional panel of 100 SNP have recently been undertaken in cattle (Strucken et al., 2014). A panel of 163 SNP has also been proposed for parentage testing of sheep (Heaton et al., 2014).

The use of molecular markers for parentage analysis has been accompanied by the development of several statistical techniques for data management, principally based on exclusion, categorical, or fractional allocation and full probability (Jones et al., 2010).

A goat 53K SNP array has been recently developed by Illumina (San Diego, CA), in collaboration with the International Goat Genome Consortium (http://www. goatgenome.org/; Tosser-Klopp et al., 2014). In spite of the high potential of such a tool in goat breeding, a few technical issues still partially hamper its full exploitation. One of the main issues concerns the early stage of the reference goat genome assembly (Dong et al., 2013). One of the main parameters used to select SNP for parentage assessment (**PA**) in other species (e.g., cattle) is physical distance and linkage disequilibrium (**LD**) among markers (Strucken et al., 2014). Although recent studies have characterized overall LD in some goat breeds (Brito et al., 2015), inaccurate SNP positioning on the goat genome makes traditional methods based on physical distance less effective for PA purposes (Benjelloun et al., 2015; Bickhart et al., 2015).

This paper evaluates a new 3-step procedure to select a small number of SNP for PA in goat. Advantages of this procedure are particularly relevant for species without an accurate SNP positioning, such as the goat. Our selection procedure of informative SNP markers is independent from SNP positioning, and it is based on parent–offspring identification by assessment of Mendelian errors (**MDE**), canonical discriminant analysis, and backward stepwise regression.

At present, the official MS-based method for parentage analysis in goats is generally too expensive for this species, considering the limited economic value of a single animal (Strucken et al., 2014). As already observed in sheep, the introduction of low-density SNP panels coupled with a cost-effective DNA-based technique would allow a 40% reduction of the analysis cost compared with MS (Heaton et al., 2014). This decrease in price would likely increase the use of DNA-based parentage determination in goats and help overcome the major constraints in control of inbreeding and implementing genetic improvement strategies.

Material and Methods

Animal Sampling and Genotyping

Blood samples were collected from 154 animals be- longing to the Alpine (n = 109) and the Saanen (n = 22) goat breeds, reared in 3 flocks in Northern Italy, and the Teramana goat breed (n = 23), reared in one flock in central Italy. Samples were collected according to the recommendations of the European Council (Council of Europe, 1986) on animal care. The DNA was extracted from whole blood using a commercial kit (NucleoSpin Blood, Macherey-Nagel, Düren, Germany) following the manufacturer's instructions. The DNA samples were genotyped using the Illumina GoatSNP50 BeadChip (Tosser-Klopp et al., 2014).

All Alpine goats (8 males and 101 females) were used as the reference data set (**REF**), whereas 22 Saanen females (**VAL1**) and 23 (2 males, 21 females) Teramana individuals (**VAL2**) were used as the validation data set.

According to pedigree data, 46 out of 109 individuals in REF formed 20 trios (father, mother, and offspring), and in total 58 animals belonged to 50 parent–offspring (**PO**) pairs.

Group VAL1 included 5 PO pairs. Group VAL2 included one trio and 5 PO pairs. Group VAL2 animals were from the only existing flock of Teramana, an endangered breed consisting of about 80 animals recorded by the National Breeder Association (http:// www.assonapa.com/) that was chosen specifically for its small size and difficult PA. The latter is due to high inbreeding, which increases the number of cryptic relationships, leading to PO classification instead of full-sib.

To confirm our results, 53 animals belonging to all trios in the REF data set were also analyzed with the official MS parentage test (13 different markers: *HSC*, *ILSTS19*, *INRA005*, *INRA063*, *MAF65*, *SRCRSP5*, *SRCRSP8*, *SRCRSP24*, *ILSTS23*, *INRA023*, *MCM527*, *CSRD247*, *SRCRSP23*). This test was performed in outsourcing at the official Laboratory of the Italian Breeders Association (AIA-LGS, http://www.lgscr.it/ it/chi.htm).

Data Set Preparation

The REF genotype data were quality checked according to the following thresholds: SNP call rate ≥0.95; minor allele frequency (**MAF**) ≥0.01; individual genotype call rate ≥0.90; and in Hardy-Weinberg equilibrium (Bonferroni corrected threshold, $P \leq 1 \times 10^{-7}$). The resulting data set was used in the initial genetic population analysis and for PA. A characterization of the animals included in the REF data set was performed by PLINK 1.07 Multidimensional Scaling (Purcell et al., 2007) to verify the absence of sampling errors. For the selection of a low-density SNP panel, all markers with MAF ≤0.3, unknown chromosomal assignment or placement on the sexual chromosomes were excluded.

SNP Selection and PA

The 3-step procedure to select candidate SNP for the PA panel consisted of (1) the identification of real PO by MDE, (2) the identification of informative SNP by canonical discriminant analysis, and (3) reduction of the number of markers by MDE and stepwise regression.

Mendelian error was calculated using an in-house script. Mendelian error is a simple comparison among the genotypes of 2 animals, checking for incompatible homozygosity. A Mendelian error occurs when opposing

homozygotes are present at the same locus (e.g., AA for the sire and BB for the offspring). Although MDE for 2 PO pairs should be 0 on a 50K panel, a low number of MDE are allowed to account for genotyping errors (Hayes et al., 2009a). Pairwise comparisons were performed among all 109 REF animals, for a total of 5,886 comparisons. Each pair of individuals was classified as PO or not PO (**NPO**) based on MDE. Considering the SNP that remained after quality control check (MAF ≥1%, SNP call rate ≥95%, individual call rate ≥90%, and not in Hardy-Weinberg equilibrium), we classified as PO all pairs with <1,000 MDE (Hayes et al., 2009a).

A canonical discriminant analysis (**CANDISC**), included in the R package MASS (Venables and Ripley, 2002), was then performed on the REF data set. A CANDISC was applied as a dimensionality reduction method that uses a linear combination of variables to separate 2 or more known groups. To characterize each group, the method gives a linear score to each variable (i.e., in our case markers) to best assign each observation to its group (Fisher, 1936). We performed CANDISC one chromosome at a time to keep the number of markers lower than the number of pairwise comparisons. A CANDISC was performed using pairwise individuals' genotype comparison as predictive variables and PO-NPO classification as discriminating groups. The SNP with extreme discriminant linear scores (mean ± 2 SD) were retained, whereas all redundant markers ($|r| > 0.7$ in pairwise comparisons) were discarded. It is worth noting that, although no direct LD measure was used in this work, the exclusion of redundant markers indirectly accounts for local LD among markers. We then reduced the panel size to 200 SNP, retaining only the markers that showed the highest number of MDE in NPO group and lowest in the PO group. This size was chosen ac- cording to the ISAG panel for PA in cattle (http:// www.isag.us/Docs/Cattle-SNP-ISAG-core-additional- panel-2013.xlsx) and is an intermediate step before final reduction. Finally, we further reduced the panel size using backward stepwise regression, excluding one marker at a time by using MDE as response vector, until we found the lowest number

of variables able to identify groups with 100% specificity. According to ISAG standards, parentage assignment was performed by MDE estimation on the reduced SNP panel. For each identified panel, we also calculated the following performance parameters: sensitivity, specificity, accuracy, median estimates for probability of a coincidental match between 2 animals (Heaton et al., 2014), and the fraction of potential adults excluded from parentage test (Heaton et al., 2014). Sensitivity is the ratio of true positives identified by the test on the total number of true parents; specificity is the ratio of nonparents excluded by the test to the total number of nonparents; and accuracy is the ratio of correct assignment of the test to the total number of tests performed. These parameters can range from 0 to 1 for the worst and best scenario, respectively. The probability of a random coincidental match at an SNP locus between random animals is $\boldsymbol{P_i}$. The P_i for locus A with SNP alleles A1 and A2 was the sum of the squares of the 3 genotype frequencies. Values were calculated with the following formula: $P_i = (f_{A1A1})^2 + (f_{A1A2})^2 + (f_{A2A2})^2$, where f_{A1A1}, f_{A1A2}, and f_{A2A2} are the relative genotypic frequencies of A_1A_1, A_1A_2, and A_2A_2, respectively. The combined Pi for multiple SNP markers is the product of the P_i values of individual markers. The probability of observing opposing homozygotes at a SNP locus between a random eligible adult and a random offspring is $\boldsymbol{P_e}$. Values for biallelic loci and with information on only one parent were calculated using the formula $P_e = 2 \times (f_{A1A1} \times f_{A2A2})$, where f_{A1A1} and f_{A2A2} are the relative genotype frequencies of A_1A_1 and A_2A_2, respectively. Combined P_e for multiple SNP was calculated as follows: $P_{e(\text{SNPn})} = R_o \times P_{e(\text{SNP}_1)} + R_1 \times P_{e(\text{SNP}_2)} + R_2 \times P_{e(\text{SNP}_3)} \ldots + R_{n-1} \times P_{e(\text{SNPn})}$, where $P_{e(\text{SNP}_1)}$ was the fraction of eligible adults excluded by the first SNP and R was the remaining fraction of unexcluded adults [$R_o = 1$, $R_1 = R_{n-1} - (R_{n-1} \times P_e) = R_o - (R_o \times P_e)$, $R_2 = R_1 - (R_1 \times P_e)$, and so on]. To improve data clearness, we report the fraction of potential adults excluded (P_e) per single SNP as a fraction of the potential adults retained ($\boldsymbol{PR}$; $PR = 1 - P_e$) from the parentage test.

Results and Discussion

After the preliminary editing, a total of 49,609 SNP were retained in the REF data set.

A multidimensional scaling plot was produced to verify the absence of population substructure. The analysis did not reveal any strong population structure, except for a mild sire effect (Figure 3.13).

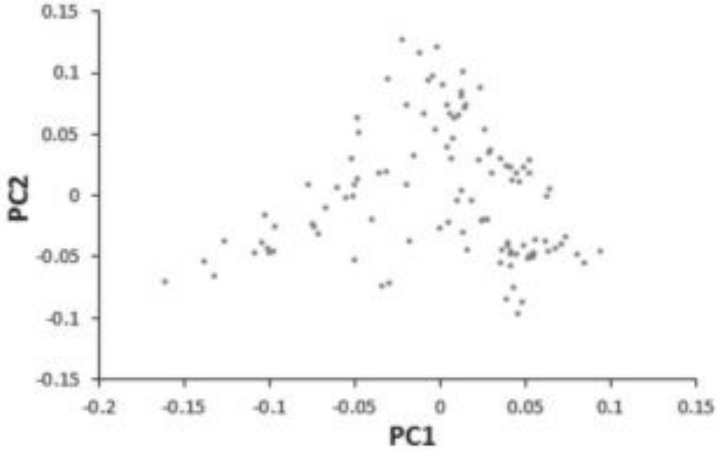


Figure 3.13 - Multidimensional scaling (MDS) plot of the reference (REF) panel, showing absence of relevant substructures among animals, PC = principal component. Modified from Talenti et al. (2016)

Mendelian errors confirmed 10 pedigree-based trios out of 20, resulting in a 50% trio pedigree error rate (Table 3.4).

	Reference	Validation 1	Validation 2

	REF			VAL1			VAL2		
	Pedigree	Confirmed	New	Pedigree	Confirmed	New	Pedigree	Confirmed	New
Trios	20	10	5	0	0	0	1	0	1
Pairs	50	30	20	5	3	2	7	5	10

Table 3.4 - Pedigree information and SNP identification for the reference and both validation data sets. For each data set, we reported the number of pedigree-based trios and pairs, the number of trios and pairs confirmed by SNP, and the number of new trios and pairs identified by SNP; modified from Talenti et al. (2016)

In addition, 5 new undeclared trios were identified, increasing the total number of related animals. All trios identified by the above analyses were confirmed by the official MS parentage test. Interestingly, MDE failed to identify PO pair of individuals identified by MS. This pair of individuals (a buck and its putative offspring) did not pass the MDE threshold with the 50K SNP panel, yielding 1,479 MDE. Mendelian error distributions for pairwise comparisons are reported on Supplemental Figure S3.17 (http://dx.doi.org/10.3168/jds.2015-10077). In conclusion, our analysis confirmed 30 out of 50 declared pedigree PO (error rate = 40%). However, PA identified 20 more pairs of PO, for a total of 50 PO pairs.

Parentage Analysis with 200 and 130 SNP Panels

After PA, a second and more stringent editing was applied on the REF data set (MAF >0.3, no markers with unknown position or on heterochromosomes). A total of 27,523 SNP were retained for further analyses.

As described above, only an indirect marker exclusion by LD was applied, because of the uncertain positioning of the SNP loci on the reference genome map that could heavily affect the results. In fact, direct LD estimation on 36 not closely related individuals (MDE ≥2,020) showed a strong smear, composed of markers with very high level of r^2 at high

distances (1–10 Mb, Figure 2a and 2b) and SNP very close with low r2 values (Figure 3.14).

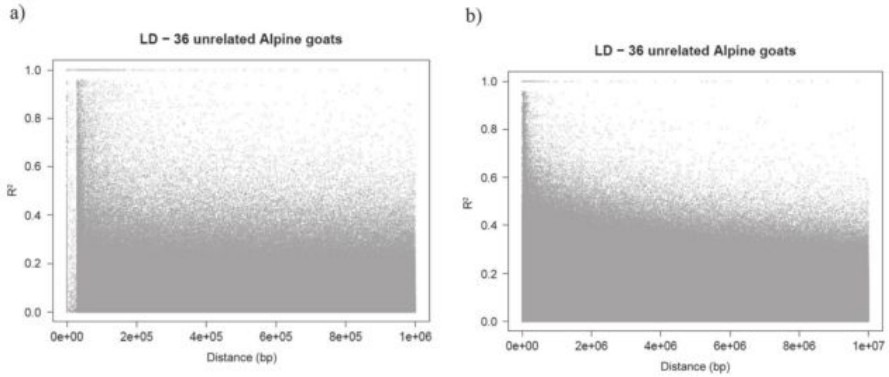


Figure 3.14 - Linkage disequilibrium (LD) plot for our data, estimated using PLINK 1.07 software (Purcell S., Boston, MA). On the x-axis, we report the distance in base pairs, and on the y-axis we report the LD level estimated as r2. The plot shows smearing at (a) distances lower than 1 Mb and (b) distances lower than 10 Mb. In addition, (a) markers shows anomalous r2 values at very close distances, as shown; modified from Talenti et al. (2016)

A CANDISC reduction identified 1,206 highly poly- morphic SNP markers, which were further thinned to 1,133 after the exclusion of collinear SNP.

The 200 SNP panel was obtained by identifying the markers with the highest MDE in the NPO group and the lowest MDE in the PO group (Supplemental Table S3.6; http://dx.doi.org/10.3168/jds.2015-10077). We chose this panel size not because it showed the best performance, but because this is the number of markers in the ISAG standard panel for PA in cattle. The thinning procedure led to an uneven distribution of the selected markers on the genome, independent from differences in marker density and length of chromosomes (Figure 3.15a).

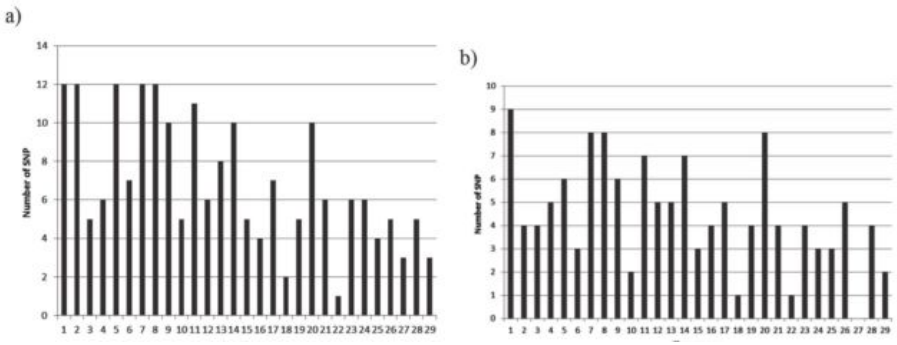


Figure 3.15 - Chromosomal distribution of (a) 200 and (b) 130 identified SNP. (a) All chromosomes have at least one marker in the 200 SNP panel. (b) All chromosomes but one (27) have at least one marker in the 130 SNP panel; modified from Talenti et al. (2016)

We checked the assessment power of the 200 SNP data set according to ISAG standards (http://www. isag.us/docs/guideline-for-cattle-snp-use-for-parentage-2012.pdf). Parentage was assigned as PO in the case of 0 to 1 MDE, unassigned and doubtful in the case of 2 to 3 MDE, and unassigned in the case of more than 3 MDE. The 200 SNP panel tested on the REF population had 100% sensitivity, specificity, accuracy; low PR; and low Pi, as reported in Table 3.5.

	Reference			Validation 1			Validation 2			Composed dataset		
	200 SNP	130 SNP	114 SNP	200 SNP	130 SNP	114 SNP	200 SNP	130 SNP	114 SNP	200 SNP	130 SNP	114 SNP
Sensitivity	1	1	1	0.8	1	1	0.9333	0.9333	0.9333	0.9714	0.9857	0.9714
Specificity	1	1	1	1	1	1	1	1	1	1	1	1
Accuracy	1	1	1	0.9957	1	1	0.9961	0.9961	0.9961	0.9998	0.9999	0.9998
PR	0	5.67E-10	1.05E-08	7.22E-09	4.07E-06	7.31E-06	5.37E-07	9.48E-05	1.39E-04	0	5.72E-10	3.96E-09
mean SNP Pe	0.1505	0.1509	0.1488	0.0873	0.0888	0.0964	0.0677	0.0669	0.0726	0.1492	0.1507	0.1558
SD Pe per SNP	0.0133	0.0142	0.0118	0.0616	0.0629	0.062	0.0592	0.0587	0.0663	0.0205	0.0227	0.022

Table 3.5 - Panels performance on Alpine, Saanen, Teramana and all three breeds together. For each dataset, we estimated values of sensitivity, specificity, accuracy, fraction of included alleged parent (PR), probability of a random coincidental match (Pi) and fraction of excluded alleged parents (Pe) mean per SNP ± Standard Deviation; modified from Talenti et al. (2016)

Mean of single SNP Pe ± standard deviation (SD) was 0.1505 ± 0.0133 (Pe values per SNP are reported in Supplemental Table S3.7; http://dx.doi.org/10.3168/ jds.2015-10077). As expected, the exclusion power of a single SNP is far lower than for a single MS (up to 0.5; Bolormaa et al., 2008), because of the higher variability of the second compared with the first (SNP are only biallelic, whereas MS are multiallelic). On average, 2 or 3 SNP are needed per MS marker to obtain equivalent cumulative exclusion power (Al-Atiyat, 2015).

The panel was further reduced to 130 SNP by applying a backward stepwise regression on the previously identified 200 SNP on the REF data set (Supplemental Table S3.6; http://dx.doi.org/10.3168/jds.2015-10077). The distribution of the 130 markers on chromosomes is reported in Figure 3.15b. The reduction process led to an unequal distribution of markers on each chromosome, and all but one (chromosome 27) were represented in the panel.

The 130 SNP panel performed similarly to the 200 SNP panel with REF, having 100% sensitivity, specificity, and accuracy but slightly larger PR and P_i (as shown in Table 3.5). Mean of single SNP P_e ± SD was 0.1509 ± 0.0142, slightly larger than in the 200 SNP panel (P_e values per SNP are reported in Supplemental Table S3.7; http://dx.doi.org/10.3168/jds.2015-10077). The reduction in P_e and increase in P_i are due to the lower number of markers, which reduces the combined probability of exclusion and increases the probability of inclusion. Also, the increase in mean P_e per SNP was expected, because of the selection of more discriminant markers. For 49,609 SNP panels, mean of single SNP P_e ± SD was 0.0886 ± 0.0437. These values are far lower than values for the low-density panels, supporting the efficacy of this markers selection method.

Validation of the 200 and 130 SNP Panels

Evaluating the power of the SNP panel for PA on the same data set used for SNP selection would likely lead to an overestimate of the performance. Thus, the 2 reduced panels were tested for PA in 2 validation data sets of different breeds, Saanen and Teramana (VAL1 and VAL2, respectively). Because MS and SNP-based PA were in close agreement on the REF population, VAL1 and VAL2 were analyzed using SNP data only.

Using the full panel in VAL1, 3 pairs of related animals were confirmed and 2 undeclared pairs were identified. In total, 5 pairs of related animals with low MDE were detected (Table 3.4). In VAL2, only one declared trio was excluded by MDE. In contrast, a total of 15 PO undeclared relationships were found. The 200 SNP panel showed high performance on VAL1, with no false positives and high values for specificity and accuracy. Only the sensitivity value was lower due to an undeclared pair of animals sharing 799 MDE. This pair of animals had borderline values for our less stringent MDE threshold of 1,000, thus we defined it as a PO pair. However, it is interesting to observe that the MDE threshold used in this study is to be considered too lenient for real case scenarios. In fact, using a more realistic threshold of 0.5% (~250 MDE), which is still high, we would have classified this couple as NPO, and sensitivity and accuracy of the panel would have risen to 100%. Therefore, the results presented here can be considered as an underestimate of the real potential of this method. Information given by pedigree registration indicated that these 2 animals were full-sibs. This kind of relationship is very difficult to diagnose by molecular markers, and prior information is useful to obtain a more accurate assignment. In fact, when pedigree reliability is low, full-sibs remain difficult to distinguish from PO. Both PR and P_i were low as shown in Table 3.5. The average $P_e \pm$ SD per SNP was 0.0873 ± 0.0616 (Pe values per SNP are reported in Supplemental Table S3.7; http://dx.doi.org/10.3168/jds.2015-10077). The same panel on VAL2 showed higher performances, with high values of specificity and accuracy and low values of PR and P_i as shown in Table 3.5. Only sensitivity was lower

due to a false negative with MDE 920. As for VAL1, this pair was classified as PO because of the high MDE threshold. Again, the adoption of the more realistic threshold of 250 MDE would have classified it as NPO. Also, in this case, sensitivity and accuracy would have increased to 100%. Average P_e ± SD per SNP was 0.0677 ± 0.0592 (Pe values per SNP are reported in Supplemental Table S3.7; http://dx.doi. org/10.3168/jds.2015-10077).

The 130 SNP panel performed differently in VAL1, still showing no false positives and 100% assignment, when no prior information was considered. In any case, the panel showed a decreased PR and increased P_i when compared with the 200 SNP panel, as shown in Table 3.5. Mean P_e per SNP was 0.0888 ± 0.0629 (P_e values per SNP are reported in Supplemental Table S3.7; http:// dx.doi.org/10.3168/jds.2015-10077).

In VAL2, the 130 SNP panel performed slightly poorer compared with the 200 SNP panel. Sensitivity, specificity, and accuracy remained the same, still showing no false positives. However, also in this case, performance rose when a more realistic threshold to identify PO was used, showing perfect sensitivity, specificity, and accuracy. In both cases, also when considering the high MDE threshold, PR and P_i increased, as shown in Table 3.5. Mean single SNP Pe ± SD was 0.0669 ± 0.0587 (P_e values per SNP are reported in Supplemental Table S3.7; http://dx.doi.org/10.3168/jds.2015-10077). Even though SNP showed a similar mean value on VAL1 and VAL2, the difference in their performance may be due to a different exclusion power of each marker. Each marker has different Pe values in the 2 data sets, with some SNP having no exclusion power in 1 of the 2 breeds. High variability of Pe values among breeds is also due to different allelic frequencies of each SNP among populations, which could negatively affect efficacy of the test. This further underlines the need to include more breeds in the discovery phase, to assess the performance of the panel on a larger set of breeds.

Final Reduction to 114 SNP Panel

A third reduced panel was constructed performing stepwise regression on the 200 SNP panel that included animals from REF, VAL1, and VAL2 data sets, for a total of 11,781 pairwise comparisons among individuals. The reduction process led to the identification of a 114 SNP panel (chromosomal distribution on Figure 3.16).

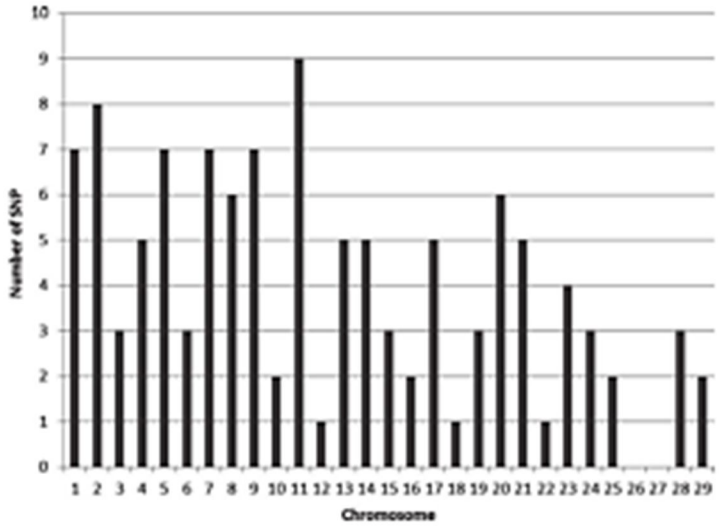


Figure 3.16 - Chromosomal distribution of 114 identified SNP. All chromosomes but 2 (26 and 27) have at least one marker in the panel; modified from Talenti et al. (2016)

These markers were distributed on all but 2 chromosomes (chromosomes 26 and 27) and are reported in Supplemental Table S3.6 (http://dx.doi.org/10.3168/ jds.2015-10077). This panel showed high sensitivity, specificity, and accuracy and low PR and P_i, as shown in Table 3.5. As in the previous panels, a more realistic MDE threshold to identify PO resulted in a panel performance matching all true PO pairs and excluding all NPO pairs. Mean single SNP $P_e \pm$ SD was 0.1558 ± 0.0220 (P_e values per SNP are reported in Supplemental Table S3.7;

http://dx.doi.org/10.3168/jds.2015-10077). Overall performance of this panel was comparable to the 2 larger panels.

Random Selection of Markers

To evaluate the efficacy of our 3-step method, 100 random data sets for each of the panel subsets (200, 130, and 114 SNP) were randomly extracted from 27,523 highly polymorphic SNP in Alpine breeds. Each random panel was tested on a data set including within- and across-breeds pairwise comparison, for a total of 11,781 pairs of animals. The 100 randomly chosen panels of comparable size for the 200 SNP panel were equivalent to our method in 66% of cases. However, when reducing the number of SNP to 130 and 114, the random choice of markers showed comparable results only on 2 and 0% of the cases, respectively (Supplemental Table S3.8; http://dx.doi.org/10.3168/jds.2015-10077).

Moreover, we also evaluated our random panels on 10,338 highly polymorphic SNP in all 3 breeds (MAF >0.3 in each breed). Again, the 100 randomly chosen panels of comparable size for the 200 SNP panel were comparable to our method in 81% of cases. However, when reducing the number of SNP to 130 and 114, the random choice of markers showed comparable results only in 9 and 0% of the cases, respectively (Supplemental Table S3.8; http://dx.doi.org/10.3168/jds.2015- 10077).

