Unravelling the evolutionary history and adaptation of European mouflon and some domestic sheep populations with special emphasis on the ovines of Sardinia

Mario Barbato

Thesis submitted to Cardiff University in candidature for the degree of Doctor of Philosophy

December 2015

Cardiff School of Biosciences

Cardiff University

Unravelling the evolutionary history and adaptation of European mouflon and some domestic sheep populations with special emphasis on the ovines of Sardinia.

Mario Barbato *

Main supervisor:

Prof. Michael W. Bruford

Co-supervisor:

Dr. Pablo Orozco-terWengel

Organisms and Environment research group, Cardiff School of Biosciences, Cardiff University, Sir Martin Evans Building, Museum Avenue, Cardiff CF10 3AX, Cardiff UK.

*Email for correspondence: barbatom@cardiff.ac.uk

To my mother Francesca and my sister Anna,

none of this would have been possible without your support…

Acknowledgements

First and foremost, I would like to thank Mike: you allowed me to do research in the most pleasant environment possible, always ready to support me, providing example and inspiration, it really was a great privilege to work with you. Pablo, since day one, we equally distributed our time together either discussing science or fighting over a great variety of topics, where I am afraid, you were mostly wrong (crazy ideas on which coffee is the best and such). It surely was fun, and I really felt privileged being the one teaching you about good food!

Pier: he is one of the reasons I am in Cardiff, the one that helped me to settle-in once I arrived, and with whom I shared many of the good and bad things that happened over the past four years; that is what friendship is all about. Mafalda and Isa, two of the most peculiar characters I have ever met (that must be why I fit with you), thanks for all the chats, laughs, advice and inspiration, you really made this adventure a special and unforgettable one. Thanks to the fellowship of C/5.15: Dave and his talent with foreign languages, Niall and the misapplication of tiger-balm, Josie (Lidl?), Silke and her extremely relaxed approach to life (not), Hannah MacBurton and her cooking (s)kills, Joana and the Portuguese crash course, Rose aka the kindest person on the blue planet, Renata and the music chats, Camilla and her geekiness, Jez and…oh…he left already, Jen and the English grammar, and Julia, Gerardo and Uxia, I am looking forward to the next Galician session! And Renatinha, Jorge, Gonçalo, Luis, Marta, XJ, Pete, Ancuta and Silvie.

Some I met later on and yet had the most profound impact on my life in Cardiff. Natalia and the ten thousand angles that she can see while commenting everything, quite in contrast with her dietary choices though. Frank, meeting you changed the tide of my PhD, thanks for all the chats and the many laughs (so unfortunate you do not understand Italian food tradition that much). Also thanks to Dr. Sula, the best-behaved post-dog at CU. Elia, quanta mestizia, ma quanta gaiezza pure, prossimo obbiettivo: un bel piatto di tortelli con la coda! And those outside CU: Anna and Nicole, the best housemates ever. Laura, for her unwavering cheerleading. Susanna, always there for me, no matter the distance.

And Morena, having the chance to know you made everything worth it.

Summary

After being transported into Europe during the Neolithic, mouflon (*Ovis aries musimon*) became extinct from mainland Europe, but remnant populations persisted and became feral on the Mediterranean islands of Corsica and Sardinia. These populations have been used for reintroductions across continental Europe during the last 200 years. This thesis aimed to investigate the global and local ancestry of European mouflon and domestic sheep, to investigate signals of artificial and natural selection in their genomes, and to develop analytical frameworks and informatic tools to aid similar analyses using SNP array data. I describe the development of software that allows rapid investigation of genome-wide SNP data to infer effective population size trajectories using patterns of linkage disequilibrium**.** I inferred the absence of widespread sheep introgression in extant European mouflon populations although signals of recent introgression were recorded in one enclosed Sardinian mouflon population. By applying a novel approach to aid the investigation of local genomic ancestry data, signals of mouflon ancestry in sheep could be inferred and were found to be related to biological functions involved with innate immunity processes with bitter taste recognition being identified in two breeds known for their broad dietary choices. By investigating signals of positive selection and local adaptation in feral and domestic sheep using novel locus-specific empirical p-value inference, traits with selection signatures such as fertility, pigmentation and behaviour were identified in sheep, while traits involved with stature - probably related to mating success - were found in mouflon. Signals of local adaptation to environmental variables were not detected, which is likely to be due to the inadequate sample available, determined by *post-hoc* analysis.

Contents

Chapter one

Mouflon hunting in Sardinia (Illustrierte Zeitung, 1893)

1 Chapter one – General introduction

1.1 Background on sheep and mouflon

The genus *Ovis* is classified in the *Bovidae* family, subfamily *Caprinae*. According to molecular studies, the *Caprinae* subfamily diverged from *Bovidae* around 15-20 million years ago (MYA) and, within *Bovidae,* the divergence between the genera *Capra* and *Ovis* dates to around 5-7 MYA (Bruford and Townsend, 2006). Wild and feral (domesticates that have returned to the wild state) sheep can be found across the majority of the northern hemisphere, and domesticated sheep are present almost all-over the world. The taxonomy of wild and feral sheep is, however, quite controversial (Rezaei et al., 2010) but they are both generally referred to with using common name "mouflon".

1.1.1 Mouflon taxonomy in general

Mouflon can be found in Europe, the near east, Asia and North America. The different sub-species of mouflon differ in terms of morphological traits such as: body size, coat colour, pelage pattern and horn length and morphology. Additionally, they differ in terms of ploidy (Nadler et al., 1973) and yet the hybrids are fertile and present intermediate chromosome numbers (Rezaei et al., 2010). These factors have contributed to a quite controversial classification of these species: in fact, the systematics of the genus have been profoundly modified during the last century (Rezaei et al., 2010) with as little as one and as many as seven different taxa classified among wild sheep (Table 1-1). Currently, taxonomic status is well-established for some species: Dall's sheep (*O. dalli*) typical of North West Canada and Alaska, Bighorn sheep (*O. canadensis*) from West US, Snow sheep (*O. nivicola*) from Siberia, Argali (*O. ammon*) from central Asia and Urial (*O. vignei*) from western central Asia.

The designation of the *O. orientalis* group remains debated. Within this group are the Armenian mouflon as *O. orientalis gmelinii* but more often defined as *O. gmelinii anatolica* (Demirci et al., 2013). The two Iranian *O. o. isphahanica* and *O. o. laristanica*, which sometimes are classified as *O. gmelinii isphahanica* and *O. g. laristanica* respectively (Rezaei et al., 2010). The Cypriot mouflon denomination changes between *O. ophion*, *O. gmelinii ophion* and *O. orientalis ophion*, and the

European mouflon, once *O. o. musimon*, is now mostly (but not exclusively) described as *O. aries musimon* (Rezaei et al., 2010; Sanna et al., 2015).

Table 1-1 Different classifications of the genus *Ovis* (Table 1 from Rezaei et al., 2010).

Authors	Tsalkin (1951)	Haltenorth (1963)	Nadler et al. (1973)	Valdez (1982) Wilson and Reeder (1993) Shackleton and Lovari (1997)	Festa-Bianchet (2000)
Dall Sheep	*O. canadensis/O. nivicola*	*O. ammon*	*O. dalli*	*O. dalli*	*O. dalli*
Bighorn	*O. canadensis/O. nivicola*	*O. ammon*	*O. canadensis*	*O. canadensis*	*O. canadensis*
Snow Sheep	*O. canadensis/O. nivicola*	*O. ammon*	*O. nivicola*	*O. nivicola*	*O. nivicola*
Argali	*O. ammon*	*O. ammon*	*O. ammon*	*O. ammon*	*O. ammon*
Asiatic Mouflon	*O. ammon*	*O. ammon*	*O. orientalis*	*O. orientalis*	*O. gmelinii*
Urial	*O. ammon*	*O. ammon*	*O. vignei*	*O. orientalis*	*O. vignei*
European Mouflon	*O. ammon*	*O. ammon*	*O. musimon*	*O. orientalis musimon*	*O. orientalis musimon*

1.1.2 Mouflon morphology

The European mouflon is one of the smallest extant sheep in the wild, with a height at withers of 65-75 cm in females and 70-80 cm in males and a weight between 25-35 kg in females and 35-55 kg in males. Males have triangular sickle-shaped perennial horns that can reach a meter in length, females instead tend to be polled (hornless) or with smaller horn size than males. The frequency of polled ewes varies depending on the population. Between the two Mediterranean island populations Corsica has up to 60% of horned ewes (Bon et al., 1991), while Sardinian ewes are mostly polled (Cetti, 1774). Unlike domestic sheep and similarly to all wild sheep, the European mouflon has short ears and the tail is also short (~10 cm). Mouflon coat has a red-brown colour with a dark area along its back, and lighter coloured side patches, often referred to as 'saddle', that can be more evident in winter when the rest of the coat becomes darker. The white saddle patch becomes larger with age. During winter mouflon grow a woolly undercoat that sheds in spring. The coat becomes darker during winter. The ventral part of its coat is white as well as the bottom half of their legs. Mouflon have a white muzzle, and the eyes are often circled in white (Figure 1-1).

Figure 1-1 European and Asiatic mouflon. A) Sardinian mouflon. B) Corsican mouflon. C) Cypriot mouflon. D) Argali wild sheep.

1.1.3 Mediterranean Colonization pattern

Archeozoological evidence indicates that the genus *Ovis* appeared in Asia during the Pleistocene around 1.8–2.4 MYA. Remains of an ancient form of Argali (*Ovis ammon antiqua*) from the Middle Pleistocene have also been found in most of southern and central Europe (Hungary, Italy, France, Spain and Portugal), with several of these remains suggesting they were hunted by humans (Santiago-Moreno et al., 2004; Rozzi et al., 2011; Rival, 2000). However, climate change, hunting pressure and the arrival of the domesticated species might have combined to reduce numbers to an extent that few or no wild mouflon survived in mainland Europe. Although the presence of remains belonging to the genus has been confirmed throughout Europe, none have been found on the islands of Corsica or Sardinia, and which could be dated to the same strata. The absence of *O. a. antiqua* records has been attributed to the lack of ideal conditions for ancient bone conservation on these islands (Santiago-Moreno et al., 2004). Corsica and Sardinia belong to the same geological microplate and land connecting the two island periodically appeared until the beginning of the Holocene (Vigne, 1992). However, no land bridge from Corsica to mainland Italy was present at the time

that the first *Ovis* species appeared in Europe in the late Pliocene (Magri et al., 2007).

1.1.4 The European mouflon goes feral

Poplin (1979) was among the firsts to suggest the origin of the Corsican and Sardinian mouflon due to feralization of semi-domesticated animals introduced in the island around 5-6,000 YA (years ago), supporting his hypothesis with the lack of paleontological evidence in older strata. The same conclusions were reached by Vigne (1992) that similarly dated the depletion of a pre-existing late Pleistocene fauna and the discovery of different terrestrial mammals (such as *Ovis*) with human colonisation at around 10,000 YA (Vigne, 1992) (Figure 1-2). The feralization hypothesis gained additional support when molecular studies were applied to the mouflon populations from Corsica and Sardinia. Analyses of the mitochondrial control region first (Hiendleder et al., 2002) and of the whole mitogenome later (Meadows et al., 2011) showed that the European mouflon clusters within the sheep haplogroup representative of the majority of European domestic sheep breeds.

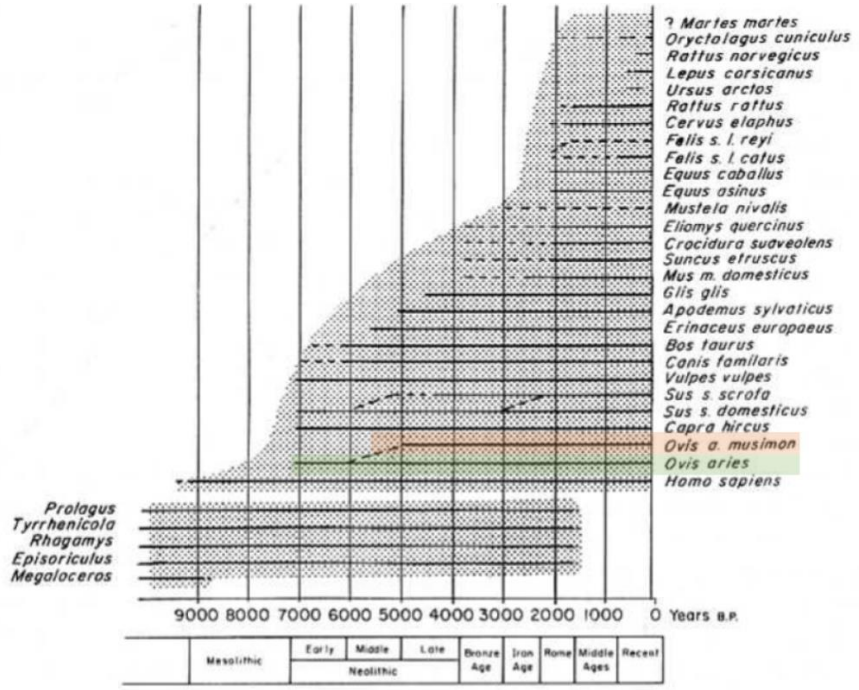


Figure 1-2 Chronological distribution of Corsico-Sardinian terrestrial mammals since the beginning of human occupation (Figure 2 from Vigne at al., 1992). Mouflon and sheep are highlighted in red and green, respectively. According to this representation sheep remains predate mouflon; however, wild and domestic remains were impossible to discriminate prior to the late Neolithic decrease in size of domestic sheep (Vigne et al., 1992).

1.1.4.1 Reintroduction in Europe

The first mouflon to be brought to mainland Europe were a few individuals captured in Corsica and moved to the Vienna royal hunting reserve around 1730 (Santiago-Moreno et al., 2004). However, it was around the end of the XVIII century that larger numbers of animals were translocated from the islands to mainland Europe to provide park animals at first and in the following century to address the game requirements of hunters (Bon et al., 1991). The first large populations were created in southern France by using Corsican mouflon (Bon et al., 1991). Animals from Austria were then used to start mouflon populations in Czechoslovakia around 1858 and animals from this latter population were then used as founders to start new mouflon populations in most central European countries. In countries like Spain, animals from both France and Germany were used to start mouflon populations in several forested areas of the country (Santiago-Moreno et al., 2004). Not much is known about the use of mouflons from Sardinia, although there are records of Sardinian and Corsican mouflon being used to repopulate the north Apennines in Italy (Guerrini et al., 2015).

The current largest mouflon population in mainland Europe (Czechoslovakia and Germany) is ~60,000 individuals (Carnevali et al., 2009) while the autochthonous populations of Corsica and Sardinia are ~1,000 and ~6,000 animals, respectively (Apollonio et al., 2005).

1.1.4.2 Current status of the Sardinian Mouflon

After becoming feral, the Sardinian mouflon diffused across the island, thanks to the very low human density and rich pastures. In 1774 Cetti wrote "*With luck, the slaughter can count hundreds [of animals]...*" (Cetti, 1774). In a book from 1841 on Sardinia and its traditions, there is a chapter on the Mouflon and it ends with this statement: "*...in the Gennargentu ravines you can see countless flocks roaming around...*" (Botta, 1841). A document from 1904 defined the mouflon as "*extremely common*" on the island (Nicolai, 1904). However, in 1911 the naturalist Gighi recorded a decline in the Sardinian mouflon population (Gighi, 1917, cited in Apollonio et al., 2005) and in a document from 1948 concerning Sardinian fauna, while describing ancient Sardinian hunting practices the author wrote about hunting sessions in which more than 500 animals were killed at once

and in the same article he complained about the smaller number of animals present during his time (Biagini, 1948). These statements give us a hint on how large the mouflon population was and at the same time about the reason why the population declined. By 1969 it was claimed that the population numbered only 300-360 animals, with a minimum of 300 declared in 1978, although a later record (from 1985) recorded an increased population of 1,500-1,600 animals (Cassola, 1985, cited in Apollonio et al., 2005).

Figure 1-3 Nuragic bronze statuettes from Sardinia representing mouflons (IX-VIII century BC). The animals here represented have curved horns and a short tail that distinguish them from sheep rams. A-B) Mouflons. C) Warrior with spear offers a small mouflon. D) Man carries a mouflon lamb on his shoulders. (Museo Nazionale Archeologico-Etnografico "G. A. Sanna", via Roma, 64, Sassari (SS), Sardegna - Italy).

The recent population increase is thought to be due to several laws enacted in 1924 (decrt. Minist. 7 aprile 1924) that reduced the hunting season along with promoting the Sardinian citizens' appreciation of preserving their native fauna (Figure 1-3). At this point several areas became natural reserves for the mouflon, where human presence and domestic sheep pastures were decreased sufficiently to give mouflon room to multiply. In 1998 (Legge Regionale n. 23 del 1998) the European mouflon was declared an endangered species and protected from hunting in both Corsica and Sardinia; moreover, its protection was registered in the II/IV EU Habitats directive 92/43/EEC. Nowadays the Sardinian mouflon comprises around 6,000 animals that can be found on the Gennargentu massif, and mouflon populations have been reintroduced to other districts of the island (Figure 1-4).

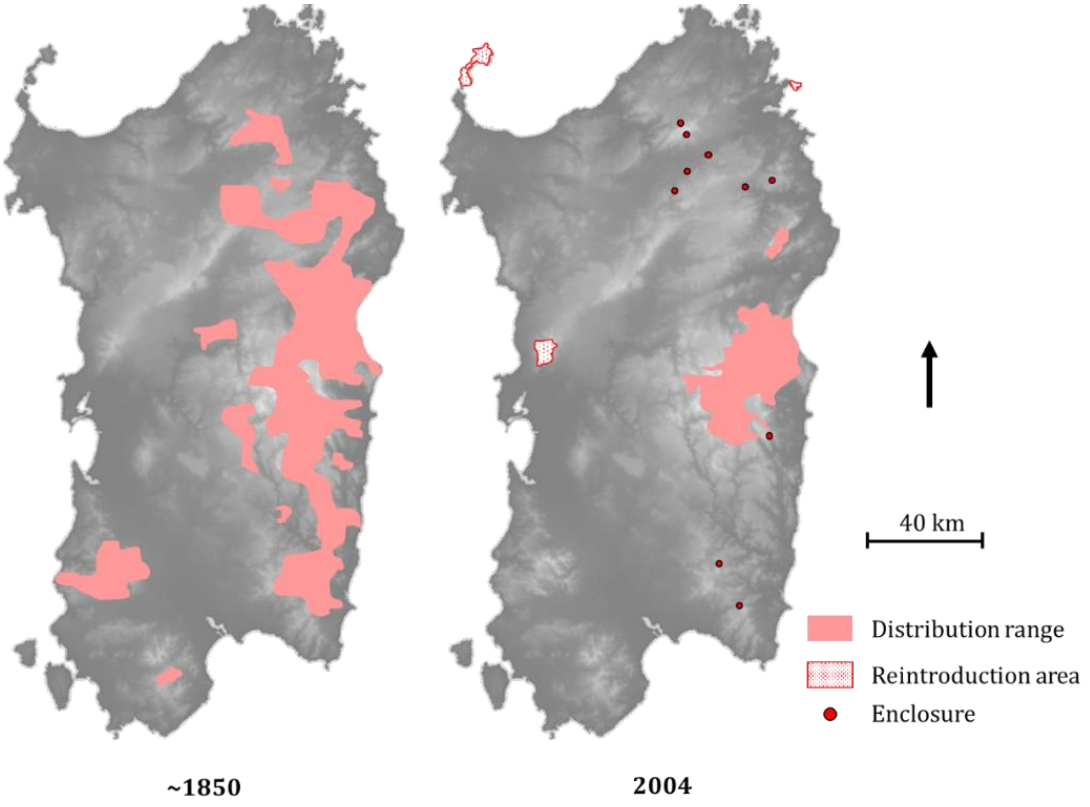


Figure 1-4 Past and current mouflon distribution in Sardinia (Apollonio et al., 2005).

1.2 Domestication

The process of domestication has been described as "the process of increasing mutual dependence between human societies and the plant and animal populations they target" (Zeder et al., 2006) and is *de facto* one of the main processes to contribute in shaping human society as we know it.

Domestication is both a biological and cultural process. The cultural aspect is due to the integration of an animal population into the social structure of a human community, therefore involving the concepts of ownership, inheritance and exchange (Clutton-Brock, 1992). Once the animal population becomes permanently isolated from the wild species and all the aspects of their life are under human control, the biological aspect of domestication takes place, genetically and morphologically shaping the tamed species into a domesticated one (Clutton-Brock, 1992; Zeder et al., 2006).

The dependence of domesticated populations on humans is reflected by species changes affecting morphological, physiological and behavioural traits. Domestication acts as an evolutionary force, with both natural and artificial selection shaping its genome. Additionally, is important to note that not all animals can be domesticated (Clutton-Brock, 1999) and the majority of the domesticated species share specific characteristics compatible with humans' necessities: they are gregarious, breed readily in captivity, have a wide home range and a short flight distance (Clutton-Brock, 1999), features often found in herbivores as in the sheep.

The effect of artificial selection is evident on the conscious choice of the herder in choosing a specific phenotype, either by controlling the mating in favour of the most desirable traits or by culling. This is the case with the absence/reduction of horns in sheep and goats. Whereas the wild counterparts (mouflon and bezoar, respectively) have large horns, the domesticated species are either polled (as for most sheep) or have smaller and twisted horns (as for the goat). This feature was - and is - advantageous for the herder who would avoid managing animals with dangerous weapons. Also, reduction in body size has been related to domestication and is visible in almost all the early domesticated animals

according to archeozoological findings (Zeder, 2008). Reduction in body size could also be due to the choice of early herders to manage smaller animals. However, both body size and horn presence could also be a by-product of adaptation by wild animals to a novel anthropogenic domain.

One of the main differences between the wild and human controlled condition is the human mediated mating scheme. Both sheep and goats are characterised by a polygamous mating system, where a strict hierarchical structure via competition among males determines the alpha-male that will control the harem (Zohary et al., 1998). Body and horn size play a dominant role in the outcome of these fights, with larger, more agile males with better horns having the greatest chances of producing a progeny (Grignolio et al., 2008). Under domestication controlled conditions, the dominant-male mating system is removed, and consequently the selective pressure to have large body and horn size is relaxed, allowing smaller males with reduced or differently shaped horns to be selected for breeding. Moreover, while carnivore predation plays a critical role as a selective force in the wild by shaping both morphology (e.g., camouflaged coat, long limbs) and behaviour (e.g., danger awareness, aggression), this factor loses its importance when humans (or other domesticated animals) provide protection from predation. Similarly, characters such as the wild type colouration do not increase fitness under human care and colour might instead be chosen by the herder for distinctiveness or utility (Zohary et al., 1998). It has been also hypothesised that the concurrence of low-predation risk and nutritional stability influences cognitive abilities, with domestics not having to remembering feeding sites or be aware of changes in their surroundings (Brust and Guenther, 2015); however, although a functional link between behaviour and cognitive traits exists its extent is still debated (Brust and Guenther, 2015).

As a consequence, deciding which features are due to selective breeding (by the first herders) or natural selection (due to the change in environment from wild to anthropogenic) is difficult, as the two factors might have likely contributed in shaping the domestic species as we know them.

1.2.1 Ovine agricultural status in Sardinia

Sheep farming in Sardinia is by far the most important agricultural activity, characterising the island economically, socially, culturally and historically. With almost 4 million sheep Sardinia holds ~45% of the ovines in Italy, producing over 60% of the national sheep milk. Despite its prominent role for the economy of the island, sheep farming in Sardinia has not kept the pace with advancements in technologies, with the vast majority of the flock management involving grazing pastures, transhumance and manual milking (Serra, 2009). One of the reasons why such high productivity has not imposed improvements in managing techniques can be found in the adaptability to different pastures of the main sheep breed of Sardinia: the Sarda sheep (Figure 1-5). The Sarda accounts for almost all the sheep in the island with only ~2,000 animals belonging to another autochthonous breed, the Nera di Arbus (Figure 1-5), and a few non-Sardinian breeds (Associazione Nazionale della Pastorizia: www.assonapa.com, 2012).

Figure 1-5 Two Sardinian autochthonous sheep breeds: the Sarda (A) and Nera di Arbus sheep (B).

1.2.1.1 Sarda sheep

Sarda is a medium size breed (65 and 45 kg on average for adult ewes and rams, respectively), polled, with a white fleece and a very pale skin. The main characteristic of the breed is milk productivity, with 200 Litres of milk per year produced on average, and over 550 Litres for the best producing animals (Le Razze Ovine e Caprine in Italia, 2002). The Sarda has an extremely good

prolificacy (~1.3 lambs per birth), whereas wool quality is poor and mostly used as insulating material (Le Razze Ovine e Caprine in Italia, 2002). Along with the high milk and meat productivity, broad dietary choices and the high adaptability of the Sarda has resulted in the diffusion of this breed overseas, with almost a million sheep present in Central and Southern Italy and other districts of the Mediterranean (Le Razze Ovine e Caprine in Italia, 2002).

1.3 Molecular tools

Ovine taxonomy (see 1.1.1) has traditionally used archaeological, morphological and kariological analyses with little incorporation of molecular data (Poplin, 1979; Rando et al., 1996). However, in the last decade several molecular studies have been carried out which have shed light on the genus' systematics. Until recently two genetic markers were dominant in genetic studies on sheep: mitochondrial DNA (mtDNA) sequences and microsatellites (SSR) (Hiendleder et al., 2002; Bruford and Townsend, 2006; Tapio et al., 2006; Meadows et al., 2007; Moioli et al., 2010). More recently, Chessa at al. (2009) used endogenous retroviruses to reveal additional patterns not accessible using the other marker types. However, in the last decade, single nucleotide polymorphisms (SNPs) played a major role in investigating the history of the genus *Ovis*, thanks to the availability of relatively inexpensive genotyping arrays.

1.3.1 SNPs

A single nucleotide polymorphism (SNP) (also known as SNV, or single nucleotide variant) is the variation at a single nucleotide at a specific genomic position. This variation can be measure between chromosomes within an individual's genotype, or among members of a biological group such as a family, population or species. SNPs mostly occur due to point mutations in the germ line, and if the mutation is not repaired and the cell produces a gamete, the mutation can be inherited by the progeny.

1.3.1.1 SNP identification

Advances in technology have transformed SNP detection from being labour intensive, time consuming, and expensive to highly automated, efficient, and relatively inexpensive (Kwok and Chen, 2003), allowing SNPs to become one of

the most widely used class of genetic markers. SNP detection includes the identification of previously unknown polymorphism (SNP calling) and screening for known polymorphism (SNP genotyping) (Kwok and Chen, 2003; Nielsen et al., 2011).

1.3.1.2 SNP arrays

DNA microarray technology represents an expanding approach in the genomic field, as it allows tens of thousands of genetic markers to be rapidly assayed at the same time. DNA microarrays use DNA probes attached to a solid surface and synthesized to match a specific genomic region. They can be used to detect both DNA and RNA (such as cDNA) allowing a wide variety of applications such as: gene expression profiling and SNP detection.

The most advanced SNP array technology has been applied to human research and SNP chips with ~2M SNPs and ~1M probes for copy number variation (CNV) are currently commercially available (e.g., Affimetrix Genome-Wide Human SNP array 6.0). Additionally, several SNP chips have been assembled in order to genotype the most important livestock species and moderate density chips (e.g., ~50k SNPs) are currently available for: sheep (*Ovis aries*), horse (*Equus caballus*), pig (*Sus scrofa*), cattle (*Bos taurus*) and chicken (*Gallus gallus*), and for some of them high density chips (e.g., >500K SNPs) have been recently developed (e.g., cattle and sheep).

There are two main phases involved in producing a SNP chip: SNP discovery, and the SNP selection. In the discovery phase, the genome of a selection of animals is obtained and then SNPs are identified by aligning the genomes to a reference. In the second phase SNPs are selected based on parameters such as allele frequency and call rate. SNPs are also selected according to their physical position in order to achieve approximately even spacing across the genome. Most of the SNPs chosen with this approach are putatively neutral. However, several SNP that are known to have a specific impact on phenotypes of interest have been included.

The most critical aspect that ultimately defines the utility of SNP chips depends on the very first phase of the SNP chip development process, namely, the composition of the SNP discovery panel. Since the set of samples used to ascertain

polymorphisms to be included in the chip is usually not representative of the entire species, the polymorphisms identified tend to be biased towards being polymorphic in samples related to the ones used for the chip development. This inherent 'ascertainment bias' results in the identified polymorphisms not necessarily reflecting the true genetic variation in other less related samples, if they are even polymorphic in such samples.

Ascertainment bias has been described as: "the systematic deviation from the expected allele frequency distribution that occurs because of the sampling processes used to find (ascertain) marker loci" (Helyar et al., 2011). There are two main elements that impact on the ascertainment bias, its width and its depth. Width describes the subset of individuals from the species' range used for the SNP discovery, and depth refers to the minor allele frequency (MAF) threshold defined for a SNP to be included in the chip. An insufficient ascertainment width and depth leads to a SNP selection that excludes rare variants and has an impact on analyses that rely on allele frequencies. This impact translates in the overestimation of the average diversity of polymorphic sites (fewer rare variants are sampled and SNPs with MAF closer to 0.5 than to 0 are more likely to be selected) and an underestimation of the average diversity across all sites (Helyar et al., 2011; Miller et al., 2012a).

1.3.1.3 How to tackle ascertainment bias

Ascertainment bias is inherent in any marker development and can influence SSRs as well as SNPs although for SSRs the higher mutation rate tends to reduce its impact on this marker type (Li and Kimmel, 2013; Albrechtsen et al., 2010).

Three main methodologies have been applied to tackle this bias: i) the use of investigation methods not dependent on single locus allele frequencies, as in the case of principal component analysis or haplotype dependent approaches (e.g., EHH-derived methods, see 1.4.3.1.1) (Nielsen et al., 2007), ii) the simulation of datasets that underwent the same ascertainment process in order to derive confidence intervals to apply to the results (Voight et al., 2006), and iii) the correction of the statistics applied through specific models (Nielsen et al., 2004). However, the latter two approaches rely on precise knowledge of the SNP discovery protocol (Nielsen et al., 2007), which may be not available.

Contrastingly, the recently developed genotype-by-sequencing (GBS) technologies might provide an ascertainment bias-free alternative to cross-species application of SNP chips for both marker discovery and genotyping (Davey et al., 2011; Elshire et al., 2011).

1.3.1.4 Illumina Ovine50SNP BeadChip

A mid-density chip able to simultaneously study 54,241 positions in the ovine genome for 12 samples was released by Illumina Inc. with the commercial name 'Ovine SNP50 BeadChip' (Kijas et al., 2012). The SNPs chosen came from three different sequencing efforts that made use of data obtained using 454 Life Technologies (33,115 SNPs), Illumina (15,427 SNPs) and Sanger (492 SNPs) sequencing for a variety of domestic sheep breeds (Table 1-1). Of the 26 sheep breeds used for SNP discovery 17 were European, five Asiatic, two were South-American and two were African. The initial pool of polymorphisms identified was pruned to discard singletons and low quality variants, resulting in 354,448 putative SNPs left. This set was further filtered for MAF (SNPs with MAF<0.2 were discarded). To ensure evenly spaced SNPs through the genome, some SNPs where further discarded due to technical problems in implementing them into the actual chip leaving 54,241 SNPs available for the chip. A high density chip able to assay approximately 685,000 evenly spaced SNPs was recently developed (J. McEwan *unpublished*) and used to investigate the decay of linkage disequilibrium in five industrial sheep breeds (Kijas et al., 2014). However, this SNP array has not been included in a commercial catalogue yet, and no information is publicly available on the specifics of its development.

Table 1-2 Sheep breeds used in the SNP discovery panel of the Ovine SNP50 Beadchip (Table S2 from Kijas et al., 2012).

Discovery Set	Breed	No. Per Breed	Breed Development
454	Poll Dorset	1	Europe and Australia
454	Merino	1	Europe and Australia
454 and Illumina GA	Awassi	1	Middle East
454	Texel	1	Europe and New Zealand
454	Romney	1	Europe and New Zealand
454	Scottish Blackface	1	Europe
Sanger	Poll Dorset	1	Europe and Australia
Sanger	Merino	1	Europe and Australia
Sanger and Illumina GA	Awassi	1	Middle East
Sanger	Lacaune	1	Europe
Sanger and Illumina GA	Red Masai	1	Africa
Sanger	Texel	1	Europe and New Zealand
Sanger	Romney	1	Europe and New Zealand
Sanger and Illumina GA	Katahdin	1	Africa and USA
Sanger and Illumina GA	Gulf Coast Native	1	Africa and USA
Illumina GA	American Suffolk	5	Europe and USA
Illumina GA	Scottish Blackface	5	Europe
Illumina GA	Indonesian Thin Tail	5	Asia
Illumina GA	Italian Sarda	5	Europe
Illumina GA	Merino	5	Europe and Australia
Illumina GA	Poll Dorset	5	Europe and Australia
Illumina GA	Romney	5	Europe and New Zealand
Illumina GA	Sumatran Thin Tail	5	Asia
Illumina GA	Texel	5	Europe and New Zealand
Illumina GA	Tibetan	5	Asia
Illumina GA	Namaqua Afrikaner	5	Africa

1.3.2 SNP array data applicability

The availability of genotyping arrays filled a perceived gap between the use of a low number of highly informative markers (e.g., SSR) and the 'gold standard' of whole-genome sequencing. While improving the resolution (e.g., reducing the variance) of statistics normally applied using a few markers (e.g., population structure or diversity indices), arrays have also enabled research previously limited by the need of expensive sequencing efforts (see Sabeti et al., 2007).

The availability of these new tool for the research community prompted the development of new informatics tools tailored for mid- to high-density SNP data (Nicolazzi et al., 2015), as well as the application of tools intended for extremely

high density data (e.g., 3M SNPs in human of the '1000 Genomes' project; Abecasis et al., 2010).

The background on the methodological approaches that benefitted the most from the advent of mid- to high-density SNP arrays and that are used in this thesis will be discussed in the following sections.

1.3.2.1 Population structure

Studies on population structure are key to understanding the evolutionary history of a species. Structuring within a population can be fuelled by a variety of drivers, e.g., inbreeding or demographic isolation, counteracted by gene-flow among sub-populations and revealed by the differentiation of allele frequencies (Hartl and Clark, 2007).

Population structure can be investigated in terms of genome ancestry starting from the assumption that any current individual genome or population is a mixture of ancestries from (one or more) past populations (Pugach and Stoneking, 2015; Wollstein and Lao, 2015). Further, genetic ancestry can be inferred at global and local level (Alexander et al., 2009). At the global level, the proportion of ancestry from each contributing population is inferred, whereas at local level the ancestry of genomic regions of each chromosome is inferred (Alexander et al., 2009).

1.3.2.1.1 Inferring global ancestry

Global ancestry estimation approaches can be separated into model-based and algorithmic approaches (Wollstein and Lao, 2015). The former evaluates ancestry coefficients as the parameters of a statistical models (e.g., as implemented in STRUCTURE, ADMIXTURE or FRAPPE) (Pritchard et al., 2000; Alexander et al., 2009; Tang et al., 2005), whereas the latter represents genetic relationship by a new set of orthogonal variables that are ordered by the amount of variation explained (e.g., principal component analysis – PCA – or multidimensional scaling methods - MDS) (Wollstein and Lao, 2015).

The algorithmic approaches (e.g., PCA) do not require any *a priori* assumptions of population structure as only those variables with the highest amount of explained variation (that maximally separate the data) are commonly considered

to infer and visualise the dataset genetic structure (Schraiber and Akey, 2015). Additionally, the depth and precision of the inference increase with the amount of data provided (Novembre et al., 2008).

The model-based methods instead work on the assumption of an expected underlying population structure (such as a number of prescribed clusters, K). Subsequently the methods infer the sample arrangements among the expected K clusters, in order to satisfy some method-dependent convergence parameters (Alexander et al., 2009). The user normally tests different K values with several replicates for each K. Several implementation-dependent methods have been proposed to estimate which of the tested K best fits the data (Schraiber and Akey, 2015). However, although considering a single value of K as correct, according to post-analysis methods seems tempting, it has been argued that no 'optimal' K exists in reality and that many clustering solutions can provide interesting insights on the species under analysis (Björnerfeldt et al., 2008; Orozco-terWengel et al., 2011). These methods, alone or in combination, allow for the identification of population sub-structure as well as individual ancestry, and are therefore capable of identifying admixture at an individual level. Along with inferring population structure, model-based methods provide an estimate of the fraction of each individual's genome that originates from each population (Schraiber and Akey, 2015), allowing identification of admixture.

1.3.2.1.2 Inferring local ancestry

Admixture can be further explored by identifying the ancestors of each genomic region at the chromosome level. These approaches are described as local ancestry deconvolution or chromosome painting. Methods and implementations that can produce profiles of local ancestry (Pasaniuc et al., 2009; Brisbin et al., 2012) are based on dividing the genome into windows, with ancestry assignment for each window defined by comparing it against a reference panel (Brisbin et al., 2012). This model-free approach allows the inference of local ancestry also in case of extremely complex or unknown underlying demographic histories; however, interpreting the results can be problematic (Schraiber and Akey, 2015). A knowledge of local ancestry can help in understanding fine scale admixture,

including the timing of the event (Brisbin et al., 2012) and can also help in determining recent targets of selection (Tang et al., 2007).

1.4 Selection

The idea that a beneficial trait could increase the survival and reproductive success (fitness) of an individual was first formally introduced by Darwin and Wallace in mid-19th century while theorising the principles of natural selection. Since then, selection became a keystone concept in evolution, along with mutation and genetic drift (Kimura, 1991). Different selection types have been described, including a distinction between natural and artificial selection. The former refers to ecological and sexual selection, and the latter often refers to selective breeding. Independent of the primary force driving selection, higher fitness can be signified at the molecular level by the detection of a non-random frequency of genetic variants that are linked to the beneficial trait.

1.4.1 Identifying selection with genomic data

Prior to the last decade the identification of selection was performed in a 'top-to-bottom' fashion with candidate traits/genes putatively under selection investigated to confirm or reject the assumption. By using this approach, notable results were achieved, e.g., identifying positive selection for the gene LCT, which allows lactose tolerance throughout adulthood in humans (Bersaglieri et al., 2004). However, the need of prior knowledge of the gene under selection (e.g., the genotype/phenotype relationship) and the difficulty of properly evaluating the confounding effect of demography reduced the impact of this approach to a surprisingly small number of cases (Akey, 2009).

The advent of next-generation sequencing and the availability of genome-wide data have helped to overcome these limitations. By scanning the whole genome, no *a priori* assumption is needed on which gene or genic system to examine, removing some bias in detecting regions under selection (Akey, 2009). Moreover, null hypotheses on the genetic drift-driven selection (e.g., standing variation) can be incorporated, allowing reduction of I type error (Akey, 2009).

For example, the full genome coverage of SNP arrays allowed the development of genome wide association study (GWAS), where a population showing specific

phenotypes for a particular trait or disease are compared with those not presenting the same phenotype. SNP variants linked to genes or regulatory elements and associated with the traits of interest can then be identified (Zhang et al., 2012).

1.4.2 Patterns of selection

Several selection patterns have been described including two major categories which are: *positive* and *negative selection* (Vitti et al., 2013). The baseline of this classification refers to the occurrence of an allele variant, which is retained by selection (positive selection) if beneficial, or removed (negative selection). Negative selection is more often described as *purifying selection*, due to the removal of disadvantageous alleles. The presence of linkage between a variant under selection and other loci produces what is described as *linked selection*. Within this latter scenario neutral variants are either selected for (*genetic hitchhiking*) or removed (*background selection*) along with the variant promoting the selection if the two are in linkage disequilibrium (Charlesworth et al., 1997).

The occurrence of either positive or negative selection towards an extreme phenotype is often described as *directional selection*, and at a molecular level, pushes the variant towards fixation or loss. However, when variation at a locus provides the greatest fitness, both positive and negative selection alternate in order to keep a locus heterozygous (heterozygous advantage or overdominance). This selection pattern is described as *balancing selection*. Further combinations of positive and balancing selection are: *diversifying* (or *disruptive*) *selection*, in which both extreme phenotypes are retained while removing the intermediate, and *stabilizing selection*, where intermediate phenotypes are kept (e.g., by balancing selection on codominant alleles).

Among these, positive selection is one of the most studied. Although not exclusively involved, positive selection is one of the main drivers of adaptation (Akey, 2009). While negative selection acts in highly conserved regions (e.g., towards genes coding for fundamental functional proteins) and balancing selection acts at a more subtle level (i.e., a locus under balancing selection can have allele frequencies within the neutral expectations), positive selection is

easier to detect and several methods have been developed to identify it in genome-wide data (Vitti et al., 2013).

In the following sections some of the most common methods recently developed to detect positive selection will be described, with a greater focus on the methodological class used in this thesis.

1.4.3 Methods to identify selection

Methods to detect selection (mostly) between species and to identify ancient selection events can be inferred using macroevolutionary approaches (e.g., K_a/K_s; Hurst, 2002). Whereas methods used to study selection within species are said to investigate microevolutionary selective events (Vitti et al., 2013).

1.4.3.1 Methods for microevolution

Among the methods to identify selection at a microevolutionary level, two main category are recognisable: frequency and linkage-disequilibrium based methods, with both approaches aiming to identify the presence of regions with reduced genetic diversity.

When a selective sweep occurs, a loss in genetic diversity is expected in the region surrounding the locus under selection due to the hitchhiking of neighbouring variants from the same linkage block (Figure 1-6). With time both recombination and new mutations occur and restore genetic diversity. Consequently, all the new mutations arising within a selective sweep will initially be rare, and several generations might be needed to restore the 'site frequency spectrum' to that expected in the absence of selection (Vitti et al., 2013).

A frequency based method like Tajima's D (Tajima, 1989), detects the excess of rare variants, and other methods as Fay and Wu's H (Fay and Wu, 2000), look at other distortions in the site frequency spectrum (e.g., the presence of numerous high frequency derived alleles) in order to identify the occurrence of positive selection (Figure 1-7).

Frequency-based methods relying on population differentiation at the locus level can also been used to detect selection. For example, Flori et al. (2009) and Kijas et al. (2012) applied a pure drift model F_{ST} (Nicholson et al., 2002) to detect loci

under selection in cattle and sheep, respectively. Nicholson's F_{ST} resembles Wright's F_{ST} but uses a method-of-moments estimator analogous to the Weir and Cockerham F_{ST} (Weir and Cockeram, 1984). For each allele it identifies the difference with the frequency of the estimated 'ancestral' allele, essentially detecting the allele frequency heterogeneity in the population.