These results, if confirmed when using a wider set of breeds in the reference population, suggest that the 200 SNP panel chosen to match the bovine ISAG SNP panel size is not the best choice in terms of PA accuracy on goats. In addition, our results would confirm that our method is able to efficiently identify small SNP panels for PA and that MAF-based selection alone is ineffective for identification of useful PA panels.

Conclusion

These results suggest that the proposed method can successfully identify a reduced number of markers able to assess parentage in species with low map resolution. The identification of small panels can easily help in assessing direct relationships among animals, enhancing the control of inbreeding and the effectiveness of genetic improvement programs. In spite of the promising results obtained, it is necessary to assess the proposed procedure independently in a wider set of breeds, to identify the smallest panel for PA in different goat breeds. This study is a first step toward the implementation of new technologies for PA in species with low genomic information, such as the goat. Such species, which are often economically relevant in marginal, rural areas of the world, would greatly benefit from the availability of low-cost tools for breeding management and genetic improvement.

Acknowledgement

This research received funding by project "Innovagen" funded by the Italian Ministry of Agriculture (MiPAAF, Rome, Italy). Talenti and Frattini were sup- ported by the Doctoral School on Veterinary and Ani- mal Sciences of the Veterinary Faculty of the University of Milan (Italy). Coizet and Nicoloso were supported by GenHome Project (Lodi, Italy).

Supplementary material

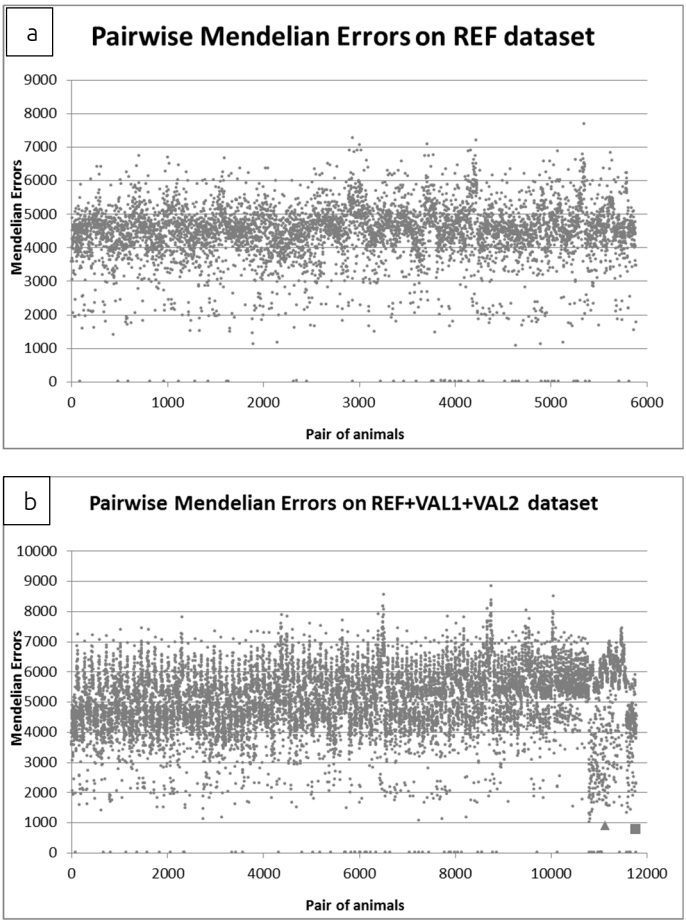


Figure S3.17 - The scatterplots represent mendelian errors values for: a) pairwise comparison among animals in REF dataset; b) pairwise comparison among animals in REF+VAL1+VAL2 dataset. Square and Triangular points in graph 2 are couple of animals (belonging to Saanen and Teramana breeds respectively) with more than 100 and less than 1,000 ME. Modified from Talenti et al. (2017).

Chromosome	rs ID	Position	130 SNPs	114 SNPs	MAF Alpine	MAF Saanen	MAF Teramana
1	rs268246894	22594761	Yes	No	0.4266	0.4545	0.2174
1	rs268277578	45318610	Yes	Yes	0.3670	0.3864	0.1304
1	rs268237133	52071174	Yes	Yes	0.4174	0.4545	0.1522
1	rs268272505	77101471	Yes	No	0.4037	0.3636	0.4783
1	rs268279562	82433467	Yes	No	0.3440	0.2045	0.0435
1	rs268280267	86943152	No	Yes	0.4862	0.3636	0.0435
1	rs268280291	87940819	No	No	0.4817	0.5000	0.4565
1	rs268258713	90873881	No	No	0.4587	0.3182	0.4783
1	rs268278510	105155775	Yes	Yes	0.4907	0.1818	0.3696
1	rs268257920	110168450	Yes	Yes	0.4862	0.1818	0.4348
1	rs268233519	113146255	Yes	Yes	0.3211	0.5000	0.3696
1	rs268286780	134540117	Yes	Yes	0.3991	0.2955	0.4565
2	rs268239713	17031741	No	No	0.4954	0.2955	0.2609
2	rs268261967	23607634	No	Yes	0.3303	0.4773	0.4783
2	rs268253020	36052714	No	Yes	0.4766	0.4318	0.3261
2	rs268267960	42902914	No	No	0.3761	0.4773	0.2826
2	rs268276840	46116901	Yes	Yes	0.4626	0.3636	0.2174
2	rs268239885	53472858	No	Yes	0.3714	0.1818	0.4348
2	rs268239859	54554491	No	Yes	0.3095	0.4091	0.4565
2	rs268261308	62556317	No	No	0.4862	0.2500	0.2826
2	rs268282666	73666511	No	Yes	0.4771	0.3182	0.2391
2	rs268285441	81115463	Yes	Yes	0.3578	0.3636	0.0217
2	rs268292755	124506565	Yes	Yes	0.3761	0.3864	0.1739
2	rs268279477	125308567	Yes	No	0.3945	0.3182	0.1739
3	rs268248233	42135080	Yes	Yes	0.3670	0.4091	0.0217
3	rs268248191	43955403	No	No	0.4817	0.4773	0.1957
3	rs268261853	49886261	Yes	No	0.3991	0.1364	0.0435
3	rs268251430	55616220	Yes	Yes	0.4954	0.1818	0.2609
3	rs268275850	92297040	Yes	Yes	0.4862	0.3182	0.0217
4	rs268278238	1327116	No	Yes	0.4676	0.4318	0.0652
4	rs268287539	23292639	Yes	Yes	0.4120	0.1818	0.0000
4	rs268274140	28916317	Yes	Yes	0.4266	0.4773	0.0870
4	rs268263470	50257096	Yes	Yes	0.4908	0.2500	0.0217
4	rs268263449	51071852	Yes	Yes	0.4725	0.4091	0.1739
4	rs268263359	55167467	Yes	No	0.4862	0.4773	0.2826
5	rs268266016	1470207	Yes	No	0.3761	0.3182	0.0870
5	rs268250051	25734084	No	Yes	0.4404	0.2273	0.1087

Table S3.6 - The table contains all information about selected SNP. The first four columns contains information about: chromosome, rs ID and physical position. Panel columns indicates if the SNP is retained in reduced panel (of 130 or 114 SNPs). Remaining three columns contain information about the minor allele frequency for each breed; modified from Talenti et al. (2016)

Chromosome	rs ID	Position	130 SNPs	114 SNPs	MAF Alpine	MAF Saanen	MAF Teramana
5	rs268275159	32063215	Yes	Yes	0.4174	0.4318	0.3696
5	rs268275167	32370118	No	No	0.3716	0.2727	0.4565
5	rs268238520	41176602	No	Yes	0.4633	0.4773	0.4348
5	rs268265695	42491486	Yes	Yes	0.4083	0.2273	0.0870
5	rs268281587	50028383	Yes	Yes	0.4174	0.1818	0.3043
5	rs268265613	52262827	Yes	Yes	0.4954	0.2045	0.4130
5	rs268264560	74401011	Yes	No	0.3364	0.2619	0.1087
5	rs268235408	77288117	No	No	0.3945	0.4318	0.1087
5	rs268280631	101356476	No	Yes	0.3624	0.4773	0.3261
5	rs268280688	104117349	No	No	0.4862	0.4545	0.2391
6	rs268263951	655359	No	No	0.3945	0.5000	0.4565
6	rs268263057	12264321	Yes	Yes	0.3211	0.4773	0.0870
6	rs268245369	31754679	Yes	No	0.3991	0.3636	0.4565
6	rs268273538	53264333	No	No	0.3578	0.3182	0.5000
6	rs268289495	63726709	No	No	0.4817	0.2727	0.4783
6	rs268260251	77174216	No	Yes	0.4398	0.4091	0.0652
6	rs268242736	93909892	Yes	Yes	0.4450	0.4773	0.3696
7	rs268257860	4192615	Yes	No	0.4817	0.2955	0.3261
7	rs268248284	8617316	Yes	Yes	0.4954	0.4318	0.2609
7	rs268247911	39668489	Yes	No	0.4541	0.4091	0.4348
7	rs268262813	42087284	Yes	Yes	0.4174	0.4091	0.4565
7	rs268262879	44894653	No	Yes	0.4633	0.4773	0.0000
7	rs268262913	46539792	Yes	No	0.3945	0.2045	0.3478
7	rs268242911	48380166	Yes	Yes	0.5000	0.3636	0.2174
7	rs268238690	83379519	Yes	Yes	0.4575	0.2955	0.0435
7	rs268252538	92013304	No	No	0.4862	0.4545	0.1739
7	rs268237682	95788706	No	Yes	0.4907	0.2273	0.2826
7	rs268279527	102250706	No	Yes	0.3394	0.4091	0.2826
7	rs268254013	103700389	Yes	No	0.4725	0.4545	0.4783
8	rs268247598	10088172	No	Yes	0.3810	0.1905	0.2174
8	rs268276094	16754774	Yes	Yes	0.4633	0.3636	0.4348
8	rs268244448	22324402	No	No	0.4861	0.2500	0.3478
8	rs268277168	26298640	No	No	0.4771	0.3409	0.1739
8	rs268236867	30307773	Yes	Yes	0.4404	0.2045	0.5000
8	rs268278187	33997194	Yes	Yes	0.4725	0.4545	0.1739
8	rs268292741	46738116	Yes	Yes	0.4679	0.4773	0.4783
8	rs268245008	55874397	No	No	0.4541	0.3182	0.3913
8	rs268286103	62828040	Yes	No	0.3073	0.4545	0.0435
8	rs268266882	89489430	Yes	No	0.3670	0.1364	0.4565
8	rs268266897	89981503	Yes	No	0.4220	0.4773	0.1304
8	rs268266898	90013365	Yes	Yes	0.4679	0.4773	0.1739
9	rs268264165	12991411	Yes	Yes	0.3654	0.4773	0.2609

Table S3.6 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	MAF Alpine	MAF Saanen	MAF Teramana
9	rs268236937	23828851	Yes	Yes	0.5000	0.2955	0.4348
9	rs268272814	26916830	Yes	Yes	0.4174	0.4318	0.3043
9	rs268275918	35120000	No	Yes	0.4541	0.2955	0.3261
9	rs268260994	44816684	Yes	No	0.4817	0.3636	0.4565
9	rs268233440	45493989	No	Yes	0.4908	0.4091	0.1739
9	rs268266688	54414444	Yes	Yes	0.4725	0.3409	0.3696
9	rs268275277	64658280	No	No	0.4266	0.3095	0.3696
9	rs268282225	66461656	Yes	Yes	0.4862	0.4773	0.0217
9	rs268251679	67146104	No	No	0.4813	0.4545	0.5000
10	rs268261737	22632227	No	No	0.4908	0.3182	0.1739
10	rs268257058	42941965	Yes	Yes	0.4404	0.3636	0.1087
10	rs268288791	43634448	No	No	0.4771	0.2045	0.2609
10	rs268240865	69364926	Yes	No	0.3119	0.1818	0.4348
10	rs268270243	89232772	No	Yes	0.4312	0.4545	0.4783
11	rs268272993	700718	No	No	0.4587	0.2955	0.3261
11	rs268291501	4600369	Yes	Yes	0.4815	0.3182	0.3043
11	rs268292075	20664777	Yes	Yes	0.4037	0.2727	0.4565
11	rs268250711	60564735	No	No	0.4495	0.3409	0.4130
11	rs268250673	62446881	Yes	Yes	0.4954	0.1818	0.3696
11	rs268250509	69480159	Yes	Yes	0.4266	0.2727	0.3043
11	rs268253243	75527471	Yes	Yes	0.4583	0.3636	0.4130
11	rs268247515	79589304	Yes	Yes	0.4238	0.3636	0.3333
11	rs268243784	83719137	Yes	Yes	0.4679	0.2955	0.1957
11	rs268271367	85418308	No	Yes	0.4266	0.4545	0.3696
11	rs268284989	97787718	No	Yes	0.4037	0.3636	0.2391
12	rs268278336	3757154	Yes	No	0.3716	0.2500	0.0000
12	rs268256530	17064107	Yes	No	0.4450	0.2045	0.3478
12	rs268262656	34769122	Yes	No	0.3716	0.4545	0.2826
12	rs268268097	49872578	Yes	Yes	0.4771	0.4545	0.0435
12	rs268251731	58797148	Yes	No	0.4633	0.3864	0.4783
12	rs268274667	64087392	No	No	0.4450	0.2727	0.1304
13	rs268280844	26389054	No	No	0.4220	0.2955	0.2391
13	rs268280831	27021612	Yes	No	0.4358	0.3409	0.3043
13	rs268280826	27284949	No	Yes	0.4587	0.2045	0.5000
13	rs268280760	30261547	Yes	Yes	0.4450	0.3409	0.2609
13	rs268290264	43206738	Yes	Yes	0.4174	0.3182	0.4130
13	rs268245797	47426348	No	No	0.4037	0.2955	0.4783
13	rs268236129	52485704	Yes	Yes	0.4495	0.2500	0.4565
13	rs268291580	61123452	Yes	Yes	0.4312	0.0227	0.4348

Table S3.6 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	MAF Alpine	MAF Saanen	MAF Teramana
14	rs268279625	917480	Yes	Yes	0.4862	0.4545	0.2826
14	rs268276667	4121438	Yes	No	0.4083	0.4545	0.3913
14	rs268259843	18784724	No	No	0.4676	0.2500	0.3913
14	rs268275082	22162099	Yes	Yes	0.3578	0.2727	0.3478
14	rs268245930	29338091	Yes	No	0.3394	0.3571	0.4130
14	rs268290362	34041200	Yes	No	0.4220	0.3409	0.4565
14	rs268246190	41740181	No	Yes	0.4817	0.4091	0.0652
14	rs268267042	56395555	Yes	Yes	0.4128	0.4545	0.3913
14	rs268234921	68650539	Yes	Yes	0.4312	0.1364	0.0000
14	rs268243931	70863732	No	No	0.5000	0.4545	0.4783
15	rs268264644	13789853	Yes	Yes	0.3578	0.4318	0.4348
15	rs268264147	39837325	Yes	No	0.5000	0.4773	0.4565
15	rs268274888	56970382	No	No	0.4537	0.3182	0.4783
15	rs268282317	69182969	Yes	Yes	0.3716	0.1136	0.2174
15	rs268272447	74585825	No	Yes	0.3532	0.2273	0.0652
16	rs268275031	4798240	Yes	Yes	0.4083	0.4091	0.2174
16	rs268253542	14825302	Yes	No	0.4769	0.4524	0.3043
16	rs268236735	60438254	Yes	Yes	0.4450	0.5000	0.0435
16	rs268236775	62199820	Yes	No	0.4541	0.4091	0.4130
17	rs268279076	5707314	No	No	0.4771	0.3636	0.3913
17	rs268246995	13653729	Yes	No	0.3899	0.4773	0.1087
17	rs268264834	14008159	Yes	Yes	0.4587	0.2273	0.5000
17	rs268253701	15963988	Yes	Yes	0.4358	0.2500	0.1522
17	rs268275250	23081958	Yes	Yes	0.4174	0.3182	0.3696
17	rs268278647	42428686	No	Yes	0.3991	0.2273	0.2391
17	rs268258030	67278960	Yes	Yes	0.4495	0.1818	0.1522
18	rs268239108	289696	No	No	0.3853	0.1818	0.4348
18	rs268247366	26965743	Yes	Yes	0.3945	0.4773	0.3913
19	rs268234495	2843167	Yes	No	0.4908	0.3864	0.3478
19	rs268256467	5286869	No	No	0.4862	0.3409	0.1739
19	rs268249505	13898134	Yes	Yes	0.4725	0.5000	0.4348
19	rs268269273	20035874	Yes	Yes	0.4037	0.5000	0.5000
19	rs268243499	23605701	Yes	Yes	0.4771	0.0682	0.4565
20	rs268249880	12542099	No	Yes	0.4725	0.1818	0.3696
20	rs268281553	18963051	Yes	No	0.3148	0.1818	0.2826
20	rs268277257	30376741	Yes	Yes	0.4725	0.3636	0.4565
20	rs268277265	30797637	Yes	Yes	0.4541	0.2955	0.2174
20	rs268276522	36929792	No	Yes	0.4908	0.4091	0.0217
20	rs268276533	37342590	Yes	Yes	0.3750	0.0909	0.0217
20	rs268269363	40345448	Yes	No	0.3440	0.4318	0.1957
20	rs268269339	41511357	Yes	No	0.3028	0.3182	0.4348
20	rs268247180	46642677	Yes	Yes	0.4862	0.3864	0.4348
20	rs268247074	57915104	Yes	No	0.4037	0.4318	0.4565
21	rs268288665	3428092	No	No	0.3486	0.4773	0.2174

Table S3.6 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	MAF Alpine	MAF Saanen	MAF Teramana
21	rs268264795	22472563	Yes	Yes	0.4633	0.5000	0.1739
21	rs268260911	29432600	Yes	Yes	0.3972	0.2955	0.0652
21	rs268268779	37301417	Yes	Yes	0.4299	0.4286	0.1304
21	rs268270896	38207636	Yes	Yes	0.4541	0.4091	0.2609
21	rs268251546	45905777	No	Yes	0.4312	0.2955	0.4348
22	rs268285828	43510684	Yes	Yes	0.4312	0.2955	0.5000
23	rs268281624	10978508	Yes	Yes	0.4954	0.3864	0.4348
23	rs268279581	14416954	No	No	0.3945	0.3182	0.4783
23	rs268279579	14482478	Yes	Yes	0.3704	0.4318	0.0000
23	rs268279227	22752605	Yes	Yes	0.4954	0.2500	0.3913
23	rs268243154	40148134	No	No	0.4633	0.2955	0.1304
23	rs268271196	45114815	Yes	Yes	0.4817	0.3636	0.0652
24	rs268255186	8913378	Yes	Yes	0.4266	0.4545	0.4565
24	rs268269245	13189959	Yes	Yes	0.3303	0.4773	0.3913
24	rs268269230	13848679	No	No	0.4358	0.2273	0.4565
24	rs268233780	20116737	No	No	0.4771	0.2273	0.1304
24	rs268272844	43893141	No	Yes	0.3578	0.2955	0.1304
24	rs268272887	45722088	Yes	No	0.3532	0.3864	0.4348
25	rs268249581	1748234	Yes	No	0.4037	0.4773	0.3261
25	rs268245219	4934933	Yes	Yes	0.4404	0.2500	0.1522
25	rs268256122	21133143	Yes	Yes	0.4450	0.2500	0.4348
25	rs268242388	28597312	No	No	0.4679	0.4091	0.4783
26	rs268273098	19736750	Yes	No	0.4404	0.1364	0.1087
26	rs268273096	19812897	Yes	No	0.4815	0.2727	0.4565
26	rs268244309	22137817	Yes	No	0.4771	0.3409	0.3696
26	rs268263926	28011282	Yes	No	0.3798	0.1591	0.4565
26	rs268287936	45605084	Yes	No	0.4771	0.4773	0.5000
27	rs268262955	2766505	No	No	0.5000	0.2273	0.2174
27	rs268251226	34359106	No	No	0.4771	0.3182	0.4348
27	rs268283580	41446346	No	No	0.4587	0.2727	0.4348
28	rs268245044	800147	Yes	Yes	0.4771	0.2727	0.0000
28	rs268233533	1729134	Yes	Yes	0.5000	0.4773	0.2391
28	rs268249295	22580645	Yes	No	0.3532	0.4091	0.2609
28	rs268271895	26524406	No	No	0.4404	0.2727	0.3261
28	rs268238065	35491803	Yes	Yes	0.3899	0.3409	0.2826
29	rs268249410	14020518	No	No	0.4771	0.4773	0.3696
29	rs268262582	15642290	Yes	Yes	0.4450	0.3864	0.2391
29	rs268275734	22452097	Yes	Yes	0.4450	0.3182	0.1522

Table S3.6 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	Alpine	Teramana	Saanen	All together
1	rs268246894	22594761	Yes	No	0.1407	0.0529	0.1446	0.1393
1	rs268277578	45318610	Yes	Yes	0.1424	0.0000	0.2066	0.1973
1	rs268237133	52071174	Yes	Yes	0.1379	0.0000	0.3306	0.1590
1	rs268272505	77101471	Yes	No	0.1704	0.1134	0.0661	0.1579
1	rs268279562	82433467	Yes	No	0.1459	0.0000	0.0579	0.1305
1	rs268280267	86943152	No	Yes	0.1364	0.0000	0.1653	0.1716
1	rs268280291	87940819	No	No	0.1409	0.1323	0.2645	0.1558
1	rs268258713	90873881	No	No	0.1532	0.0756	0.0826	0.1417
1	rs268278510	105155775	Yes	Yes	0.1646	0.2079	0.0620	0.1784
1	rs268257920	110168450	Yes	Yes	0.1667	0.1059	0.0000	0.1619
1	rs268233519	113146255	Yes	Yes	0.1481	0.0605	0.1033	0.1516
1	rs268286780	134540117	Yes	Yes	0.1414	0.0907	0.0413	0.1427
2	rs268239713	17031741	No	No	0.1367	0.0000	0.0909	0.1307
2	rs268261967	23607634	No	Yes	0.1427	0.1134	0.1240	0.1488
2	rs268253020	36052714	No	Yes	0.1722	0.1248	0.0413	0.1472
2	rs268267960	42902914	No	No	0.1364	0.0000	0.2314	0.1564
2	rs268276840	46116901	Yes	Yes	0.1651	0.0529	0.0661	0.1480
2	rs268239885	53472858	No	Yes	0.1469	0.0681	0.0000	0.1323
2	rs268239859	54554491	No	Yes	0.1497	0.1815	0.0868	0.1572
2	rs268261308	62556317	No	No	0.1667	0.0416	0.0000	0.1558
2	rs268282666	73666511	No	Yes	0.1357	0.1059	0.0826	0.1476
2	rs268285441	81115463	Yes	Yes	0.1374	0.0000	0.1116	0.1332
2	rs268292755	124506565	Yes	Yes	0.1471	0.0000	0.0579	0.1609
2	rs268279477	125308567	Yes	No	0.1448	0.0605	0.0000	0.1187
3	rs268248233	42135080	Yes	Yes	0.1424	0.0000	0.0496	0.1437
3	rs268248191	43955403	No	No	0.1409	0.0000	0.1240	0.1376

Table S3.7 - The table contains Pe values (fraction of excluded alleged parent) for each SNP. The first four columns contains information about: chromosome, SNP ID, rs ID and physical position. Panel columns indicates if the SNP is retained in reduced panel (of 130 or 114 SNPs). Remaining four columns contain Pe values for each breed, and considering all breeds together. Graph below shows Pe values for each SNP; modifed from Talenti et al. (2016).