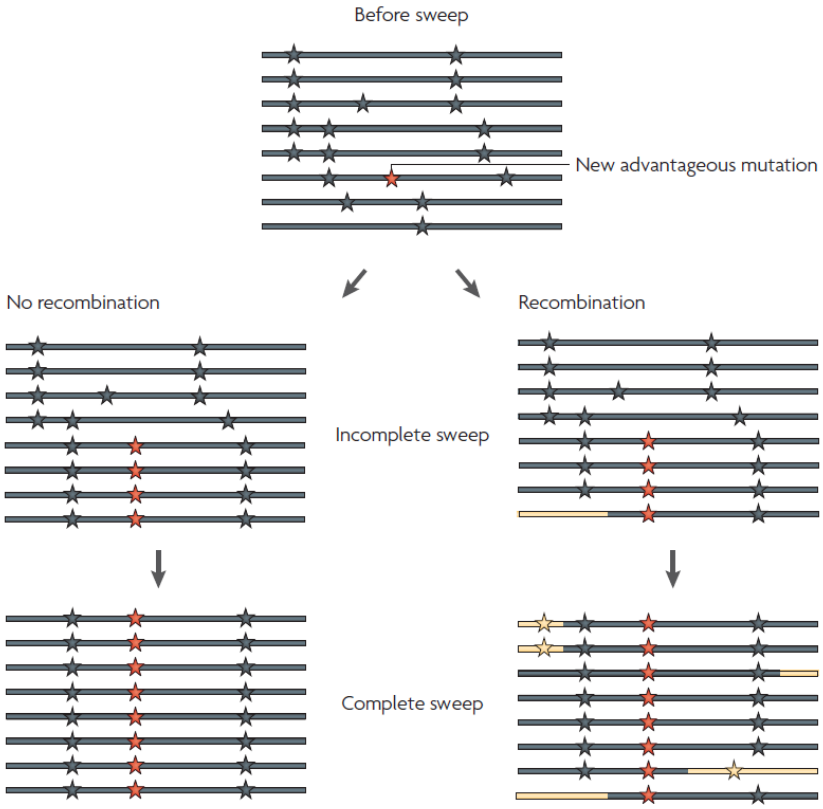


Figure 1-6 Generation of selective sweeps (Figure 1 from Nielsen et al., 2007). At first a new advantageous mutation appears in a population of 8 haploid individuals. The new haplotype will soon increase in frequency within the population until reaching fixation and reducing the variability in the neighboring genomic regions around the focal locus. In the short term, new mutations will not contribute to restoring variability in that region. However, the occurrence of recombination can break the linkage in the haplotype and bring new variants at each generation.

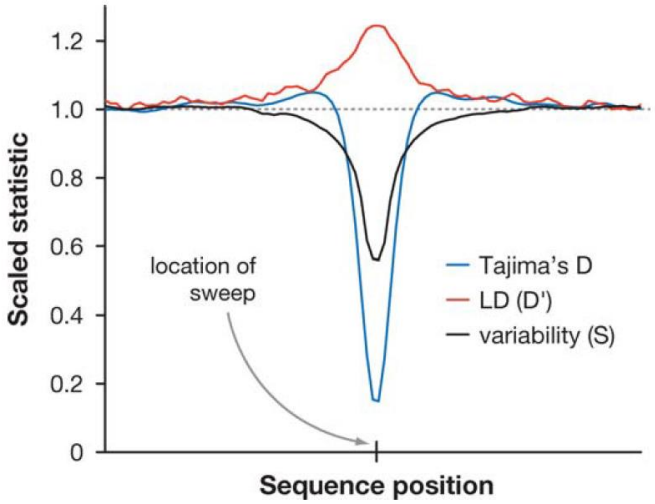


Figure 1-7 Effect of a selective sweep on genetic variation as inferred on 100 simulations of a strong selective sweep (Figure 1 from Nielsen et al. 2005). Within a selective sweep, the linkage disequilibrium (LD) is higher and variability (S) is reduced, also Tajima's D is skewed due to rare allelic variants distorting the site frequency spectrum. The three statistics are scaled so that their value under neutrality equals one.

1.4.3.1.1 Extended Haplotype Homozygosity-based methods

The loss of genetic diversity through a selective sweep can be expressed in terms of linkage. A locus under selection and all the linked neighbouring variants will persist as a long haplotype, until recombination breaks down the linkage among them. Therefore, if a long haplotype is found relative the background haplotype length, the assumption is that a [recent] selective sweep caused it (Sabeti et al., 2007).

Several methods based on investigating such long haplotypes have been developed and theorized within the Extended Haplotype Homozygosity (EHH) statistic (Sabeti et al., 2007). EHH-based methods calculate the probability that a region of extended homozygosity around a core position is identical by descent between any two haploid individuals within the population. Consequently within a selective sweep, EHH will have the highest score where the locus under selection is, and will decrease both down- and up-stream further away from the core locus.

Several EHH-based methods have being developed (e.g., Ferrer-Admetlla et al., 2014); however, two of the most widely used statistics are the integrated haplotype score (iHS) (Voight et al., 2006) and the cross population extended haplotype homozygosity (XP-EHH) (Sabeti et al., 2007). Both statistics can detect

on-going sweeps but iHS outperforms XP-EHH in detecting ancient or soft sweeps (e.g., selection on standing variation), while XP-EHH performs better with loci close to fixation (Pickrell et al., 2009). However, iHS requires the ancestral allelic state to be known, whereas XP-EHH uses a pairwise population comparison to detect linkage distortion. The latter allows EHH-based analysis in all those cases were the ancestral allelic state is not known, as in the cases of sheep, were the progenitor genome is not available.

1.4.4 Landscape genomics

When evolutionary forces act in a spatially heterogeneous environment, the selective pressure on the organisms will also be heterogeneous and generate patterns of genetic variability usually described as locally adaptive variation (Blanquart et al., 2013). Locally adaptive variation is promoted by environment and influenced by both the level of isolation of the locally adapted genotypes, and genetic drift. While the bottom-up genomic strategies for identifying signals of positive selection (e.g., EHH-based methods) rely on scanning the genome for signals mostly due to beneficial alleles that are subsequently related to a possible selection driver (provided gene annotation is available), the effect of local adaptation on the genome can be rather subtle, with small allele frequency distortions that are difficult to reliably detect (Rellstab et al., 2015). For example, in the case of polygenic effects where, following a change in the environment, a small change in frequency in many loci can lead to the optimal phenotype (Pritchard and Di Rienzo, 2010). In order to identify genetic variants related to adaptive processes, approaches that specifically included the environmental variables in the identification of loci under selection have been developed. Initially, these methodologies relied on relatively few genetic markers (Manel et al., 2005; Joost et al., 2007); more recently the availability of genome-wide data prompted the development of new methods and tools (Coop et al., 2010; Frichot et al., 2013; Stucki et al., 2014; Frichot and François, 2015), fuelling research known as 'Landscape genomics' (Rellstab et al., 2015).

Among the approaches proposed to identify local adaptation are logistic regression tests, which investigate whether an environmental factor affects the presence or absence of an allele or single-locus genotype (Rellstab et al., 2015).

Initially implemented in the spatial analysis method (SAM) (Joost et al., 2007) which tested univariate logistic correlation among genotypes and environmental variables, the method has been recently extended allowing multivariate analysis (Stucki et al., 2014), with the logistic correlation performed by considering more than one environmental variable at a time.

Landscape genomics has shown potential to identify patterns of local adaptation in both plants and animals (Eckert et al., 2010; Chen et al., 2012; Günther and Coop, 2013; Stucki et al., 2014). The sampling strategy for both individuals and landscape variables is a fundamental step of a landscape genomic experimental design and much research has been done to improve these aspects of the analysis (Manel et al., 2010; Lotterhos and Whitlock, 2015; Rellstab et al., 2015).

1.5 History and status of sheep genetics

1.5.1 Constructing the phylogeny of *O. aries*

Studies on *O. aries* mtDNA have revealed five haplogroups, named: HA, HB (Hiendleder et al., 1998), HC (Pedrosa et al., 2005; Tapio et al., 2006; Bruford and Townsend, 2006), HD and HE (Meadows et al., 2005; Demirci et al., 2013). HA and HB are the most frequent identified, being found in almost every area where *O. aries* samples have been collected. HC has a more restricted but still wide distribution having been found in Asia, the Fertile Crescent, Caucasus and the Iberian Peninsula. HD and HE are the rarest as they have been found only in sheep from the Caucasus and Turkey. These five haplogroups were identified using short portions of mtDNA, the D-loop (1,246 bp) or the cytochrome *b* gene (1,272 bp). Although such data-sets, comprising samples from different regions allowed the identification of different haplogroups, the use of short genomic portions gave poor statistical robustness to the phylogenetic trees (Meadows et al. 2007; Tapio et al. 2006). The analysis of the complete mtDNA sequence belonging to the five haplogroups of domestic sheep and samples of two species of wild sheep: the Argali and the Urial, plus the European mouflon, was carried out by Meadows et al. (2011).The results summarized in Figure 1-8 shows that the ancestry of the domestic sheep remains unsolved as neither of the two candidates could reliably be assigned as ancestor, although the Argali appeared to be closer to the domestic

sheep than the Urial. These observations are in agreement with the karyology of these species (Domestic sheep 2n = 54, Argali 2n = 56 and Urial 2n = 58). The two rarer haplogroups, HD and HE, that could have been ancestral, instead grouped perfectly in the domestic branch. The European mouflon clustered as part of the domestic haplogroups HB, supporting the hypothesis of European mouflon being a remnant from early domestication events then went back to feral life (Clutton-Brock 1999; Chessa et al. 2009).

Available cytochrome *b* sequence data of the extinct *Myotragus balearicus* was used to date the divergences of *Ovis*. *Myotragus* was a goat-like species of the subfamily *Caprinae,* which were isolated in the Balearic Island at the end of the Messinian salinity crisis, when the desiccation of the Mediterranean ended around 5.53 MYA (Lalueza-Fox et al., 2005). This date was used to calibrate the *Bovidae* cytochrome *b* tree, resulting in the node joining *O. ammon* to the rest of the tree dating to 2.13 ± 0.29 MYA. Previous estimates from other authors are highly discordant and range from 3.21–3.62 MYA (Hiendleder et al. 2002) to 0.8 MYA (Hernández Fernández and Vrba 2005), but this reflects the few *Ovis* fossils found that could have been used to perform successful DNA extractions. Nonetheless the use of the *Myotragus* sequence is likely to underestimate the dates involved, as this must be a minimum estimate for *Ovis-Myotragus* divergence.

With the support of this calibration point the authors defined the radiation point for the haplogroups HA-HE at around 0.92 MYA. As the HC and HE haplogroups had the lowest genetic distance (number of substitutions per site K = 0.36) and both of them were represented by a low number of samples, the chance of bias due to using samples from the same haplogroup is high. The divergence time of the pair HC-HE was evaluated as 0.26 MYA, much older that the domestication event (around 11,000 YA), and still older than the last pre-domestication population expansion, estimated at ca. 20,000 YA, an event that could have accelerated the generation of new haplotypes within haplogroups. These observations led the authors to assess the minimum number of domestication events as five according to the haplogroups found.

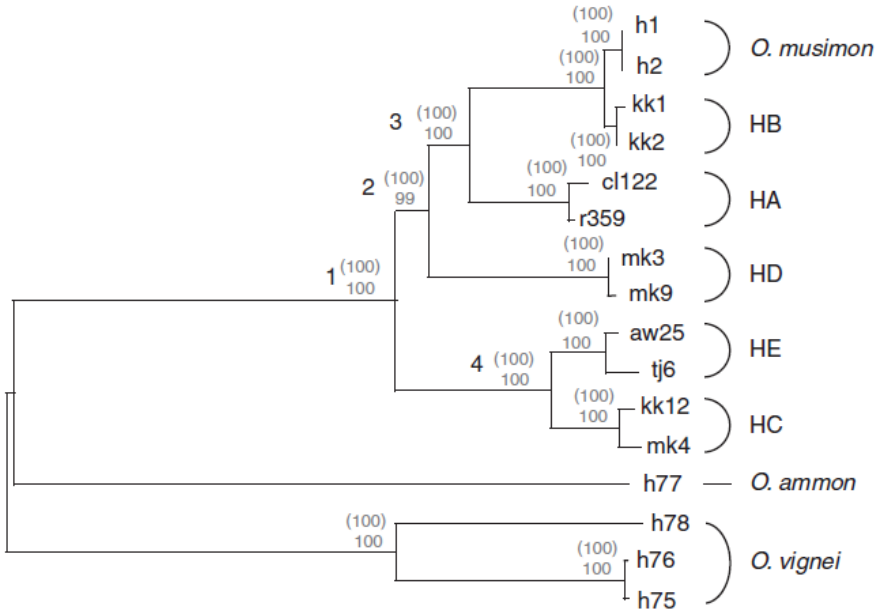


Figure 1-8 Phylogenetic relationship between six wild and ten domestic sheep inferred using the complete mitochondrial genome sequence (Figure 1a from Meadows et al., 2011). On the nodes of the neighbor joining tree are the bootstrap support value and the clade credibility (in brackets).

1.5.2 Dating sheep divergence

The first Asian mouflon mitogenome (from a Cyprus mouflon individual, *O. o. ophion*) was recently sequenced and compared with other *Ovis* species (Appendix A1 - Sanna et al. 2015). Analyses were carried out comparing both the concatenated sequence of 28 mitochondrial genes located in the heavy strand (28H) and the 1,179 bp long D-loop sequence of the Cypriot mouflon against those obtained from samples representative of the domestic sheep main haplogroups and samples of the extant wild and feral sheep (Figure 1-9).

Phylogenetic relationships among individuals were investigated by means of Bayesian and Maximum Likelihood analysis using the 28H molecular marker and the split time dated using three calibration points (CP). Two based on fossil records: the separation between artiodactyls and cetaceans (60 MYA) and the emergence of the *Bovidae* (18.5 MYA) and a third obtained from a molecular approach and used to define the split between *O. ammon* and domestic sheep (2.31 MYA) (Meadows et al., 2011). The analyses shown the same tree topology obtained by previous studies (e.g., Meadows et al., 2011), identifying two main clusters, an 'Asiatic' one where the Cypriot mouflon and sheep haplotypes C and E could be found, and the 'European' one with the sheep haplotype A, D and B and

the European mouflon radiating from the latter. The results suggests the demographic expansion of extant sheep haplogroups occurred at around 5–35 kYA, about 25k years before domestication. The divergence between Urial and the European/Asiatic clusters was dated much earlier, at around 300 kYA.

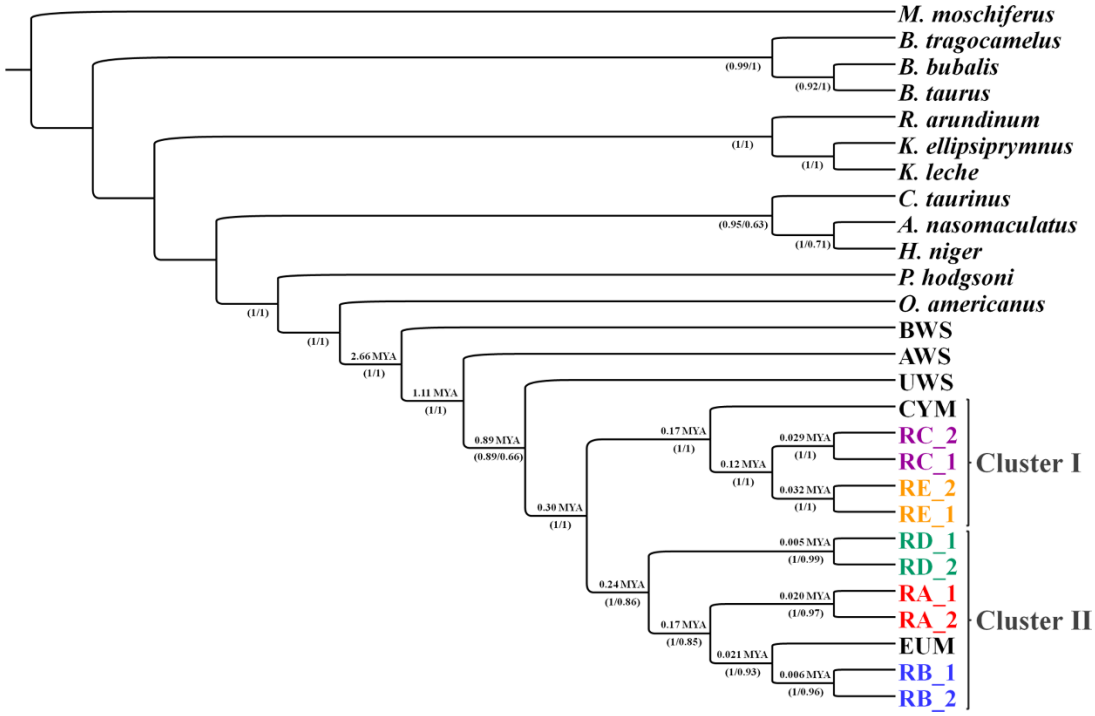


Figure 1-9 Rooted tree obtained by Bayesian inference (BI) for 28H dataset showing two clusters of sheep haplogroups (Figure 3 from Sanna et al., 2015). Nodal supports are indicated below the nodes (posterior probability for BI/bootstrap values for Maximum likelihood). Molecular dating in million years are indicated above the nodes. See Appendix A1 for sample codes.

1.5.3 The sheep colonization of Europe

Signals of two separate migratory events were revealed by an analysis of 1,362 animals belonging to 133 domestic sheep breeds and wild relatives, divided in 65 distinct groups selected by geographical location (Chessa et al., 2009). The samples were tested for the presence/absence of six independently inherited endogenous retroviruses belonging to the Jaagsiekte sheep pathogen family (enJSRVs). Retroviruses can be used as genetic markers because once their genome inserts itself into the host, it will remain there permanently and will be vertically inherited in a Mendelian fashion. Typical of retrovirus integrations are long terminal repeats (LTR), caused as the virus inserts into the host genome. Upon insertion the LTR regions are identical, originating from the same genomic portion of the insertion point, yet, over time, polymorphisms can arise as in for

every other non-coding sequence in the genome allowing researchers to date the insertion order. Here, EnJSRV-7 was identified as the oldest retrovirus among those analysed. The data obtained revealed that in some samples the most common enJSRV-18, found almost everywhere and with high frequencies in the old world breeds, had very low frequency in some Mediterranean island and peripheral regions in the northwest Europe. At the same time, the highest frequencies of enJSRV-7 was found in the Mediterranean mouflon and in the Soay sheep, a primitive breed now confined in the Scottish St. Kilda Island (Scotland, UK), which shows several ancestral morphological traits such as a dark coat and ewes with horns (Figure 1-10).

The observation of insertional polymorphism in enJSRVs combined in 'retrotypes' (i.e., haplotypes of retrovirus insertions) led to other interesting considerations. The 'R2' retrotype, representative of the enJSRV-18 (Figure 1-10), confirmed the former observation, as it was present in almost all the population tested, while the 'R4', representative of the combination of enJSRV-18 and enJSRV-7 was present in southern Europe and in those territories once homeland of the Phoenicians and within their trading routes in the Mediterranean basin.

The analysis of the genetic distances between these retrotypes revealed a marked separation between two groups, one containing the majority of the domestic breed analysed and a second one comprising European Mouflon, Soay sheep, Hebridean, Orkney, Icelandic and Nordic breeds. The animals in this last group share morphologically 'ancestral' traits, and additionally, they show very low frequency, or absence of enJSRV-18 and the presence of the ancient marker enJSRV-7 at very high frequency.

These observations are consistent with the possibility of two different waves of migration. The first one brought the first domesticated sheep all over the Old World; these animals were probably selected for behaviour and used primarily as meat. Then second wave of migration characterized by the presence of enJSRV-18 and dated around 6-7,000 YA involved sheep with improved traits for wool and milk production that overlapped and mostly replaced the first sheep. The results supported the hypothesis that this primitive kind survived returning to a

feral or semi-feral condition like the Sardinian, Corsican and Cypriot mouflon, in general finding refuge in the steepest mountains of the island (Poplin, 1979; Vigne, 1992).

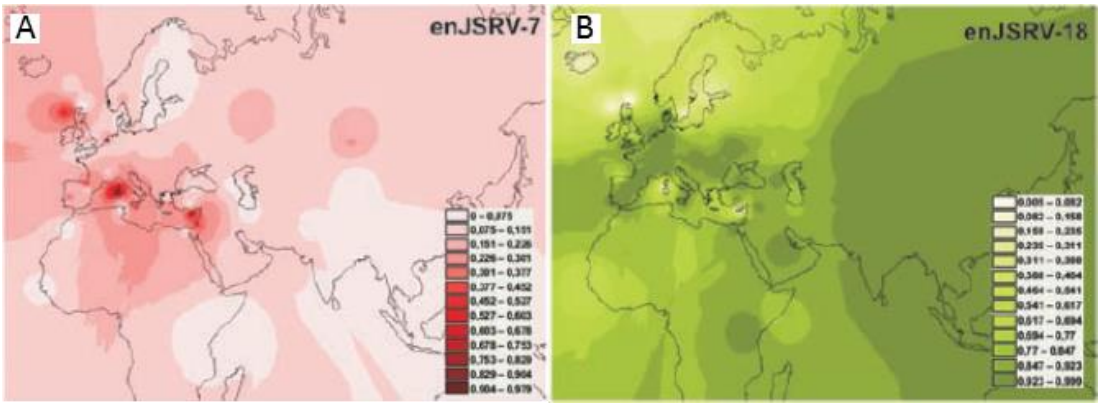


Figure 1-10 Interpolation maps displaying the spatial distribution of estimated enJSRV frequencies (Figure 1-C,D from Chessa et al., 2009). A) Distribution of enJSRV-7, found in mouflon and 'primitive' domestic breeds. B) Distribution of enJSRV-18, typical of all the modern sheep breeds.

1.5.4 From genetics to genomics

As previously stated, the domestication process took place potentially as a way to have a fresh meat supply, and then, around 4,000-5,000 YA the first herders started to perform selection in order to enhance the production of wool and milk (Chessa et al. 2009). Until recently the majority of the research on sheep history were performed using: autosomal SSR to estimate the levels of genetic diversity (Lawson Handley et al. 2007), mtDNA was used to provide a better understanding of sheep phylogeny and gave several clues on the process of domestication (Hiendleder et al. 2002; Pedrosa et al. 2005; Meadows, Hiendleder, and Kijas 2011), Y-chromosomal markers have revealed patterns of male mediated introgression (Meadows et al. 2006) and a relatively small panel of autosomal SNP (Kijas et al. 2009) were also developed.

In March 2012, Kijas and collaborators published a landmark paper describing the application of the high density Ovine SNP50 BeadChip (Illumina©) to a wide range of sheep samples collected across the world (Kijas et al. 2012) with the aim of examining genetic diversity in global sheep populations and of characterising the genetic legacy of selection and adaptation in the sheep genome.

SNP genotyping was performed on 2,819 animals from 74 breeds sampled from each continent: six breeds from both Africa and America, seven from South-West Asia, eight from Asia and the rest from Europe. The initial set of SNPs was pruned according to five quality control filters: markers with <0.99 call rate (3,612 SNPs pruned), markers that showed assay abnormalities (4,101 SNPs pruned), markers with MAF <0.01 (1,165 SNPs pruned), SNPs discordant after crosschecking with genotyping results from different laboratories (125 SNPs pruned), and positions that showed Mendelian inconsistencies (11 SNPs pruned). After pruning, 5,207 markers (9.6 % of the total) were discarded leaving 49,034 SNPs for further analyses.

On the basis of the polymorphism observed in this SNP panel set, the effective population size (*Ne*) of 75% of modern sheep breeds appears to have retained high values (>300): the lowest value was 100 for the Wiltshire (an old British native breed). These *Ne* values are consistently higher than other domestic animals such as cattle, where the majority of breeds have a current *Ne* of 150 or less (Gibbs et al., 2009). These *Ne* high values could be due to the existence of a very heterogeneous population before the domestication event, so that the possible bottleneck associated to the domestication event captured a sufficient number of genotypes to generate the extant diversity, as also recorded in other species (Naderi et al., 2008). Additionally, the authors suggest the possibility of an on-going cross-breeding after the domestication event with the wild population that enriched the gene pool of the first flocks.

The data seem to confirm that after the domestication event there was a high level of gene flow between sheep populations as previously suggested by mtDNA analysis (Meadows et al. 2005), and the authors suggest three lines of molecular evidence: firstly, the high degree of conservation in haplotypes sharing across short chromosomal distances independent of geographic origin, secondly the lack of a strong association between genetic diversity and physical distance from the domestication centre, which usually shows the highest values of heterozygosity decreasing with distance (Bruford et al., 2003), and lastly Principal Component Analysis (PCA) suggested weak population structure with the first four components explaining less than the 7% of the variation.

This large genome-wide dataset for several breeds from around the world enabled the authors to estimate signatures of selection reflecting the domestication process. Signatures of selection were investigated using a global F_{ST} approach (see 1.4.3.1), and 47 SNPs (the 0.1 % of the total SNPs-set) were identified by unusually high F_{ST} values. These SNPs were distributed among 31 genomic regions, which contained genes regulating pigmentation, body size and reproduction. As an example the highest F_{ST} value (F_{ST} = 0.682) was reached by a SNP (OAR10) located in chromosome 10 and associated to a relaxin/insulin-like family receptor 2 (RXFP2). This gene is related with horn presence as already revealed from other studies (Johnston et al. 2011) (Figure 1-11).

When a pairwise comparison was performed for the same position between breeds defined as horned or polled the F_{ST} signal was absent for horned-horned and polled-polled comparison, but strongly present when horned breeds were compared to polled breeds, confirming the validity of the selection signal. A strong signal of balancing selection (identified by extremely low F_{ST} values) was revealed on chromosome 20 in the MHC region, a result previously observed in other species including cattle (Gautier et al., 2009). Additionally, using the F_{ST} index the authors were able to identify breed specific genomic regions under selection, as they did with the Texel, a breed raised as a meat animal with massive muscles, that when compared with the other breeds showed a strong peak on chromosome 2 around the GDF8, a gene already described in Texel for carrying a mutation responsible for muscle hypertrophy (Clop et al. 2006).

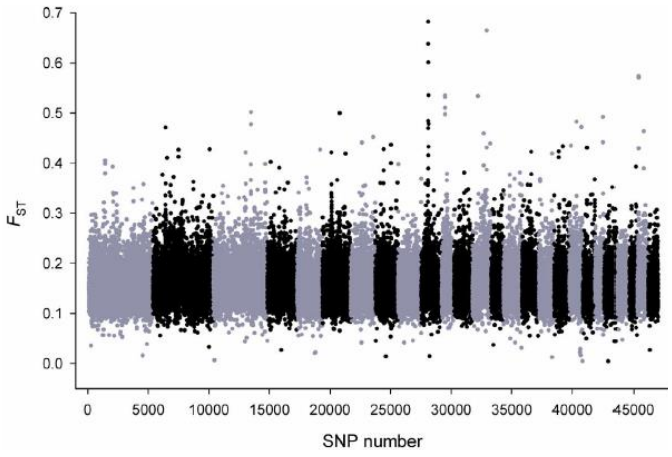


Figure 1-11 Genome-wide distribution on global F_{ST} (Figure 5 from Kijas et al., 2012) calculated across 74 domestic sheep breeds. The highest peak is on chromosome 10 and has been associated to horn presence/absence in ungulates.

1.6 Aims of this PhD

The general aim of this study was to investigate the signature of selection involved with sheep domestication process, and local adaptation to the environment. By analysing feral and domestic sheep populations simultaneously with genomic tools, the following topics were investigated:

1) What are the levels of admixture between the European mouflon and domestic sheep?
2) Can signatures of selections be related to domestication?
3) What is the occurrence of selection signatures due to environmental drivers?
4) These questions were facilitated by the development of analytical frameworks and informatic tools using SNP array data.

Each chapter in this thesis aims at investigating one or more of these topics. The following section introduces each of the chapters, while providing additional background information or justification for the study.

Although the main focus of this thesis is the identification of selection signatures in domestic and feral sheep, the methodological frameworks applied and the informatic tools developed are independent from the species studied. Instead, they add to the current analytical tool set available for low- to mid-density SNP arrays but also can be extended to every kind of SNP based dataset within a genomics framework.

1.6.1 Second Chapter

In this chapter the development of an informatic tool to infer historical effective population size trends from SNP data is described.

Effective (N_e) and census population size (N_C) are two of the most important indicators used to describe a population's viability, conservation status and to make decisions for breed management. While N_C is "The number of individuals in a population; the head count size of a population", N_e is "the size of an ideal Wright–Fisher population that maintains as much genetic variation or

experiences as much genetic drift as an actual population regardless of census size" (Wright, 1931).

Recent N_e values and their trends in the past can be used to better understand the level of genetic variation within a population and its change over time. Retrospectively, N_e can be used to explain the observed pattern of genetic variation, while prospectively it can be used to model its decline due to genetic drift (Wang, 2005). The management of genetic diversity is clearly essential in conservation genetics. An endangered species has the opportunity to overcome the effect of genetic drift only if its gene-pool is sufficiently wide; whereas the utility of prioritising funds and energy to manage genetically depauperate populations could be debatable.

Genetic diversity is fundamental for domestic species. Even though often characterized by large N_C, a domestic species or breed's typical N_e is usually low (Taberlet et al., 2008). This can be due to the genetic bottleneck usually related with the domestication event and with the assortative mating used to improve those traits of interest for farmers. Thus, knowledge of N_e facilitates the design of efficient selection schemes in animal breeding and the effective management of endangered species as well as the understanding and modelling of the genetic architecture underlying complex traits.

N_e has been calculated using a number of methods and approaches and can be summarized into three categories: demographic, pedigree based or marker based (Flury et al., 2010). Pedigree data have been traditionally used to obtain N_e estimates. But to produce a reliable estimate, a well-constructed and complete pedigree is necessary, and this is rarely available.

1.6.1.1 A marker based approach

One solution to overcome the limitation of an incomplete pedigree, is to estimate the trend in N_e on the basis of genomic data. In the 1980's several authors recognised that N_e could be estimated from linkage disequilibrium (LD) data (Sved, 1971; Hill, 1981). LD describes the non-random association of alleles for different loci as a function of the recombination rate between their physical positions in the genome. However, LD signatures can also result from admixture,

direct or indirect selection (e.g., hitchhiking, Smith and Haigh, 1974 or background selection, Charlesworth et al., 1997), and genetic drift in finite populations (Wang, 2005; Wright, 1943), independently of whether the loci studied occur in close proximity in the genome. Assuming that a population is isolated and panmictic, the LD value calculated between neutral unlinked loci will depend exclusively on genetic drift (Sved, 1971; Hill, 1981). This occurrence can be used to predict N_e thanks to the known relationship between the variance in LD (that depends on gene frequencies) and the effective population size.

However, the application of this methods requires a large number of loci (Sved, 1971) that only the most recent advances in NGS technologies were able to provide. Since then the LD method to infer N_e has been investigated in several livestock species using commercially available SNP arrays; however, a software tool that enables estimation of N_e from LD is lacking, and researchers currently rely on a combination of tools to manipulate data, infer LD, and tend to use bespoke scripts to perform the appropriate calculations and estimate N_e (Mbole-Kariuki et al., 2014; Corbin et al., 2010; Flury et al., 2010; Burren et al., 2014).

In this chapter the development of *SNeP* is described: a software tool that allows the estimation of N_e trends across generations using SNP data that corrects for sample size, phasing and recombination rate. This program is a valuable tool for recent demographic inference on genomic data. *SNeP* was used at various stages of the analyses performed in this thesis to infer N_e trajectories of the species under scrutiny and help in the interpretation of results from other analytical frameworks under the light of species demography.

1.6.2 Third Chapter

In this chapter the level of crossbreeding between the extant European mouflon populations and their local domestic sheep populations was first investigated and signals of selection within regions of local and global ancestry were subsequently explored.

Since the arrival of the second wave of domestication in Europe ~7-6,000 YA the sheep population brought to Corsica, Sardinia and Cyprus have lived in sympatry with the already established mouflon population residing in the islands. The

Roman author and naturalist Pliny the Elder (first century B.C.E) in his *Naturalis historia* (VIII-LXXV) mentioned the presence of sheep and mouflon hybrids called 'umbri', and Cetti, an Italian naturalists from the XIX century, also mentioned the occasional occurrence of crossbreeding among sheep and mouflons in Sardinia (Cetti, 1774). Recent research also provided evidence of admixed individuals in a population sampled nearby the same central east mountainous area of Sardinia where mouflon samples used in this thesis were collected (Lorenzini et al., 2011). Additionally, the use of *mouflon x sheep* hybrids was documented to have been used to restock some mainland mouflon populations (Petit et al., 1997). However, little is known about the amplitude and direction of gene flow between the two sub-species. The presence of admixture has to be assessed in order to identify selection signatures involved in domestication and adaptation. With admixture, large genomic portions carrying potentially different demographic histories of selection are inherited, potentially harbouring selection sweeps that incurred in one of the parental lineages exclusively and having not had the time for subsequent evolution, would affect the selection analysis with background noise and false and conflicting signals. Signals of crossbreeding with domestic sheep were hypothesised to be present in all European mouflon populations, with potential bidirectional introgression in the Sardinian mouflon populations, in accordance with the historical records.

1.6.3 Fourth Chapter

Here the signatures of selection related to the second wave of sheep domestication and those involved with local adaptation to the environment in Sardinian feral and domestic sheep population were investigated.

Sardinian Mouflon and Sarda sheep are extant representatives of the first and second wave of domestication, respectively (see 1.5.3). Previous research investigated signatures of selection in sheep using SNP array data and provided evidence for gene associated with sexual weaponry (Kijas et al., 2012), fat-deposition (Moradi et al., 2012), coat pattern (Gratten et al., 2012) and adaptation to arid environments (Kim et al., 2015) among others.

In this chapter two approaches were used to investigate the footprint of selection in feral and domestic Sardinian sheep populations. Firstly, EHH (see 1.4.3.1.1) was used along with novel methodological pipelines developed to aid in the identification of signals of selection. Secondly, a preliminary study on the contribution that the landscape had in shaping the genome in both Sardinian mouflon and domestic sheep was performed within a Landscape Genomics framework (see 1.4.4). The two species lived in sympatry for thousands of years since the arrival of the domestic sheep in Sardinia. By using Landscape genomic tools, we aimed to assess the association of sheep and mouflon genotypes with a collection of topo-climatic variables. Additionally, we then applied statistical tools to assess the power of the method applied and to infer the ideal sampling to maximise the discovery of positive signals if they exist.

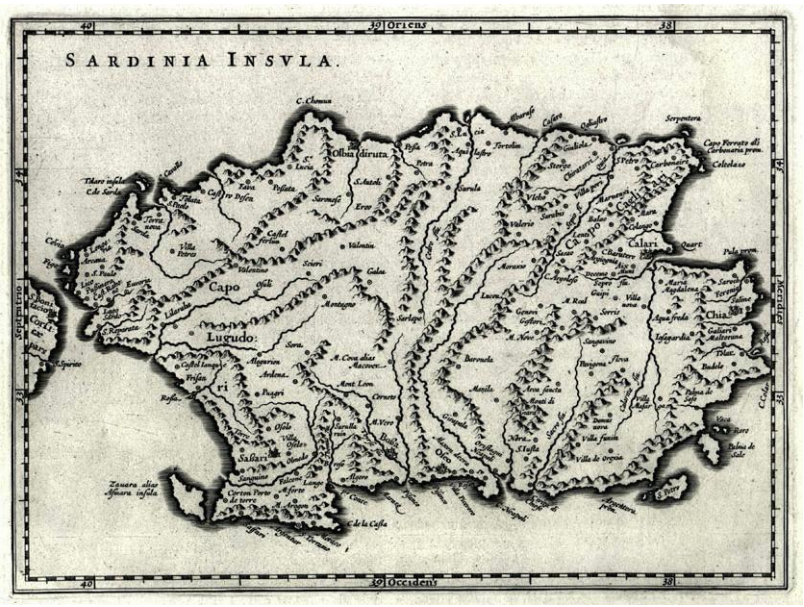
1.6.4 Fifth Chapter

This chapter provides an overview of the results obtained in the previous section in relation to the main questions that this thesis wanted to address. Firstly, the contribution of the novel methodological approaches described in the previous chapter was further discussed in terms of their applicability of mid-density SNP array data. Secondly, the conservation implications of crossbreeding between European mouflon and domestic sheep were evaluated, with a focus on the autochthonous feral and domestic Sardinian sheep populations, and the across-species implications of the identified selection signatures were discussed. Lastly, some future directions that can be taken to expand the results obtained in this thesis are described.

1.6.5 Appendix

In this section, the web links to two published studies to which I contributed as co-author and whose topic is relevant to this thesis are provided.

Chapter two



Map of Sardinia (Blaeu, 1638)

2 Chapter two - *SNeP*: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data

Mario Barbato[1*], Pablo Orozco-terWengel[1], Miika Tapio[2] and Michael W. Bruford[1]

[1] School of Biosciences, Cardiff University, Cardiff, UK

2 MTT Agrifood Research Finland, Biotechnology and Food Research, Jokioinen, Finland

** Lead author contribution: conceived and wrote the software, performed analyses, wrote the first manuscript draft, and modified the manuscript based on co-author comments.*

Published: (2015) *Frontiers in Genetics*, 6:109. doi: 10.3389/fgene.2015.00109

Keywords: effective population size, linkage disequilibrium, SNPChip, demography, large scale genotyping Introduction

2.1 Abstract

Effective population size (N_e) is a key population genetic parameter that describes the amount of genetic drift in a population. Estimating N_e has been subject to much research over the last 80 years. Methods to estimate N_e from linkage disequilibrium (LD) were developed ~40 years ago but depend on the availability of large amounts of genetic marker data that only the most recent advances in DNA technology have made available. Here we introduce *SNeP,* a multithreaded tool to perform the estimate of N_e using LD using the standard PLINK input file format (.ped and .map files) or by using LD values calculated using other software. Through *SNeP* the user can apply several corrections to take account of sample size, mutation, phasing and recombination rate. Each variable involved in the computation such as the binning parameters or the chromosomes to include in the analysis can be modified. When applied to published datasets, *SNeP* produced results closely comparable with those obtained in the original studies. The use of *SNeP* to estimate N_e trends can improve understanding of population demography in the recent past, provided a sufficient number of SNPs and their physical position in the genome are available.

Binaries for the most common operating systems are available at https://sourceforge.net/projects/snepnetrends/.

2.2 Introduction

Effective population size (N_e) is an important genetic parameter that estimates the amount of genetic drift in a population, and has been described as the size of an idealized Wright–Fisher population expected to yield the same value of a given genetic parameter as in the population under study (Crow and Kimura, 1970). N_e sizes can be influenced by fluctuations in census population size (N_c), by the breeding sex ratio and the variance in reproductive success.

N_e estimation can be achieved using approaches that fall into three methodological categories: demographic, pedigree-based, or marker-based (Flury et al., 2010). Pedigree data have been traditionally used to obtain N_e estimates in livestock. However, reliable estimates of N_e depend on the pedigree being complete. This state of knowledge is feasible in some domestic populations, the demographic parameters of which have been accurately monitored for a sufficiently large number of generations. However, in practice, the applicability of this approach remains limited to a few cases involving highly managed breeds (Flury et al., 2010; Uimari and Tapio, 2011).

One solution to overcome the limitation of an incomplete pedigree is to estimate the recent trend in N_e using genomic data. Several authors have recognized that N_e could be estimated from information on linkage disequilibrium (LD) (Sved, 1971; Hill, 1981). LD describes the non-random association of alleles in different loci as a function of the recombination rate between the physical positions of the loci in the genome. However, LD signatures can also result from demographic processes such as admixture and genetic drift (Wang, 2005; Wright, 1943), or through processes such as 'hitchhiking' during selective sweeps (Smith and Haigh, 1974) or background selection (Charlesworth et al., 1997). In such scenarios alleles at different loci become associated independently of their proximity in the genome. Assuming that a population is closed and panmictic, the LD value calculated between neutral unlinked loci depends exclusively on genetic drift (Sved, 1971; Hill, 1981). This occurrence can be used to predict N_e due to the

known relationship between the variance in LD (calculated using allele frequencies) and effective population size (Hill, 1981).

Recent advances in genotyping technology (e.g., using SNP bead arrays with tens of thousands of DNA probes) have enabled the collection of vast amounts of genome-wide linkage data ideal for estimating N_e in livestock and humans among others (e.g., Tenesa *et al.* 2007; de Roos *et al.* 2008; Corbin *et al.* 2010; Uimari and Tapio 2011; Kijas *et al.* 2012). However, a software tool that enables estimation of N_e from LD is lacking, and researchers currently rely on a combination of tools to manipulate data, infer LD, and tend to use bespoke scripts to perform the appropriate calculations and estimate N_e.

Here we describe *SNeP*, a software tool that allows the estimation of N_e trends across generation using SNP data that corrects for sample size, phasing and recombination rate.

2.3 Material and Methods

The method *SNeP* uses to calculate LD depends on the availability of phased data. When the phase is known the user can select Hill and Robertson (1968) squared correlation coefficient that makes use of haplotype frequencies to define LD between each pair of loci (Equation 1). However, in the absence of a known phase, squared Pearson's product-moment correlation coefficient between pairs of loci can be selected. While these two approaches are not the same, they are highly comparable (McEvoy et al., 2011):

$$(1) \qquad r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}$$

$$(2) \qquad r^2_{X,Y} = \frac{\left[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where p_A and p_B are respectively the frequencies of alleles A and B at two separate loci (X, Y) measured for n individuals, p_{AB} is the frequency of the haplotype with alleles A and B in the population studied, $\bar{X}$ and $\bar{Y}$ are the mean genotype frequencies for the first and second locus respectively, X_i is the genotype of individual i at the first locus and Y_i is the genotype of individual i at the second

locus. Equation (2) correlates the genotypic allele counts instead of the haplotype frequencies and is not influenced by double heterozygotes (this approach results in the same estimates as the --r2 option in PLINK).

SNeP estimates the historic effective population size based on the relationship between r^2, N_e and c (recombination rate), (Equation 3 – Sved, 1971), and enabling users to include corrections for sample size and uncertainty of the gametic phase (Equation 4 - Weir and Hill 1980):

$$(3) \qquad E(r^2) = (1 + 4N_e c)^{-1}$$

$$(4) \qquad r^2_{adj} = r^2 - (\beta n)^{-1}$$

*w*here n is the number of individual sampled, $\beta = 2$ when the gametic phase is known and $\beta = 1$ if instead the phase is not known.

Several approximations are used to infer the recombination rate using the physical distance (δ) between two loci as a reference and translating it into linkage distance (d), which is usually described as $Mb(\delta) \approx cM(d)$. For small values of d the latter approximation is valid, but for larger values of d the probability of multiple recombination events and interference increases, moreover the relationship between map distance and recombination rate is not linear, as the maximum recombination rate possible is 0.5. Thus, unless using very short δ, the approximation $d \approx c$ is not ideal (Corbin et al., 2012). We therefore implemented mapping functions to translate the estimated d into c, following Haldane (Haldane, 1919), Kosambi (Kosambi, 1943), Sved (1971) and Sved and Feldman (Sved and Feldman, 1973). Initially *SNeP* infers d for each pair of SNPs as directly proportional to δ according to $d = k\delta$ where k is a user defined recombination rate value (default value is 10^{-8} as in $Mb = cM$). The inferred value of δ can then be subjected to one of the available mapping functions if required by the user.

Solving Equation (3) for Ne and including all the corrections described, allows the prediction of Ne from LD data using (Corbin et al., 2012):

$$(5) \qquad N_{T(t)} = \left(4f(c_t)\right)^{-1}\left(E\left[r^2_{adj}\middle|c_t\right]^{-1} - \alpha\right)$$

where N_t is the effective population size t generations ago calculated as $t = (2f(c_t))^{-1}$(Hayes et al., 2003), c_t is the recombination rate defined for a specific physical distance between markers and optionally adjusted with the mapping functions mentioned above, r^2_{adj} is the LD value adjusted for sample size and $\alpha := \{1, 2, 2.2\}$ is a correction for the occurrence of mutations (Ohta and Kimura, 1971). Therefore, LD over greater recombinant distances is informative on recent N_e while shorter distances provide information on more distant times in the past. A binning system is implemented in order to obtain averaged r^2 values that reflect LD for specific inter-locus distances. The binning system implemented uses the following formula to define the minimum and maximum values for each bin:

$$b_i^{min} = minD + (maxD - minD)\left(\frac{b_i - 1}{totBins}\right)^x \quad (6a)$$

$$b_i^{max} = minD + (maxD - minD)\left(\frac{b_i}{totBins}\right)^x \quad (6b)$$

Where b_i ($\mathbb{N}^1$) is the i^{th} bin of the total number of bins (*totBins*), *minD* and *maxD* are respectively the minimum and the maximum distance between SNPs and x is a positive real number ($\mathbb{R}^0$). When x equals 1, the distribution of distances between the bins is linear and each bin has the same distance range. For larger values of x the distribution of distances changes allowing a larger range on the last bins and a smaller range on the first bins. Varying this parameter allows the user to have a sufficient number of pairwise comparisons to contribute to the final N_e estimate for each bin.