Chromosome	rs ID	Position	130 SNPs	114 SNPs	Alpine	Teramana	Saanen	All together
3	rs268261853	49886261	Yes	No	0.1520	0.0000	0.0000	0.1417
3	rs268251430	55616220	Yes	Yes	0.1465	0.0000	0.0000	0.1345
3	rs268275850	92297040	Yes	Yes	0.1562	0.0000	0.0372	0.1670
4	rs268278238	1327116	No	Yes	0.1471	0.0000	0.1653	0.1685
4	rs268287539	23292639	Yes	Yes	0.1440	0.0000	0.0000	0.1982
4	rs268274140	28916317	Yes	Yes	0.1510	0.0000	0.0826	0.1704
4	rs268263470	50257096	Yes	Yes	0.1513	0.0000	0.1074	0.1725
4	rs268263449	51071852	Yes	Yes	0.1709	0.0000	0.1322	0.1741
4	rs268263359	55167467	Yes	No	0.1562	0.0416	0.1240	0.1446
5	rs268266016	1470207	Yes	No	0.1471	0.0000	0.0826	0.1328
5	rs268250051	25734084	No	Yes	0.1394	0.0000	0.1157	0.1813
5	rs268275159	32063215	Yes	Yes	0.1379	0.0605	0.0744	0.1223
5	rs268275167	32370118	No	No	0.1503	0.1323	0.0992	0.1511
5	rs268238520	41176602	No	Yes	0.1488	0.1059	0.0496	0.1279
5	rs268265695	42491486	Yes	Yes	0.1778	0.0000	0.1860	0.2299
5	rs268281587	50028383	Yes	Yes	0.1587	0.1361	0.0620	0.1548
5	rs268265613	52262827	Yes	Yes	0.1367	0.1210	0.0579	0.1474
5	rs268264560	74401011	Yes	No	0.1425	0.0000	0.0499	0.1193
5	rs268235408	77288117	No	No	0.1448	0.0000	0.1157	0.1376
5	rs268280631	101356476	No	Yes	0.1454	0.1248	0.1736	0.1535
5	rs268280688	104117349	No	No	0.1774	0.0491	0.0992	0.1634
6	rs268263951	655359	No	No	0.1448	0.1323	0.1033	0.1391
6	rs268263057	12264321	Yes	Yes	0.1481	0.0000	0.0826	0.1990
6	rs268245369	31754679	Yes	No	0.1414	0.0567	0.1116	0.1228
6	rs268273538	53264333	No	No	0.1599	0.1361	0.0372	0.1425
6	rs268289495	63726709	No	No	0.1611	0.1134	0.0992	0.1560

Table S3.7 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	Alpine	Teramana	Saanen	All together
6	rs268260251	77174216	No	Yes	0.1523	0.0000	0.0868	0.1767
6	rs268242736	93909892	Yes	Yes	0.1557	0.1512	0.1240	0.1585
7	rs268257860	4192615	Yes	No	0.1939	0.0340	0.0909	0.1744
7	rs268248284	8617316	Yes	Yes	0.1367	0.0454	0.1157	0.1366
7	rs268247911	39668489	Yes	No	0.1682	0.1059	0.0207	0.1322
7	rs268262813	42087284	Yes	Yes	0.1481	0.1815	0.1860	0.1677
7	rs268262879	44894653	No	Yes	0.1806	0.0000	0.0826	0.1971
7	rs268262913	46539792	Yes	No	0.1555	0.0681	0.0579	0.1356
7	rs268242911	48380166	Yes	Yes	0.1543	0.0000	0.0661	0.1461
7	rs268238690	83379519	Yes	Yes	0.1513	0.0000	0.0413	0.1482
7	rs268252538	92013304	No	No	0.1364	0.0000	0.1446	0.1403
7	rs268237682	95788706	No	Yes	0.1440	0.0907	0.0000	0.1507
7	rs268279527	102250706	No	Yes	0.1374	0.0416	0.1322	0.1391
7	rs268254013	103700389	Yes	No	0.1931	0.0756	0.0992	0.1585
8	rs268247598	10088172	No	Yes	0.1517	0.1134	0.0635	0.1447
8	rs268276094	16754774	Yes	Yes	0.1488	0.2042	0.1116	0.1596
8	rs268244448	22324402	No	No	0.1389	0.0681	0.0000	0.1323
8	rs268277168	26298640	No	No	0.1454	0.0000	0.0744	0.1370
8	rs268236867	30307773	Yes	Yes	0.1394	0.0945	0.1240	0.1559
8	rs268278187	33997194	Yes	Yes	0.1603	0.0605	0.0992	0.1660
8	rs268292741	46738116	Yes	Yes	0.1757	0.1134	0.1240	0.1585
8	rs268245008	55874397	No	No	0.1473	0.0907	0.0000	0.1280
8	rs268286103	62828040	Yes	No	0.1688	0.0000	0.1983	0.1707
8	rs268266882	89489430	Yes	No	0.1535	0.0567	0.0000	0.1356
8	rs268266897	89981503	Yes	No	0.1656	0.0000	0.0248	0.1337
8	rs268266898	90013365	Yes	Yes	0.1444	0.0000	0.0826	0.1359
9	rs268264165	12991411	Yes	Yes	0.1531	0.0983	0.2314	0.1932
9	rs268236937	23828851	Yes	Yes	0.1416	0.1512	0.1488	0.1552

Table S3.7 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	Alpine	Teramana	Saanen	All together
9	rs268272814	26916830	Yes	Yes	0.1481	0.1361	0.1157	0.1625
9	rs268275918	35120000	No	Yes	0.1374	0.1248	0.0413	0.1442
9	rs268260994	44816684	Yes	No	0.1508	0.0907	0.1116	0.1387
9	rs268233440	45493989	No	Yes	0.1616	0.0000	0.1322	0.1634
9	rs268266688	54414444	Yes	Yes	0.1500	0.1021	0.1240	0.1513
9	rs268275277	64658280	No	No	0.1407	0.1021	0.1497	0.1511
9	rs268282225	66461656	Yes	Yes	0.1364	0.0000	0.1736	0.1642
9	rs268251679	67146104	No	No	0.1565	0.0605	0.0992	0.1316
10	rs268261737	22632227	No	No	0.1414	0.0000	0.0000	0.1251
10	rs268257058	42941965	Yes	Yes	0.1394	0.0000	0.0289	0.1391
10	rs268288791	43634448	No	No	0.1357	0.0983	0.0000	0.1447
10	rs268240865	69364926	Yes	No	0.1535	0.1059	0.0000	0.1316
10	rs268270243	89232772	No	Yes	0.1370	0.0454	0.0992	0.1164
11	rs268272993	700718	No	No	0.1431	0.0756	0.0000	0.1207
11	rs268291501	4600369	Yes	Yes	0.1536	0.1966	0.0372	0.1581
11	rs268292075	20664777	Yes	Yes	0.1380	0.0567	0.0455	0.1154
11	rs268250711	60564735	No	No	0.1414	0.0794	0.1240	0.1376
11	rs268250673	62446881	Yes	Yes	0.1670	0.0605	0.0620	0.1660
11	rs268250509	69480159	Yes	Yes	0.1725	0.0832	0.0000	0.1668
11	rs268253243	75527471	Yes	Yes	0.1560	0.0794	0.1653	0.1517
11	rs268247515	79589304	Yes	Yes	0.1741	0.0363	0.1653	0.1726
11	rs268243784	83719137	Yes	Yes	0.1444	0.0567	0.1488	0.1707
11	rs268271367	85418308	No	Yes	0.1407	0.1021	0.1983	0.1438
11	rs268284989	97787718	No	Yes	0.1380	0.1059	0.0661	0.1548
12	rs268278336	3757154	Yes	No	0.1394	0.0000	0.0000	0.1214
12	rs268256530	17064107	Yes	No	0.1663	0.0302	0.0579	0.1619
12	rs268262656	34769122	Yes	No	0.1394	0.0416	0.0620	0.1214
12	rs268268097	49872578	Yes	Yes	0.1357	0.0000	0.1446	0.1586

Table S3.7 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	Alpine	Teramana	Saanen	All together
12	rs268251731	58797148	Yes	No	0.1591	0.0756	0.0992	0.1387
12	rs268274667	64087392	No	No	0.1885	0.0000	0.0455	0.1663
13	rs268280844	26389054	No	No	0.1549	0.0491	0.0909	0.1403
13	rs268280831	27021612	Yes	No	0.1535	0.1361	0.0331	0.1370
13	rs268280826	27284949	No	Yes	0.1532	0.1853	0.0579	0.1606
13	rs268280760	30261547	Yes	Yes	0.1557	0.0983	0.0000	0.1474
13	rs268290264	43206738	Yes	Yes	0.1481	0.1701	0.0826	0.1443
13	rs268245797	47426348	No	No	0.1380	0.0756	0.0909	0.1275
13	rs268236129	52485704	Yes	Yes	0.1414	0.1323	0.1074	0.1475
13	rs268291580	61123452	Yes	Yes	0.1576	0.1512	0.0000	0.1967
14	rs268279625	917480	Yes	Yes	0.1364	0.2117	0.0992	0.1476
14	rs268276667	4121438	Yes	No	0.1555	0.0907	0.0992	0.1412
14	rs268259843	18784724	No	No	0.1471	0.0529	0.0000	0.1316
14	rs268275082	22162099	Yes	Yes	0.1838	0.1134	0.1612	0.2028
14	rs268245930	29338091	Yes	No	0.1488	0.0794	0.1814	0.1714
14	rs268290362	34041200	Yes	No	0.1444	0.0567	0.0744	0.1336
14	rs268246190	41740181	No	Yes	0.1409	0.0000	0.1322	0.1623
14	rs268267042	56395555	Yes	Yes	0.1414	0.0907	0.1983	0.1529
14	rs268234921	68650539	Yes	Yes	0.1909	0.0000	0.1488	0.2514
14	rs268243931	70863732	No	No	0.1416	0.1134	0.1983	0.1452
15	rs268264644	13789853	Yes	Yes	0.1374	0.0681	0.0744	0.1164
15	rs268264147	39837325	Yes	No	0.1416	0.1323	0.0826	0.1316
15	rs268274888	56970382	No	No	0.1500	0.1134	0.0372	0.1324
15	rs268282317	69182969	Yes	Yes	0.1394	0.1134	0.0000	0.1809
15	rs268272447	74585825	No	Yes	0.1402	0.0000	0.1157	0.1330
16	rs268275031	4798240	Yes	Yes	0.1555	0.0000	0.0496	0.1498
16	rs268253542	14825302	Yes	No	0.1382	0.0832	0.1088	0.1377
16	rs268236735	60438254	Yes	Yes	0.1557	0.0000	0.2025	0.1967

Table S3.7 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	Alpine	Teramana	Saanen	All together
16	rs268236775	62199820	Yes	No	0.1791	0.0189	0.1322	0.1467
17	rs268279076	5707314	No	No	0.1555	0.0529	0.0661	0.1342
17	rs268246995	13653729	Yes	No	0.2062	0.0000	0.0248	0.1629
17	rs268264834	14008159	Yes	Yes	0.1431	0.1853	0.1157	0.1577
17	rs268253701	15963988	Yes	Yes	0.1641	0.0000	0.1074	0.1548
17	rs268275250	23081958	Yes	Yes	0.1379	0.1021	0.0826	0.1272
17	rs268278647	42428686	No	Yes	0.1414	0.0491	0.0537	0.1606
17	rs268258030	67278960	Yes	Yes	0.1414	0.0000	0.0620	0.1391
18	rs268239108	289696	No	No	0.1515	0.1512	0.0000	0.1511
18	rs268247366	26965743	Yes	Yes	0.1555	0.1361	0.1240	0.1594
19	rs268234495	2843167	Yes	No	0.1831	0.0302	0.0248	0.1414
19	rs268256467	5286869	No	No	0.1364	0.0000	0.0331	0.1251
19	rs268249505	13898134	Yes	Yes	0.1500	0.0378	0.0372	0.1106
19	rs268269273	20035874	Yes	Yes	0.1592	0.0945	0.2645	0.1653
19	rs268243499	23605701	Yes	Yes	0.1357	0.1323	0.0000	0.1579
20	rs268249880	12542099	No	Yes	0.1401	0.2722	0.0620	0.1642
20	rs268281553	18963051	Yes	No	0.1415	0.0000	0.0620	0.1094
20	rs268277257	30376741	Yes	Yes	0.1401	0.1323	0.1116	0.1397
20	rs268277265	30797637	Yes	Yes	0.1374	0.0529	0.0000	0.1376
20	rs268276522	36929792	No	Yes	0.1414	0.0000	0.0868	0.1586
20	rs268276533	37342590	Yes	Yes	0.1389	0.0000	0.0000	0.1307
20	rs268269363	40345448	Yes	No	0.1818	0.0567	0.1653	0.1688
20	rs268269339	41511357	Yes	No	0.1465	0.1059	0.0000	0.1170
20	rs268247180	46642677	Yes	Yes	0.1461	0.1512	0.0992	0.1452
20	rs268247074	57915104	Yes	No	0.1592	0.0907	0.0744	0.1434
21	rs268288665	3428092	No	No	0.1431	0.0000	0.1736	0.1397
21	rs268264795	22472563	Yes	Yes	0.1488	0.0605	0.1488	0.1558
21	rs268260911	29432600	Yes	Yes	0.1577	0.0000	0.0909	0.1473

Table S3.7 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	Alpine	Teramana	Saanen	All together
21	rs268268779	37301417	Yes	Yes	0.1635	0.0681	0.1270	0.1712
21	rs268270896	38207636	Yes	Yes	0.1374	0.0983	0.0868	0.1321
21	rs268251546	45905777	No	Yes	0.1576	0.1059	0.0413	0.1370
22	rs268285828	43510684	Yes	Yes	0.1370	0.1853	0.1488	0.1625
23	rs268281624	10978508	Yes	Yes	0.1367	0.1512	0.0992	0.1383
23	rs268279581	14416954	No	No	0.1555	0.1588	0.1364	0.1595
23	rs268279579	14482478	Yes	Yes	0.1420	0.0000	0.0744	0.1346
23	rs268279227	22752605	Yes	Yes	0.1465	0.1361	0.1736	0.1660
23	rs268243154	40148134	No	No	0.1389	0.0000	0.0909	0.1389
23	rs268271196	45114815	Yes	Yes	0.1508	0.0000	0.0661	0.1687
24	rs268255186	8913378	Yes	Yes	0.1407	0.1323	0.1446	0.1457
24	rs268269245	13189959	Yes	Yes	0.1427	0.0529	0.1736	0.1356
24	rs268269230	13848679	No	No	0.1641	0.0907	0.1157	0.1704
24	rs268233780	20116737	No	No	0.1454	0.0000	0.0537	0.1464
24	rs268272844	43893141	No	Yes	0.1485	0.0000	0.0413	0.1218
24	rs268272887	45722088	Yes	No	0.1402	0.1059	0.0579	0.1222
25	rs268249581	1748234	Yes	No	0.1380	0.0756	0.0248	0.1117
25	rs268245219	4934933	Yes	Yes	0.1394	0.0000	0.1736	0.1447
25	rs268256122	21133143	Yes	Yes	0.1454	0.1059	0.1074	0.1417
25	rs268242388	28597312	No	No	0.1545	0.0454	0.1322	0.1346
26	rs268273098	19736750	Yes	No	0.1394	0.0718	0.0702	0.1518
26	rs268273096	19812897	Yes	No	0.1435	0.1323	0.0455	0.1382
26	rs268244309	22137817	Yes	No	0.1555	0.0605	0.0331	0.1281
26	rs268263926	28011282	Yes	No	0.1546	0.0907	0.0000	0.1347
26	rs268287936	45605084	Yes	No	0.1768	0.0605	0.0826	0.1414
27	rs268262955	2766505	No	No	0.1515	0.0529	0.0537	0.1538
27	rs268251226	34359106	No	No	0.1660	0.1512	0.0372	0.1552
27	rs268283580	41446346	No	No	0.1431	0.1059	0.0455	0.1316

Table S3.7 – Continuing.

Chromosome	rs ID	Position	130 SNPs	114 SNPs	Alpine	Teramana	Saanen	All together
28	rs268245044	800147	Yes	Yes	0.1357	0.0000	0.0455	0.1480
28	rs268233533	1729134	Yes	Yes	0.1416	0.1059	0.1240	0.1498
28	rs268249295	22580645	Yes	No	0.1402	0.0000	0.1860	0.1280
28	rs268271895	26524406	No	No	0.1495	0.0340	0.0000	0.1336
28	rs268238065	35491803	Yes	Yes	0.1481	0.0907	0.0000	0.1370
29	rs268249410	14020518	No	No	0.1660	0.1512	0.0826	0.1559
29	rs268262582	15642290	Yes	Yes	0.1454	0.1059	0.0579	0.1518
29	rs268275734	22452097	Yes	Yes	0.1557	0.0000	0.0372	0.1359

Table S3.7 – Continuing.

	Random selection from highly polymorphic in Alpine breed						Random selection from highly polymorphic in all breeds					
	Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel		Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel	
Set	FP 200 SNP 27k	Work	FP 130 SNP 27k	Work	FP 114 SNP 27k	Work	FP 200 SNP 10k	Work	FP 130 SNP 10k	Work	FP 114 SNP 10k	Work
1	0	Y	4	N	12	N	0	Y	4	N	5	N
2	0	Y	5	N	11	N	0	Y	4	N	6	N
3	1	N	6	N	5	N	0	Y	3	N	2	N
4	1	N	2	N	5	N	0	Y	1	N	3	N
5	0	Y	8	N	5	N	0	Y	4	N	1	N
6	1	N	1	N	5	N	0	Y	0	Y	2	N
7	0	Y	1	N	6	N	0	Y	1	N	8	N
8	0	Y	9	N	8	N	0	Y	5	N	6	N
9	0	Y	1	N	10	N	0	Y	0	Y	6	N
10	1	N	3	N	16	N	0	Y	2	N	6	N
11	0	Y	6	N	11	N	0	Y	1	N	5	N
12	0	Y	5	N	5	N	0	Y	1	N	7	N
13	0	Y	8	N	9	N	0	Y	1	N	4	N
14	0	Y	4	N	13	N	0	Y	1	N	4	N
15	0	Y	2	N	9	N	0	Y	4	N	2	N
16	0	Y	4	N	10	N	0	Y	2	N	3	N
17	1	N	5	N	7	N	0	Y	2	N	10	N
18	0	Y	4	N	4	N	1	N	4	N	4	N
19	0	Y	3	N	6	N	0	Y	4	N	5	N
20	0	Y	5	N	7	N	0	Y	3	N	2	N
21	0	Y	3	N	4	N	0	Y	1	N	6	N

Table S3.8 -Results of random analysis of markers on the dataset post quality check. For each panel size, 300 random dataset were selected out of 27,523 markers highly polymorphic in Alpine breed and 300 random dataset were selected out of 10,338 markers highly polymorphic in all breeds. For each dataset, we report the number of false positive and if the panel is 'working'. We consider a panel as 'working' if it have no false positive. Below, we classify each panel of each size by the number of False Positive; Modified from Talenti et al. (2016).

	Random selection from highly polymorphic in Alpine breed						Random selection from highly polymorphic in all breeds					
	Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel		Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel	
Set	FP 200 SNP 27k	Work	FP 130 SNP 27k	Work	FP 114 SNP 27k	Work	FP 200 SNP 10k	Work	FP 130 SNP 10k	Work	FP 114 SNP 10k	Work
22	0	Y	3	N	6	N	2	N	2	N	9	N
23	2	N	11	N	10	N	0	Y	4	N	5	N
24	1	N	4	N	4	N	0	Y	0	Y	4	N
25	3	N	6	N	5	N	1	N	1	N	4	N
26	2	N	6	N	8	N	1	N	3	N	5	N
27	0	Y	5	N	7	N	0	Y	1	N	7	N
28	0	Y	4	N	11	N	0	Y	1	N	2	N
29	0	Y	2	N	9	N	1	N	3	N	3	N
30	1	N	3	N	2	N	0	Y	1	N	4	N
31	1	N	10	N	9	N	0	Y	1	N	2	N
32	3	N	6	N	3	N	0	Y	1	N	2	N
33	0	Y	5	N	10	N	1	N	1	N	1	N
34	1	N	5	N	11	N	0	Y	0	Y	4	N
35	1	N	4	N	10	N	0	Y	0	Y	4	N
36	0	Y	0	Y	14	N	1	N	2	N	4	N
37	0	Y	3	N	5	N	0	Y	5	N	4	N
38	0	Y	7	N	16	N	0	Y	0	Y	4	N
39	1	N	4	N	11	N	0	Y	5	N	7	N
40	0	Y	10	N	10	N	0	Y	4	N	5	N
41	0	Y	4	N	2	N	0	Y	1	N	4	N
42	0	Y	3	N	4	N	0	Y	3	N	5	N
43	1	N	6	N	6	N	0	Y	5	N	3	N
44	1	N	1	N	16	N	0	Y	3	N	5	N
45	1	N	6	N	10	N	0	Y	3	N	2	N
46	0	Y	4	N	4	N	1	N	1	N	2	N

Table S3.8 – Continuing.

	Random selection from highly polymorphic in Alpine breed						Random selection from highly polymorphic in all breeds					
	Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel		Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel	
Set	FP 200 SNP 27k	Work	FP 130 SNP 27k	Work	FP 114 SNP 27k	Work	FP 200 SNP 10k	Work	FP 130 SNP 10k	Work	FP 114 SNP 10k	Work
47	0	Y	4	N	13	N	0	Y	3	N	5	N
48	0	Y	6	N	7	N	1	N	3	N	1	N
49	0	Y	2	N	10	N	0	Y	3	N	4	N
50	1	N	4	N	5	N	0	Y	3	N	4	N
51	0	Y	7	N	2	N	0	Y	3	N	4	N
52	0	Y	4	N	1	N	0	Y	0	Y	4	N
53	2	N	3	N	11	N	0	Y	4	N	9	N
54	0	Y	4	N	6	N	0	Y	2	N	2	N
55	0	Y	6	N	9	N	0	Y	1	N	4	N
56	0	Y	2	N	7	N	0	Y	0	Y	6	N
57	0	Y	4	N	6	N	0	Y	0	Y	2	N
58	0	Y	4	N	8	N	0	Y	1	N	6	N
59	1	N	8	N	8	N	0	Y	4	N	6	N
60	0	Y	2	N	5	N	0	Y	2	N	6	N
61	1	N	5	N	5	N	0	Y	3	N	9	N
62	0	Y	3	N	8	N	0	Y	2	N	8	N
63	0	Y	2	N	6	N	0	Y	3	N	3	N
64	0	Y	4	N	9	N	0	Y	4	N	3	N
65	1	N	4	N	11	N	0	Y	1	N	8	N
66	0	Y	9	N	18	N	0	Y	1	N	4	N
67	0	Y	3	N	15	N	0	Y	2	N	4	N
68	0	Y	4	N	4	N	0	Y	2	N	4	N
69	0	Y	3	N	8	N	0	Y	3	N	3	N
70	1	N	4	N	5	N	0	Y	1	N	3	N

Table S3.8 – Continuing.

	Random selection from highly polymorphic in Alpine breed						Random selection from highly polymorphic in all breeds					
	Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel		Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel	
Set	FP 200 SNP 27k	Work	FP 130 SNP 27k	Work	FP 114 SNP 27k	Work	FP 200 SNP 10k	Work	FP 130 SNP 10k	Work	FP 114 SNP 10k	Work
71	1	N	8	N	4	N	0	Y	5	N	8	N
72	0	Y	5	N	14	N	0	Y	3	N	8	N
73	0	Y	3	N	11	N	0	Y	1	N	4	N
74	0	Y	1	N	11	N	0	Y	4	N	7	N
75	2	N	3	N	12	N	0	Y	2	N	3	N
76	0	Y	4	N	7	N	1	N	1	N	6	N
77	0	Y	4	N	7	N	0	Y	3	N	4	N
78	0	Y	6	N	11	N	0	Y	4	N	7	N
79	1	N	5	N	2	N	0	Y	2	N	3	N
80	1	N	6	N	9	N	1	N	2	N	6	N
81	1	N	2	N	11	N	1	N	1	N	1	N
82	0	Y	8	N	6	N	0	Y	3	N	5	N
83	1	N	3	N	6	N	0	Y	6	N	7	N
84	0	Y	3	N	10	N	1	N	1	N	3	N
85	0	Y	6	N	7	N	0	Y	3	N	5	N
86	1	N	13	N	13	N	0	Y	1	N	5	N
87	0	Y	7	N	9	N	1	N	1	N	10	N
88	0	Y	4	N	4	N	0	Y	2	N	5	N
89	0	Y	3	N	13	N	0	Y	1	N	4	N
90	0	Y	2	N	6	N	0	Y	1	N	10	N
91	1	N	2	N	6	N	0	Y	4	N	3	N
92	0	Y	4	N	11	N	1	N	2	N	5	N
93	1	N	6	N	7	N	1	N	2	N	1	N
94	0	Y	4	N	8	N	0	Y	2	N	6	N

Table S3.8 – Continuing.

	Random selection from highly polymorphic in Alpine breed						Random selection from highly polymorphic in all breeds					
	Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel		Random 200 SNPs Panel		Random 130 SNPs Panel		Random 114 SNPs Panel	
Set	FP 200 SNP 27k	Work	FP 130 SNP 27k	Work	FP 114 SNP 27k	Work	FP 200 SNP 10k	Work	FP 130 SNP 10k	Work	FP 114 SNP 10k	Work
95	0	Y	0	Y	9	N	1	N	4	N	10	N
96	0	Y	2	N	11	N	1	N	3	N	4	N
97	0	Y	2	N	5	N	0	Y	2	N	6	N
98	0	Y	8	N	11	N	0	Y	1	N	5	N
99	1	N	5	N	10	N	1	N	1	N	7	N
100	0	Y	8	N	6	N	0	Y	0	Y	2	N

Table S3.8 – Continuing.

3.3 Breed Characterization

3.3.1 Aim

Selection signatures discovery described in chapter 2.3.1 and applied in chapter 3.1 could be used also to disentangle the genetic basis of different phenotypes such as coat color, a trait important in the standardization of modern breeds and even in field applications as breed traceability.

In this chapter, for example, I studied selection signatures in 929 animals of 41 Italian and Pakistani goat populations to discover signals associated to a peculiar coat color pattern, the roan, shown by the Barri and Beetal Muki Cheni Pakistani goat breeds. This pattern is characterized by the intermingling of white and pigmented hairs at different extent, ranging from almost white to almost pigmented animals. In this case, the two breeds shown this peculiar pattern, but different pigmentation: black for the Barri and red for the Beetal Muki Cheni, making it an interesting case to study this phenotype.

Our core analyses included two methods to detect regions with a strong decrease in variability, which could be therefore associated with the presence of a strong, recent positive selective pressure. Our findings highlighted one gene in particular, the KITLG, known as associated to the roan coat color in cattle. These findings are important to disentangle the genetic basis of the complex coat color pathway in the goat species, an important trait in the standardization of modern breeds .

3.3.2 *Brief Communication*: Genomic analysis suggests KITLG is responsible for a roan pattern in two Pakistani goat breeds

Accepted on Journal of Heredity on scheduled print issue 109-3, 2018.

Authors and Affiliations

Andrea Talenti[1,*], Francesca Bertolini[2,*], Jamie Williams[2], Muhammad Moaeen-ud-Din[3], Stefano Frattini[1], Beatrice Coizet[1], Giulio Pagnacco[1], James Reecy[2], Max F. Rothschild[2], Paola Crepaldi[1] and the Italian Goat Consortium

1 Department of Veterinary Medicine, Università degli Studi di Milano, Milan, Italy.

2 Department of Animal Science, Iowa State University, Ames, Iowa, USA.

3 Laboratories of Animal Breeding & Genetics, PMAS-Arid Agriculture University, Rawalpindi, Pakistan.

* These authors contributed equally to the work

Abstract

The roan coat color pattern is described as the presence of white hairs intermixed with pigmented hairs. This kind of pigmentation pattern has been observed in many domestic species, including the goat. The molecular mechanisms and inheritance that underlie this pattern are known for some species and the KITLG gene has been shown associated with this phenotype. To date, no research effort has been done to find the gene(s) that controls roan coat color pattern in goats. In the present study, after genotyping with the GoatSNP50 BeadChip, 35 goats that showed a roan pattern and that belonged to two Pakistan breeds (Group A) were analyzed and then compared to 740 goats of 39 Italian and Pakistan goats breeds that did not have the same coat color pattern (Group B). Run of homozygosity-based and XP-EHH analyses were used to identify unique genomic regions potentially associated with the roan pattern. A total of 3 regions on chromosomes 5, 6 and 12 were considered unique among the group A versus B group comparisons. The A region > 1.7 Mb on chromosome 5 was the most divergent between the two groups. This region contains six genes, including the KITLG gene. Our findings support the hypothesis that the KITLG gene may be associated with the roan phenotype in goats.

Introduction

Genetic studies of coat color and color patterns have been conducted in many domestic animal species. The genetic basis of pigmentation is complex and may involve several genes that can act on several steps of melanogenesis. One of the results of the modulation of the melanocyte activity is the roan pattern, described as presence of white hairs intermixed with pigmented hairs. This pattern gives rise to individual phenotypes that range from almost white to almost completely pigmented. The roan pattern can be confused with ticking pattern, which is characterized by flecks or spot of color on white area. So far, for the roan pattern, both the *KITLG* gene in Belgian blue cattle (Seitz et al., 1999; Li et al., 2016) and the *KIT* gene in horses (Dürig et al., 2017; Brooks and Bailey, 2005; Marklund et al., 1999) and pigs (Fontanesi et al., 2010; Cho et al., 2011) have been reported, while no genes have been detected as candidate genes for the ticking pattern.

Goats are one of the most important livestock species in developing countries because of their adaptability to different environments. Goats also display a high degree of genetic variability in coat color patterns. To date, only a few SNPs in 20 genes involved in coat color have been considered for their ability to discriminate among goat breeds (Nicoloso et al., 2012) and associations of SNPs and CNVs (Copy Number Variations) to red, black and white colors phenotypes have been reported for the *ASIP* and *MC1R* genes (Fontanesi et al., 2009a; b). No study has been tailored to detect genomic regions associated with the roan pattern in goats.