2.4 Example application

We tested *SNeP* with two published datasets that had been previously used to describe trends in N_e over time using LD, *Bos indicus* (54,436 SNPs of 423 East African Shorthorn Zebu (SHZ) - Mbole-Kariuki et al., 2014, data available at Dryad Digital Repository: doi:10.5061/ dryad.bc598.) and *Ovis aries aries* (49,034 SNPs genotyped in 24 Swiss White Alpine (SWA), 24 Swiss Black-Brown Mountain sheep (SBS), 24 Valais Blacknose sheep (VBS), 23 Valais Red sheep (VRS), 24 Swiss Mirror sheep (SMS) and 24 Bundner Oberländer sheep (BOS) - Burren et al., 2014). The r^2 estimates for the cattle datasets were obtained by the authors

using GenABLE (Aulchenko et al., 2007) using a minimum allele frequency (MAF) <0.01 and adjusting the recombination rate using Haldane's mapping function (Haldane, 1919). The r^2 estimates of the sheep data were calculated by the authors using PLINK-1.07 (Purcell et al., 2007), with a MAF < 0.05 and no further corrections. For both autosomal datasets r^2 estimates where corrected for sample size using eq. 4 with $\beta = 2$. For these comparative analyses the *SNeP* command line included the same parameters used for the published data apart from the r^2 estimates, calculated through genotype count and the use of *SNeP*'s novel binning strategy.

2.5 Results

SNeP is a multithreaded application developed in C++ and binaries for the most common operating systems (Windows, OSX and Linux) can be downloaded from https://sourceforge.net/projects/snepnetrends/. The binaries are accompanied by a manual describing the step-by-step use of *SNeP* to infer trends in N_e as described here. *SNeP* produces an output file with tab delimited columns showing the following for each bin that was used to estimate N_e: the number of generations in the past that the bin corresponds to (e.g., 50 generations ago), the corresponding N_e estimate, the average distance between each pair of SNPs in the bin, the average r^2 and the standard deviation of r^2 in the bin, and the number of SNPs used to calculate r^2 in the bin. This file can be easily imported in Microsoft Excel, R or other software to plot the results. The plots shown here (Figure 2-1Figure 2-3) correspond to the columns of generations ago and N_e from the output file. The column with the r^2 standard deviation is provided for users to inspect the variance in the N_e estimate in each bin, particularly for those bins reflecting older time estimates and which are less reliable as the number of SNPs used to estimate r^2 becomes smaller.

The format required for the input files is the standard PLINK format (ped and map files; Purcell et al., 2007). *SNeP* allows the users to either calculate LD on the data as described above, or use a custom precalculated LD matrix to estimate N_e using Equation (5).

The software interface allows the user to control all parameters of the analysis, e.g., the distance range between SNPs in bp, and the set of chromosomes used in the analysis (e.g., 20-23). Additionally, *SNeP* includes the option to choose a MAF threshold (default 0.05), as it has been shown that accounting for MAF results in unbiased r^2 estimates irrespective of sample size (Sved et al., 2008). *SNeP*'s multithreaded architecture allows fast computation of large datasets (we tested up to ~100k SNPs for a single chromosome), for example the BOS data described here was analysed with one processor in 2'43", the use of two processors reduced the time to 1'43", four processors reduced the analysis time to 1'05".

2.5.1 Zebu example

For the zebu analysis, the shapes of the N_e curves obtained with *SNeP* and their published data trends showed the same trajectory with a smooth decline until around 150 generations ago, followed by an expansion with a peak around 40 generations ago and ending in a steep decline on the most recent generations (Figure 2-1). However, while the trends in both curves were the same, the two approaches resulted in different N_e estimates, with *SNeP*'s values being approximately three-fold larger than those in the original paper. While we attempted to use the authors' parameters in our analyses, some differences were inevitable, i.e., the original publication of the cattle data estimated r^2 with a different approach to that implemented in *SNeP*. Analyses with *SNeP* were based on genotypes, while the original analysis was based on inferred two locus haplotypes, which results in the published data showing an expected r^2 of 0.32 at the minimum distance, while our estimates was 0.23. Similarly, Mbole-Kariuki et al. (2014) obtained a background level $r^2 = 0.013$ around 2 Mb, while our estimate at the same distance was 0.0035 (data not shown). Consequently, as our estimates of LD were consistently smaller than Mbole-Kariuki et al. (2014) it is expected that our N_e estimates should be larger. While this observation highlights the importance of a careful choice of the parameters and their thresholds, it is important to highlight that although the absolute magnitude of the N_e values is different, the trends are almost identical.

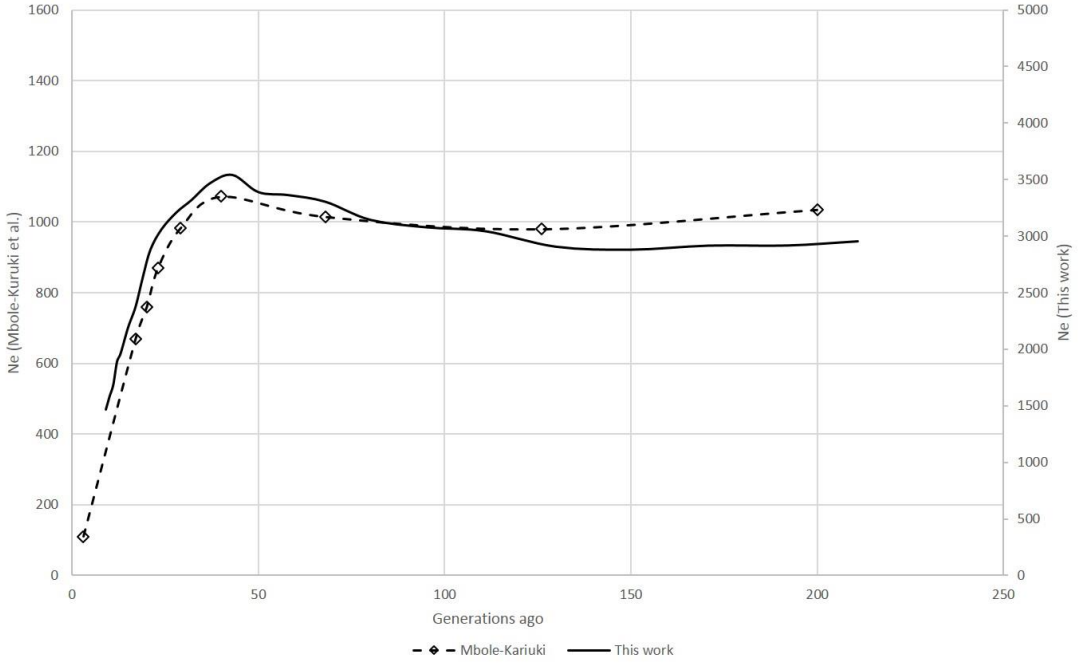


Figure 2-1 Comparison of N_e trends for the last 250 generations in the SHZ data obtained by Mbole-Kariuki et al. (2014) (dashed line) and using *SNeP* (solid line).

2.5.2 Swiss sheep example

The six Swiss sheep breeds analysed with *SNeP* produced comparable results with those from the original paper (Figure 2-2), with mostly overlapping N_e trend curves (Figure 2-3). However, the general trend in N_e showed a decline toward the present. *SNeP* produced slightly larger values of N_e for the more distant past (700-800 generations). This is due to the different binning system used in *SNeP*, which allows the user to obtain a more even distribution of pairwise comparisons within each bin (i.e., the number of SNP pairwise comparisons within each bin is comparable). For the time span extending beyond 400 generations ago, Burren et al. (2014) used only three bins in their analysis (centered at 400, 667, and 2000 generations ago) while for the same time span *SNeP* used five bins with a number of pairwise comparisons dependent to the range defined with formulae 6a,b. Consequently, Burren and colleagues' approach ends with a higher density of data describing the most recent generations than describing the oldest generations. Therefore, the use of fewer bins tends to increase the presence of smaller values of N_e in each bin, consequently lowering the average N_e value for each bin. The N_e values for the recent past, compared at the 29th generation in the

past, gave very similar results. The largest difference (50) was obtained for the SBS breed.

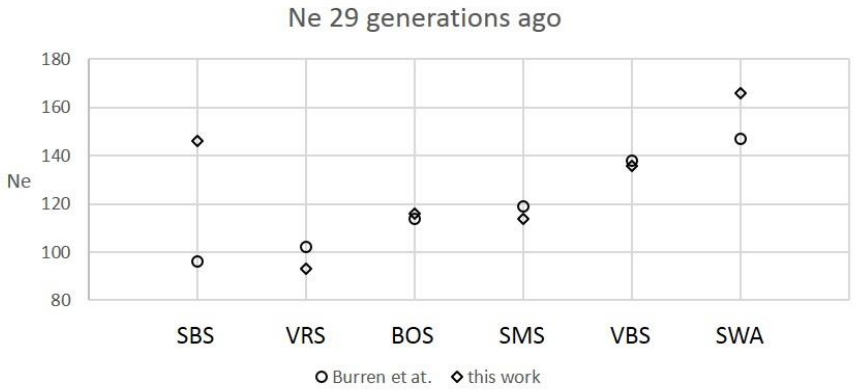


Figure 2-2 Comparison between recent N_e values calculated at the 29th generation in this work and Burren et al. (2014) for 6 Swiss sheep breeds.

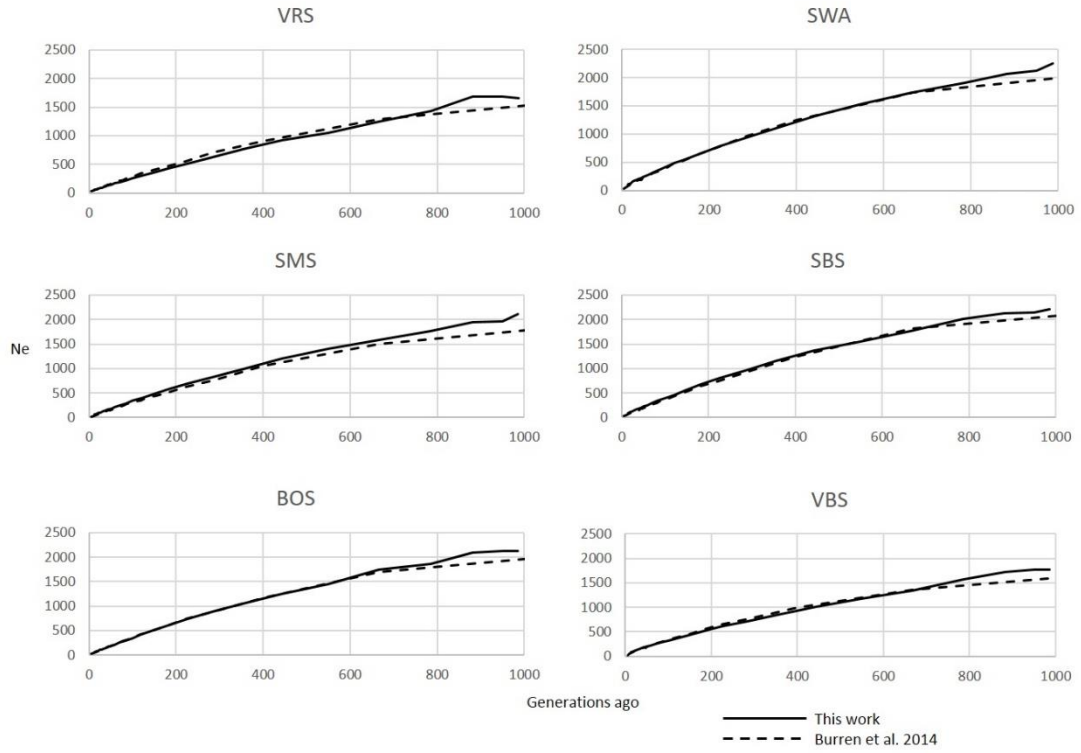


Figure 2-3 Comparison of N_e trends of six Swiss sheep breeds according to Burren et al. (2014) (dashed lines) and this work (solid lines).

2.6 Discussion

Analysis of N_e using LD data was first demonstrated 40 years ago, and has been applied, developed and improved since (Corbin et al., 2012; de Roos et al., 2008; Hayes et al., 2003; Sved, 1971; Sved et al., 2013; Tenesa et al., 2007). The traditionally small number of SNPs analysed is no longer a limitation, since SNP Chips comprise an extremely large number of SNPs, available in a short time and at a reasonable price. This has boosted the use of the method, which has been applied to humans (Tenesa et al., 2007; McEvoy et al., 2011) as well as to several domesticated species (Corbin et al., 2012; England et al., 2006; Kijas et al., 2012; Uimari and Tapio, 2011). Along with these improvements, methodological limitations have become apparent and have been addressed here, with the majority of the efforts pointing to the correct estimation of recent N_e. Yet, the quantitative value of the estimate is highly dependent on sample size, the type of LD estimation and the binning process (Waples and Do, 2008; Corbin et al., 2012), while its qualitative pattern depends more on the genetic information than on data manipulation.

So far this method has been applied using a variety of software, no standardized approach exists to bin the results and each study has applied a more or less arbitrary approach, e.g., binning for generation classes in the past (Corbin et al., 2012), binning for distance classes with a constant range for each bin (Kijas et al., 2012) or binning per distance classes in a linear fashion but with larger bins for the more recent time points (Burren et al., 2014). To our knowledge the only software available that estimate N_e through LD is NeEstimator (Do et al., 2014), an upgraded version of the former LDNE (Waples and Do, 2008) allowing the analysis of large dataset (as 50k SNPChip). Importantly, while *SNeP* focuses on estimating historical N_e trends, NeEstimator's aim is to produce contemporary unbiased N_e estimates, the latter should therefore be considered as a complementary tool while investigating demography through LD.

We used *SNeP* to analyze two datasets where the method was previously applied. The results we obtained for the sheep data were both quantitatively and qualitatively comparable with those obtained by Burren and colleagues (2014), while for the Zebu data we obtained a N_e trend estimate that closely matched that

of Mbole-Kariuki and colleagues (2014) although our point estimates of N_e were larger than those described for the data (Mbole-Kariuki et al., 2014). The discrepancy between these two results reflects that Burren and colleagues produced their r^2 estimates using PLINK (the standard software for large scale SNP data manipulation) which uses the same approach used to estimate r^2 by *SNeP*, while Mbole-Kariuki et al. followed Hao and colleagues (Hao et al., 2007) for r^2 estimation. The use of different estimates for LD is critical for the quantitative aspect of the N_e curve, where due to the hyperbolic correlation between N_e and r^2, a decrease in r^2 on its range closer to 0 can lead to a very large change in N_e estimates, while differences in estimates are less significant when the r^2 value is high, i.e., closer to 1. Therefore, although in one of the datasets the N_e values where substantially different, in both cases the N_e curves overlapped with those originally published.

As already suggested by other authors, the reliability of the quantitative estimates obtained with this method must be taken with caution, especially for N_e values related to the most recent and the oldest generations (Corbin et al., 2012) because for recent generations, large values of c are involved, not fitting the theoretical implications that Hayes proposed to estimate a variable N_e over time (Hayes et al., 2003). Estimates for the oldest generations might also be unreliable as coalescent theory shows that no SNP can be reliably sampled after $4N_e$ generations in the past (Corbin et al., 2012). Further, N_e estimates, and especially those related to generations further in the past, are strongly affected by data manipulation factors, such as the choice of MAF and alpha values. Additionally, the binning strategy applied can interfere with the general precision of the method, for example where an insufficient number of pairwise comparisons are used to populate each bin.

One of the applications of method is to compare breed demographies. In this case the shape of the N_e curves would be the optimal tool to differentiate different demographic histories, more than their numerical values, by using them as a potential demographic fingerprint for that breed or species, yet taking into consideration that mutation, migration, and selection can influence the N_e estimation through LD (Waples and Do, 2008). Additionally, careful

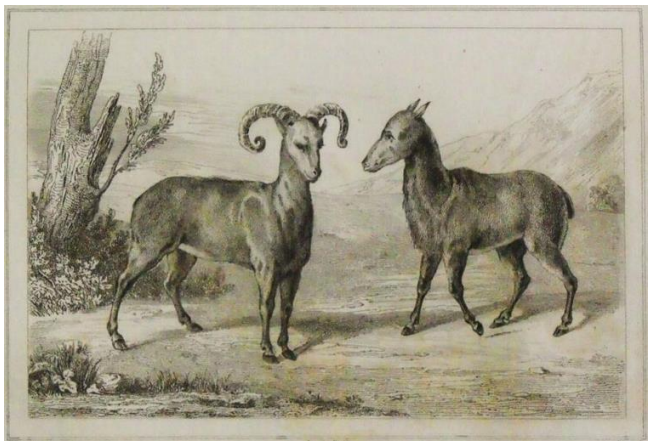
consideration of the data analysed with *SNeP* (and other software to estimate N_e) is very important, as the presence of confounding factors such as admixture, may result in biased estimates of N_e (Orozco-Terwengel and Bruford, 2014).

The aim of *SNeP* is therefore to provide a fast and reliable tool to apply LD methods to estimate N_e using high throughput genotypic data in a more consistent way. It allows two different r^2 estimation approaches plus the option of using r^2 estimates from external software. The use of *SNeP* does not overcome the limits of the method and the theory behind it, yet it allows the user to apply the theory using all corrections suggested to date.

2.7 Acknowledgments

We thank Christine Flury for providing the sheep data and for useful discussion. We also thank the two reviewers for useful suggestions to improve this paper.

Chapter three



Sardinian mouflon (Le Chevalier, 1839)

3 Chapter three – Genomic signatures of adaptive introgression between European mouflon and domestic sheep

Mario Barbato*[1], Frank Hailer[1], Pablo Orozco-terWengel[1], James Kijas[2], Paolo Mereu[3], Pierangela Cabras[4], Raffaele Mazza[5], Monica Pirastru[3] and Michael W. Bruford[1]

[1] School of Biosciences, Cardiff University, Cardiff CF10 3AX, Wales, UK.

[2] CSIRO Agriculture, St Lucia, Brisbane 4067, QLD, Australia.

[3] Department of Biomedical Sciences, and Centre for Biotechnology Development and Biodiversity Research, University of Sassari, Italy.

[4] Istituto Zooprofilattico Sperimentale della Sardegna, Tortolí, Ogliastra, Italy.

[5] Laboratorio Genetica e Servizi (LGS) - Associazione Italiana Allevatori (AIA), Cremona, Italy.

** Lead author contribution: conceived the study, developed the analytical and informatics methods, performed all the analysis and wrote the first draft, modified the manuscript based on co-author comments.*

Keywords: European mouflon, sheep domestication, adaptive introgression, *Ovis aries*, local ancestry, SNP array.

3.1 Abstract

Mouflon (*Ovis aries musimon*) became extinct from mainland Europe after the Neolithic, but remnant populations persisted on the Mediterranean islands of Corsica and Sardinia and have been used for reintroductions across continental Europe since the 19th-century. Documented mouflon *x* sheep hybrids are larger-bodied than mouflon, with potential impacts on male reproductive success, but little is known about genomic levels of admixture, or about any adaptive significance of introgression between resident mouflon and local sheep breeds. Here we analysed Ovine medium-density SNP array genotypes of 92 mouflon from six geographic regions, along with data from 330 individuals of 16 domestic sheep breeds. We found lower levels of genetic diversity in mouflon populations than in domestic sheep, consistent with past bottlenecks in mouflon and with unusually high variability in sheep compared to other domesticates. Most analysed mouflon populations appeared largely unaffected by admixture.

However, mouflon enclosures in Italy harboured admixed individuals, highlighting the need for careful genetic monitoring of these populations. Conversely, we found clear signals of mouflon introgression into domestic sheep, affecting most breeds. Using a novel approach to identify consistent signals of introgression, we infer (a) that breeds from the second wave of sheep domestication show significant introgression from mouflon, and (b) that genomic regions involved are enriched for the function of triggering Neutrophil Extracellular Traps (NETs), an innate immunity mechanism. Further, we infer that Soay and Sarda sheep, two breeds with broad dietary preferences, carry introgressed mouflon genomic regions involved in bitter taste perception and/or innate immunity. Our results provide evidence of adaptive introgression from mouflon in domestic sheep, and our analytical framework can be applied to any low- to mid-density SNP array data.

3.2 Introduction

Introgression is increasingly documented as a potentially adaptive evolutionary force (Hedrick, 2013), with recent developments in high-throughput sequencing facilitating the detection of even small genomic regions that have been passed from one taxon to another (Harrison and Larson, 2014). Since Darwin's early work on domesticates, evolutionary biologists have devoted much attention to the relationships between domesticates and their wild ancestors. While much work has focused on describing the evolutionary consequences of domestication on modern sheep breeds (Bruford et al., 2015), less work has focused on their wild counterpart, the mouflon.

Sheep have a complex evolutionary history shaped by widespread extinctions in the wild, domestication and feralization. The European mouflon (*O. aries musimon*) is the only wild sheep currently occurring in Europe. Present in the archeozoological record in Europe since the middle Pleistocene (Rozzi et al., 2011), European mouflon went extinct across mainland Europe after the Neolithic (Santiago-Moreno et al., 2004). Early agricultural societies then brought domesticated sheep into Europe during the Neolithic transition (Pedrosa et al., 2005). Analysis of retroviral genomic markers in wild and domestic sheep breeds has provided evidence for two main domestication

events: a first wave of domestication at around 11,000 years ago (YA), and a second wave around 6,000 YA (Chessa et al., 2009). The European mouflon, along with the Cypriot mouflon (*O. orientalis ophion*) and some primitive domestic breeds present in Northern Europe such as the Soay and Spael sheep are considered remnants of sheep from the first wave of domestication (Chessa et al., 2009).