The present study focuses on the roan pattern seen in the Barri and Beetal Muki Cheni Pakistani breeds. The primary purpose of this investigation was to uncover the genomic region(s) that could be responsible for this pattern and to determine the gene(s) likely responsible for these phenotypes.

Material and Methods

Sample collection and genotyping

No animals were sacrificed for this study. Blood samples for Italian goats were collected according to the recommendations of the European Council (1986) concerning animal care. Animals were sampled from different farms in Italy and Pakistan. Italian goats were sampled in order to be as much unrelated as possible and offer a wider sample. Pakistani goats were sampled by PMAS University staff lead by Dr. Mouaeen-ud-din within the villages of Pakistanuder the approved Iowa State University animal care protocol. A total of 929 animals belonging to 41 different breeds were considered in this study (Table S3.9): 1) 369 animals and 15 breeds from Italy, described by Nicoloso et al. (Nicoloso et al., 2015). 2) 560 animals of 26 breeds from Pakistan including the small roan like spotted coat pattern Barri and Beetal Muki Cheni breeds (Figure 3.18a and 3.18b).

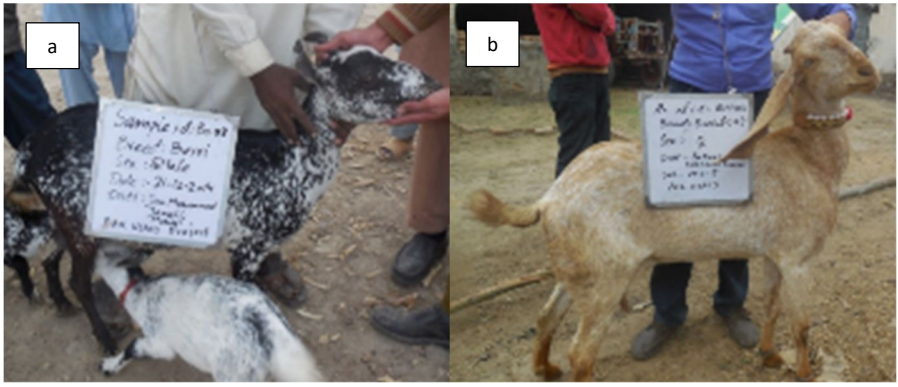


Figure 3.18 – Pictures of a) Barri and b) Beetal Muki Cheni goat breeds. The distribution of roan ranged from almost colored to almost white with only few spots.

DNA was extracted using commercially available kits and DNA samples were genotyped using the GoatSNP50 BeadChip (Illumina Inc., San Diego, CA) (Tosser-Klopp et al., 2014). SNPs with low call rate (< 95 %), non-polymorphic in all populations (minor allele frequency of 0 %) and not mapped to the assembly or on the X chromosome were excluded from subsequent analyses using Plink v1.9 (Chang et al., 2015). After the exclusion of low quality markers, all animals with call rates < 95 % were excluded from the dataset. Multi-dimensional scaling (MDS) was calculated using the cluster algorithm of Plink v1.9 (Chang et al., 2015). Duplicated and related animals were identified using an in-house script and removed if Identity by State (IBS) > 99% and Discordant Homozygotes (also known as Mendelian Errors, ME) <= 100. Analyses were performed by dividing all breeds into two groups: Group A, which was composed of the two breeds that showed the roan pattern (Barri and Beetal Muki Cheni), and the Group B, which was composed of Italian and Pakistan breeds with different coat color patterns. Genotypes in the resulting dataset were imputed and phased using Beagle v3.3.2 software considering one breed at a time (Browning, 2011).

Runs of Homozygosity and ROH comparison

The screening for the low heterozygosity regions across the genome was conducted for the two groups using the plink V1.9 software (Chang et al., 2015). The analysis was performed using 20 SNPs sliding windows (command --homozyg-snp), allowing no heterozygotes in each (--het 0). The number of homozygous animals at each marker was then normalized dividing that number by the total number of goats included in the analysis, obtaining a locus homozygosity score (H) ranging from 0 (0 %) to 1 (100 %). . The H score at each SNP calculated for the Group A was then compared with the H scores of all other populations calculated separately following the same parameters previously described (dROH analysis). This

comparison was performed applying the formula reported in Bertolini et al., 2016, obtaining a dH score that was then used to find the markers that most distinguish the Group A against all other breeds in Group B. Only regions with an H score value > 0.46 for the group A and a dH score value > 3.39 for the ROH comparison, which represented the top 0.2% of all SNP, were considered highly homologous for the ROH analysis and highly divergent for the ROH comparison respectively. These regions were then investigated for genes using most recent gene annotation available for *Capra hircus* genome (ARS1; Bickhart et al., 2017) and screened for the region of interest using the Bedtools software (Quinlan and Hall, 2010).

Haplotype analysis and gene annotation

Single SNP Cross-Population Extended Haplotype Homozygosity (XP-EHH) was calculated using the Selscan software (Szpiech and Hernandez, 2014). This method compared the integrated EHH at each marker between two different populations, allowing the detection of strong, directional selection of one allele in one of the two populations while remaining polymorphic in the other (Sabeti et al., 2007). Ancestral alleles were determined considering 15 wild animals of other *Capra* species including 8 *Capra ibex* and 7 *Capra Aegagrus* genotyped with the same SNPchip (data not shown).

All markers with normalized XP-EHH score above 3.09 (top 0.2 %) were considered as relevant and only the regions above the threshold that were concordant with the regions detected through the ROH analyses were considered for further investigations.

Results

The initial dataset contained 53,347 SNPs: among the 50,619 autosomal SNPs, 3,213 did not pass the quality-filtering steps. Animal filtering removed 72 individuals due to low call rate, 1 animal was excluded because it was duplicated and 82 animals were excluded because of high relatedness with other animals. After data editing, a total of 47,406 autosomal SNPs, 35 goats in Group A and 740 goats in Group B were retained. The MDS plot, produced on the imputed dataset (Figure S3.20), showed a clear separation between breeds raised in Italy and Pakistan, with no overlap between the two major clusters.

Run of homozygosity

Using the selected thresholds previously mentioned, six ROH with normalized H scores ≥ 3.4 were considered regions of high homozygosity in the group A (Figure 3.19, Supplementary Figures S3.21-S3.24). Two regions were located on chromosome 5 with lengths of 308.49 Kb and 1.75 Mb respectively (16,323,819-16,632,308 bp and 17,885,772-19,634,050 bp). Two other regions were located on chromosome 6, with lengths of 1.31Mb and 230.0 Kb (13,341,998-14,648,705 bp and 69,566,293-69,796,504 bp).

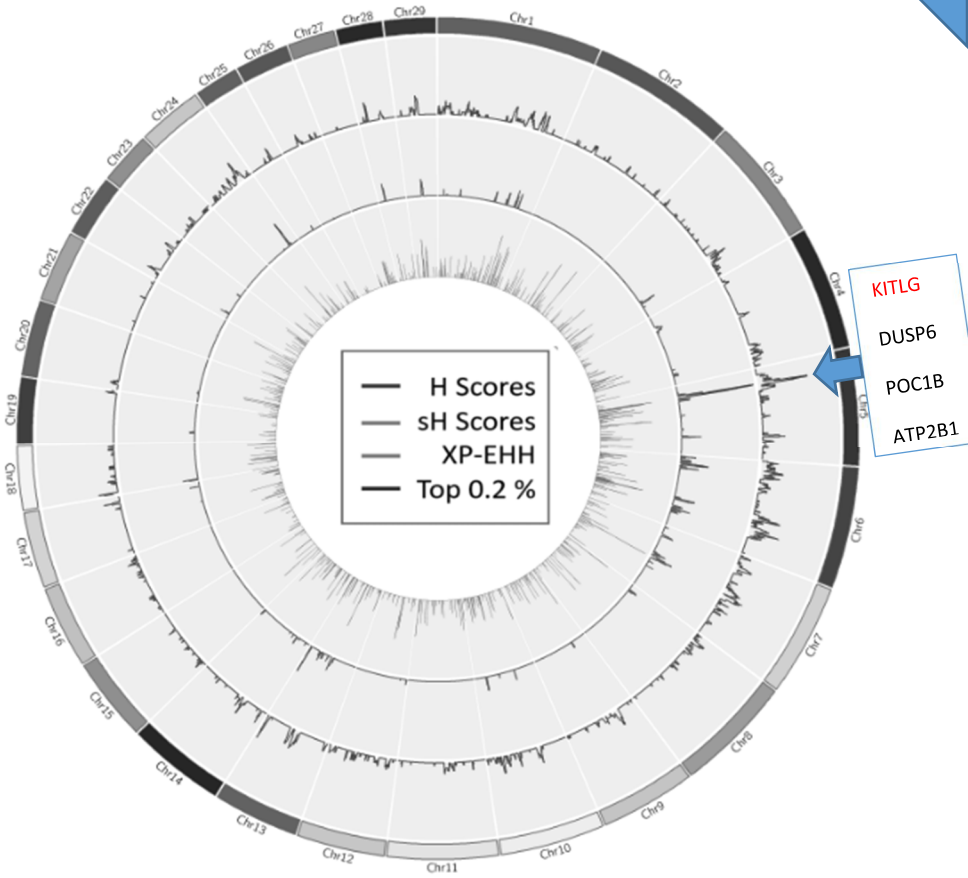


Figure 3.19 - Circos plot of the ROH H scores and dH score (outer circles, grey and blue tracks, respectively), XP-EHH (inner circle, green track); for each analysis the top 0.2% SNPs are marked in red.

One region was located on chromosome 13, with length of 308.19 Kb (77,687,773-77,995,965) and on chromosome 24, with length 706.19Kb (46,684,650-47,390,842 bp). The region with the highest H score was the 1.7 Mb region on chromosome 5, with the max H value of 0.71 reached within the sub region between 18,127,496 and 18,846,626 bp. The remaining regions showed lower H values that ranged from 0.46 to 0.5. Among these

regions, three were also above the dROH analysis threshold that compared ROH of group A with all ROH in each breed of group B (Figure 3.19): the 1.7Mb region on chromosome 5, the 230.0 Kb region chromosome 6, and the 308.19 Kb region on chromosome 13. The region on chromosome 5 contained 9 genes, the region on chromosome 6 contained 1 gene and the region on chromosome 13 contained 4 genes. The full list of the 14 genes included in these 3 regions is reported in Table S3.10. Among these genes, a marker on chromosome 5, which had the max H value of 0.71 (snp14289-scaffold157-1998233, 18,127,496 bp), was located within the KITLG gene (from 18,044,632 to 18,151,924 bp).

Haplotype analysis

The XP-EHH analysis was performed to discover regions of the genome that were under strong directional selection. Normalized XP-EHH analysis identified a total of 18 regions of the genome with more than 1 consecutive SNP that were under strong directional selection (Figure 3.19, Supplementary Table S3.11). Among these, the longest region was located on chromosome 5, from 18,127,496 to 19,478,621 bp. This region overlaps the region identified by ROH-based analyses and contained the KITLG gene. These results indicate that this region (and gene) was under strong selection. In addition, no other regions were in common with the ROH identified region, which, highlights the importance of this region.

Discussion

Runs of homozygosity analyses identified several regions of the genome that might be associated with roan coat color. The region with the highest H score for Group A was on chromosome 5, in a 1.7 Mb region from 17,448,053 to 19,198,567 bp, which included a genetic marker within the *KITLG* gene sequence. This region was also identified by standardized H

score analyses, which compared the ROH in the two populations pooled together with all other ROH in all other populations. Furthermore, this same region was identified by XP-EHH. This approach also identified several other regions that were not found using ROH, which indicates that those signals were either non-specific or the result of different selection forces. Finally, even after removing animals from Italy and the most divergent from Pakistan based on the MDS plot, the results still confirm this region.

An additional clue of the potential involvement of this gene in this pigmentation pattern is given by the minor allele frequency of the marker within the *KITLG* gene sequence (snp14289-scaffold157-1998233; 18,127,496 bp). This marker was completely fixed in the Group A (MAF = 0), whereas it was variable in Group B individuals (MAF = 0.294).

The *KIT* ligand (*KITLG*) gene is known as Mast Cell Grow Factor (*MGF*) or Stem Cell Factor (*SCF*). and it is involved in and can affect many biological processes, such as hematopoiesis, gametogenesis and melanogenesis. The *KITLG* gene affects pigmentation in both human and mice (Guenther et al., 2014). A missense mutation in the 7th exon of this gene was associated with the roan phenotype in cattle (Seitz et al., 1999). The genetic basis of the roan pattern has also been studied in the horse and pig, but no mutation in the *KITLG* gene has been associated with the roan phenotype, whereas the *KIT* gene has been found as the major candidate for these species (Fontanesi et al., 2010; Cho et al., 2011; Marklund et al., 1999). Cattle and goats belong to the same *Bovidae* family, have the same number of chromosomes and few rearrangements at the genomic level (Schibler et al., 2009; Cribiu et al., 2001; Dong et al., 2013). Therefore, it is not unusual to find similarities both at the genetic and biological level. Polymorphisms in the *KITLG* gene have already been associated with litter size in goat (An et al., 2012, 2015). No phenotypic information about litter size was available in our dataset, and therefore this possible association could not be tested with this present dataset.

Conclusion

The results of this study indicate that a region that includes *KITLG* gene has undergone a loss of variability in two caprine populations that exhibit the roan pattern. Particularly, a SNP located within the gene is monomorphic in the roan group compared to many other breeds. These results indicate that *KITLG* gene is the likely a strong gene candidate for this roan pattern phenotype in goats and should be investigated further.

Funding

This work was supported by State of Iowa funds, Iowa State University, the Ensminger International Animal Agriculture fund; the Italian Ministry of Agriculture (grant INNOVAGEN); the European Union's Seventh Framework Programme (FP7/2010-2014) "NEXTGEN", and by the U.S. Agency for International Development through the Pakistan – U.S. Science & Technology Cooperation Program. The opinions expressed herein are those of the author(s) and do not necessarily reflect the views of the U.S. Agency for International Development.

Acknowledgements

Supplementary materials

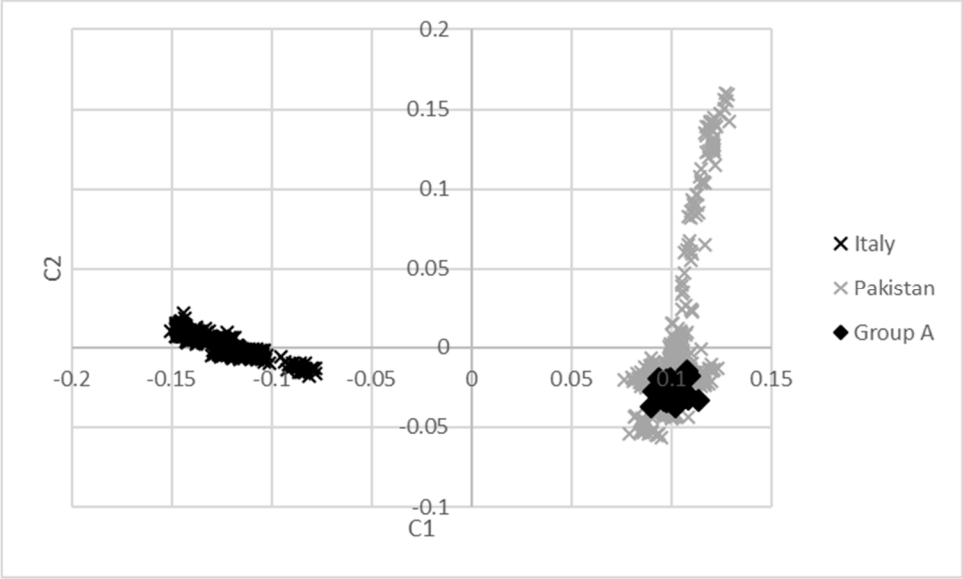


Figure S3.20 - Multidimensional Scaling of Pakistani and Italian breeds, which shows a strong differentiation between Pakistani (right, grey crosses) and Italian (left, black crosses) goats. The Group A breeds (black squares) clearly cluster with other Pakistani goats; modified from Talenti et al. (accepted on Journal of Heredity).

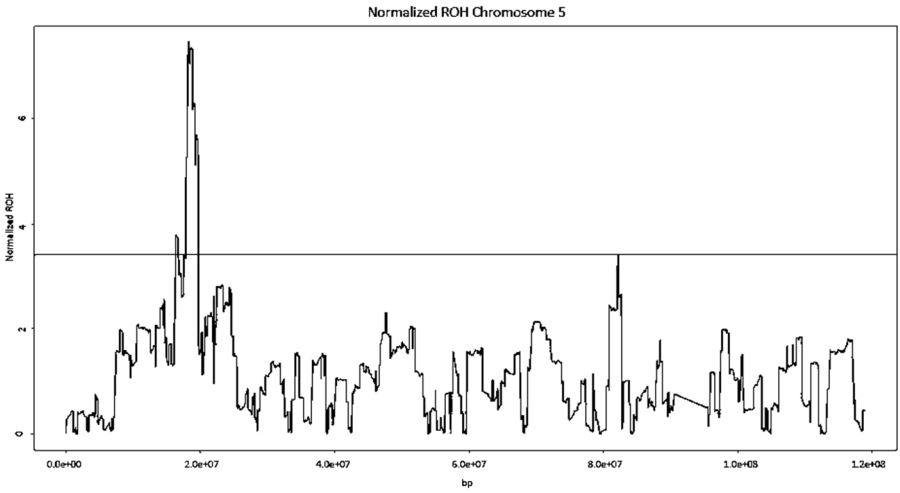


Figure S3.21 - Details of runs of homozygosity on chromosome 5 (chromosome length is reported on the X axis, while normalized H scores are shown on the Y axis). The horizontal line indicates H = 3.4 threshold; modified from Talenti et al. (accepted on Journal of Heredity).

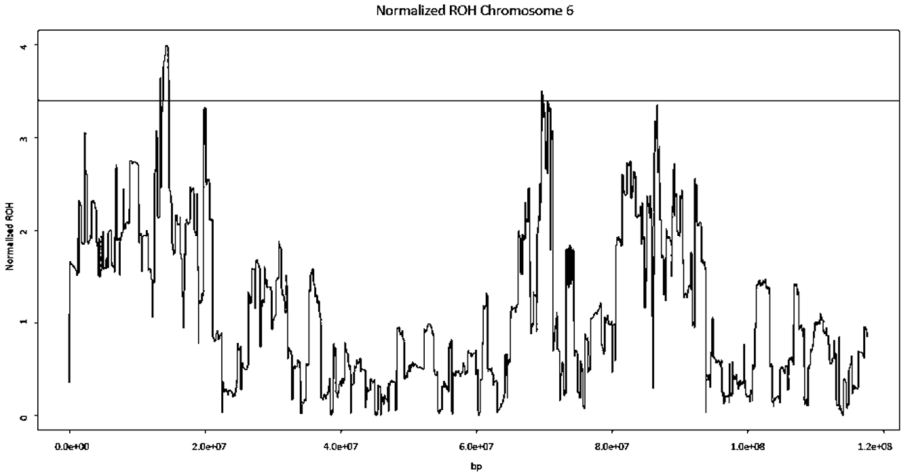


Figure S3.22 - Details of runs of homozygosity on chromosome 6 (chromosome length is reported on the X axis, while normalized H scores are shown on the Y axis). The red line indicates H = 3.4 threshold; modified from Talenti et al. (accepted on Journal of Heredity).

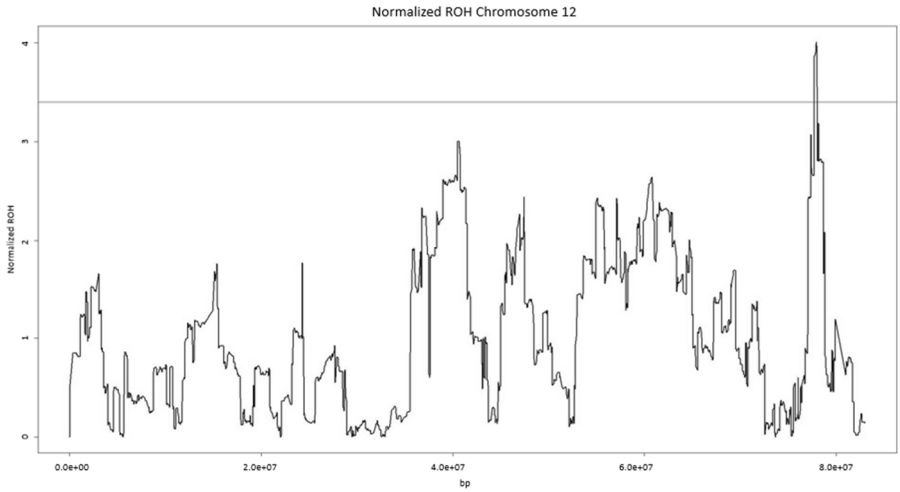


Figure S3.23 - Details of runs of homozygosity on chromosome 13 (chromosome length is reported on the X axis, while normalized H scores are shown on the Y axis). The horizontal line indicates H = 3.4 threshold; modified from Talenti et al. (accepted on Journal of Heredity).

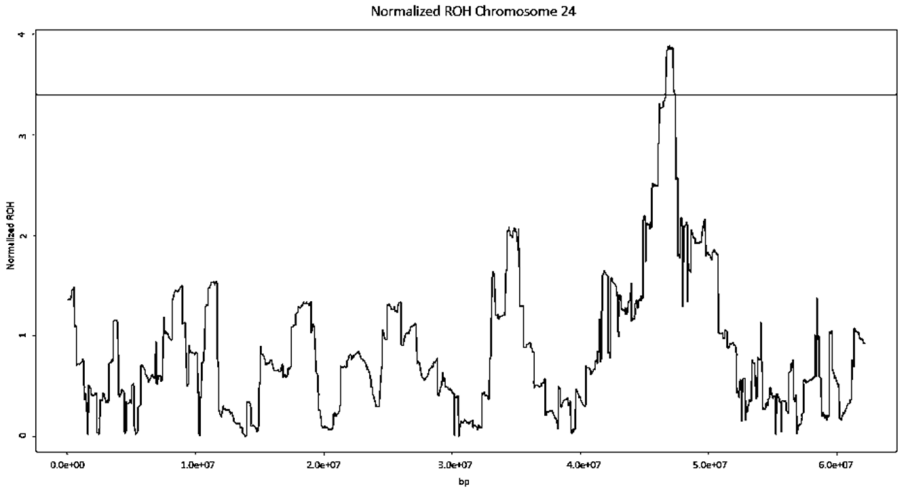


Figure S3.24 - Details of runs of homozygosity on chromosome 24 (chromosome length is reported on the X axis, while normalized H scores are shown on the Y axis). The horizontal line indicates H = 3.4 threshold; modified from Talenti et al. (accepted on Journal of Heredity).

Country	Breed name	N.
Pakistan	Beetal Muki Cheni	12
Pakistan	Barri	35
Italy	Argentata dell'Etna	24
Italy	Aspromontana	24
Italy	Bionda dell'Adamello	24
Italy	Alpine	30
Italy	Ciociara Grigia	19
Italy	Girgentana	24
Italy	Maltese	31
Italy	Nicastrese	24
Italy	Nera di Verzasca	19
Italy	Orobica	23
Italy	Saanen	24
Italy	Sarda	32
Italy	Di Teramo	23
Italy	Valdostana	24
Italy	Valpassiria	24
Pakistan	Balochi	1
Pakistan	Barbari	23
Pakistan	Beetal Fasalabadi	25
Pakistan	Beetal Gujrati	3
Pakistan	Beetal Nokri	25
Pakistan	Spotted Beetal	8
Pakistan	Bugitori	40
Pakistan	Chappar	14
Pakistan	Dera-Din Panah	22
Pakistan	Jattan	24
Pakistan	Jera Kali	25
Pakistan	Kachan	24
Pakistan	Kamori	42
Pakistan	Koh-e-Sulaiman	18
Pakistan	Kurasani	4
Pakistan	Lohri	25
Pakistan	Potohari	16
Pakistan	Maric	3
Pakistan	Nachi	24
Pakistan	Pahari	19
Pakistan	Pateri	37
Pakistan	Tapri	24
Pakistan	Teddi	51
Pakistan	Tharri	16

Table S3.9 - Name and number of animals for each breed utilized in these analyses. All Italian animals, with the exception of Nera di Verzasca, have been previously described in Nicoloso et al. 2015. In red, the breeds with the roan like pattern phenotype; modified from Talenti et al. (accepted on Journal of Heredity).

Chromosome	Start	End	Gene Symbol	Gene Name
5	18670099	18670170	*TRNAC-GCA*	Transfer RNA Cysteine (Anticodon GCA)
5	18684663	18685616	*LOC102174095*	60S ribosomal protein L7-like 1
5	18931313	18931385	*TRNAK-UUU*	Transfer RNA Lysine (Anticodon UUU)
5	18983256	18987813	*DUSP6*	Dual specificity phosphatase 6
5	19164275	19169931	*LOC102172790*	Polypeptide N-acetylgalactosaminyltransferase 4
5	18044632	18151924	*KITLG*	Kit Ligand
5	19064736	19172498	*POC1B*	POC1 Centriolar Protein B
5	19246076	19379946	*ATP2B1*	ATPase Plasma Membrane Ca2+ Transporting 1
5	19528043	19536089	*LOC108636126*	Uncharacterized
6	69547545	69738030	*LNX1*	Ligand Of Numb-Protein X 1
13	77854739	77855626	*LOC108637376*	Uncharacterized
13	77685549	77711085	*LOC102173207*	Transmembrane protein 189
13	77740304	77742226	*CEBPB*	CCAAT/Enhancer Binding Protein Beta
13	77839515	77842199	*LOC108637375*	Uncharacterized

Table S3.10 - List of genes located in the highly homozygous regions specific for the Group A. In red, highlights the genes in the region with the highest H score (H = 0.71) ; modified from Talenti et al. (accepted on Journal of Heredity).

Chromosome	Initial BP	Ending BP	SNP Number
5	18127496	19478621	13
8	2097338	2377431	5
7	55375129	55520204	4
1	138469469	138519003	3
2	114987790	115050138	3
5	33072091	33139827	3
5	53683109	53902299	3
2	84599212	84672518	2
3	38014156	38063070	2
5	28195308	28296662	2
6	69993237	70048354	2
6	70834068	70959770	2
6	86335395	86403884	2
8	60811802	60870729	2
12	34015668	34071415	2
14	83968097	84038953	2
15	16703161	16866787	2
21	57660136	57698228	2

Table S3.11 - Details of the regions with more than 1 consecutive SNP identified by XP-EHH on all the chromosomes. The largest window, located on chromosome 5 that include the KITLG gene is marked in red; modified from Talenti et al. (accepted) ; modified from Talenti et al. (accepted on Journal of Heredity).

Section 4 - Discussion

4.1 General summary

In this thesis, I addressed three main topics related to the application of genomics in ruminant species, with a particular attention to goats.