Humans translocated mouflon onto the Mediterranean island of Cyprus ~10,000 YA, and to Corsica and Sardinia ~6-7,000 YA (Vigne et al., 2011), where feral populations became established (Vigne, 1992). Since the late 18th century, Corsican and Sardinian mouflon have been used to repopulate several regions of mainland Europe to provide game for hunting (Poplin, 1979; Bon et al., 1991). Corsican mouflon were introduced as game and park animals in Southern France, and subsequently into other European countries (Lorenzini et al., 2011), whereas both Sardinian and Corsican animals were moved to central and northern Italy and Austria (Guerrini et al., 2015; Apollonio et al., 2005). Currently, continental European mouflon are distributed from the Iberian Peninsula to the Caucasus (Figure 3-1).

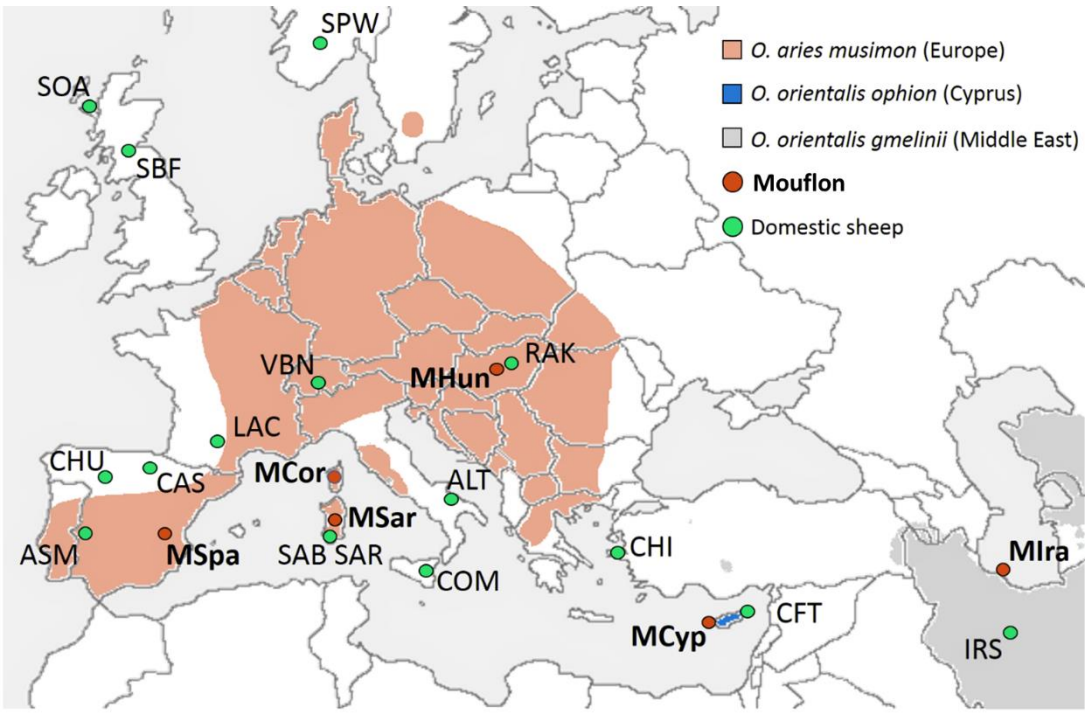


Figure 3-1 Sampling location and mouflon range. Approximate current range of wild and domestic sheep across Europe and area of geographic origins of analysed samples. The sampling sites shown are coloured according to the species.

Since the arrival of the second wave of domesticated sheep, the feral mouflon populations of Sardinia, Corsica and Cyprus have coexisted with sheep populations. Even today, sheep herding practices in Sardinia involve seasonal transhumance from lower towards higher-altitude pastures, where mouflon reside and farmers habitually allow sheep to graze in the wild (Lorenzini et al., 2011). Records from ancient Rome and more recently from the 18th century (both based on Cetti 1774) describe interbreeding between wild and domestic sheep in Europe. Mouflon and domestic sheep have therefore occurred in sympatry on several Mediterranean islands for millennia, and historical records indicate that admixture may be common. Furthermore, mouflon *x* sheep hybrids tend to be larger than mouflon (Hess et al., 2006), and larger-bodied mouflon males have higher reproductive success (Grignolio et al., 2008). Despite these historical records of mouflon *x* sheep admixture, and although sexual selection might act to enhance introgression into mouflon, little information is available on the scale, impact and adaptive significance of admixture between feral and domestic sheep.

Molecular approaches have been used to investigate the genetic structure of European and Mediterranean mouflon populations, including several phylogenetic studies with datasets comprising mouflon and domestic sheep (Bruford and Townsend, 2006; Hiendleder et al., 2002; Meadows et al., 2011). Results suggest that domestic Sardinian sheep and local mouflon show varying levels of admixture (Lorenzini et al., 2011; Ciani et al., 2013), and that mtDNA cannot be used to effectively infer gene flow between them, as both mouflon and domestic breeds belong to the same mitochondrial haplogroup (B) (Meadows et al., 2011). The OvineSNP50 BeadChip (Illumina Inc.), includes 54,241 domestic sheep polymorphisms and while originally developed to assess genetic diversity (Kijas et al., 2012) and perform genome wide association studies (Johnston et al., 2015), it can be used for other purposes such as studying conservation genetics (Allendorf et al., 2010), domestication, local adaptation (Orozco-terWengel et al., 2015) and admixture in wild/feral populations (Miller et al., 2012b; Iacolina et al., 2015).

Here we used the OvineSNP50 BeadChip to analyse a comprehensive European mouflon sample from the European mainland and the Mediterranean, including

Corsican, Corsican-derived (Spain, Hungary) and three Sardinian mouflon populations, providing the most comprehensive genome-scale dataset to date of European mouflon. Mouflon samples from Cyprus and Iran were also included, along with adjacent domestic breeds. We used this dataset to investigate the form, extent and adaptive significance of admixture between feral and domestic populations. We explored signals of local ancestry along sheep chromosomes, applying a novel approach to identify chromosomal regions of consistent ancient ancestry and to infer the direction of introgression (feral to domestic or vice versa). In Sardinia, we analysed sympatric mouflon and sheep populations to infer the adaptive significance of introgression, hypothesizing that introgression from feral mouflon into more recently imported local sheep breeds could have greater adaptive significance with regard to local environmental conditions, than introgression from local (Sarda) sheep into resident mouflon.

3.3 Materials and Methods

3.3.1 Samples, DNA extraction and genotyping

We analysed 92 mouflon from eight populations across continental Europe, remnant populations on Mediterranean islands including Cyprus and three subpopulations from Sardinia, and from Iran. For comparison, we collected as part of this study data from 330 individuals from domestic European sheep breeds that either live in sympatry with or adjacent to mouflon, or have been generally described as old/autochthonous breeds. Other domestic sheep data were available from the Sheep HapMap project (Kijas et al., 2012) (Table 3-1; for additional details see supplementary Text S1). Genomic DNA was obtained from blood and muscle tissue using phenol/chloroform extraction. Sample quality and concentration was determined via spectrophotometry using a ND-8,000 (NanoDrop Technologies, Thermo Fisher Scientific Inc., Wilmington, DE).

Samples were genotyped using the OvineSNP50 BeadChip in the 'Laboratorio Genetica e Servizi' (Cremona) or as part of the ISGC HapMap experiment described previously (Kijas et al 2012). Markers with a call rate <0.99 and minor allele frequency (MAF) <0.05 were excluded from all analyses. The use of SNP array data can be affected by ascertainment bias (Nielsen et al., 2004). For the

Ovine 50k SNP array, no mouflon were included in the discovery panel (Kijas et al., 2012). We therefore pruned SNPs on the basis of linkage disequilibrium (LD) in our dataset, as this approach has been shown to reduce the impact of ascertainment bias, allowing more unbiased comparisons among populations by preferentially reducing mean heterozygosity within the populations used during SNP discovery (Kijas et al., 2012). LD pruning was performed using the --indep-pairwise function in PLINK 1.7 (Purcell et al., 2007), where SNPs with r^2 >0.5 were removed from sliding windows of 50 SNPs and with 10 SNPs of overlap. Only autosomal markers were kept for analysis. After pruning for MAF and LD, 36,961 SNPs distributed across 26 chromosomes were retained for analysis.

Table 3-1 Sample information and diversity indexes. For each sheep and mouflon population used in this manuscript, the breed/population name, the acronym (Acr) used throughout the manuscript, the Country of origin and the number of individuals analysed in this work are shown in the first four columns along with the observed heterozygosity (H_o) and its standard deviation (SD, within brackets), the effective population size (N_e) and the inbreeding coefficient (F). Genotypes collected as part of this study are indicated in the last column as 'T', those collected as part of the HapMap project as 'HM', and those provide by the NextGen consortium as 'NG'.

	Breed/population	***Acr***	***Origin***	***N***	***H_o (SD)***	***N_e***	***F***	***Origin***
mouflon	Sardinian mouflon	MSar1	Sardinia	19	0.22 (0.19)	261	0.45	T
	Sardinian mouflon	MSar2	Sardinia	8	0.22 (0.24)	130	0.46	T
	Sardinian mouflon	MSar3	Sardinia	28	0.34 (0.19)	273	0.16	HM
	Spanish mouflon	MSpa	Spain	21	0.20 (0.19)	96	0.51	HM
	Hungarian mouflon	MHun	Hungary	8	0.24 (0.21)	282	0.42	T
	Corsican mouflon	MCor	Corsica	3	0.24 (0.27)	259	0.41	T
	Cypriot mouflon	MCyp	Cyprus	3	0.09 (0.20)	244	0.78	T
	Iranian mouflon	MIra	Iran	2	0.25 (0.31)	-	0.35	NG
	total			*92*				
domestic sheep	Altamurana	ALT	Italy	24	0.37 (0.16)	628	0.06	HM
	Australian Merino	ASM	Spain	24	0.37 (0.15)	920	0.06	HM
	Castellana	CAS	Spain	23	0.38 (0.16)	813	0.02	HM
	Chios	CHI	Greece	23	0.33 (0.17)	391	0.15	HM
	Churra	CHU	Spain	24	0.37 (0.16)	617	0.05	HM
	Comisana	COM	Italy	24	0.38 (0.16)	1028	0.03	HM
	Cyprus Fat Tail	CFT	Cyprus	24	0.34 (0.19)	186	0.13	HM
	Iranian sheep	IRS	Iran	6	0.37 (0.22)	412	0.05	NG
	Milk Lacaune	LAC	France	24	0.37 (0.16)	607	0.06	HM
	Nera di Arbus sheep	SAB	Sardinia	20	0.36 (0.18)	366	0.08	HM
	Racka	RAK	Hungary	8	0.35 (0.21)	327	0.11	T
	Sarda sheep	SAR	Sardinia	10	0.37 (0.19)	755	0.07	T
	Scottish Blackface	SBF	UK	24	0.37 (0.17)	428	0.05	HM
	Soay	SOA	UK	24	0.27 (0.20)	179	0.32	HM
	Spael-white	SPW	Norway	24	0.34 (0.18)	367	0.14	HM
	Valais Blacknose sheep	VBN	Switzerland	24	0.31 (0.18)	306	0.22	HM
	total			*330*				

3.3.1.1 Genetic diversity and population structure

Heterozygosity values were calculated using custom scripts and the inbreeding coefficient (F) was estimated using PLINK. Effective population size (N_e) was estimated with the software *SNeP* v1.1 (Barbato et al., 2015). The software uses linkage disequilibrium to infer N_e at different t generations in the past where $t = 1/2c$ and c is the distance between SNPs in Morgans (in this case assuming 100 Mb = 1 Morgan; Kijas *et al.* 2012). The following options were used: sample size correction for unphased genotypes, correction to account for mutation, and Sved & Feldman's mutation rate modifier. The most recent estimate of N_e was taken for c calculated at 1 Mb (Kijas et al., 2012).

Maximum likelihood analysis of population structure was conducted using ADMIXTURE v1.23 (Alexander et al., 2009). Clustering solutions for the whole dataset were calculated for K values from 2 to 24, the latter corresponding to the total number of sampled populations/breeds in our study. A principal component analysis (PCA) was performed to investigate the ordinal relationships between populations and individuals, using flashpca v1.2 (Abraham and Inouye, 2014) with default settings. Neighbour-net graphs based on Reynolds' distances, calculated with a custom script, were generated using Splitstree v4.13.1 (Huson and Bryant, 2006). The occurrence of admixture was further investigated using Treemix v1.12 (Pickrell and Pritchard, 2012). This software models the relationship among the sample populations with their ancestral population using genome-wide allele frequency data and a Gaussian approximation of genetic drift (Pickrell and Pritchard, 2012). The f index representing the fraction of the variance in the sample covariance matrix ($\widehat{W}$) accounted for by the model covariance matrix (W) was used to identify the information contribution of each migration vector added to the tree. Up to 20 possible migration vertices were computed.

3.3.1.2 Inference of local genomic ancestry (PCAdmix)

We used PCAdmix v1.0 (Brisbin et al., 2012) to infer local genomic ancestry. PCAdmix utilizes haplotypes from ancestral representatives to infer ancestry of focal individuals. The software performs the inference chromosome-wide through PCA, via short windows along each chromosome. Using a hidden Markov

Model, PCAdmix then returns the posterior probability (PP) of ancestry from each reference population for each haploid individual for each window. Additional information on PCAdmix parameters is available in supplementary Text S1. PCAdmix requires phased genotypes, which we obtained using fastPHASE v1.2 (Scheet and Stephens, 2006). Default parameters were used in fastPHASE, except that we allowed for the incorporation of subpopulation labels, this has been shown to significantly improve the imputation accuracy (Hayes et al., 2012).

To perform the local genomic ancestry analyses we used three reference populations: one population representative of the Sardinian mouflon lineage, one of the Corsican mouflon lineage, and one domestic sheep breed (see supplementary Text S1 for details). Analyses were conducted on a range of domestic breeds, to assess how this choice affected the results. When using the MSar1 and SAB samples as reference populations, we removed individuals that showed evidence of introgression (introgressed ancestry component >0.01) based on the Admixture results (see below). These subsampled populations were renamed MSar1_p and SAB_p with the "p" denoting "pure".

3.3.1.2.1 A novel approach to identify consistent genomic windows of introgression

Given relatively high variation among PCAdmix results for different breed combinations, we developed a pipeline that uses a sliding-window approach to identify genomic regions that show consistent signals of introgression across different population comparisons in PCAdmix. Specifically, several PCAdmix analyses are performed, each utilizing different reference populations, and the results are filtered for highly concordant introgression signals: the sliding-window approach assigns a concordance score along chromosomes. Next, regions exhibiting concordance scores higher than a certain percentile of the genome-wide concordance score distribution are identified as Consistently Introgressed Windows of Interest (CIWI). Here we conducted the analysis based on both the 95th and 99th percentile. Additional information on the CIWI approach is available in supplementary Text S1.

3.3.1.2.2 GO terms identification

To identify gene ontology (GO) terms significantly overrepresented in the CIWI, all genes located inside or within 20kbp (half median distance between 2 SNPs in the Ovine 50k SNP chip) from the endpoints of each CIWI were compared against a background set of 11,089 genes, each containing or being in close proximity with a SNPs present in the 50k SNP chip (Kijas et al., 2012). The comparisons were performed using GOrilla (Eden et al., 2009), employing a false discovery rate (FDR) threshold of 0.05.

All C++ code and R scripts used in this research are available upon request.

3.4 Results

In total 422 individuals from seven wild and 16 domestic sheep populations were analysed at 36,961 SNP positions after pruning for MAF and LD. Observed heterozygosity (Table 3-1) ranged from 0.09 to 0.34 for mouflon populations, with MSar3 showing the highest value and MCyp the lowest. Heterozygosity for domestic sheep breeds was generally higher (range: 0.30 to 0.38), with most values overlapping with those reported previously for the same populations (Ciani et al., 2013; Kijas et al., 2012). The N_e values of most mouflon populations were around 250, with the highest value recorded for MHun (282) and the lowest for MSpa (96). N_e values for domestic sheep were generally comparable with those from previous studies (Kijas et al., 2012; Ciani et al., 2013), showing differences in N_e <50, with the exception of ALT, COM, CAS where the difference in N_e was >100. The inbreeding values estimated ranged around 0.40-0.45 in most mouflon populations, although a lower value was recorded for MSar3 (0.16) and the largest was recorded for MCyp (0.78). The inbreeding values for most domestic sheep populations were low (0.02-0.15); larger values were recorded for VBN (0.22) and SOA (0.32).

3.4.1 Population structure and genome-wide signals of admixture

Results from Admixture analysis at K=2 separated the samples into largely distinct domestic sheep and mouflon clusters (Figure 3-2). However, extensive signals of admixture were discernible in 24 individuals of the MSar3 population

showing between the 21-51% of sheep assignment, as well as in one individual of the MCor population. Otherwise, this, low admixture proportions (~5%; Fig. 2) were discernible in most surveyed mouflon populations. Additionally, the eastern MCyp and MIra populations showed ~74% cluster membership consistent with domestic assignment. Domestic breeds showed on average 14% of mouflon cluster membership, except Eastern Mediterranean and SW Asiatic breeds (IRS, CHI and CFT) for which the mouflon component was <5%. At K=5, a cluster restricted to mouflons derived from Corsican stocks (MCor, MHun and MSpa) was detected, which was absent from other mouflon populations in Sardinia and other regions. At K=11, MCyp was detected as a distinct cluster, as was MSar2 at K=12, while the Corsican and Hungarian mouflon formed a distinct cluster at K=18.

In the PCA analysis (Figure S 3-1), the first PC accounted for 7.3% of the variance and discriminated sheep and mouflon, basically mirroring the admixture results obtained at K=2. The second PC accounted for 3.4% of the variance and resembled Admixture results for K=3, discriminating northern sheep breeds from other domesticates. The third and fourth PCs split the Sardinian and Corsican mouflon and the Asiatic and European sheep breeds respectively. The Neighbour-net analysis of pairwise Reynolds' distances between sampled populations (Figure S 3-2) clearly differentiated mouflon from domestic sheep. The European mouflon occupied a separate branch and was further split into the Corsican and Sardinian lineages. Separate branches differentiated the North European domestic sheep breeds, the two Sardinians and the breeds from the East Mediterranean.

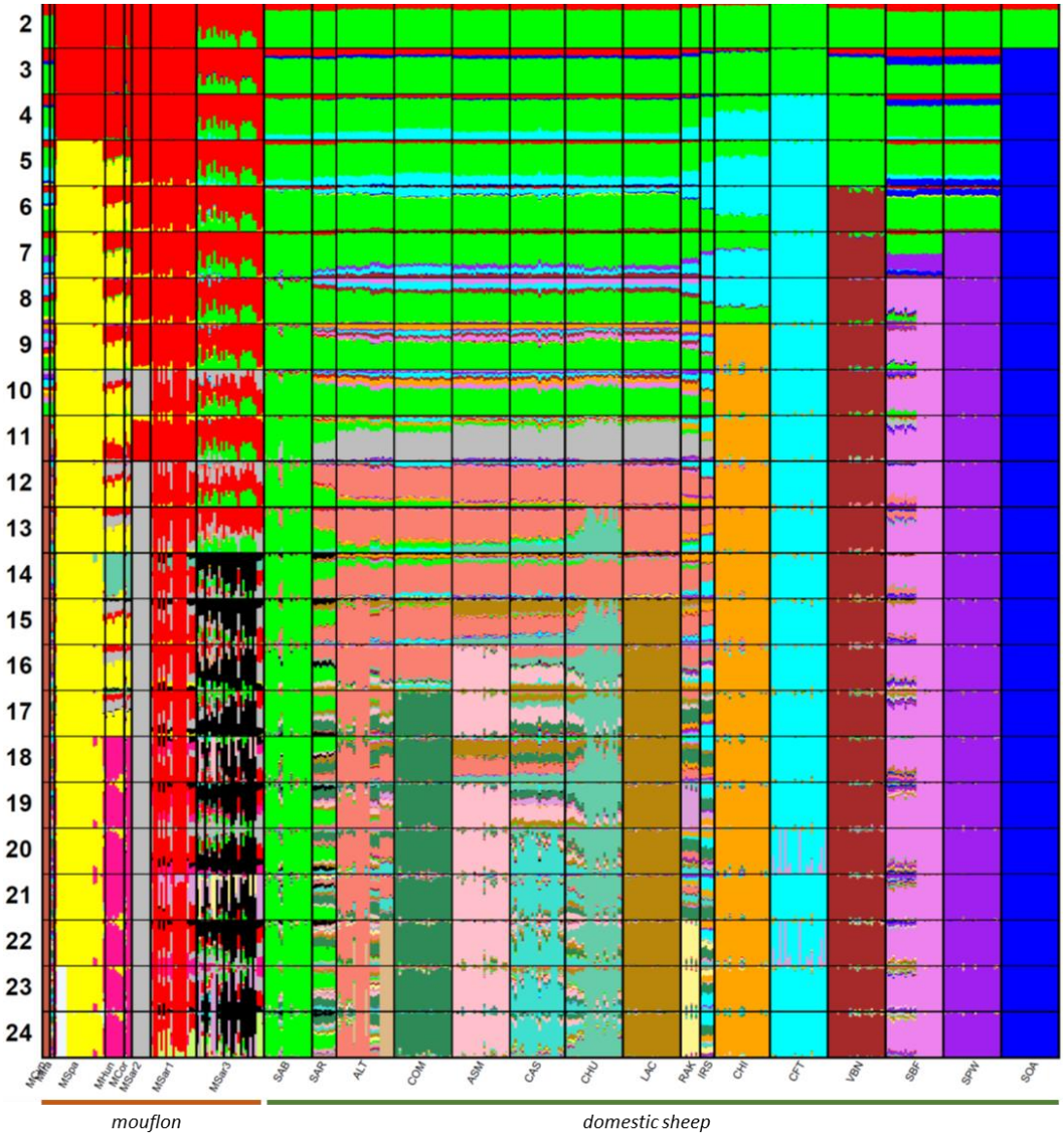


Figure 3-2 Admixture (Alexander et al., 2009) plot comprising the first 24 clustering solutions (numbers on the left) of all the 422 individuals analysed in this work. The analysis is based on 36,961 SNPs from the Ovine SNP50BeadChip. For population abbreviations see Table 1.