1. In chapter 3.1 I studied selection signature in 369 animals of 16 Italian goat populations, which identified an interesting signal in the genome of the Valdostana goat. This breed is typical of the Valle d'Aosta, an Italian region where it is raised for milk production and is also used for competitive purposes. Using runs of homozygosity (ROH), Fst, XP-EHH, Rsb and a Bayesian genome-wide association analyses we investigated the genome of this population and identified a 4 Mb region on chromosome 7 that undergone to a strong decrease in variability. This region, that includes a high number of genes related to immune system and response, is unique of this breed. The presence of such a high number of immune-related genes let us assume two main hypotheses: i) the involvement of these animals in non-cruel fighting events also selected these genes playing a role in aggressiveness or ii) the population may have undergone to some disease outbreaks, that boosted the selective pressure on these genes.
2. In chapter 3.2 I developed a new method that could be used to define small sets of markers to assess parentage. This new procedure is not assembly-biased, which is a major pitfall of available methods based on the choice of markers for parentage by a combination of minor allele frequency and linkage disequilibrium. This three-step methodology use a first selection of polymorphic markers, followed by the selection of discriminant SNPs by the multivariate technique Canonical Discriminant Analysis, and finally shrinks the size of the panel by a stepwise regression. This method,

which was tested on three Italian goat breeds, could be of great help not only for the caprine species, but also for these animals that could benefit of cheap genomic tools but present a low reliability of the genome assembly.

3. In chapter 3.3 I studied selection signatures in the genome of 929 animals of 41 Italian and Pakistani goat populations. I identified an interesting signal in the genome of two Pakistani goat breeds, the Barri and the Beetal Muki Cheni, presenting the roan coat color pattern with black and red pigmentation, respectively. The detection have been performed using two different approaches, the Runs of Homozygosity (ROH) and the XP-EHH, two methods that detect events of strong recent selection on the genome. Using these approaches, we identified a 1.7 Mb region on chromosome 7 that includes the *KITLG*. This gene has been previously associated with roan in Belgian Blue cattle, where it is recognized as the causative locus in heterozygous state. Even if not in heterozygosity, our analyses suggest the *KITLG* as candidate gene for the roan pattern in goat species. These results helped in defining the genetic basis of this coat color pattern. In the future, this gene could further be studied to investigate whether it leads to deafness or lethality, similarly to other species. Moreover, in highly standardized breeds it could be evaluated as a possible indicator for breeds or products from breeds that present this peculiar pattern.

4.2 Perspectives

The availability of highthroughput genomic technologies allowed the production of vast amount of data for different species. These new information could be used first to understand the biodiversity of populations but also for other purposes, providing benefits to breeders, breeders' association and customers.

The amount of data produced by projects and consortia is increasing on a daily basis. Moreover, the availability of new technologies providing an even greater amount of data, reduce the senescence time for older approaches. Nonetheless, these data could still be a great source of information, and used in a myriad of fast studies that aims to explore the genetic diversity and the basis of the different phenotypes.

With this in mind, my thesis' aims was to explore the genome by analyzing the genomic data produced by SNP arrays in the past few years using different approaches with different goals. In chapter 3.1, for example, I provided evidences of the genetic uniqueness of the Valdostana breed. These findings give value to this small breed reared in a small Alpine region and to Valdostana breeders association, that could state the genetic heritage of this peculiar population as an important source of biodiversity. Understanding the genetic uniqueness of a population by genomic tools could help in defining the animals with an homogeneous genetic background. However these methods should always account for a measure of the relatedness among animals. In fact, inbreeding depression is a major drawback in several highly standardized breeds. In chapter 3.2, we provided a new method to select markers for parentage assessment. If properly developed, it could be a useful supplement to the standard pedigree information. After the method development, we worked on the selection of markers for parentage assessment in more than 100 worldwide breeds and population (Talenti et al., in preparation). Finally, I tried to unveil the

genetic basis of a coat color pattern, the roan, using two different Pakistani goat breeds as a model. This study identified a gene, the *KITLG*, as candidate for this phenotype. This promising signal need to be further evaluated by new analysis to screen the causal mutation of the phenotype, to define both the allelic variant and the unexpected presence of the mutation at high level of homozygosity. If confirmed, this gene will become a major locus for this coat color pattern. This would help unveiling the genetic basis of pigmentation patterns, one of the first trait standardized in modern and ancient breeds and that rely on a really complex group of metabolic pathways. For their role in breed formation, the of coat color genes proven to be extremely important in product tracing.

The studies presented in this thesis was obtained by the use of mid-density SNP arrays available for the goat species. Despite its affordability, this tool will likely be replaced by the adoption of panels with a higher density once released and by new species-independent technologies already available. This is the case of Next Generation Sequencing, that provides data for an extremely high number of variants, including structural and newly discovered variants. However, due to cost restraints of this technology, the data produced from previous projects will still keep their value thanks to their high number of genotyped animals.

In this context, several research groups are working on integrating genomic data from different technologies (i.e. low density SNP arrays with high density SNP arrays, and then Next Generation Sequencing data) to create enlarged panels of animals with the highest number of markers possible and the lowest waste of resources. This solution is possible thanks to the efficient imputation algorithms that are getting available, allowing reliable estimates of missing genotypes, and therefore allowing the expansion of data from low to high density arrays and then from high density to whole genome information. Consequently, future works will have to be technologically up-to-date, but at the same looking at the past

to not lose all data and efforts spent in producing genotyping data, important heritage of the past consortia.

4.3 Final remarks

Despite its great diffusion, goat remain a species economically relevant mainly in marginal rural areas of the world thanks to its adaptability, and have the potential to become a pivotal player in exploiting the economy of these areas. In this thesis, I provided new evidences that genomic information, if properly analyzed, could efficiently identify breeds' uniqueness and define tools that can help in planning better management strategies in local and marginal farming. Therefore, the implementation of the presented results in cheap genomic tools could be a new challenge to achieve genetic improvement of animal uniqueness and welfare, together with product valorization.

Section 5 - Other Articles

5.1 Published articles

5.1.1 Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip

Published on Genetic Selection Evolution 2015, 47:1–10; DOI: 10.1186/s12711-015-0140-6.

Authors

Nicoloso L, Bomba L, Colli L, Negrini R, Milanesi M, Mazza R, Sechi T, Frattini S, **Talenti A**, Coizet B, Chessa S, Marletta D, Andrea MD, Bordonaro S, Ptak G, Carta A, Pagnacco G, Valentini A, Pilla F, Ajmone-Marsan P, Crepaldi P, Consortium G, D'Andrea M, Bordonaro S, Ptak G, Carta A, Pagnacco G, Valentini A, Pilla F, Ajmone-Marsan P, Crepaldi P and the Italian Goat Consortium

Abstract

Background: Among the European countries, Italy counts the largest number of local goat breeds. Thanks to the recent availability of a medium-density SNP (single nucleotide polymorphism) chip for goat, the genetic diversity of Italian goat populations was characterized by genotyping samples from 14 Italian goat breeds that originate from different geographical areas with more than 50 000 SNPs evenly distributed on the genome.

Results: Analysis of the genotyping data revealed high levels of genetic polymorphism and an underlying North-south geographic pattern of genetic diversity that was highlighted by both the first dimension of the multi-dimensional scaling plot and the Neighbour network reconstruction. We observed a moderate and weak population structure in Northern and Central-Southern breeds, respectively, with pairwise FST values between breeds ranging from 0.013 to 0.164 and 7.49 % of the total variance assigned

to the between-breed level. Only 2.11 % of the variance explained the clustering of breeds into geographical groups (Northern, Central and Southern Italy and Islands).

Conclusions: Our results indicate that the present-day genetic diversity of Italian goat populations was shaped by the combined effects of drift, presence or lack of gene flow and, to some extent, by the consequences of traditional management systems and recent demographic history. Our findings may constitute the starting point for the development of marker-assisted approaches, to better address future breeding and management policies in a species that is particularly relevant for the medium- and long-term sustainability of marginal regions.

5.1.2 Fonni's dog: morphological and genetic characteristics for a breed standard definition

Published on Italian Journal of Animal Science 2016, 16(1):22-30; DOI: 10.1080/1828051X.2016.1248867 PAPER.

Authors

Sechi S, Polli M, Marelli SP, **Talenti A**, Crepaldi P, Fiore F, Spissu N, Dreger DL, Zedda M, Dimauro C, Ostrander EA, Di Cerbo A and Cocco R.

Abstract

Italy is home to several populations of native dogs that reside only in certain demographic regions. Such dog populations have not been under tight selection by humans and, as such, have never been officially recognised as breeds. One such population is the 'Cane Fonnese' or Fonni's Dog, which features uniform morphologic and behavioural traits that reproduce across generations, thus qualifying Fonni's Dog as a true breed eligible for recognition by national or international breed registries. The Fonni's Dog population examined in the present work is native to Sardinia, where they are used as property or livestock guardian dogs. As such, they are greatly appreciated by the local populace. We have carried out morphological evaluations on 200 Fonni's Dogs with the aim of developing a standard breed definition upon which the foundation of the Fonni's Dog breed can be based. We have also reported genetic data of the Fonni's Dog compared to four other established breeds sampled from the same geographic area.

5.1.3 Genome-wide analysis of DNA methylation in hypothalamus and ovary of *Capra hircus*

Published on BMC Genomics 2017, 16(1):22-30; DOI: 10.1080/1828051X.2016.1248867.

Authors

Frattini S, Capra E, Lazzari B, McKay SD, Coizet B, **Talenti A**, Groppetti D, Riccaboni P, Pecile A, Chessa S, Castiglioni B, Williams JL, Pagnacco G, Stella A and Crepaldi P

Abstract

Background: DNA methylation is a frequently studied epigenetic modification due to its role in regulating gene expression and hence in biological processes and in determining phenotypic plasticity in organisms. Rudimentary DNA methylation patterns for some livestock species are publically available: among these, goat methylome deserves to be further explored.

Results: Genome-wide DNA methylation maps of the hypothalamus and ovary from Saanen goats were generated using Methyl-CpG binding domain protein sequencing (MBD-seq). Analysis of DNA methylation patterns indicate that the majority of methylation peaks found within genes are located gene body regions, for both organs. Analysis of the distribution of methylated sites per chromosome showed that chromosome X had the lowest number of methylation peaks. The X chromosome has one of the highest percentages of methylated CpG islands in both organs, and approximately 50% of the CpG islands in the goat epigenome are methylated in hypothalamus and ovary. Organ-specific Differentially Methylated Genes (DMGs) were correlated with the expression levels.

Conclusions: The comparison between transcriptome and methylome in hypothalamus and ovary showed that a higher level of methylation is not accompanied by a higher gene suppression. The genome-wide DNA methylation map for two goat organs produced here is a valuable starting point for studying the involvement of epigenetic modifications in regulating goat reproduction performance.

5.1.4 Polymorphism of the STAT5A , MTNR1A and TNF α genes and their effect on dairy production in *Bubalus bubalis*

Published on Italian Journal of Animal Science 2017; DOI: 10.1080/1828051X.2017.1335181.

Authors

Coizet B, Frattini S, Nicoloso L, Iannuzzi L, Coletta A, **Talenti A**, Minozzi G, Pagnacco G and Crepaldi P

Abstract

The water buffalo is a fundamental resource, especially in developing countries, however, differently from other species, its genetic potential is still poorly investigated. In this work, we per- formed a candidate gene association study for milk composition in 491 female buffaloes. Animals were from four farms located in Southern Italy, where the Out-of-Breeding-Season-Mating technique is usually performed. We analysed three genes: (1) the signal transducer and activator of transcription 5A (STAT5A), (2) the tumour necrosis factor alpha (TNFa)and (3)the melatonin receptor 1A (MTNR1A). We confirmed the mutation at the MTNR1A gene and we found five novel single nucleotide polymorphisms (SNPs): one in the TNFa and four in the STAT5A. No associations were found for the SNPs in the MTNR1A and TNFa genes, while we identified a marked association with milk protein % for a C>T substitution at the STAT5A gene. At this locus, the TT buffaloes showed significantly higher protein percentage in milk. Conversely, this genotype class was the less frequent in the population. Moreover, an A>G substitution at the STAT5A showed an influence on reproductive seasonality, with the advantageous allele most frequent in the population, suggesting a possible effect of selection for this trait. The C>T substitution on STAT5A detected in present study could be used in marker assisted selection of Mediterranean Italian buffalo, and should be

monitored to understand the reasons behind the low frequency of the favourable genotype at this locus and to stop this unfavourable trend in the population.

5.1.5 Investigating the population structure and genetic differentiation of livestock guard dog breeds from Italy

Published on animal 2018; DOI: 10.1017/S1751731117003573.

Authors

Bigi D, Marelli SP, Liotta L, Frattini S, **Talenti A**, Pagnacco G, Polli M and Crepaldi P

Abstract

Livestock guarding dogs are a valuable adjunct to the pastoral community. Having been traditionally selected for their working ability, they fulfil their function with minimal interaction or command from their human owners. In this study, the population structure and the genetic differentiation of three Italian livestock guardian breeds (Sila's Dog, Maremma and Abruzzese Sheepdog and Mannara's Dog) and three functionally and physically similar breeds (Cane Corso, Central Asian Shepherd Dog and Caucasian Shepherd Dog), totalling 179 dogs unrelated at the second generation, were investigated with 18 autosomal microsatellite markers. Values for the number of alleles per locus, observed and expected heterozygosity, Hardy–Weinberg Equilibrium, F stats, Nei's and Reynold's genetic distances, clustering and sub-population formation abilities and individual genetic structures were calculated. Our results show clear breed differentiation, whereby all the considered breeds show reasonable genetic variability despite small population sizes and variable selection schemes. These results provide meaningful data to stakeholders in specific breed and environmental conservation programmes.

5.2 Accepted articles

:. Accepted on animal.

Talenti A, Dreger DL, Frattini S, Polli M, Marelli SP, Harris A, Liotta L, Cocco R, Hogan A, Bigi D, Caniglia R, Parker HG, Pagnacco G, Ostrander EA and Crepaldi P: Studies of modern Italian dog populations reveals multiple patterns for domestic breed evolution. Accepted on Ecology and Evolution.

5.3 Submitted articles

Capra E, Lazzari B, Frattini S, Ajmone-Marsan P, Castiglioni B, Chessa S, Coizet B, Crepaldi P, Pagnacco G, **Talenti A**, Williams J, and Stella A: Distribution of ncRNAs expression across hypothalamic-pituitary-gonadal axis in Capra hircus. Submitted to BMC Genomics.

Talenti A, Palhière I, Tortereau F, Pagnacco G, Stella A, Nicolazzi EL, Crepaldi P, Tosser-Klopp G and ADAPTmap Consortium: Functional SNP panel for parentage assessment and assignment in worldwide goat breeds. Submitted to Genetic Selection Evolution.

Bertolini F, Servin B, **Talenti A**, Rochat E, Kim ES, Oget C, Palhière I, Crisà A, Catillo G, Steri R, Amills M, Colli L, Marras G, Milanesi M, Nicolazzi EL, Rosen BD, Van Tassell CP, Guldbrandtsen B, Sonstegard TS, Tosser-Klopp G, Stella A, Rothschild MF, Joost S, Crepaldi P and the ADAPTmap consortium: Signatures of selection and environmental adaptation across the goat genome post domestication. Submitted to Genetic Selection Evolution.

Colli L, Milanesi M, **Talenti A**, Bertolini F, Chen M, Crisà A, Daly K, Del Corvo M, Guldbrandtsen B, Lenstra JA, Rosen BD, Vajana E, Catillo G, Joost S, Nicolazzi EL, Rochat E, Rothschild MF, Servin B, Sonstegard T, Steri R, Van Tassell CP, Ajmone-Marsan P, Crepaldi P, Stella A, the ADAPTmap Consortium: Drawing up worldwide goat diversity and post-domestication history. Submitted to Genetic Selection Evolution.

Section 6 - References

1000 Bulls Genome Consortium, T. 2013. 1000 Bulls Genome Project.

3SR. 2014. 3SR.

Abdi, H., and L.J. Williams. 2010. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* 2:433–459. doi:10.1002/wics.101.

Achilli, A., A. Olivieri, M. Pellecchia, C. Uboldi, L. Colli, N. Al-Zahery, M. Accetturo, M. Pala, B.H. Kashani, U.A. Perego, V. Battaglia, S. Fornarino, J. Kalamati, M. Houshmand, R. Negrini, O. Semino, M. Richards, V. Macaulay, L. Ferretti, H.J. Bandelt, P. Ajmone-Marsan, and A. Torroni. 2008. Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr. Biol.* 18:157–158. doi:10.1016/j.cub.2008.01.019.

ADAPTmap. 2014. ADAPTmap.

Afgan, E., D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, and J. Goecks. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44:W3–W10. doi:10.1093/nar/gkw343.

Ajmone-Marsan, P., E. Milanesi, and R. Negrini. 2004. Breed traceability using molecular methods. *Proc. seventh world Conf. Brown Swiss cattle breeders*. 7–10.

Akey, J.M., A.L. Ruhe, D.T. Akey, A.K. Wong, C.F. Connelly, J. Madeoy, T.J. Nicholas, and M.W. Neff. 2010. Tracking footprints of artificial selection in the dog genome. *Proc. Natl. Acad. Sci.* 107:1160–1165. doi:10.1073/pnas.0909918107.

Al-Atiyat, R.M. 2015. The power of 28 microsatellite markers for parentage testing in sheep. *Electron. J. Biotechnol.* 18:116–121. doi:10.1016/j.ejbt.2015.01.001.

Alexander, D.H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1655–1664. doi:10.1101/gr.094052.109.

An, X.P., J.X. Hou, T.Y. Gao, Y.N. Lei, Y.X. Song, J.G. Wang, and B.Y. Cao. 2015. Association analysis between variants in KITLG gene and litter size in goats. *Gene*. 558:126–130. doi:10.1016/j.gene.2014.12.058.

An, X.P., J.X. Hou, G. Li, Y.X. Song, J.G. Wang, Q.J. Chen, Y.H. Cui, Y.F. Wang, and B.Y. Cao. 2012. Polymorphism identification in the goat KITLG gene and association analysis with litter size. *Anim. Genet.* 43:104–107. doi:10.1111/j.1365-2052.2011.02219.x.

Anderson, E.C. 2005. The Power of Single-Nucleotide Polymorphisms for Large-Scale Parentage Inference. *Genetics*. 172:2567–2582. doi:10.1534/genetics.105.048074.

Anderson, R. 2014. Development of a High Density (600K) Illumina Ovine SNP Chip and Its Use to Fine Map the Yellow Fat Locus.

Andersson, L., and M. Georges. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat. Rev. Genet.* 5:202–212. doi:10.1038/nrg1294.

Arsenos, G. 2016. iSAGE Project.

Association Comité Régional Batailles des Chèvres. 2016. Batailles de Chevre.

ASSONAPA. 2014. Valdostana Breed Standard.

Aulchenko, Y. 2009. Genome-wide association analysis with GenABEL : quck start guide for the impatient. 1–19.

Baran, Y., I. Quintela, A. Carracedo, B. Pasaniuc, and E. Halperin. 2013. Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am. J. Hum. Genet.* 92:882–94. doi:10.1016/j.ajhg.2013.04.023.

Barcos, L.O. 2001. Recent developments in animal identification and the traceability of animal products in international trade. *Rev. Sci. Tech.* 20:640–51.

Becker, D., M. Otto, P. Ammann, I. Keller, C. Drögemüller, and T. Leeb. 2015. The brown coat colour of Coppernecked goats is Asso. with a non-synonymous variant at the TYRP1 locus on chromosome 8. *Anim. Genet.* 46:50–54. doi:10.1111/age.12240.

Benjelloun, B., F.J. Alberto, I. Streeter, F. Boyer, E. Coissac, S. Stucki, M. BenBati, M. Ibnelbachyr, M. Chentouf, A. Bechchari, K. Leempoel, A. Alberti, S. Engelen, A. Chikhi, L. Clarke, P. Flicek, S. Joost, P. Taberlet, and F. Pompanon. 2015. Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Front. Genet.* 6:107. doi:10.3389/fgene.2015.00107.

Bertolini, F., G. Galimberti, D.G. Calò, G. Schiavo, D. Matassino, and L. Fontanesi. 2015. Combined use of principal component analysis and random forests identify population-informative single nucleotide polymorphisms: application in cattle breeds. *J. Anim. Breed. Genet.* 132:346–356. doi:10.1111/jbg.12155.

Bertolini, F., B. Gandolfi, E.S. Kim, B. Haase, L.A. Lyons, and M.F. Rothschild. 2016. Evidence of selection signatures that shape the Persian cat breed. *Mamm. Genome*. 27:144–155. doi:10.1007/s00335-016-9623-1.

Bhatia, G., N. Patterson, S. Sankararaman, and A.L. Price. 2013. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* 23:1514–21. doi:10.1101/gr.154831.113.

Bhatt, S., and A. Siegel. 2006. Potentiating role of interleukin 2 (IL-2) receptors in the

midbrain periaqueductal gray (PAG) upon defensive rage behavior in the cat: Role of neurokinin NK1 receptors. *Behav. Brain Res.* 167:251–260. doi:10.1016/j.bbr.2005.09.011.

Bickhart, D., S. Koren, A. Phillippy, T.P.L. Smith, and and T.S. 2015 Liachko, Ivan, J. N. Burton, B. Sayre, H. J. Huson, S. G. Schroeder, C. P. VanTassell, Sonstegard. The use of PacBio and Hi-C data in de novo assembly of the goat genome. - Pacific Biosciences.

Bickhart, D., S. Koren, A. Phyllippy, T. Smith, J. Burton, I. Liachko, T. Sonstegard, C. Van Tassell, H. Huson, and S. Schroeder. 2015. The use of PacBio and Hi-C data in denovo assembly of the goat genome Selecting an animal for sequencing Current CHI reference genome. PAG, San Diego. 2–4.

Bickhart, D.M., B.D. Rosen, S. Koren, B.L. Sayre, A.R. Hastie, S. Chan, J. Lee, E.T. Lam, I. Liachko, S.T. Sullivan, J.N. Burton, H.J. Huson, J.C. Nystrom, C.M. Kelley, J.L. Hutchison, Y. Zhou, J. Sun, A. Crisà, F. Abel Ponce de León, J.C. Schwartz, J.A. Hammond, G.C. Waldbieser, S.G. Schroeder, G.E. Liu, M.J. Dunham, J. Shendure, T.S. Sonstegard, A.M. Phillippy, C.P. Van Tassell, and T.P. L Smith. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Publ. Gr.* 49:4–3. doi:10.1038/ng.3802.

Biffani, S., N. Morandi, V. Locatelli, D. Pravettoni, A. Boccardo, A. Stella, E.L. Nicolazzi, and F. Biscarini. 2014. Adding evidence for a role of the SLITRK gene family in the pathogenesis of left displacement of the abomasum in Holstein-Friesian dairy cows. *Livest. Sci.* 167:104–109. doi:10.1016/j.livsci.2014.05.002.

Bjørnstad, G., and K.H. Røed. 2001. Breed demarcation and potential for breed allocation of horses assessed by microsatellite markers. *Anim. Genet.* doi:10.1046/j.1365-2052.2001.00705.x.

Bollongino, R., J. Elsner, J.-D. Vigne, and J. Burger. 2008. Y-SNPs Do Not Indicate Hybridisation between European Aurochs and Domestic Cattle. *PLoS One*. 3:e3418. doi:10.1371/journal.pone.0003418.

Bolormaa, S., B.J. Hayes, K. Savin, R. Hawken, W. Barendse, P.F. Arthur, R.M. Herd, and M.E. Goddard. 2011. Genome-wide association studies for feedlot and growth traits in cattle. *J. Anim. Sci.* doi:10.2527/jas.2010-3079.

Bolormaa, S., a. Ruvinsky, S. Walkden-Brown, and J. van der Werf. 2008. DNA-based parentage verification in two Australian goat herds. *Small Rumin. Res.* 80:95–100. doi:10.1016/j.smallrumres.2008.08.005.

Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, and M. Sancristobal. 2010. Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics*. 186:241–62. doi:10.1534/genetics.104.117275.

Bovine HapMap Consortium, T.B.H., R.A. Gibbs, J.F. Taylor, C.P. Van Tassell, W. Barendse, K.A. Eversole, C.A. Gill, R.D. Green, D.L. Hamernik, S.M. Kappes, S. Lien, L.K. Matukumalli, J.C. McEwan, L. V Nazareth, R.D. Schnabel, G.M. Weinstock, D.A. Wheeler, P. Ajmone-Marsan, P.J. Boettcher, A.R. Caetano, J.F. Garcia, O. Hanotte, P. Mariani, L.C. Skow, T.S. Sonstegard, J.L. Williams, B. Diallo, L. Hailemariam, M.L. Martinez, C.A. Morris, L.O.C. Silva, R.J. Spelman, W. Mulatu, K. Zhao, C.A. Abbey, M. Agaba, F.R. Araujo, R.J. Bunch, J. Burton, C. Gorni, H. Olivier, B.E. Harrison, B. Luff, M.A. Machado, J. Mwakaya, G. Plastow, W. Sim, T. Smith, M.B. Thomas, A. Valentini, P. Williams, J. Womack, J.A. Woolliams, Y. Liu, X. Qin, K.C. Worley, C. Gao, H. Jiang, S.S. Moore, Y. Ren, X.-Z. Song, C.D. Bustamante, R.D. Hernandez, D.M. Muzny, S. Patil, A. San Lucas, Q. Fu, M.P. Kent, R. Vega, A. Matukumalli, S. McWilliam, G. Sclep, K. Bryc, J. Choi, H. Gao, J.J. Grefenstette, B. Murdoch, A. Stella, R. Villa-Angulo, M. Wright, J. Aerts, O. Jann, R. Negrini, M.E. Goddard, B.J. Hayes, D.G. Bradley, M. Barbosa da Silva, L.P.L. Lau, G.E. Liu, D.J. Lynn, F. Panzitta, and K.G. Dodds. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*. 324:528–32. doi:10.1126/science.1167936.

Boyko, A.R. 2011. The domestic dog: Man's best friend in the genomic era. *Genome Biol.* 12:216. doi:10.1186/gb-2011-12-2-216.

Boyko, A.R., P. Quignon, L. Li, J.J. Schoenebeck, J.D. Degenhardt, K.E. Lohmueller, K. Zhao, A. Brisbin, H.G. Parker, B.M. vonHoldt, M. Cargill, A. Auton, A. Reynolds, A.G. Elkahloun, M. Castelhano, D.S. Mosher, N.B. Sutter, G.S. Johnson, J. Novembre, M.J. Hubisz, A. Siepel, R.K. Wayne, C.D. Bustamante, and E.A. Ostrander. 2010. A Simple Genetic Architecture Underlies Morphological Variation in Dogs. *PLoS Biol.* 8:e1000451. doi:10.1371/journal.pbio.1000451.

Brito, L.F., M. Jafarikia, D.A. Grossi, J.W. Kijas, L.R. Porto-Neto, R. V. Ventura, M. Salgorzaei, and F.S. Schenkel. 2015. Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats. *BMC Genet.* 16:67. doi:10.1186/s12863-015-0220-1.