Maximum likelihood assessment of population history with overlaid admixture events using Treemix (Figure 3-3, Figure S 3-3) confirmed several aspects already detected by Admixture (Figure 3-2). The first four migration edges (gene flow events) accounted for more than half of the total model significance explained by the *f* statistic, with the first migration edge having an *f* value of 0.98. Vectors from 9 to 20 brought only a small increase in *f* value (<0.001) and migration weights close to 0 (Figure 3-3, inset). The first four vectors all indicated gene flow between sympatric mouflon and sheep (Figure 3-3): vectors 1-3 denote gene flow from domestic sheep into mouflon on Sardinia, Iran and Cyprus,

mirroring K= 2 results from Admixture (Figure 3-2). The fourth Treemix vector connected European mouflon to Sardinian domestic sheep breeds (SAR and SAB), indicating bidirectional gene flow.

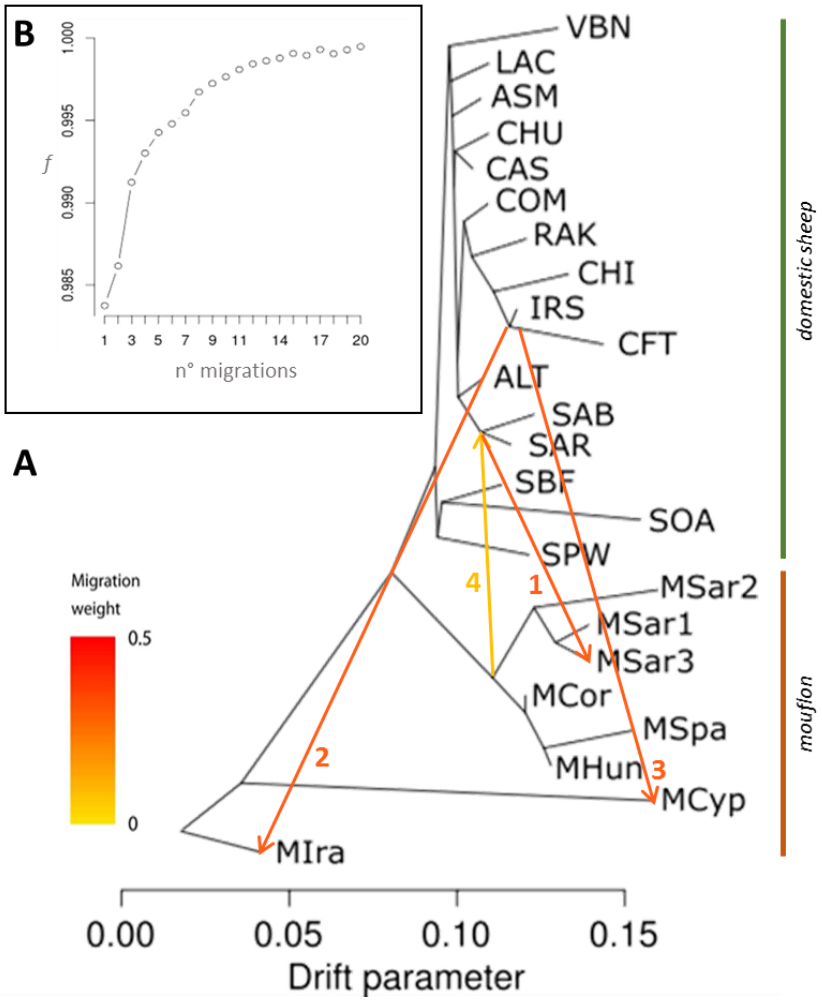


Figure 3-3 A) The phylogenetic network inferred by Treemix (Pickrell and Pritchard, 2012) of the relationships between mouflon and domestic sheep populations. The first four migration edges between breeds are shown with arrows pointing in the direction toward the recipient breed, and coloured according to the ancestry percent received from the donor breed. The number associated to each vector identify the identification order. B) The inset shows the f index representing the fraction of the variance in the sample covariance matrix ($\widehat{W}$) accounted for by the model covariance matrix (W), as a function of the number of modelled migration events. For population abbreviations see Table 1.

3.4.2 Inferring sheep versus mouflon ancestry of specific genomic locations

The observed widespread signals of genomic admixture prompted analysis using PCAdmix to identify specific genomic regions showing signals of introgression. Using either MSar1_p or MSar2 as representatives of the Sardinian mouflon, the obtained results consistently overlapped, regardless of the domestic population compared (data not shown). All subsequent analyses were therefore performed using three putatively admixture-free reference populations in PCAdmix: MSar2 and MHun represented the mouflon gene pool, while we tested four different domestic sheep reference populations for each focal population to be investigated for local genomic ancestry. Graphical representations of all results are in supplementary results (Figure S 3-4).

Table 3-2 Genome-wide local ancestry assignment values from PCAdmix. Shown is the average proportion of genome assigned by PCAdmix with posterior probability ≥ 0.95 to Sardinian mouflon, Corsican mouflon and domestic sheep, respectively. PP < 0.95 denotes the cumulative genome proportion remaining unassigned with posterior probability < 0.95. Averages were calculated across four reference sets (detailed in Table S5), each comprising the same mouflon references (MSar2 and MHun for Sardinian and Corsican mouflon, respectively) and four different domestic sheep breeds (CAS, ASM, LAC and SAB_p). Light grey background highlights the mouflon population (MSar3) that presents a higher sheep genetic component than mouflon. For population abbreviations see Table 1-1.

		Reference populations			
		Sardinian mouflon	***Corsican mouflon***	***Domestic sheep***	***PP < 0.95***
	Mouflon				
	MSar1	26.3 ± 0.40	20.4 ± 3.26	10.8 ± 1.74	42.6 ± 1.84
	MSar3	16.5 ± 0.59	11.5 ± 2.24	30.6 ± 3.49	41.4 ± 0.75
	MSpa	16.8 ± 2.19	48.3 ± 5.50	4.0 ± 0.29	30.9 ± 3.21
	Domestic sheep				
Focal populations	ALT	1.5 ± 0.85	2.6 ± 2.69	82.3 ± 11.43	13.6 ± 7.89
	CHI	1.4 ± 1.16	2.2 ± 2.87	85.8 ± 12.47	10.6 ± 8.45
	CHU	1.7 ± 1.01	2.8 ± 3.03	82 ± 12.68	13.5 ± 8.68
	COM	1.7 ± 0.92	2.5 ± 2.69	82.3 ± 11.47	13.5 ± 7.91
	CTF	1.0 ± 0.73	2.3 ± 3.17	87.7 ± 11.06	9.0 ± 7.16
	RAK	1.3 ± 0.60	2.3 ± 2.82	86.2 ± 10.74	10.2 ± 7.35
	SAR	1.5 ± 0.67	2.1 ± 2.21	85.5 ± 8.22	11.0 ± 5.34
	SBF	1.6 ± 0.84	3.3 ± 3.61	80.8 ± 12.86	14.3 ± 8.42
	SOA	2.3 ± 0.63	4.1 ± 3.37	80.0 ± 10.24	13.6 ± 6.26
	SPW	1.7 ± 1.05	3.3 ± 3.62	81.0 ± 12.76	14.1 ± 8.10
	VBN	1.5 ± 0.88	2.6 ± 3.17	84.4 ± 12.07	11.5 ± 8.03

The highest assignment of MSar1 was to Sardinian mouflon (26.3% assigned to MSar2; Table 3-2), while Corsican mouflon assigned with 48.3% to MHun. MSar3 was the mouflon population with the highest proportion of its genome assigned to domestic sheep (30.6%), close to the mean estimate obtained from Admixture at K=2 (29%). The proportion of non-assigned (PP < 0.95) regions were similar for MSar1 and MSar3 (42.6% and 41.1% respectively), and the proportions of Sardinian and Corsican ancestry in MSar3 were each ~10 percent lower than for MSar1.

For domestic sheep, the average proportion of the genome assigned with PP ≥ 0.95 was 83.5% (range 61.9-94.7%; Table S 3-3), a value comparable to Admixture results for K=2, showing an average of 88% for the same populations. The highest and lowest mouflon admixture proportions from PCAdmix were consistently recorded by SOA and CFT respectively (Table 3-2). A marked difference between mouflon and domestic sheep in the certainty of assignment was recorded, with mouflon showing a much higher proportion (~39%) of non-assigned regions (PP < 0.95) than domestic sheep (~12%).

PCAdmix therefore provided estimates of genome-wide admixture proportions for mouflon and domestic sheep that were consistent with results from Admixture. However, PCAdmix results for particular genomic regions – the main goal of using this approach - showed inconsistency with regard to the location and length of inferred admixture regions (Figure 3-4-D, Figure S 3-4): with the same genomic region being assigned to either mouflon or sheep ancestry, depending on the reference population utilized in PCAdmix. Consequently, we developed a consensus approach that jointly analysed the results obtained with the four different domestic sheep reference populations, and across all analysed individuals from the focal population, highlighting genomic regions (CIWIs) that, independently of the reference population used, showed highly consistent signals of introgression (Figure 3-4-C, Figure S 3-5). The CIWI regions obtained were then crosschecked against all genes covered by (or adjacent to) SNPs on the ovine 50k BeadChip. On average across mouflon populations, our method identified introgression signals for 524 genes associated with 2,044 CIWI when using the 95th percentile confidence threshold, and 116 genes associated with

397 CIWI for the 99th percentile threshold. Across domestic sheep, a larger number of CIWIs were identified, with an averages of 996 genes located in 3,205 CIWI, or 253 genes in 625 CIWI, based on the confidence thresholds of the 95th and 99th percentile, respectively (Table S 3-1).

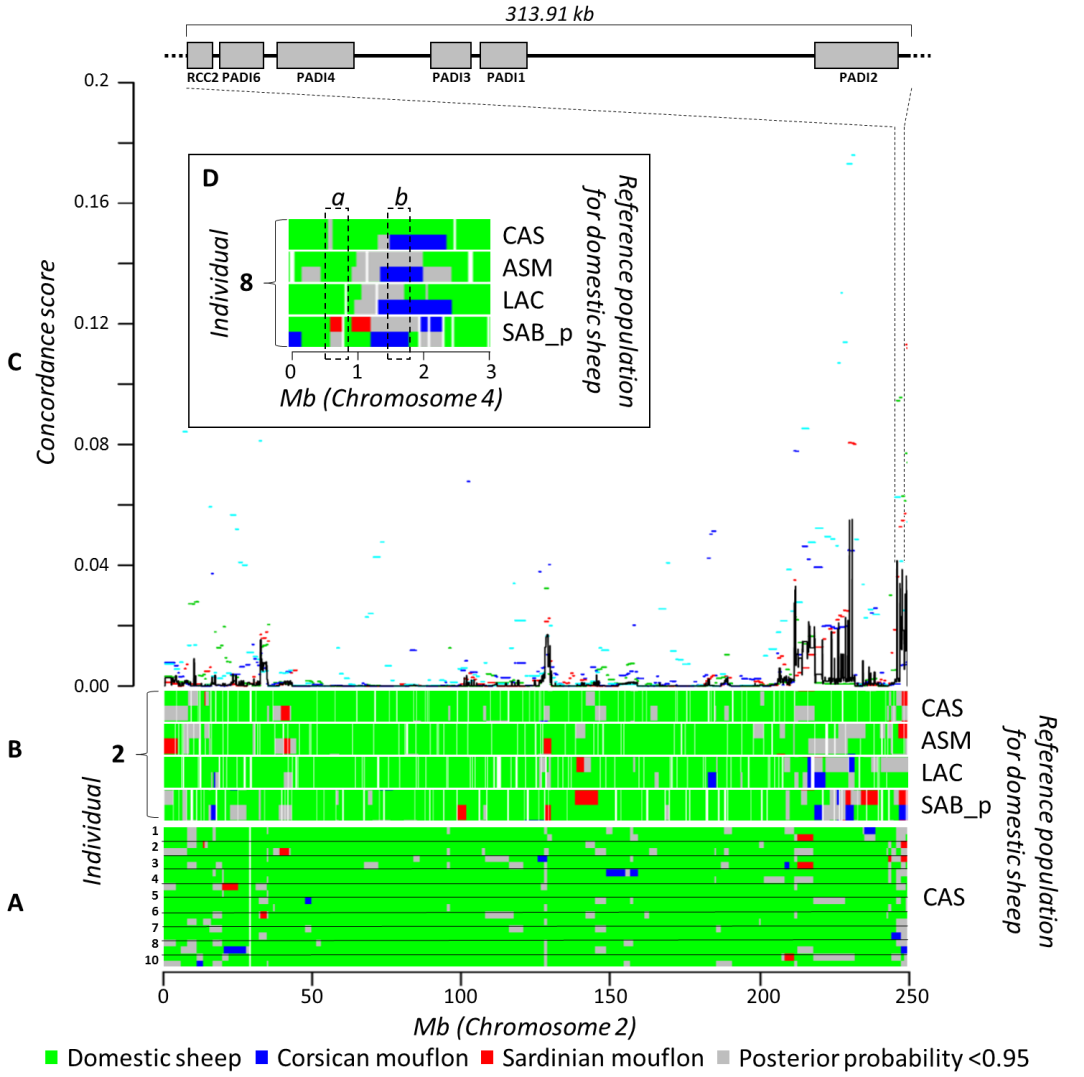


Figure 3-4 A) Graphical representation of the inferred local ancestry for a domestic sheep breed (SAR) according to PCAdmix (Brisbin et al., 2012). The 10 diploid individual belonging to SAR are represented by the 10 numbered lines. Each line represents a diploid individual of the SAR population, and extends for the total length of the ovine chromosome 2 (249.99 Mb). The colour scheme indicates the assignment of each block to one of the three reference populations. B) PCAdmix results for the same individual (2) of the SAR populations obtained with 4 different reference populations for domestic sheep. Genomic regions not analysed by the software due to the absence of SNPs are visible as white gaps. C) A-Scores are calculated along the chromosome to represent the concordance of ancestry assignment between the individuals of the population under scrutiny (SAR). The A-scores relative to the 4 reference population are represented by the coloured segments within the graph. The CIWI score is then calculated according to the A-scores and is represented by the black solid line. The inset expands the genomic region within the chromosome where genes related with the citrullination function can be found. D) The comparison between PCAdmix results for a portion of the fourth chromosome in the same individual produced using four different domestic sheep references is shown. Regions of discordance (a) and concordance (b) are highlighted by dashed boxes.

3.4.3 GO term analysis of introgressed loci

Genes identified within CIWIs in MSar1, MSar3 and MSpa showed no significant enrichment of any GO terms (FDR ≈ 1), independent of the percentile threshold used. Conversely, seven GO terms associated to mouflon ancestry were identified in domestic breeds. All domestic breeds except SOA and SPW produced GO terms involved with protein citrullination (Table S 3-2) whereas in both SAR and SOA, GO terms involved with the perception of bitter taste were present with significant support (Table S 3-2). The genes identifying the protein citrullination were found in four distinct chromosomal regions (two on chromosome 2 and one each on chromosomes 3 and 25; Table 3-3). The identified genes related to bitter taste perception were located on chromosomes 2, 4 and 16 (Table 3-3).

Table 3-3 Genes involved with the GO terms identified. The genes identified by the CIWI approach and involved in either Citrullination or Bitter-taste detection are shown. The chromosome (Chr) and physical position are shown.

GO class	*Gene*	*Chr*	*Position (Mb)*
Citrullination	CPS1 - carbamoyl-phosphate synthase 1	2	211.16-211.3
	ATIC - 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/imp cyclohydrolase	2	216.27-216.3
	PADI6 - peptidyl arginine deiminase, type vi	2	248.06-248.07
	PADI4 - peptidyl arginine deiminase, type iv	2	248.08-248.11
	PADI3 - peptidyl arginine deiminase, type iii	2	248.15-248.16
	PADI1 - peptidyl arginine deiminase, type i	2	248.17-248.19
	PADI2 - peptidyl arginine deiminase, type ii	2	248.31-248.35
	MAT2A - methionine adenosyltransferase ii, alpha	3	57.23-57.24
	MAT1A - methionine adenosyltransferase i, alpha	25	35.29-35.33
Bitter taste detection	TAS1R2 - taste receptor, type 1, member 2	2	246.63-246.65
	TAS2R3 - taste receptor, type 2, member 3	4	104.79-104.79
	TAS2R4 - taste receptor, type 2, member 4	4	104.81-104.81
	TAS2R38 - taste receptor, type 2, member 38	4	104.95-104.96
	PIP - prolactin-induced protein	4	105.91-105.92
	TAS2R39 - taste receptor, type 2, member 39	4	106.01-106.01
	TAS2R40 - taste receptor, type 2, member 40	4	106.07-106.07
	TAS2R1 - taste receptor, type 2, member 1	16	63.38-63.38

3.5 Discussion

3.5.1 Genetic diversity of European mouflon

Estimates of genetic diversity were similar for most previously studied mouflon populations, although we observed two apparent outliers, MCyp and MSar3. Lower variability and higher inbreeding levels were found for mouflon on Cyprus, which also showed a long branch in the Neighbour-net analysis (Figure S 3-2). This result is likely attributable to strong genetic drift and inbreeding in this lineage. While this inference could be biased by small sample size for MCyp (n=3), the Iranian mouflon population had fewer samples (n=2), but showed higher heterozygosity, comparable to that of the other mouflon populations (albeit with a larger standard deviation as expected). In contrast, one mouflon population from Sardinia (MSar3) showed higher genetic variability and lower inbreeding than the others (Table 1-1). The H_o value recorded was the highest among the surveyed mouflon populations, falling within the range of domestic sheep heterozygosity, while both the PCA and Admixture analyses showed clear signals of gene flow between this mouflon population and Sardinian domestics. Given the consistent signals of introgression from domestic sheep into MSar3, the elevated diversity and reduced inbreeding levels found in this particular Sardinian mouflon population likely results from the introgression of domestic sheep alleles.

Mouflon showed lower heterozygosity, higher inbreeding and lower N_e than domestic sheep breeds. The observed heterozygosity and inbreeding coefficient obtained for domestic breeds were comparable with those obtained previously (Kijas et al., 2012; Ciani et al., 2013). Similarly, the effective population size estimates were comparable with those obtained by others who used the same theoretical approach with different sample sizes and applying different corrections (Kijas et al., 2012; Ciani et al., 2013).

Ascertainment bias may contribute to our observation of lower variability in mouflon than in sheep (Albrechtsen et al., 2010; McTavish and Hillis, 2015), given that mouflon were not part of the panel of individuals included when selecting the markers for the OvineSNP50 BeadChip. While this complicates direct

comparisons of observed heterozygosity between mouflon and domestic sheep, comparisons within groups (e.g., among mouflon populations) are affected to a lower extent. Furthermore, ascertainment bias can be partly alleviated by pruning data for high levels of linkage disequilibrium (Kijas et al., 2012), and by using multilocus or haplotype-dependent analyses that are less affected by ascertainment bias than single locus statistics (VonHoldt et al., 2011; McTavish and Hillis, 2015; Miller et al., 2011). Hence, here we removed loci with high levels of LD and employed both multilocus (e.g., Admixture) and haplotype-based (e.g., PCAdmix) analysis approaches.

Previous studies using microsatellites on both domestic sheep and European mouflon show the same trend as our SNP chip data, with mouflon populations generally showing lower heterozygosity than domestic sheep (Lawson Handley et al., 2007; Stahlberger-Saitbekova et al., 2001; Calvo et al., 2011). Microsatellites can also be affected by ascertainment bias (Frascaroli et al., 2013) but their high mutation rate reduces the effect of the bias in comparison to that in SNP data. These observations suggest that, despite impacts of SNP chip ascertainment bias, European mouflon harbour clearly lower levels of genetic diversity than their domestic counterparts. We suggest two complementary explanations for this:

Bottlenecks in mouflon: The mouflon populations analysed from the Mediterranean represent survivors of several past population bottlenecks (Apollonio *et al.* 2005; Sanna *et al.* 2015). Except for the Iranian population, for which no detailed population history information is available, all the studied mainland mouflon populations derived from reintroductions of Corsican or Sardinian animals (Apollonio *et al.* 2005). These bottlenecks have likely reduced genetic variability in mouflon.

Unusually high genetic diversity in sheep breeds compared with other domesticates: Domestic sheep breeds have previously been shown to contain higher genetic diversity than breeds of other domesticates (Taberlet et al., 2008). This has been attributed to the recruitment from highly heterogeneous pre-domestication populations, as evidenced by mitochondrial DNA (Tapio et al., 2006; Meadows et al., 2011; Kijas et al., 2012). The persistence of this high diversity may result from

cross-breeding presumably occurred in the centre of domestication with wild populations after the initial domestication (Kijas et al., 2012), although the latter seems similarly true for other domesticates (Vilà et al., 2005) and thus seems unlikely to explain the particularly high variability within sheep breeds. One main difference between sheep and other domesticates is that sheep breeds are typically managed in a way that promotes the maintenance of genetic variability. Specifically, while sheep farmers uses one or more rams to sire the flocks, breeding practices in dogs, cattle and horses often involve the use of popular sires that are bred to father a large number of offspring (Vilà et al., 2001; Morelli et al., 2014; Lindgren et al., 2004; Sundqvist et al., 2005; Taberlet et al., 2011). Sheep breeding practices therefore likely contribute to the observed higher variability in sheep than in mouflon.