Brito, L.F., J.W. Kijas, R. V. Ventura, M. Sargolzaei, L.R. Porto-Neto, A. Cánovas, Z. Feng, M. Jafarikia, and F.S. Schenkel. 2017. Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. *BMC Genomics*. 18:229. doi:10.1186/s12864-017-3610-0.

Brooks, S.A., and E. Bailey. 2005. Exon skipping in the KIT gene causes a Sabino spotting pattern in horses. *Mamm. Genome*. 16:893–902. doi:10.1007/s00335-005-2472-y.

Browning, B.L. 2011. Beagle 3.3.2. 1–30.

Browning, B.L., and S.R. Browning. 2008. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84:210–223. doi:10.1016/j.ajhg.2009.01.005.

Browning, B.L., and S.R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 194:459–471. doi:10.1534/genetics.113.150029.

Browning, B.L.L., and S.R.R. Browning. 2016. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* 98:116–126. doi:10.1016/j.ajhg.2015.11.020.

Browning, S.R., and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81:1084–97. doi:10.1086/521987.

Burren, A., M. Neuditschko, H. Signer-Hasler, M. Frischknecht, I. Reber, F. Menzi, C. Drögemüller, and C. Flury. 2016. Genetic diversity analyses reveal first insights into breed-specific selection signatures within Swiss goat breeds. *Anim. Genet.* 47:727–739. doi:10.1111/age.12476.

Cai, Z., N.J. Camp, L. Cannon-Albright, and A. Thomas. 2011. Identification of regions of positive selection using Shared Genomic Segment analysis. *Eur. J. Hum. Genet.* 19:667–71. doi:10.1038/ejhg.2010.257.

Capomaccio, S., M. Milanesi, L. Bomba, K. Cappelli, E.L. Nicolazzi, J.L. Williams, P. Ajmone-Marsan, and B. Stefanon. 2015. Searching new signals for production traits through gene-based association analysis in three Italian cattle breeds. *Anim. Genet.* 46:n/a-n/a. doi:10.1111/age.12303.

Cattle Genome Sequencing International Consortium, T. 2015. Btau 5.0.1 Genome Assembly.

Cavalli-Sforza, L.L., and A.W.F. Edwards. 1967. Phylogenetic Analysis Models and Estimation Procedures. *Am. J. Hum. Genet.* 19:233–257.

Caye, K., T.M. Deist, H. Martins, O. Michel, and O. François. 2016. TESS3: Fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* 16:540–548. doi:10.1111/1755-0998.12471.

Chang, C.C., C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, and J.J. Lee. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 4:7. doi:10.1186/s13742-015-0047-8.

Chen, E.Y., C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, N.R. Clark, and A. Ma'ayan. 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 14:128. doi:10.1186/1471-2105-14-128.

Chen, H., N. Patterson, and D. Reich. 2010. Population differentiation as a test for selective sweeps. *Genome Res.* 20:393–402. doi:10.1101/gr.100545.109.

Chen, X., and H. Ishwaran. 2013. Random Forests for Genomic Data Analysis. *Genomics*.

99:323–329. doi:10.1016/j.ygeno.2012.04.003.Random.

Cho, I.-C., T. Zhong, B.-Y. Seo, E.-J. Jung, C.-K. Yoo, J.-H. Kim, J.-B. Lee, H.-T. Lim, B.-W. Kim, J.-H. Lee, M.-S. Ko, and J.-T. Jeon. 2011. Whole-genome association study for the roan coat color in an intercrossed pig population between Landrace and Korean native pig. *Genes Genomics*. 33:17–23. doi:10.1007/s13258-010-0108-4.

Ciani, E., P. Crepaldi, L. Nicoloso, E. Lasagna, F.M. Sarti, B. Moioli, F. Napolitano, a. Carta, G. Usai, M. D'Andrea, D. Marletta, R. Ciampolini, V. Riggio, M. Occidente, D. Matassino, D. Kompan, P. Modesto, N. Macciotta, P. Ajmone-Marsan, and F. Pilla. 2014. Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. *Anim. Genet.* 45:256–266. doi:10.1111/age.12106.

Colli, L., H. Lancioni, I. Cardinali, A. Olivieri, M.R. Capodiferro, M. Pellecchia, M. Rzepus, W. Zamani, S. Naderi, F. Gandini, S.M.F. Vahidi, S. Agha, E. Randi, V. Battaglia, M.T. Sardina, B. Portolano, H.R. Rezaei, P. Lymberakis, F. Boyer, E. Coissac, F. Pompanon, P. Taberlet, P. Ajmone Marsan, and A. Achilli. 2015. Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability. *BMC Genomics*. 16:1115. doi:10.1186/s12864-015-2342-2.

Colussi, S., P. Sacchi, I. Cristoferi, M.G. Maniaci, S. Maione, M. V. Riina, R. Orusa, S. Peletto, M. Caramelli, R. Rasero, and P.L. Acutis. 2008. Genetic variability of the PRNP gene in Piemonte region goat breeds and in Valdostana breed. *Large Anim. Rev.* 14:11–14.

Commission of the European Coomunities. 2000. White Paper on Food Safety.

Conolly, J., K. Manning, S. Colledge, K. Dobney, and S. Shennan. 2012. Species distribution modelling of ancient cattle from early Neolithic sites in SW Asia and Europe. *The Holocene*. 22:997–1010. doi:10.1177/0959683612437871.

Corander, J., P. Waldmann, P. Marttinen, and M.J. Sillanpaa. 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*. 20:2363–2369. doi:10.1093/bioinformatics/bth250.

Cornuet, J.M., S. Piry, G. Luikart, A. Estoup, and M. Solignac. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*. doi:10.1038/368455a0.

Council of Europe. 1986. European Convention for the Protection of Vertebrate Animals used for Experime ntal and Other Scientific Purposes CETS 123. *Strasbourg*.

Cozzi, G., and E. Ragno. 2003. Meat Production and Market in Italy. 68:71–77.

Cribiu, E.P., D. Di Berardino, G.P. Di Meo, A. Eggen, D.S. Gallagher, I. Gustavsson, H. Hayes, L. Iannuzzi, C.P. Popescu, J. Rubes, S. Schmutz, G. Stranzinger, A. Vaiman,

and J. Womack. 2001. International System for Chromosome Nomenclature of Domestic Bovids (ISCNDB 2000). *Cytogenet. Cell Genet.* 92:283–99. doi:56917.

Daetwyler, H.D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R.F. Brøndum, X. Liao, A. Djari, S.C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P.J. Bowman, D. Coote, A.J. Chamberlain, C. Anderson, C.P. VanTassell, I. Hulsegge, M.E. Goddard, B. Guldbrandtsen, M.S. Lund, R.F. Veerkamp, D. a Boichard, R. Fries, and B.J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* 46:858–865. doi:10.1038/ng.3034.

Dalvit, C., M. De Marchi, and M. Cassandro. 2007. Genetic traceability of livestock products: A review. *Meat Sci.* 77:437–449. doi:10.1016/j.meatsci.2007.05.027.

Decker, J.E., S.D. McKay, M.M. Rolf, J. Kim, A. Molina Alcalá, T.S. Sonstegard, O. Hanotte, A. Götherström, C.M. Seabury, L. Praharani, M.E. Babar, L. Correia de Almeida Regitano, M.A. Yildiz, M.P. Heaton, W.-S. Liu, C.-Z. Lei, J.M. Reecy, M. Saif-Ur-Rehman, R.D. Schnabel, and J.F. Taylor. 2014. Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLoS Genet.* 10:e1004254. doi:10.1371/journal.pgen.1004254.

Dekkers, J. 2012. Application of Genomics Tools to Animal Breeding. *Curr. Genomics*. 13:207–212. doi:10.2174/1389202212800543057.

Demirci, S., E. Koban Baştanlar, N.D. Da•taş, E. Pişkin, A. Engin, F. Özer, E. Yüncü, Ş.A. Do•an, and I. Togan. 2013. Mitochondrial DNA diversity of modern, ancient and wild sheep (Ovis gmelinii anatolica) from Turkey: New insights on the evolutionary history of sheep. *PLoS One*. 8. doi:10.1371/journal.pone.0081952.

Dimauro, C., M. Cellesi, R. Steri, G. Gaspa, S. Sorbolini, a. Stella, and N.P.P. Macciotta. 2013. Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. *Anim. Genet.* 44:377–382. doi:10.1111/age.12021.

Dimauro, C., L. Nicoloso, M. Cellesi, N.P.P. Macciotta, E. Ciani, B. Moioli, F. Pilla, and P. Crepaldi. 2015. Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. *Small Rumin. Res.* 128:27–33. doi:10.1016/j.smallrumres.2015.05.001.

Dong, Y., M. Xie, Y. Jiang, N. Xiao, X. Du, W. Zhang, G. Tosser-Klopp, J. Wang, S. Yang, J. Liang, W. Chen, J. Chen, P. Zeng, Y. Hou, C. Bian, S. Pan, Y. Li, X. Liu, W. Wang, B. Servin, B. Sayre, B. Zhu, D. Sweeney, R. Moore, W. Nie, Y. Shen, R. Zhao, G. Zhang, J. Li, T. Faraut, J. Womack, Y. Zhang, J. Kijas, N. Cockett, X. Xu, S. Zhao, J. Wang, and W. Wang. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat. Biotechnol.* 31:135–141. doi:10.1038/nbt.2478.

Druet, T., N. Ahariz, N. Cambisano, N. Tamma, C. Michaux, W. Coppieters, C. Charlier, and M. Georges. 2014. Selection in action : dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle. *BMC Genomics*. 15:1–12. doi:10.1186/1471-2164-15-796.

Du, X.Y., J.E. Womack, K.E. Owens, J.S. Elliott, B. Sayre, P.J. Bottcher, D. Milan, M. Garcia Podesta, S.H. Zhao, and M. Malek. 2012. A whole-genome radiation hybrid panel for goat. *Small Rumin. Res.* 105:114–116. doi:10.1016/j.smallrumres.2011.11.023.

Dürig, N., R. Jude, H. Holl, S.A. Brooks, C. Lafayette, V. Jagannathan, and T. Leeb. 2017. Whole genome sequencing reveals a novel deletion variant in the *KIT* gene in horses with white spotted coat colour phenotypes. *Anim. Genet.* 48:483–485. doi:10.1111/age.12556.

Eck, S.H., A. Benet-Pagès, K. Flisikowski, T. Meitinger, R. Fries, and T.M. Strom. 2009. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. *Genome Biol.* 10:R82. doi:10.1186/gb-2009-10-8-r82.

Edea, Z., H. Dadi, T. Dessie, I.H. Kim, and K.S. Kim. 2017. Association of MITF loci with coat color spotting patterns in Ethiopian cattle. *Genes and Genomics*. 39:285–293. doi:10.1007/s13258-016-0493-4.

Edwards, C.J., R. Bollongino, A. Scheu, A. Chamberlain, A. Tresset, J.-D. Vigne, J.F. Baird, G. Larson, S.Y.. Ho, T.H. Heupink, B. Shapiro, A.R. Freeman, M.G. Thomas, R.-M. Arbogast, B. Arndt, L. Bartosiewicz, N. Benecke, M. Budja, L. Chaix, A.M. Choyke, E. Coqueugniot, H.-J. Dohle, H. Goldner, S. Hartz, D. Helmer, B. Herzig, H. Hongo, M. Mashkour, M. Ozdogan, E. Pucher, G. Roth, S. Schade-Lindig, U. Schmolcke, R.J. Schulting, E. Stephan, H.-P. Uerpmann, I. Voros, B. Voytek, D.G. Bradley, and J. Burger. 2007. Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proc. R. Soc. B Biol. Sci.* 274:1377–1385. doi:10.1098/rspb.2007.0020.

Elhaik, E., T. Tatarinova, D. Chebotarev, I.S. Piras, C. Maria Calò, A. De Montis, M. Atzori, M. Marini, S. Tofanelli, P. Francalacci, L. Pagani, C. Tyler-Smith, Y. Xue, F. Cucca, T.G. Schurr, J.B. Gaieski, C. Melendez, M.G. Vilar, A.C. Owings, R. Gómez, R. Fujita, F.R. Santos, D. Comas, O. Balanovsky, E. Balanovska, P. Zalloua, H. Soodyall, R. Pitchappan, A. Ganeshprasad, M. Hammer, L. Matisoo-Smith, R.S. Wells, and T.G. Genographic Consortium. 2014. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat. Commun.* 5:3513. doi:10.1038/ncomms4513.

Elsik, C.G., R.L. Tellam, and K.C. Worley. 2009. The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution The Bovine Genome Sequencing and Analysis Consortium *. doi:10.1126/science.1169588.

Ewens, W.J. 1972. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*

3:87–112. doi:10.1016/0040-5809(72)90035-4.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*. 164:1567–87.

FAO. 2013. The State of Food and Agriculture 2012. 2.

Fariello, M.-I., B. Servin, G. Tosser-Klopp, R. Rupp, C. Moreno, M.S. Cristobal, S. Boitard, and S. Boitard. 2014. Selection Signatures in Worldwide Sheep Populations. *PLoS One*. 9:e103813. doi:10.1371/journal.pone.0103813.

Fay, J.C., and C.I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics*. 155:1405–1413.

Felius, M., M.L. Beerling, D.S. Buchanan, B. Theunissen, P.A. Koolmees, and J.A. Lenstra. 2014. On the history of cattle genetic resources. *Diversity*. 6:705–750. doi:10.3390/d6040705.

Fernández, M.E., D.E. Goszczynski, J.P. Lirón, E.E. Villegas-Castagnasso, M.H. Carino, M. V Ripoli, A. Rogberg-Muñoz, D.M. Posik, P. Peral-García, and G. Giovambattista. 2013. Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. *Genet. Mol. Biol.* 36:185–91. doi:10.1590/S1415-47572013000200008.

Fernando, R.L., and D.J. Garrick. 2009. GenSel – User Manual for a Portfolio of Genomic Selection Related Analyses. Animal Breeding and Genetics.

Finlay, E.K., D.P. Berry, B. Wickham, E.P. Gormley, and D.G. Bradley. 2012. A Genome Wide Association Scan of Bovine Tuberculosis Susceptibility in Holstein-Friesian Dairy Cattle. *PLoS One*. 7:e30545. doi:10.1371/journal.pone.0030545.

Fisher, P.J., B. Malthus, M.C. Walker, G. Corbett, and R.J. Spelman. 2009. The number of single nucleotide polymorphisms and on-farm data required for whole-herd parentage testing in dairy cattle herds. *J. Dairy Sci.* 92:369–374. doi:10.3168/jds.2008-1086.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7:179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

Fleming, D.S., J.E. Koltes, A.D. Markey, C.J. Schmidt, C.M. Ashwell, M.F. Rothschild, M.E. Persia, J.M. Reecy, and S.J. Lamont. 2016. Genomic analysis of Ugandan and Rwandan chicken ecotypes using a 600 k genotyping array. *BMC Genomics*. 17:407. doi:10.1186/s12864-016-2711-5.

Flicek, P., and E. Birney. 2010. Sense from sequence reads: methods for alignment and assembly. *Nat. Methods*. 7:479–479. doi:10.1038/nmeth0610-479b.

Fontanesi, L., F. Beretti, V. Riggio, S. Dall'Olio, E.G. González, R. Finocchiaro, R. Davoli, V. Russo, and B. Portolano. 2009a. Missense and nonsense mutations in melanocortin 1 receptor (MC1R) gene of different goat breeds: association with red and black coat colour phenotypes but with unexpected evidences. *BMC Genet.* 10:47. doi:10.1186/1471-2156-10-47.

Fontanesi, L., F. Beretti, V. Riggio, E. Gómez González, S. Dall'Olio, R. Davoli, V. Russo, and B. Portolano. 2009b. Copy Number Variation and Missense Mutations of the Agouti Signaling Protein (*ASIP)* Gene in Goat Breeds with Different Coat Colors. *Cytogenet. Genome Res.* 126:333–347. doi:10.1159/000268089.

Fontanesi, L., E. D'Alessandro, E. Scotti, L. Liotta, A. Crovetti, V. Chiofalo, and V. Russo. 2010. Genetic heterogeneity and selection signature at the KIT gene in pigs showing different coat colours and patterns. *Anim. Genet.* 41:478–492. doi:10.1111/j.1365-2052.2010.02054.x.

Frichot, E., F. Mathieu, T. Trouillon, G. Bouchard, and O. François. 2014. Fast and efficient estimation of individual ancestry coefficients. *Genetics*. 196:973–983. doi:10.1534/genetics.113.160572.

Galinsky, K.J., G. Bhatia, P.R. Loh, S. Georgiev, S. Mukherjee, N.J. Patterson, and A.L. Price. 2016. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* 98:456–472. doi:10.1016/j.ajhg.2015.12.022.

Gautier, M., A. Klassmann, and R. Vitalis. 2017. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* 17:78–90. doi:10.1111/1755-0998.12634.

Gautier, M., D. Laloë, and K. Moazami-Goudarzi. 2010. Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds. *PLoS One*. 5:e13038. doi:10.1371/journal.pone.0013038.

Gautier, M., and R. Vitalis. 2012. Rehh An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*. 28:1176–1177. doi:10.1093/bioinformatics/bts115.

Glowatzki-Mullis, M.-L., C. Gaillard, G. Wigger, and R. Fries. 1995. Microsatellite-based parentage control in cattle. *Anim. Genet.* 26:7–12. doi:10.1111/j.1365-2052.1995.tb02612.x.

Goffaux, F., B. China, L. Dams, A. Clinquart, and G. Daube. 2005. Development of a genetic traceability test in pig based on single nucleotide polymorphism detection. doi:10.1016/j.forsciint.2005.02.013.

Goldstein, D.B., and D.D. Pollock. 1997. Launching Microsatellites: A Review of Mutation

Processes and Methods of Phylogenetic Inference. *J. Hered.* 88:335–342.

Goldstein, D.B., A. Ruiz bares, L. Luca Cavalli-Sforzaf, and M.W. Feldman. 1995. An Evaluation of Genetic Distances for Use With Microsatellite Loci. *Genetics*. 139:463–471.

Gossner, A.G., and J. Hopkins. 2015. The effect of PrP(Sc) accumulation on inflammatory gene expression within sheep peripheral lymphoid tissue. *Vet. Microbiol.* 181:204–11. doi:10.1016/j.vetmic.2015.10.013.

Groeneveld, L.F., J. a. Lenstra, H. Eding, M. a. Toro, B. Scherf, D. Pilling, R. Negrini, E.K. Finlay, H. Jianlin, E. Groeneveld, and S. Weigend. 2010. Genetic diversity in farm animals - a review. *Anim. Genet.* 41:6–31. doi:10.1111/j.1365-2052.2010.02038.x.

Grossman, S.R., I. Shylakhter, E.K. Karlsson, E.H. Byrne, S. Morales, G. Frieden, E. Hostetter, E. Angelino, M. Garber, O. Zuk, E.S. Lander, S.F. Schaffner, and P.C. Sabeti. 2010. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science (80-.).* 327:883–886. doi:10.1126/science.1183863.

Guenther, C.A., B. Tasic, L. Luo, M.A. Bedell, and D.M. Kingsley. 2014. A molecular basis for classic blond hair color in Europeans. *Nat. Genet.* 46:748–52. doi:10.1038/ng.2991.

Hall, J.G., and J. Clutton-Brock. 1989. Two hundred years of British farm livestock. *Two hundred years Br. farm livestock.*

Hanchard, N.A., K.A. Rockett, C. Spencer, G. Coop, M. Pinder, M. Jallow, M. Kimber, G. McVean, R. Mott, and D.P. Kwiatkowski. 2006. Screening for recently selected alleles by analysis of human haplotype similarity. *Am. J. Hum. Genet.* 78:153–9. doi:10.1086/499252.

Hayes, B.J. 2011. Technical note: Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. *J. Dairy Sci.* 94:2114–2117. doi:10.3168/jds.2010-3896.

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41:51. doi:10.1186/1297-9686-41-51.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92:433–43. doi:10.3168/jds.2008-1646.

Heaton, M.P., G.P. Harhay, G.L. Bennett, R.T. Stone, W.M. Grosse, E. Casas, J.W. Keele, T.P.L. Smith, C.G. Chitko-McKown, and W.W. Laegreid. 2002. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle.

Mamm. Genome. 13:272–281. doi:10.1007/s00335-001-2146-3.

Heaton, M.P., K. a. Leymaster, T.S. Kalbfleisch, J.W. Kijas, S.M. Clarke, J. McEwan, J.F. Maddox, V. Basnayake, D.T. Petrik, B. Simpson, T.P.L. Smith, and C.G. Chitko-McKown. 2014. SNPs for Parentage Testing and Traceability in Globally Diverse Breeds of Sheep. *PLoS One*. 9:e94851. doi:10.1371/journal.pone.0094851.

Hellenthal, G., G.B.J. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, and S. Myers. 2014. A genetic atlas of human admixture history. *Science (80-.).* 343:747–751. doi:10.1126/science.1243518.

ICAR. International Committee for Animal Recording.

IGGC. International Goat Genome Consortium.

International Sheep Genomics Consortium, T. 2009. Sheep HapMap Project.

ISAG. ISAG cattle core+additional SNP panels.

ISAG. 2012. International Society of Animal Breeding.

ISCG. 2009. International Sheep Genomics Consortium.

Jakobsson, M., and N.A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*. 23:1801–6. doi:10.1093/bioinformatics/btm233.

Jiang, Y., M. Xie, W. Chen, R. Talbot, J.F. Maddox, T. Faraut, C. Wu, D.M. Muzny, Y. Li, W. Zhang, J.-A. Stanton, R. Brauning, W.C. Barris, T. Hourlier, B.L. Aken, S.M.J. Searle, D.L. Adelson, C. Bian, G.R. Cam, Y. Chen, S. Cheng, U. DeSilva, K. Dixen, Y. Dong, G. Fan, I.R. Franklin, S. Fu, P. Fuentes-Utrilla, R. Guan, M.A. Highland, M.E. Holder, G. Huang, A.B. Ingham, S.N. Jhangiani, D. Kalra, C.L. Kovar, S.L. Lee, W. Liu, X. Liu, C. Lu, T. Lv, T. Mathew, S. McWilliam, M. Menzies, S. Pan, D. Robelin, B. Servin, D. Townley, W. Wang, B. Wei, S.N. White, X. Yang, C. Ye, Y. Yue, P. Zeng, Q. Zhou, J.B. Hansen, K. Kristiansen, R.A. Gibbs, P. Flicek, C.C. Warkup, H.E. Jones, V.H. Oddy, F.W. Nicholas, J.C. McEwan, J.W. Kijas, J. Wang, K.C. Worley, A.L. Archibald, N. Cockett, X. Xu, W. Wang, and B.P. Dalrymple. 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. *Science (80-.).* 344.

Jombart, T., and I. Ahmed. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*. 27:3070–3071. doi:10.1093/bioinformatics/btr521.

Jones, A.G., C.M. Small, K. a. Paczolt, and N.L. Ratterman. 2010. A practical guide to methods of parentage analysis. *Mol. Ecol. Resour.* 10:6–30. doi:10.1111/j.1755-0998.2009.02778.x.

Kadarmideen, H.N. 2014. Genomics to systems biology in animal and veterinary sciences:

Progress, lessons and opportunities. *Livest. Sci.* 166:232–248. doi:10.1016/j.livsci.2014.04.028.

Karlsson, E.K., I. Baranowska, C.M. Wade, N.H.C. Salmon Hillbertz, M.C. Zody, N. Anderson, T.M. Biagi, N. Patterson, G.R. Pielberg, E.J. Kulbokas, K.E. Comstock, E.T. Keller, J.P. Mesirov, H. von Euler, O. Kämpe, A. Hedhammar, E.S. Lander, G. Andersson, L. Andersson, and K. Lindblad-Toh. 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* 39:1321–8. doi:10.1038/ng.2007.10.

Kemper, K.E., S.J. Saxton, S. Bolormaa, B.J. Hayes, and M.E. Goddard. 2014. Selection for complex traits leaves little or no classic signatures of selection. *BMC Genomics*. 15:246. doi:10.1186/1471-2164-15-246.

Kerry, J., J. Kerry, and D. Ledward. 2001. Meat processing Improving quality.

Khatkar, M.S., P.C. Thomson, I. Tammen, and H.W. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genet. Sel. Evol.* 36:163. doi:10.1186/1297-9686-36-2-163.

Kijas, J.W., T. Hadfield, M. Naval Sanchez, and N. Cockett. 2016. Genome-wide association reveals the locus responsible for four-horned ruminant. *Anim. Genet.* 47:258–262. doi:10.1111/age.12409.

Kijas, J.W., J.A. Lenstra, B. Hayes, S. Boitard, L.R. Porto Neto, M. San Cristobal, B. Servin, R. McCulloch, V. Whan, K. Gietzen, S. Paiva, W. Barendse, E. Ciani, H. Raadsma, J. McEwan, and B. Dalrymple. 2012. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLoS Biol.* 10:e1001258. doi:10.1371/journal.pbio.1001258.

Kijas, J.W., J.S. Ortiz, R. McCulloch, A. James, B. Brice, B. Swain, and G. Tosser-Klopp. 2013. Genetic diversity and investigation of polledness in divergent goat populations using 52 088 SNPs. *Anim. Genet.* 44:325–335. doi:10.1111/age.12011.

Kijas, J.W., L. Porto-Neto, S. Dominik, A. Reverter, R. Bunch, R. McCulloch, B.J. Hayes, R. Brauning, and J. McEwan. 2014. Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. *Anim. Genet.* 45:754–757. doi:10.1111/age.12197.

Kijas, J.W., D. Townley, B.P. Dalrymple, M.P. Heaton, J.F. Maddox, A. McGrath, P. Wilson, R.G. Ingersoll, R. McCulloch, S. McWilliam, D. Tang, J. McEwan, N. Cockett, V.H. Oddy, F.W. Nicholas, H. Raadsma, W. Barris, S.C. Bishop, D. Coltman, A. Crawford, A. Eggen, G. Erhardt, R. Forage, O. Hanotte, P. Hunt, H. Jianlin, K. Li, P.A. Marsan, J.E. Miller, J. Pemberton, and L. Schibler. 2009. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PLoS One.* 4. doi:10.1371/journal.pone.0004668.

Kim, E.-S., A.R. Elbeltagy, A.M. Aboul-Naga, B. Rischkowsky, B. Sayre, J.M. Mwacharo, and M.F. Rothschild. 2015. Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. *Heredity (Edinb).* 116:1–10. doi:10.1038/hdy.2015.94.

Kim, E.S., J.B. Cole, H. Huson, G.R. Wiggans, C.P. Van Tassel, B.A. Crooker, G. Liu, Y. Da, and T.S. Sonstegard. 2013. Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle. *PLoS One*. 8:1–14. doi:10.1371/journal.pone.0080813.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*. 160:765–77.

King, G.D., and R.S. Turner. 2004. Adaptor protein interactions: Modulators of amyloid precursor protein metabolism and Alzheimer's disease risk? *Exp. Neurol.* 185:208–219. doi:10.1016/j.expneurol.2003.10.011.

Klei, L., B.P. Kent, N. Melhem, B. Devlin, and K. Roeder. 2011. GemTools: A fast and efficient approach to estimating genetic ancestry.

Korte, A., and A. Farlow. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*. 9:29. doi:10.1186/1746-4811-9-29.

Krawczak, M. 1999. Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis*. 20:1676–1681. doi:10.1002/(SICI)1522-2683(19990101)20:8<1676::AID-ELPS1676>3.0.CO;2-D.

Kruglyak, L. 1997. The use of a genetic map of biallelic markers in linkage studies. *Nat. Genet.* 17:21–24. doi:10.1038/ng0997-21.

Kuleshov, M. V., M.R. Jones, A.D. Rouillard, N.F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S.L. Jenkins, K.M. Jagodnik, A. Lachmann, M.G. McDermott, C.D. Monteiro, G.W. Gundersen, and A. Ma'ayan. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* gkw377. doi:10.1093/nar/gkw377.

Lao, O., F. Liu, A. Wollstein, and M. Kayser. 2014. GAGA: a new algorithm for genomic inference of geographic ancestry reveals fine level population substructure in Europeans. *PLoS Comput. Biol.* 10:e1003480. doi:10.1371/journal.pcbi.1003480.

Lashmar, S., C. Visser, and E. Van Marle-Köster. 2015. Validation of the 50k Illumina goat SNP chip in the South African Angora goat. *S. Afr. J. Anim. Sci.* 45:56. doi:10.4314/sajas.v45i1.7.

Lashmar, S.F., C. Visser, and E. va. Marle-Köster. 2016. SNP-based genetic diversity of South African commercial dairy and fibre goat breeds. *Small Rumin. Res.* 136:65–71. doi:10.1016/j.smallrumres.2016.01.006.

Lawler, R.R. 2017. Genetic Distance. *Int. Encycl. Primatol.* 1–2.

doi:10.1002/9781119179313.wbprim0030.

Lawson, D.J., G. Hellenthal, S. Myers, D. Falush, and F. Zhang. 2012. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet.* 8:e1002453. doi:10.1371/journal.pgen.1002453.

Leaché, A.D., B.L. Banbury, J. Felsenstein, A. nieto-M. de Oca, A. Stamatakis, S. K., W. G.D., B. A.K., Z. P., M. G.D., Y. N.D., T. P., H. W., W. Z., F. W., X. X., M. C.C., D. A., O. M.J., M. J.O., B. E.W., D. L., D. R., O. L., S. C., Z. G., N. R., W. E., and W. J. 2015. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Syst. Biol.* 64:1032–1047. doi:10.1093/sysbio/syv053.

Lee, A.B., D. Luca, L. Klei, B. Devlin, and K. Roeder. 2010. Discovering genetic ancestry using spectral graph theory. *Genet. Epidemiol.* 34:51–9. doi:10.1002/gepi.20434.

Li, W., A. Sartelet, N. Tamma, W. Coppieters, M. Georges, and C. Charlier. 2016. Reverse genetic screen for loss-of-function mutations uncovers a frameshifting deletion in the melanophilin gene accountable for a distinctive coat color in Belgian Blue cattle. *Anim. Genet.* 47:110–113. doi:10.1111/age.12383.

Lindblad-Toh, K., C.M. Wade, T.S. Mikkelsen, E.K. Karlsson, D.B. Jaffe, M. Kamal, M. Clamp, J.L. Chang, E.J. Kulbokas, M.C. Zody, E. Mauceli, X. Xie, M. Breen, R.K. Wayne, E.A. Ostrander, C.P. Ponting, F. Galibert, D.R. Smith, P.J. deJong, E. Kirkness, P. Alvarez, T. Biagi, W. Brockman, J. Butler, C.-W. Chin, A. Cook, J. Cuff, M.J. Daly, D. DeCaprio, S. Gnerre, M. Grabherr, M. Kellis, M. Kleber, C. Bardeleben, L. Goodstadt, A. Heger, C. Hitte, L. Kim, K.-P. Koepfli, H.G. Parker, J.P. Pollinger, S.M.J. Searle, N.B. Sutter, R. Thomas, C. Webber, J. Baldwin, A. Abebe, A. Abouelleil, L. Aftuck, M. Ait-zahra, T. Aldredge, N. Allen, P. An, S. Anderson, C. Antoine, H. Arachchi, A. Aslam, L. Ayotte, P. Bachantsang, A. Barry, T. Bayul, M. Benamara, A. Berlin, D. Bessette, B. Blitshteyn, T. Bloom, J. Blye, L. Boguslavskiy, C. Bonnet, B. Boukhgalter, A. Brown, P. Cahill, N. Calixte, J. Camarata, Y. Cheshatsang, J. Chu, M. Citroen, A. Collymore, P. Cooke, T. Dawoe, R. Daza, K. Decktor, S. DeGray, N. Dhargay, K. Dooley, K. Dooley, P. Dorje, K. Dorjee, L. Dorris, N. Duffey, A. Dupes, O. Egbiremolen, R. Elong, J. Falk, A. Farina, S. Faro, D. Ferguson, P. Ferreira, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*. 438:803–819. doi:10.1038/nature04338.

Liu, Y., X. Qin, X.-Z.H. Song, H. Jiang, Y. Shen, K.J. Durbin, S. Lien, M.P. Kent, M. Sodeland, Y. Ren, L. Zhang, E. Sodergren, P. Havlak, K.C. Worley, G.M. Weinstock, and R.A. Gibbs. 2009. Bos taurus genome assembly. *BMC Genomics*. 10:180. doi:10.1186/1471-2164-10-180.

Liu, Z., Z. Ji, G. Wang, T. Chao, L. Hou, and J. Wang. 2016. Genome-wide analysis reveals signatures of selection for important traits in domestic sheep from different ecoregions. *BMC Genomics*. 17:863. doi:10.1186/s12864-016-3212-2.

Lush, J.L. 1943. Animal Breeding Plans. The Iowa State College Press, Ames. 457 pp.

Lyons, E., and M. Freeling. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53:661–673. doi:10.1111/j.1365-313X.2007.03326.x.

Manunza, A., A. Noce, J.M. Serradilla, F. Goyache, A. Martínez, J. Capote, J.V. Delgado, J. Jordana, E. Muñoz, A. Molina, V. Landi, A. Pons, V. Balteanu, A. Traoré, M. Vidilla, M.S. Rodríguez, A. Sànchez, T.F. Cardoso, and M. Amills. 2016. A genome - wide perspective about the diversity and demographic history of seven Spanish goat breeds. *Genet. Sel. Evol.* 1–9. doi:10.1186/s12711-016-0229-6.

De Marchi, M., C. Dalvit, C. Targhetta, and M. Cassandro. 2006. Assessing genetic diversity in indigenous Veneto chicken breeds using AFLP markers. *Anim. Genet.* doi:10.1111/j.1365-2052.2005.01390.x.

Marklund, S., M. Moller, K. Sandberg, and L. Andersson. 1999. Close association between sequence polymorphism in the KIT gene and the roan coat color in horses. *Mamm. Genome*. 10:283–288. doi:10.1007/s003359900987.

Marras, G., G. Gaspa, S. Sorbolini, C. Dimauro, P. Ajmone-Marsan, A. Valentini, J.L. Williams, and N.P.P. Macciotta. 2015. Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Anim. Genet.* 46:110–121. doi:10.1111/age.12259.

Mason-Suares, H., W. Kim, L. Grimmett, E.S. Williams, V.L. Horner, D. Kunig, I.S. Goldlust, B.-L. Wu, Y. Shen, D.T. Miller, C.L. Martin, and M. Kathar. 2013. Density matters: comparison of array platforms for detection of copy-number variation and copy-neutral abnormalities. doi:10.1038/gim.2013.36.

Mastrangelo, S., P. Ajmone-Marsan, A. Bagnato, L.M. Battaglini, R. Bozzi, A. Carta, G. Catillo, M. Cassandro, S. Casu, R. Ciampolini, E. Ciani, P. Crepaldi, M. D'Andrea, R. Di Gerlando, L. Fontanesi, M. Longeri, N.P.P. Macciotta, R. Mantovani, D. Marletta, D. Matassino, M. Mele, G. Pagnacco, C. Pieramati, B. Portolano, F.M. Sarti, and F. Pilla. 2017. BOVITA: a first overview on genome-wide genetic diversity of Italian autochtonous cattle breeds. *In* ASPA 22nd Congress. Perugia. 38–39.

Matukumalli, L.K., C.T. Lawley, R.D. Schnabel, J.F. Taylor, M.F. Allan, M.P. Heaton, J. O'Connell, S.S. Moore, T.P.L. Smith, T.S. Sonstegard, and C.P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLoS One*. 4:e5350. doi:10.1371/journal.pone.0005350.

McKean, J.D. 2001. The importance of traceability for public health and consumer protection. *Rev. Sci. Tech.* 20:363–71.

Mctavish, E.J., J.E. Decker, R.D. Schnabel, J.F. Taylor, and D.M. Hillis. New World cattle

show ancestry from multiple independent domestication events.

Menze, B.H., B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F.A. Hamprecht. 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*. 10:213. doi:10.1186/1471-2105-10-213.

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating Improvement of Livestock with Genomic Selection. *Annu. Rev. Anim. Biosci.* 1:221–237. doi:10.1146/annurev-animal-031412-103705.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. doi:11290733.

Mucha, S., R. Mrode, I. MacLaren-Lee, M. Coffey, and J. Conington. 2015. Estimation of genomic breeding values for milk yield in UK dairy goats. *J. Dairy Sci.* 98:8201–8. doi:10.3168/jds.2015-9682.

Muigai, A.W.T., and O. Hanotte. 2013. The Origin of African Sheep: Archaeological and Genetic Perspectives. *African Archaeol. Rev.* doi:10.1007/s10437-013-9129-0.

Negrini, R., E. Milanesi, L. Colli, M. Pellecchia, L. Nicoloso, P. Crepaldi, J.A. Lenstra, and P. Ajmone-Marsan. 2007. Breed assignment of Italian cattle using biallelic AFLP® markers. *Anim. Genet.* 38:147–153. doi:10.1111/j.1365-2052.2007.01573.x.

Nei, M. 1972. Genetic Distance between Populations. *Am. Nat.* 106:283–292. doi:10.1086/282771.

NEXTGEN. 2009. NEXTGEN.

Nicolazzi, E.L., S. Biffani, F. Biscarini, P. Orozco Ter Wengel, A. Caprera, N. Nazzicari, and A. Stella. 2015. Software solutions for the livestock genomics SNP array revolution. *Anim. Genet.* 46:343–353. doi:10.1111/age.12295.

Nicoloso, L., L. Bomba, L. Colli, R. Negrini, M. Milanesi, R. Mazza, T. Sechi, S. Frattini, A. Talenti, B. Coizet, S. Chessa, D. Marletta, M.D. Andrea, S. Bordonaro, G. Ptak, A. Carta, G. Pagnacco, A. Valentini, F. Pilla, P. Ajmone-Marsan, P. Crepaldi, and G. Consortium. 2015. Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip. *Genet. Sel. Evol.* 47:1–10. doi:10.1186/s12711-015-0140-6.

Nicoloso, L., R. Negrini, P. Ajmone-Marsan, and P. Crepaldi. 2012. On the way to functional agro biodiversity: coat colour gene variability in goats. *Animal*. 6:41–9. doi:10.1017/S175173111100139X.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.* 39:197–218. doi:10.1146/annurev.genet.39.073003.112420.

Nordström, E.K., K.M. Luhr, C. Iba, and K. Kristensson. 2005. Inhibitors of the Mitogen-Activated Protein Kinase Kinase 1 / 2 Signaling Pathway Clear Prion-Infected Cells from PrP Sc. *Neurobiol. Dis.* 25:8451–8456. doi:10.1523/JNEUROSCI.2349-05.2005.

Noyes, H., A. Brass, I. Obara, S. Anderson, A.L. Archibald, D.G. Bradley, P. Fisher, A. Freeman, J. Gibson, M. Gicheru, L. Hall, O. Hanotte, H. Hulme, D. McKeever, C. Murray, S. Jung Oh, C. Tate, K. Smith, M. Tapio, J. Wambugu, D.J. Williams, M. Agaba, S.J. Kemp, and by C. William Clark. Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection. doi:10.1073/pnas.1013486108.

Olsen, H.G., B.J. Hayes, M.P. Kent, T. Nome, M. Svendsen, A.G. Larsgard, and S. Lien. 2011. Genome-wide association mapping in Norwegian Red cattle identifies quantitative trait loci for fertility and milk production on BTA12. *Anim. Genet.* doi:10.1111/j.1365-2052.2011.02179.x.

Onteru, S.K., D.M. Gorbach, J.M. Young, D.J. Garrick, J.C.M. Dekkers, and M.F. Rothschild. 2013. Whole Genome Association Studies of Residual Feed Intake and Related Traits in the Pig. *PLoS One.* 8. doi:10.1371/journal.pone.0061756.

Opara, L.U., and F. Mazaud. 2001. Food Traceability from Field to Plate. *Outlook Agric.* 30:239–247. doi:10.5367/000000001101293724.

Ostrander, E.A., R.K. Wayne, A.H. Freedman, and B.W. Davis. 2017. Demographic history, selection and functional diversity of the canine genome. *Nat. Rev. Genet.* doi:10.1038/nrg.2017.67.

Óvilo, C., M.T. Cervera, C. Castellanos, and J.M. Martínez-Zapater. 2000. Characterisation of Iberian pig genotypes using AFLP markers. *Anim. Genet.* doi:10.1046/j.1365-2052.2000.00603.x.

Package, T. 2013. Package " GenABEL ."

Paetkau, D.W., W. Calvert, I. Stirling, and C. Strobeck. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.*

Park, S.D.E., D.A. Magee, P.A. McGettigan, M.D. Teasdale, C.J. Edwards, A.J. Lohan, A. Murphy, M. Braud, M.T. Donoghue, Y. Liu, A.T. Chamberlain, K. Rue-Albrecht, S. Schroeder, C. Spillane, S. Tai, D.G. Bradley, T.S. Sonstegard, B.J. Loftus, and D.E. MacHugh. 2015. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. 16:234. doi:10.1186/s13059-015-0790-2.

Parker, H.G., D.L. Dreger, M. Rimbault, B.W. Davis, A.B. Mullen, G. Carpintero-Ramirez, and E.A. Ostrander. 2017. Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. *Cell Rep.*

19:697–708. doi:10.1016/j.celrep.2017.03.079.

Parker, H.G., A.L. Shearin, and E.A. Ostrander. 2010. Man's best friend becomes biology's best in show: genome analyses in the domestic dog. *Annu. Rev. Genet.* 44:309–36. doi:10.1146/annurev-genet-102808-115200.

Pedrosa, S., M. Uzun, J.-J. Arranz, B. Gutiérrez-Gil, F. San Primitivo, and Y. Bayón. 2005. Evidence of three maternal lineages in near eastern sheep supporting multiple domestication events. *Proc. R. Soc. B Biol. Sci.* 272:2211–2217. doi:10.1098/rspb.2005.3204.

Peripolli, E., D.P. Munari, M.V.G.B. Silva, A.L.F. Lima, R. Irgang, and F. Baldi. 2016. Runs of homozygosity: current knowledge and applications in livestock. *Anim. Genet.* doi:10.1111/age.12526.

Petitto, J.M., D.T. Lysle, J.-L. Gariepy, and M.H. Lewis. 1994. Association of Genetic Differences in Social Behavior and Cellular Immune Responsiveness: Effects of Social Experience. 8:111–122. doi:doi:10.1006/brbi.1994.1011.

Phillips, C., M. García-Magariños, A. Salas, Á. Carracedo, and M.V. Lareu. 2012. SNPs as Supplements in Simple Kinship Analysis or as Core Markers in Distant Pairwise Relationship Tests: When Do SNPs Add Value or Replace Well-Established and Powerful STR Tests? *Transfus. Med. Hemotherapy*. 39:202–210. doi:10.1159/000338857.

Pickrell, J.K., J.K. Pritchard, T. Mikkelsen, E. Karlsson, and D. Jaffe. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* 8:e1002967. doi:10.1371/journal.pgen.1002967.

Plassais, J., M. Rimbault, F.J. Williams, B.W. Davis, J.J. Schoenebeck, and E.A. Ostrander. 2017. Analysis of large versus small dogs reveals three genes on the canine X chromosome associated with body weight, muscling and back fat thickness. doi:10.1371/journal.

Pontius, J.U., J.C. Mullikin, D.R. Smith, A.S. Team, K. Lindblad-Toh, S. Gnerre, M. Clamp, J. Chang, R. Stephens, B. Neelam, N. Volfovsky, A.A. Schäffer, R. Agarwala, K. Narfström, W.J. Murphy, U. Giger, A.L. Roca, M. Menotti-Raymond, N. Yuhki, J. Pecon-Slattery, W.E. Johnson, G. Bourque, G. Tesler, A. Brand, H. Ebling, D.J. Saranga, M. Rubenfield, M.J. Parisi, W. Tao, N. Tusneem, R. David, E. Gustafson, J. Tsolas, and K. Mckernan. Initial sequence and comparative analysis of the cat genome. doi:10.1101/gr.6380007.

Porto-Neto, L.R., S.H. Lee, H.K. Lee, and C. Gondro. 2013. Detection of signatures of selection using Fst. *Methods Mol. Biol.* 1019:423–36. doi:10.1007/978-1-62703-447-0_19.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006.

Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–909. doi:10.1038/ng1847.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. a. R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81:559–575. doi:10.1086/519795.

Purdie, A.C., K.M. Plain, D.J. Begg, K. de Silva, and R.J. Whittington. 2011. Candidate gene and genome-wide association studies of Mycobacterium avium subsp. paratuberculosis infection in cattle and sheep: A review. *Comp. Immunol. Microbiol. Infect. Dis.* 34:197–208. doi:10.1016/j.cimid.2010.12.003.

Quinlan, A.R., and I.M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26:841–2. doi:10.1093/bioinformatics/btq033.

Raj, A., M. Stephens, and J.K. Pritchard. 2014. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*. 197:573–589. doi:10.1534/genetics.114.164350.

Randhawa, I.A.S., M.S. Khatkar, P.C. Thomson, and H.W. Raadsma. 2016. A meta-assembly of selection signatures in cattle. *PLoS One*. 11:1–30. doi:10.1371/journal.pone.0153013.

Rannala, B., and J.L. Mountain. 1997. Detecting immigration by using multilocus genotypes. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.94.17.9197.

Raoul, J., A.A. Swan, and J.-M. Elsen. 2017. Using a very low-density SNP panel for genomic selection in a breeding program for sheep. *Genet. Sel. Evol.* 49:76. doi:10.1186/s12711-017-0351-0.

Reber, I., I. Keller, D. Becker, C. Flury, M. Welle, and C. Drögemüller. 2015. Wattles in goats are associated with the FMN1/GREM1 region on chromosome 10. *Anim. Genet.* 46:316–320. doi:10.1111/age.12279.

Reeves, P.A., and C.M. Richards. 2009. Accurate inference of subtle population structure (and other genetic discontinuities) using principal coordinates. *PLoS One*. 4:e4269. doi:10.1371/journal.pone.0004269.

Reverter, A., and M.R.S. Fortes. 2013. Genome-Wide Association Studies and Genomic Prediction. 1019. 437-447 pp.

Reynolds, J., B.S. Weir, and C.C. Cockerham. 1983. Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. *Genetics*. 105:767–779.

Rosenberg, N.A. 2003. distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes*. 4:137–138. doi:10.1046/j.1471-8286.2003.00566.x.

Sabeti, P.C., D.E. Reich, J.M. Higgins, H.Z.P. Levine, D.J. Richter, S.F. Schaffner, S.B. Gabriel, J. V. Platko, N.J. Patterson, G.J. McDonald, H.C. Ackerman, S.J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E.S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 419:832–837. doi:10.1038/nature01027.1.

Sabeti, P.C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E.H. Byrne, S.A. Mccarroll, S.F. Schaffner, E.S. Lander, and T.I. Hapmap. 2007. Genome-wide detection and characterization of positive selection in human populations. *October*. 449:913–918. doi:10.1038/nature06250.Genome-wide.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–25.

Sanchez, M.-P., A. Govignon-Gion, P. Croiseau, S. Fritz, C. Hozé, G. Miranda, P. Martin, A. Barbat-Leterrier, R. Letaïef, D. Rocha, M. Brochard, M. Boussaha, and D. Boichard. 2017. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genet. Sel. Evol.* 49:68. doi:10.1186/s12711-017-0344-z.

Sandmann, S., A.O. de Graaf, M. Karimi, B.A. van der Reijden, E. Hellström-Lindberg, J.H. Jansen, and M. Dugas. 2017. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci. Rep.* 7:43169. doi:10.1038/srep43169.

Schibler, L., G.P. Di Meo, E.P. Cribiu, and L. Iannuzzi. 2009. Molecular cytogenetics and comparative mapping in goats (Capra hircus, 2n = 60). *Cytogenet. Genome Res.* 126:77–85. doi:10.1159/000245908.

Schnabel, R.D., T.J. Ward, and J.N. Derr. 2000. Validation of 15 microsatellites for parentage testing in North American bison, Bison bison and domestic cattle. *Anim. Genet.* 31:360–366. doi:10.1046/j.1365-2052.2000.00685.x.

Seitz, J.J., S.M. Schmutz, T.D. Thue, and F.C. Buchanan. 1999. A missense mutation in the bovine MGF gene is associated with the roan phenotype in Belgian Blue and Shorthorn cattle. *Mamm. Genome*. 10:710–712. doi:10.1007/s003359901076.

Shannon, L.M., R.H. Boyko, M. Castelhano, E. Corey, J.J. Hayward, C. Mclean, M.E. White, M. Abi, B.A. Anita, N. Ikombe, J. Calero, and A. Galov. 2015. Genetic structure in village dogs reveals a Central Asian domestication origin. 112:3–8. doi:10.1073/pnas.1516215112.

Sharma, A., J.S. Lee, C.G. Dang, P. Sudrajad, H.C. Kim, S.H. Yeon, H.S. Kang, and S.-H. Lee. 2015. Stories and Challenges of Genome Wide Association Studies in Livestock - A Review. *Asian-Australasian J. Anim. Sci.* 28:1371–9. doi:10.5713/ajas.14.0715.

Shriver, M.D., G.C. Kennedy, E.J. Parra, H.A. Lawson, V. Sonpar, J. Huang, J.M. Akey, and K.W. Jones. 2004. The genomic distribution of population substructure in four

populations using 8,525 autosomal SNPs. *Hum. Genomics*. 1:274–86. doi:10.1186/1479-7364-1-4-274.

de Simoni Gouveia, J.J., M.V.G.B. da Silva, S.R. Paiva, and S.M.P. de Oliveira. 2014. Identification of selection signatures in livestock species. *Genet. Mol. Biol.* 37:330–342. doi:10.1590/S1415-47572014000300004.

Sims, D., I. Sudbery, N.E. Ilott, A. Heger, and C.P. Ponting. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15:121–32. doi:10.1038/nrg3642.

Singh, S., S. Kumar Jr, A.P. Kolte, and S. Kumar. 2013. Extensive Variation and Sub-Structuring in Lineage A mtDNA in Indian Sheep: Genetic Evidence for Domestication of Sheep in India. *PLoS One*. 8:e77858. doi:10.1371/journal.pone.0077858.

Sorbolini, S., G. Gaspa, D. Marletta, P. Crepaldi, B. Moioli, A. Carta, B. Portolano, E. Lasagna, E. Ciani, M. D'Andrea, F. Pilla, N.P.P. Macciotta, and T. Italian Sheep Consortium. 2017. Detection of signatures of selection in Italian sheep breeds. *In* 22nd Animal Science and Production Association (ASPA) Congress. Perugia. 73.

Stanford, K., J. Stitt, J. a Kellar, and T. a McAllister. 2001. Traceability in cattle and small ruminants in Canada. *Rev. Sci. Tech.* 20:510–522.

Stanley, P. 1995. Robert Bakewell and the Longhorn breed of cattle. *Robert Bakewell and the Longhorn breed of cattle.*

Stella, A., P. Ajmone-Marsan, B. Lazzari, and P. Boettcher. 2010. Identification of selection signatures in cattle breeds selected for dairy production. *Genetics*. 185:1451–61. doi:10.1534/genetics.110.116111.

Streit, M., F. Reinhardt, G. Thaller, and J. Bennewitz. 2013. Genome-wide association analysis to identify genotype × environment interaction for milk protein yield and level of somatic cell score as environmental descriptors in German Holsteins. *J. Dairy Sci.* doi:10.3168/jds.2013-7133.

Strucken, E.M., B. Gudex, M.H. Ferdosi, H.K. Lee, K.D. Song, J.P. Gibson, M. Kelly, E.K. Piper, L.R. Porto-Neto, S.H. Lee, and C. Gondro. 2014. Performance of different SNP panels for parentage testing in two East Asian cattle breeds. *Anim. Genet.* 45:572–575. doi:10.1111/age.12154.

Swine Genome Seqencing Consortium, T. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. doi:10.1038/nature11622.

Szpiech, Z.A., and R.D. Hernandez. 2014. Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31:2824–2827. doi:10.1093/molbev/msu211.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 123:585–595. doi:PMC1203831.

Talenti, A., E.L. Nicolazzi, S. Chessa, S. Frattini, R. Moretti, B. Coizet, L. Nicoloso, L. Colli, G. Pagnacco, A. Stella, P. Ajmone-Marsan, G. Ptak, and P. Crepaldi. 2016. A method for single nucleotide polymorphism selection for parentage assessment in goats. *J. Dairy Sci.* 99:3646–3653. doi:10.3168/jds.2015-10077.

Tanahashi, H., and T. Tabira. 1999. X11L2, a new member of the X11 protein family, interacts with Alzheimer's β-amyloid precursor protein. *Biochem. Biophys. Res. Commun.* 255:663–667. doi:10.1006/bbrc.1999.0265.

Tang, H., J. Peng, P. Wang, and N.J. Risch. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* doi:10.1002/gepi.20064.

Tang, K., K.R. Thornton, and M. Stoneking. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5:1587–1602. doi:10.1371/journal.pbio.0050171.

Taye, M., W. Lee, K. Caetano-Anolles, T. Dessie, O. Hanotte, O.A. Mwai, S. Kemp, S. Cho, S.J. Oh, H.K. Lee, and H. Kim. 2017a. Whole genome detection of signature of positive selection in African cattle reveals selection for thermotolerance. *Anim. Sci. J.* doi:10.1111/asj.12851.

Taye, M., W. Lee, S. Jeon, J. Yoon, T. Dessie, O. Hanotte, O.A. Mwai, S. Kemp, S. Cho, S.J. Oh, H.-K. Lee, and H. Kim. 2017b. Exploring evidence of positive selection signatures in cattle breeds selected for different traits. *Mamm. Genome*. 1–14. doi:10.1007/s00335-017-9715-6.

Thalmann, O., B. Shapiro, P. Cui, V.J. Schuenemann, S.K. Sawyer, D.L. Greenfield, M.B. Germonpré, M. V. Sablin, F. López-Giráldez, X. Domingo-Roura, H. Napierala, H.-P.H.-P. Uerpmann, D.M. Loponte, A.A. Acosta, L. Giemsch, R.W. Schmitz, B. Worthington, J.E. Buikstra, A. Druzhkova, A.S. Graphodatsky, N.D. Ovodov, N. Wahlberg, A.H. Freedman, R.M. Schweizer, K.-P.P. Koepfli, J.A. Leonard, M. Meyer, J. Krause, S. Pääbo, R.E. Green, R.K. Wayne, M.B. Germonpre, M. V. Sablin, F. Lopez-Giraldez, X. Domingo-Roura, H. Napierala, H.-P.H.-P. Uerpmann, D.M. Loponte, A.A. Acosta, L. Giemsch, R.W. Schmitz, B. Worthington, J.E. Buikstra, A. Druzhkova, A.S. Graphodatsky, N.D. Ovodov, N. Wahlberg, A.H. Freedman, R.M. Schweizer, K.-P.P. Koepfli, J.A. Leonard, M. Meyer, J. Krause, S. Paabo, R.E. Green, R.K. Wayne, M.B. Germonpré, M. V. Sablin, F. López-Giráldez, X. Domingo-Roura, H. Napierala, H.-P.H.-P. Uerpmann, D.M. Loponte, A.A. Acosta, L. Giemsch, R.W. Schmitz, B. Worthington, J.E. Buikstra, A. Druzhkova, A.S. Graphodatsky, N.D. Ovodov, N. Wahlberg, A.H. Freedman, R.M. Schweizer, K.-P.P. Koepfli, J.A. Leonard, M. Meyer, J. Krause, S. Pääbo, R.E. Green, and R.K. Wayne. 2013. Complete Mitochondrial Genomes of Ancient Canids Suggest a European Origin of

Domestic Dogs. *Science (80-.).* 342:871–874. doi:10.1126/science.1243650.

The Bactrian Camels Genome Sequencing and Analysis Consortium. 2012. Genome sequences of wild and domestic bactrian camels The Bactrian Camels Genome Sequencing and Analysis Consortium*. *Nat. Commun.* 3:1202. doi:10.1038/ncomms2192.

Tiwari, S., K. SL, V. Kumar, B. Singh, A. Rao, A. Mithra SV, V. Rai, A.K. Singh, and N.K. Singh. 2016. Mapping QTLs for Salt Tolerance in Rice (Oryza sativa L.) by Bulked Segregant Analysis of Recombinant Inbred Lines Using 50K SNP Chip. *PLoS One*. 11:e0153610. doi:10.1371/journal.pone.0153610.

Toro, M.Á., B. Villanueva, and J. Fernández. 2014. Genomics applied to management strategies in conservation programmes. *Livest. Sci.* 166:48–53. doi:10.1016/j.livsci.2014.04.020.

Toro, M., C. Barrag, C. Ovilo, J. Rodrigañez, C. Rodriguez, and L. Sili. 2002. Estimation of coancestry in Iberian pigs using molecular markers. 309–320.

Tosser-Klopp, G., P. Bardou, O. Bouchez, C. Cabau, R. Crooijmans, Y. Dong, C. Donnadieu-Tonon, A. Eggen, H.C.M. Heuven, S. Jamli, A.J. Jiken, C. Klopp, C.T. Lawley, J. McEwan, P. Martin, C.R. Moreno, P. Mulsant, I. Nabihoudine, E. Pailhoux, I. Palhière, R. Rupp, J. Sarry, B.L. Sayre, A. Tircazes, Jun Wang, W. Wang, W. Zhang, and and the I.G.G. Consortium. 2014. Design and Characterization of a 52K SNP Chip for Goats. *PLoS One*. 9:e86227.

Tosser-Klopp, G., P. Bardou, C. Cabau, a Eggen, T. Faraut, H. Heuven, S. Jamli, C. Klopp, C.T. Lawley, J. McEwan, P. Martin, C. Moreno, P. Mulsant, I. Nabihoudine, E. Pailhoux, I. Palhière, R. Rupp, J. Sarry, B. Sayre, a Tircazes, J. Wang, W. Wang, T.-P. Yu, and W. Zhang. 2012. Goat genome assembly, Availability of an international 50K SNP chip and RH panel: An update of the International Goat Genome Consortium projects. *Plant Anim. Genome Conf.* 1–14.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92:16–24. doi:10.3168/jds.2008-1514.

VarGoats. 2017.

Venables, W.N., and B.D. Ripley. 2002. MASS: modern applied statistics with S. *Springer, New York*. doi:10.1198/tech.2003.s33.

Vignal, A., D. Milan, M. Sancristobal, and A. Eggen. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* 34:275–305. doi:10.1051/gse:2002009.

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. doi:10.1016/J.CRVI.2010.12.009.

Visscher, P.M., M.A. Brown, M.I. McCarthy, and J. Yang. 2012. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* 90:7. doi:10.1016/J.AJHG.2011.11.029.

Vitti, J.J., S.R. Grossman, and P.C. Sabeti. 2013. Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet.* 47:97–120. doi:10.1146/annurev-genet-111212-133526.

Van Vleck, L.D., E.J. Pollak, and E.A.B. Oltenacu. 1987. Genetics for the animal sciences. W.H. Freeman, editor. New York.

Voight, B.F., S. Kudaravalli, X. Wen, and J.K. Pritchard. 2006. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol.* 4:e72. doi:10.1371/journal.pbio.0040072.

Wade, C.M., E. Giulotto, S. Sigurdsson, M. Zoli, S. Gnerre, F. Imsland, T.L. Lear, D.L. Adelson, E. Bailey, R.R. Bellone, H. Blöcker, O. Distl, R.C. Edgar, M. Garber, T. Leeb, E. Mauceli, J.N. MacLeod, M.C.T. Penedo, J.M. Raison, T. Sharpe, J. Vogel, L. Andersson, D.F. Antczak, T. Biagi, M.M. Binns, B.P. Chowdhary, S.J. Coleman, G. Della Valle, S. Fryc, G. Guérin, T. Hasegawa, E.W. Hill, J. Jurka, A. Kiialainen, G. Lindgren, J. Liu, E. Magnani, J.R. Mickelson, J. Murray, S.G. Nergadze, R. Onofrio, S. Pedroni, M.F. Piras, T. Raudsepp, M. Rocchi, K.H. Røed, O.A. Ryder, S. Searle, L. Skow, J.E. Swinburne, A.C. Syvänen, T. Tozaki, S.J. Valberg, M. Vaudin, J.R. White, M.C. Zody, B.I.G.S. Broad Institute Genome Sequencing Platform, B.I.W.G.A. Broad Institute Whole Genome Assembly Team, E.S. Lander, and K. Lindblad-Toh. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science.* 326:865–7. doi:10.1126/science.1178158.

Wallis, J.W., J. Aerts, M.A.M. Groenen, R.P.M.A. Crooijmans, D. Layman, T.A. Graves, D.E. Scheer, C. Kremitzki, M.J. Fedele, N.K. Mudd, M. Cardenas, J. Higginbotham, J. Carter, R. McGrane, T. Gaige, K. Mead, J. Walker, D. Albracht, J. Davito, S.-P. Yang, S. Leong, A. Chinwalla, M. Sekhon, K. Wylie, J. Dodgson, M.N. Romanov, H. Cheng, P.J. de Jong, K. Osoegawa, M. Nefedov, H. Zhang, J.D. McPherson, M. Krzywinski, J. Schein, L. Hillier, E.R. Mardis, R.K. Wilson, and W.C. Warren. 2004. A physical map of the chicken genome. *Nature.* 432:761–764. doi:10.1038/nature03030.

Wang, E.T., G. Kodama, P. Baldi, and R.K. Moyzis. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc. Natl. Acad. Sci. U. S. A.* 103:135–40. doi:10.1073/pnas.0509691102.

Werner, F. a. O., G. Durstewitz, F. a. Habermann, G. Thaller, W. Krämer, S. Kollers, J. Buitkamp, M. Georges, G. Brem, J. Mosner, and R. Fries. 2004. Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. *Anim. Genet.* 35:44–49. doi:10.1046/j.1365-2052.2003.01071.x.

Wiedemar, N., and C. Drögemüller. 2015. A 1.8-kb insertion in the 3′-UTR of RXFP2 is associated with polledness in sheep. *Anim. Genet.* 46:457–61. doi:10.1111/age.12309.

Williams, J.L., S.J.G. Hall, M. Del Corvo, K.T. Ballingall, L. Colli, P. Ajmone Marsan, and F. Biscarini. 2016. Inbreeding and purging at the genomic Level: The Chillingham cattle reveal extensive, non-random SNP heterozygosity. *Anim. Genet.* 47:19–27. doi:10.1111/age.12376.

Wollstein, A., and O. Lao. 2011. An upper limit for macromolecular crowding effects. doi:10.1186/s13323-015-0019-x.

Yang, J., S.H. Lee, M.E. Goddard, and P.M. Visscher. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* 88:76–82. doi:10.1016/j.ajhg.2010.11.011.

Yang, W.-Y., J. Novembre, E. Eskin, and E. Halperin. 2012. A model-based approach for analysis of spatial structure in genetic data. *Nat. Genet.* 44:725–731. doi:10.1038/ng.2285.

Yang, Z., and B. Rannala. 2012. Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* 13:303–314. doi:10.1038/nrg3186.

Zalcman, S.S., and A. Siegel. 2006. The neurobiology of aggression and rage: Role of cytokines. *Brain. Behav. Immun.* 20:507–514. doi:10.1016/j.bbi.2006.05.002.

Zeder, M.A. 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci.* 105:11597–11604. doi:10.1073/pnas.0801317105.

Zeder, M.A., E. Emshwiller, B.D. Smith, and D.G. Bradley. 2006. Documenting domestication: The intersection of genetics and archaeology. *Trends Genet.* 22:139–155. doi:10.1016/j.tig.2006.01.007.

Zeder, M.A., and B. Hesse. 2000. The initial domestication of goats (Capra hircus) in the Zagros mountains 10,000 years ago. *Science*. 287:2254–7. doi:10.1126/SCIENCE.287.5461.2254.

Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*. 174:1431–9. doi:10.1534/genetics.106.061432.

Zhang, J., P. Niyogi, and M.S. McPeek. 2009. Laplacian eigenfunctions learn population structure. *PLoS One*. 4:e7928. doi:10.1371/journal.pone.0007928.

Zhang, Q., M.P. Calus, B. Guldbrandtsen, M.S. Lund, and G. Sahana. 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. *BMC Genet.* 16:88. doi:10.1186/s12863-015-0227-7.

Zhao, X., S.K. Onteru, K.E. Dittmer, K. Parton, H.T. Blair, M.F. Rothschild, and D.J. Garrick. 2012. A missense mutation in AGTPBP1 was identified in sheep with a lower motor neuron disease. *Heredity (Edinb).* 109:156–162. doi:10.1038/hdy.2012.23.

Zhou, X., P. Carbonetto, and M. Stephens. 2013. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet.* doi:10.1371/journal.pgen.1003264.

Zhou, X., and M. Stephens. 2012. Genome-wide Efficient Mixed Model Analysis for Association Studies. *Nat. Genet.* 44:821–824. doi:10.1038/ng.2310.

Zhou, X., and M. Stephens. 2014. Efficient Algorithms for Multivariate Linear Mixed Models in Genome-wide Association Studies. *Nat. Methods*. doi:10.1038/nmeth.2848.

Zimin, A. V, A.L. Delcher, L. Florea, D.R. Kelley, M.C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C.P. Van Tassell, T.S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J.A. Yorke, and S.L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol.* 10:R42. doi:10.1186/gb-2009-10-4-r42.

Section 7 - Final Report

7.1 Courses and seminars

7.1.1 First year

I took part to all mandatory courses organized for PhD students of the XXX cycle.

In addition, I attended the following courses and congress:

- STAGES - Scuola di Pubblicazione in Riviste internazionali, October 2014 to January 2015 (25 hours), Università degli Studi di Milano, Milano (MI), Italy.

- GEN2PHEN - Winter School 2015, "An introduction to Bayesian Analysis and MCMC", February 4th to 6th 2015, University Cattolica del Sacro Cuore, Piacenza (PC), Italy.

- Association for Animal Science and Production - XXI ASPA Congress, June 9th to 12th 2015, University of Milan, Milan (MI), Italy.

- International Symposium on Animal Functional Genomics – 6th ISAFG congress, July 27th to 29th 2015, University Cattolica del Sacro Cuore, Piacenza (PC), Italy.

7.1.2 Second year

I took part to all mandatory courses organized for PhD students of the XXX cycle.

In addition, I attended the following courses and congress:

- GEN2PHEN - Winter School 2015, "Genomic data analysis… and beyond!", November 9th to 13th 2015, University Cattolica del Sacro Cuore, Piacenza (PC), Italy.

- Exome analysis using Galaxy, September 19th-20th 2016, University of Milan Bicocca, Milan (MI), Italy.

- Illumina presentation of HiSeq 3000, May 27th 2016, Iowa State University, Ames (IA), United States of America.

7.1.3 Third year

I took part to all mandatory courses organized for PhD students of the XXX cycle.

In addition, I attended the following congresses:

- 2016 National Human Genome Research Institute Symposium, November 3rd-4th 2016, National Institutes of Health, Bethesda (MD), United States.

- 22nd Congress of the Animal Science and Production Association (ASPA), June 13th-16th 2017, University of Perugia, Perugia (PG), Italy.

- 36th International Society for Animal Genetics (ISAG) Congress, July 16th-21st 2017, University College of Dublin, Dublin, Ireland.

7.2 Foreign Exchange

- April the 15th 2016 to June the 15th 2016 at Iowa State University (Ames, IA, United States of America) at Professor Max F. Rothschild lab.

- October the 3rd 2016 to December the 2nd 2016 at the department of Cancer Genetics and Comparative Genomics of National Institute of Health (NIH) (Bethesda, MD, United States of America) at Professor Elaine Ostrander lab.

7.3 Publications

Papers

- Nicoloso L, Bomba L, Colli L, Negrini R, Milanesi M, Mazza R, Sechi T, Frattini S, **Talenti A**, Coizet B, Chessa S, Marletta D, D'Andrea M, Bordonaro S, Ptak G, Carta A, Pagnacco G, Valentini A, Pilla F, Ajmone-Marsan P, Crepaldi P, Andrea MD, Bordonaro S, Ptak G, Carta A, Pagnacco G, Valentini A, Pilla F, Ajmone-Marsan P, Crepaldi P and the Italian Goat Consortium: Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip. Genet Sel Evol 2015, 47:1–10.

- **Talenti A**, Nicolazzi EL El, Chessa S, Frattini S, Moretti R, Coizet B, Nicoloso L, Colli L, Pagnacco G, Stella A, Ajmone-Marsan P, Ptak G, Crepaldi P: A method for single nucleotide polymorphism selection for parentage assessment in goats. J Dairy Sci 2016, 99:3646–3653.

- Sechi S, Polli M, Marelli S, **Talenti A**, Crepaldi P, Fiore F, Spissu N, Dreger DL, Zedda M, Dimauro C, Ostrander EA, Di Cerbo A, Cocco R: Fonni's dog: morphological and genetic characteristics for a breed standard definition. Ital J Anim Sci 2016, 16:22–30.

- **Talenti A**, Bertolini F, Pagnacco G, Pilla F, Ajmone P, Max M, Paola FR, Ajmone-Marsan P, Rothschild MF, Crepaldi P: The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep regions. Mamm Genome 2017, 28:114–128.

- Frattini S, Capra E, Lazzari B, Mckay SD, Coizet B, **Talenti A**, Groppetti D, Riccaboni P, Pecile A, Chessa S, Castiglioni B, Williams JL, Pagnacco G, Stella A, Crepaldi P: Genome-wide analysis of DNA methylation in hypothalamus and ovary of Capra hircus. BMC Genomics 2017, 18:476.

- Coizet B, Frattini S, Nicoloso L, Iannuzzi L, **Talenti A**, Minozzi G, Pagnacco G, Crepaldi P: Polymorphism of the STAT5A, MTNR1A and TNF α genes and their effect on dairy production in Bubalus bubalis. Ital J Anim Sci 2017, 0:1–7.

- **Talenti A**, Bertolini F, Williams J, Moaeen-ud-Din M, Frattini S, Coizet B, Pagnacco G, Reecy J, Rothschild MF, Crepaldi P: Genomic analysis suggests KITLG is responsible for a roan pattern in two Pakistani goat breeds. Journal of Heredity 2017.

- Bigi D, Marelli SP, Liotta L, Frattini S, **Talenti A**, Pagnacco G, Polli M, Crepaldi P: Investigating the population structure and genetic differentiation of livestock guard dog breeds from Italy. Animal 2018, 0:8.

Accepted

- **Talenti A**, Dreger DL, Frattini S, Polli M, Marelli SP, Harris A, Liotta L, Cocco R, Hogan A, Bigi D, Caniglia R, Parker HG, Pagnacco G, Ostrander EA, Crepaldi P: Studies of modern Italian dog populations reveals multiple patterns for domestic breed evolution. In preparation.

Submitted and in preparation

- **Talenti A**, Palhière I, Tortereau F, Pagnacco G, Stella A, Nicolazzi EL, Crepaldi P, Tosser-Klopp G and ADAPTmap Consortium: Functional SNP panel for parentage assessment and assignment in worldwide goat breeds. Submitted to Genetic Selection Evolution.

- Bertolini F, Servin B, **Talenti A**, Rochat E, Kim ES, Oget C, Palhière I, Crisà A, Catillo G, Steri R, Amills M, Colli L, Marras G, Milanesi M, Nicolazzi EL, Rosen BD, Van Tassell CP,

Guldbrandtsen B, Sonstegard TS, Tosser-Klopp G, Stella A, Rothschild MF, Joost S, Crepaldi P and the ADAPTmap consortium: Signatures of selection and environmental adaptation across the goat genome post domestication. In preparation.

- Colli L, Milanesi M, **Talenti A**, Bertolini F, Chen M, Crisà A, Daly K, Del Corvo M, Guldbrandtsen B, Lenstra JA, Rosen BD, Vajana E, Catillo G, Joost S, Nicolazzi EL, Rochat E, Rothschild MF, Servin B, Sonstegard T, Steri R, Van Tassell CP, Ajmone-Marsan P, Crepaldi P, Stella A, the ADAPTmap Consortium: Drawing up worldwide goat diversity and post-domestication history. Submitted to Genetic Selection Evolution.

7.4 Conference proceeding

Presentations

- A. Talenti, E.L. Nicolazzi, L. Nicoloso, S. Frattini, B. Coizet, S. Chessa, G. Pagnacco, F. Pilla, P. Ajmone-Marsan, P. Crepaldi and the Italian Goat Consortium: Parentage assessment with 200 single nucleotide polymorphisms on 15 Italian goat breeds. Italian Journal of Animal Science (2015), 14, suppl. 1, p.52.

- A. Talenti, E.L. Nicolazzi, L. Nicoloso, S. Frattini, B. Coizet, S. Chessa, G. Pagnacco, F. Pilla, P. Ajmone-Marsan, P. Crepaldi and the Italian Goat Consortium: Development of a 200 single nucleotide polymorphism panel for parentage assessment for 14 Italian goat breeds. International Journal of Health, Animal Science & Food Safety, 2015.

- Talenti A, Dreger DL, Danelli F, Frattini S, Coizet B, Marelli SP, Pagnacco G, Gandini G, Polli M, Caniglia R, Galaverni M, Ostrander EA, Crepaldi P: Pedigree and genomic-based relationships in a dog population. In 36th International Society for Animal Genetics (ISAG) congress. Dublin; 2017:44–45.

- Talenti A, Frattini S, Mastrangelo S, Di Gerlando R, Portolano B, Lasagna E, Sarti FM, Ceccobelli S, Milanesi M, Colli L, Ciani E, Soglia D, Sartore S, Ciampolini R, Crisà A, Steri R, Catillo G, Marletta D, Bordonaro S, D'Andrea M, Chessa S, Castiglioni B, Loi P, Sechi T, Carta A, Negrini R, Stella A, Valentini A, Panella F, Pagnacco G, et al.: Italian Goat Consortium: a collaborative project to study the Italian caprine biodiversity. In 22nd Animal Science and Production Association (ASPA) Congress; 2017.

- Talenti A, Dreger DL, Frattini S, Coizet B, Danelli F, Marelli SP, Picchi A, Riva J, Moretti E, Cocco R, Bigi D, Liotta L, Polli M, Gandini G, Pagnacco G, Ostrander EA, Crepaldi

P: Genomic landscape and biodiversity of Italian dogs. In 22nd Animal Science and Production Association (ASPA) Congress. Perugia; 2017:119.

Contribution to presentation

1. Colli L, Milanesi M, Del Corvo M, Talenti A, Bertolini F, Chen M, Crisà A, Daly K, Guldbrandtsen B, Joost S, Lenstra JA, Nicolazzi EL, Rochat E, Rosen BD, Rothschild MF, Servin B, Sonstegard TS, Steri R, Vajana E, Van Tassel CP, Ajmone-Marsan P, Crepaldi P, Stella A: Drawing up worldwide goat diversity and post-domestication history: update from ADAPTmap project. In 22nd Animal Science and Production Association (ASPA) Congress. Perugia; 2017:74–75.

Posters

- A. Talenti, M. Milanesi, E.L. Nicolazzi, L. Nicoloso, S. Frattini, B. Coizet, G. Pagnacco, J.L. Williams, P. Ajmone-Marsan, P. Crepaldi: Birth date regression to identify genomic signatures of recent selection in Italian Holstein. Italian Journal of Animal Science (2015), 14, suppl. 1, p.28.

- A. Talenti, L. Nicoloso, S. Frattini, B. Coizet, M. D'Andrea, F. Pilla, P. Ajmone-Marsan, G. Pagnacco, P. Crepaldi and the Italian Goat Consortium: Analysis of Single Nucleotide Polymorphisms in Alpine Ibex using the GoatSNP50 BeadChip. Italian Journal of Animal Science (2015), 14, suppl. 1, p.124.

- S. Frattini, E. Capra, B. Lazzari, B. Coizet, D. Groppetti, P. Riccaboni, A. Pecile, S. Arrighi, S. Chessa, B. Castiglioni, A. Giordano, D. Pravettoni, A. Talenti, L. Nicoloso, J.L. Williams, P. Crepaldi, A. Stella, G. Pagnacco: The analysis of the methylome of Capra hircus. Italian Journal of Animal Science (2015), 14, suppl. 1, p.53.

- E. Capra, S. Frattini, B. Lazzari, B. Coizet, D. Groppetti, P. Riccaboni, A. Pecile, S. Arrighi, S. Chessa, B. Castiglioni, A. Giordano, D. Pravettoni, A. Talenti, L. Nicoloso, P. Crepaldi, J.L. Williams, G. Pagnacco, A. Stella: MicroRNAs expression in hypothalamus and pituitary of Saanen goat. Italian Journal of Animal Science (2015), 14, suppl. 1, p.53.

- T. Sechi, M.G. Usai, L. Nicoloso, A. Talenti, B. Coizet, S. Frattini, G. Pagnacco, S. Casu, A. Carta, P. Crepaldi and the Italian Goat Consortium: Genetic variability of the

Sardinian goat population by the GoatSNP50 BeadChip. Italian Journal of Animal Science (2015), 14, suppl. 1, p.112.

- B. Coizet, S. Frattini, L. Nicoloso, A. Talenti, A. Tamiozzo-Calligarich, G. Pagnacco, P. Crepaldi: Study of the duoblesex and Mab-3 related transcription factor 3 gene in Italian trotters. Large Animal Review (2014), 4, suppl. 1, p.105.

- A. Talenti, M. Milanesi, E.L. Nicolazzi, S. Frattini, B. Coizet, G. Pagnacco, J.L. Williams, A. Valentini, A. Nardone, J.V. Kaam, P. Ajmone-Marsan, P. Crepaldi: Genomic retrospective evaluation of 20 years of selection in Italian Holstein bulls for feet and legs trait. Proceedings, 6th International Symposium on Animal Functional Genomic ISAFG, Piacenza, 27-29 July 2015.

- S. Frattini, E. Capra, B. Lazzari, B. Coizet, D. Groppetti, P. Riccaboni, A. Pecile, S. Arrighi, S. Chessa, B. Castiglioni, A. Giordano, D. Pravettoni, A. Talenti, L. Nicoloso, J.L. Williams, P. Crepaldi, A. Stella, G. Pagnacco: DNA methylation pattern of hypothalamus and ovary in Capra hircus. Proceedings, 6th International Symposium on Animal Functional Genomic ISAFG, Piacenza, 27-29 July 2015.

- E. Capra, S. Frattini, B. Lazzari, B. Coizet, D. Groppetti, P. Riccaboni, A. Pecile, S. Arrighi, S. Chessa, B. Castiglioni, A. Talenti, L. Nicoloso, A. Giordano, D. Pravettoni, P. Crepaldi, J.L. Williams, G. Pagnacco, A. Stella: MicroRNAs expression in Hypothalamic-Pituitary-Gonadal axis in goat. Proceedings, 6th International Symposium on Animal Functional Genomic ISAFG, Piacenza, 27-29 July 2015.

- A. Talenti, Rota G., Frattini S., Coizet B., Minozzi G., Pagnacco G., Crepaldi P. and the Italian Goat Consortium: Genomic study of horn morphology in Italian goat populations. Proceedings to EAAP (2016).

- S. Frattini, Lazzari B., Capra E., Talenti A., Coizet B., McKay S.D., Stella A., Pagnacco G., Crepaldi P.: DNA Methylation and Gene Expression Levels in Hypothalamus and Ovary of Capra hircus Across the Genome. Large Animal Genetic Engineering Summit (2016).

- Talenti A, Cortellari M, Milanesi M, Frattini S, Colli L, Pagnacco G, Ajmone-Marsan P, Crepaldi P: Identification of genomic regions of recent selection for productive and reproductive traits in Italian Holstein bulls. In 22nd Animal Science and Production Association (ASPA) Congress. Perugia; 2017:164–165.

- Talenti A, Dreger DL, Danelli F, Frattini S, Coizet B, Marelli SP, Pagnacco G, Gandini G, Polli M, Caniglia R, Galaverni M, Ostrander EA, Crepaldi P: Whole genome

analysis of the Lupo Italiano. In 22nd Animal Science and Production Association (ASPA) Congress. Perugia: Italian Journal of Animal Science; 2017:165–166.

- Frattini S, Marelli SP, Picchi A, Danelli F, Riva J, Moretti E, Talenti A, Gandini G, Pagnacco G, Polli M, Crepaldi P: Genetic trend of the junctional epidermolysis bullosa (JEB) in the German Shorthaired Pointer in Italy. In 36th International Society for Animal Genetics (ISAG) congress. Dublin; 2017:167.