3.5.2 Limited signals of sheep introgression into wild European mouflon

Despite the records of crossbreeding occurring since Roman times (Cetti, 1774), the majority of the analysed European mouflon populations showed no obvious signs of introgression from domestic sheep. However, marginal levels of sheep cluster membership consistent with domestic assignment were identified by Admixture analysis for some mouflon individuals. However, similar to the findings of Lorenzini et al. (2011) who documented the presence of second-generation crossbred or backcrossed Sardinian mouflons sampled in the "Ogliastra" and "Nuoro" provinces, we found concordant signals of recent introgression for MSar3 in several analyses. Concordant results from Admixture and Treemix suggest that most MSar3 individuals represent first and second generation crossbreeding or backcrossing with Sardinian sheep as putative introgression source. In fact, the MSar3 population has been sampled in an enclosure in which mouflon were reared together with crossbred animals (Tiziana Sechi, *personal communication*). Our results indicate that this population, which has been used so far as reference for Sardinian mouflon (Ciani et al., 2013, 2015) should be used with caution as a reference for Sardinian mouflon.

A strong sheep component was also present in the two *O. orientalis* populations MCyp and MIra according to Admixture results. In fact, although at K=2 the dataset could be split between mouflon and domestic sheep these populations had ~75% of sheep component assigned (Figure 3-2) and the same pattern was visible in both the PCA and Neighbour-net (Figure S 3-1 and Figure S 3-2, respectively). However, these results are most likely an effect of ascertainment bias, as *O. orientalis* are the most divergent populations with respect to the domestic sheep (i.e., the species for which the SNP chip was designed)(Sanna et al., 2015). We anticipate that genome sequencing and ascertainment bias-free characterization of genetic diversity in mouflon will be important in lineages that are more divergent from domestic sheep, such as mouflon from Cyprus and Iran.

While no obvious explanation is available for the overall absence of contemporary introgression in Sardinian mouflon (except the enclosure-held population MSar3), the same observation in mainland European mouflon may be explained by current mouflon management. On the mainland, mouflon populations derive from recent introductions (e.g., Hungary, Romania and Spain). These populations are kept as game for hunters and therefore either confined in enclosures, partially managed or monitored, potentially reducing the chances of crossbreeding. Additionally, sheep *x* mouflon hybrids tend to deviate phenotypically from mouflon (e.g., white fleece patches, woolly coat) and are in some areas actively removed from the wild gene pool by the hunting community (Frisina and Frisina, 2013). In conclusion, mouflon-sheep hybrids are rare throughout Europe, and might be actively selected against by humans.

Historically large effective population size of mouflon may have limited the impact of any introgression from sheep. Cetti (1774) described Sardinian mouflon flocks each comprising hundreds of animals, much larger than the currently typical group sizes of less than ten individuals (Pipia et al., 2009). Furthermore, rarity of sheep introgression into mouflon could result from natural selection, with hybrid fitness being reduced in feral conditions. Indeed, none of the regions in the mouflon genome that we infer to have domestic sheep ancestry shows any significant enrichment of GO terms, whether in the wild or in enclosures (MSar3). This is consistent with Corsican and Sardinian mouflon

being adapted to local environmental conditions when sheep from the second wave of domestication arrived on the islands. Under this scenario, resident mouflon might not have benefited from introgression of non-resident alleles from domestic sheep. Notably, our results also imply that hybrids – despite their larger body size and hence potentially increased reproductive success (Hess et al., 2006) (Grignolio et al., 2008) - do not seem to have a larger reproductive success than purebred mouflon.

3.5.3 Adaptive introgression in domestic sheep

Most analysed sheep breeds showed a small percentage of contribution from the mouflon cluster at K=2 (Figure 3-2). Among these, all SAR individuals showed similar amounts of Sardinian mouflon ancestry (at K=24; Figure 3-2). Additional support came from the Treemix analysis, which shows a migration vector starting from the root of Sardo-Corsican mouflon and ending at the root of the two Sardinian sheep breeds (Figure 3-3). These results likely represent ancient admixture events.

Applying our newly developed consensus approach to identify CIWIs and associated genes in domestic sheep, we found mouflon ancestry in genomic regions related to the citrullination process. Citrullination enzymes such as PAD4 are essential in triggering the antibacterial innate immunity response known as neutrophil extracellular traps (NETs) (Brinkmann et al., 2004), with histone citrullination as the first step leading to NET assembly (Wang et al., 2009; Li et al., 2010). Citrullination also plays a major role in bacteria-dependent inflammatory response in livestock, and enzymes responsible for this process are overrepresented in mastitic Sarda sheep milk (Pisanu et al., 2015). Interestingly, the ancestry signature on the citrullination function was found in all analysed sheep breeds, except SOA and SPW, which are regarded as primitive breeds that derive from the first wave of domestication (along with other Northern sheep breeds and the Mediterranean mouflon subspecies; Chessa *et al.* 2009). This finding suggests that introgression of these genomic regions into sheep does not stem from recent admixture, but rather results from introgression relatively early during the second wave of the sheep domestication process. We hence hypothesize that adaptive introgression from mouflon could have been positively

selected in domestic sheep, possibly because of fitness effects of higher plasticity of the antibacterial innate immunity provided by NETs. This adaptive introgression from mouflon into sheep could have helped translocated - and thus not necessarily locally adapted - domestic breeds cope in their novel environments.

We also found evidence of enrichment in GO terms related to bitter taste perception in Soay and Sarda sheep. Bitter taste perception in ruminants is associated with sensitivity to toxins and with food choices that avoid dangerous substances in the diet (Liu et al., *in review*). However, in domestic sheep, perception of bitter taste in a food item does not correlate with its rejection, probably reflecting a trade-off between toxicity avoidance and dietary plasticity (Hernández-Orduño et al., 2015; Favreau et al., 2010). Although taste receptors are mostly located in the tongue, they can also be expressed in other tissues (Lee and Cohen, 2014). The function of the majority of these extraoral receptors is unknown, although bitter and sweet receptors in the airways have been linked to innate immunity functions (Lee and Cohen, 2014). Soay sheep live under particularly low management conditions (Ryder, 1981), which could explain the selective advantage of an introgressed genetic material from mouflon related to bitter taste perception. Similarly, the Sarda sheep is known to have broad dietary preferences and is characterized by primitive management practices. These aspects could explain why mouflon-derived bitter taste genes proved adaptive in these breeds, regardless whether the involved genes are ultimately related to bitter taste perception or to innate immunity derived functions.

3.5.4 Conclusions and outlook

Here we developed an approach to identify consistent regions showing introgression signals, based on mid-density SNP array data from multiple reference populations and focal individuals. This strategy is prone to generate false negative results, but it reduces the occurrence of false positives. We note that our approach focuses on introgression signals that are close to fixation in the focal population, so situations where introgressed material is relatively rare in the population would not be detected (Sankararaman et al., 2014).

Despite millennia of coexistence of mouflon and domestic sheep, our findings indicate that only very limited introgression from domestics in the wild has occurred. Given expected better local adaptation of mouflon compared to recently domesticated and typically geographically translocated domestic breeds, this finding is not unexpected. Conversely, introgression from wild mouflon into sympatric domestics seems to have been more common and targeted by positive selection. Adaptive introgression of mouflon alleles may be explained by limited local adaptation in domestics. Specifically, we here show that genes with functions related to innate immunity and bitter taste have been introgressed into numerous sheep breeds, putatively allowing them adapt to local parasite/disease pressures, and perhaps aiding in the utilization of local food resources.

3.6 Acknowledgments

We want to thank Levente Czeglédi for providing the MHun samples, Salvatore Naitana and Giovanni Leoni for providing MSar2 and MCor and Eleftherios Hadjisterkotis for MCyp. Antonello Carta and Tiziana Sechi sampled MSar3. The MIra samples were provided by the NextGen Consortium (FP7/2010-2014, grant agreement no 244356 - "NextGen").

3.7 Supplementary materials

3.7.1 Text S1. Supplementary materials and methods.

3.7.1.1 Sample information

The samples denoted MSar1 were collected in the mountainous Central-Eastern part of Sardinia belonging to the 'Ogliastra' and 'Supramonte' toponyms partially enclosed in the Gennargentu National park. This area has been historically sparsely settled, featuring harsh terrain, deep ravines and cold winters. Blood (preserved in EDTA) or tissue samples (preserved in ethanol) from poached animals found by the forestry department in this area were collected and stored at -20°C. Sampling for MSar2, MCor and MCyp was carried out between 2005 and 2012. Peripheral blood was collected from captured animals and stored in EDTA vacutainers. The MSar2 and MCor samples came from animals captured in the Montes area, a protected area within the Gennargentu national park, and the Cinto massif (North-Central Corsica) respectively. The MCyp samples were obtained from captured specimens belonging to mouflons living in the Paphos forest (Cyprus). Data from a third group of Sardinian mouflon (MSar3) subpopulation were available from the Sheep HapMap project (Kijas et al., 2012). Data from a Spanish mouflon population were also available from the Sheep HapMap project. Two blood samples of Iranian mouflons (*Ovis orientalis gmelinii*) from North-West Iran were also genotyped.

Ten Sarda sheep individuals (SAR) were sampled in the same area overlapping with MSar1 population. The Sarda sheep is an autochthonous breed comprising the vast majority of the Sardinian sheep population and is almost exclusively bred for milk production. Eight Hungarian Racka sheep (RAK) were also sampled, as Racka is an autochthonous sheep from Hungary (Ryder, 1981) used in the past to improve trophy size in mouflons used to repopulate the Caucasus (Tomiczek and Türcke, 2003).

Apart from SAR individuals all the domestic sheep data were available from the Sheep HapMap project (Kijas ar al., 2012). Among these are 20 individuals of Pecora Nera di Arbus (SAB), the only other autochthonous Sardinian domestic sheep. This breed presents a mixture of modern and ancestral traits (e.g., non-

shedding and black fleece (Piras et al., 2009)) and breeds from the first wave of domestication (Chessa et al., 2009) such as Soay from the island of St Kilda (UK) and Spael from Norway

3.7.1.2 PCAdmix parameters

When a linkage map is not available, the window size used by PCAdmix is defined by a fixed number of SNPs, otherwise different approaches are possible by using physical and genetic distances. Also, LD pruning is encouraged by the authors to "prevent high-LD blocks from having excessive influence on the inferred ancestry of a region, while retaining a dense, informative set of SNPs" (Brisbin et al., 2012). For our dataset a linkage map was not available, therefore a SNP based window size definition was used. The default value for window size in PCAdmix is set to 20 SNPs and refers to high-density datasets (1 SNP every ~5kbp). Given the lower density of the Ovine SNP chip (1 SNP every ~40kbp) we applied 5 as SNPs/window parameter as a lower number of SNPs/window is best suited for sparser datasets (Brisbin et al., 2012).

We used MSar1_p and MSar2 as representatives of the Sardinian mouflon (MSar3 was not used, as it showed strong signs of domestic sheep introgression, Figure 3-2). For the Corsican mouflon, as MCor and MHun showed highly similar signals of ancestry in all 24 clustering solutions produced by Admixture, we used the more extensively sampled MHun. The domestic breeds selected as reference for this analysis were SAB, a sympatric breed with both SAR and Sardinian mouflon populations, CAS and LAC, two European breeds from Spain and southern France respectively, areas where mouflon are still present: southern France was one of the first area where mouflon were reintroduced at the end of the 18th century. We also included ASM that has been widely used to improve several European breeds (Ciani et al., 2015; Kijas et al., 2012; Bon et al., 1991). These breeds were chosen as they live in the same areas where mouflon are present and in order to provide different ancestry assignment for each combination. These reference populations were used to infer local ancestry in the following sheep populations breeds: ALT, CHI, CHU, COM, CFT, RAK, SAR, SBF, SOA, SPW, VBN and the feral MSar1, MSar3 and MSpa populations (populations with less than eight individuals were excluded).

3.7.1.3 Consensus approach

The limitation given by applying a chromosome painting method as PCAdmix to such a sparse dataset resulted in high noise in the single results (i.e., in the analyses where only two populations were compared). Despite of the large number of markers in the SNP chip (~54k), the SNPs in this array are distanced from each other on average by ~60k base pairs leaving large gaps that may harbour additional and valuable information. Additionally, SNP array data are not in phase by definition, implying that the correct coupling of alleles needs to be inferred statistically in order to have haplotypes for multiple types of analyses like the ones implemented here. The accuracy of ancestry assignment depends on the correctness of the inferred haplotypes (Brisbin et al. 2012), and it has been shown that it is easier to achieve with denser SNP panels (Hayes et al. 2012). Therefore, as our SNP array is of medium density, we developed a consensus approach that uses several domestic breeds as reference populations and minimize the effect of results due to poor haplotype imputation.

For one or more reference populations (see main text) an assignment score (A-Score) was calculated for each window as defined by PCAdmix. The score was calculated averaging the PP assigned to the reference population of interest by root mean square, hence reducing the weight of low PP values in the final score. As using different reference populations resulted in the software defining windows with different sizes and genomic locations, the arrays of A-Scores from different analyses could not be compared directly because the windows would not perfectly overlap. A sliding window method was therefore implemented to normalize the window sizes among different sets of A-Scores by transferring the A-Score information into a window frame common to all of the reference populations. The A-Score information was transferred to a new set of windows of equal size the common for all the reference populations. The size of the windows was defined as half the size of the smallest window defined by PCAdmix among the data under analysis. Subsequently, we defined a window concordance score as the harmonic mean of the A-scores within the window. We repeated this analysis for each chromosome, enabling us to define genomic regions that exhibit concordance scores outside the 95th and 99th percentiles of the genome-wide

concordance scores and defined in the main text as Constantly Introgressed Windows of Interest (CIWIs).

3.7.2 Supplementary tables and figures

Table S 3-1 Number of Consistently Introgressed Windows of Interest (CIWI) and genes associated to those windows identified applying the consensus approach described in this work. Two genome-wide consensus score thresholds at the 99th and 95th percentile were used. For population abbreviations see Table 1-1.

Breed	***CIWI***	***genes***	***CIWI***	***genes***
	95th percentile		*99th percentile*	
mouflon				
MSar1	1784	641	347	167
MSar3	1740	488	344	57
MSpa	2609	443	501	125
domestic sheep				
ALT	1264	796	252	164
CHI	2678	993	497	183
CHU	5346	1077	1003	299
COM	5057	1722	1014	466
CTF	6170	1225	1199	390
RAK	1823	451	356	74
SAR	1499	769	291	232
SBF	5310	1168	1037	366
SOA	1794	926	359	94
SPW	1776	781	355	222
VBN	2543	1051	507	297

Table S 3-2 GO terms analysis results. The two sets of genes identified through the CIWI approach (top 1 and 5 percentile used as threshold) were screened for GO term enrichment. The population acronym, percentile threshold used, GO term identifier, description of the Go term, P-value and false discovery rate (FDR q-value), are shown. GO terms with FDR < 0.05 are displayed in bold. For population abbreviations see Table 1-1.

Population	*percentile threshold*	*GO term identifier.*	*Description*	*P-value*	*FDR q-value*
MSar1	*99*	-	*none*	-	-
	95	GO:0042157	*lipoprotein metabolic process*	5.36E-04	1.00E+00
		GO:0010885	*regulation of cholesterol storage*	6.10E-04	1.00E+00
		GO:2000177	*regulation of neural precursor cell proliferation*	7.59E-04	1.00E+00
		GO:2000179	*positive regulation of neural precursor cell proliferation*	8.89E-04	1.00E+00
MSar3	*99*	GO:0019935	cyclic-nucleotide-mediated signaling	1.08E-04	1.00E+00
		GO:0019933	cAMP-mediated signaling	2.82E-04	1.00E+00
		GO:0001654	eye development	7.64E-04	1.00E+00
		GO:0048598	embryonic morphogenesis	8.14E-04	1.00E+00
	95	GO:0051972	regulation of telomerase activity	4.49E-04	1.00E+00
		GO:0045216	cell-cell junction organization	6.03E-04	1.00E+00
MSpa	*99*	GO:0050819	negative regulation of coagulation	9.03E-05	1.00E+00
		GO:0050818	regulation of coagulation	1.29E-04	7.75E-01
		GO:0010543	regulation of platelet activation	3.39E-04	1.00E+00
	95	GO:2000009	negative regulation of protein localization to cell surface	6.65E-04	1.00E+00
SAR	*99*	GO:0018101	protein citrullination	9.39E-06	1.11E-01
		GO:0050912	detection of chemical stimulus involved in sensory perception of taste	2.02E-05	1.20E-01
		GO:0019240	citrulline biosynthetic process	7.98E-05	3.14E-01
		GO:0001580	detection of chemical stimulus involved in sensory perception of bitter taste	1.26E-04	3.73E-01
		GO:0036413	histone H3-R26 citrullination	1.79E-04	4.24E-01
		GO:0036414	histone citrullination	1.79E-04	3.53E-01
		GO:2001199	negative regulation of dendritic cell differentiation	1.79E-04	3.03E-01
		GO:0018195	peptidyl-arginine modification	2.66E-04	3.92E-01
		GO:0000052	citrulline metabolic process	2.66E-04	3.49E-01
		GO:0045909	positive regulation of vasodilation	4.77E-04	5.64E-01

		GO:0042398	cellular modified amino acid biosynthetic process	4.96E-04	5.33E-01
		GO:2001198	regulation of dendritic cell differentiation	5.33E-04	5.25E-01
		GO:0007028	cytoplasm organization	5.33E-04	4.85E-01
	95	**GO:0018101**	**protein citrullination**	**3.99E-07**	**4.72E-03**
		GO:0019240	**citrulline biosynthetic process**	**5.32E-07**	**3.15E-03**
		GO:0001580	**detection of chemical stimulus involved in sensory perception of bitter taste**	**6.81E-07**	**2.69E-03**
		GO:0050912	**detection of chemical stimulus involved in sensory perception of taste**	**8.97E-07**	**2.65E-03**
		GO:0000052	**citrulline metabolic process**	**7.66E-06**	**1.81E-02**
		GO:0042398	**cellular modified amino acid biosynthetic process**	**2.07E-05**	**4.09E-02**
		GO:0018195	peptidyl-arginine modification	1.41E-04	2.39E-01
		GO:0006997	nucleus organization	5.60E-04	8.28E-01
		GO:1901607	alpha-amino acid biosynthetic process	8.13E-04	1.00E+00
ALT	*99*		none		
	95	GO:0019240	citrulline biosynthetic process	6.46E-4	1E0
CHI	*99*	GO:0000956	nuclear-transcribed mRNA catabolic process	4.18E-4	1E0
		GO:0006402	mRNA catabolic process	7.06E-4	1E0
		GO:0019058	viral life cycle	8.92E-4	1E0
		GO:0060669	embryonic placenta morphogenesis	9.18E-4	1E0
	95	**GO:0018101**	**protein citrullination**	**2.18E-6**	**2.64E-2**
		GO:0019240	citrulline biosynthetic process	1.01E-4	6.13E-1
		GO:0071442	positive regulation of histone H3-K14 acetylation	4.03E-4	1E0
		GO:0018195	peptidyl-arginine modification	6.91E-4	1E0
		GO:0000052	citrulline metabolic process	6.91E-4	1E0
CHU	*99*	**GO:0019240**	**citrulline biosynthetic process**	**2.04E-9**	**2.46E-5**
		GO:0018101	**protein citrullination**	**3.72E-9**	**2.24E-5**
		GO:0000052	**citrulline metabolic process**	**3.2E-8**	**1.28E-4**
		GO:0018195	**peptidyl-arginine modification**	**1.55E-6**	**4.67E-3**
		GO:0042398	cellular modified amino acid biosynthetic process	2.41E-5	5.79E-2
		GO:1901607	alpha-amino acid biosynthetic process	1.61E-4	3.22E-1
		GO:0008652	cellular amino acid biosynthetic process	3.98E-4	6.85E-1

		GO:0036413	histone H3-R26 citrullination	4.32E-4	6.49E-1
		GO:0036414	histone citrullination	4.32E-4	5.77E-1
	95	**GO:0018101**	**protein citrullination**	**2.23E-6**	**2.69E-2**
		GO:0019240	**citrulline biosynthetic process**	**4.04E-6**	**2.44E-2**
		GO:0000052	citrulline metabolic process	5.48E-5	2.21E-1
		GO:0018195	peptidyl-arginine modification	7.04E-4	1E0
		GO:0018130	heterocycle biosynthetic process	9.78E-4	1E0
COM	*99*	**GO:0018101**	**protein citrullination**	**4.92E-08**	**5.84E-04**
		GO:0019240	**citrulline biosynthetic process**	**2.53E-06**	**1.50E-02**
		GO:0018195	**peptidyl-arginine modification**	**1.91E-05**	**7.56E-02**
		GO:0000052	citrulline metabolic process	1.91E-05	5.67E-02
		GO:0006824	cobalt ion transport	3.95E-04	9.37E-01
		GO:0015889	cobalamin transport	3.95E-04	7.81E-01
	95	GO:0018101	protein citrullination	3.28E-05	3.93E-01
		GO:0021756	striatum development	5.51E-04	1.00E+00
		GO:0035637	multicellular organismal signaling	6.97E-04	1.00E+00
VBN	*99*	**GO:0018101**	**protein citrullination**	**1.80E-09**	**2.13E-05**
		GO:0019240	**citrulline biosynthetic process**	**9.63E-08**	**5.72E-04**
		GO:0018195	**peptidyl-arginine modification**	**7.60E-07**	**3.01E-03**
		GO:0000052	**citrulline metabolic process**	**7.60E-07**	**2.26E-03**
		GO:0042398	cellular modified amino acid biosynthetic process	1.50E-04	3.57E-01
		GO:0036413	histone H3-R26 citrullination	3.23E-04	6.40E-01
		GO:0036414	histone citrullination	3.23E-04	5.48E-01
		GO:0060448	dichotomous subdivision of terminal units involved in lung branching	3.23E-04	4.80E-01
		GO:1901607	alpha-amino acid biosynthetic process	5.57E-04	7.35E-01
		GO:0008652	cellular amino acid biosynthetic process	8.91E-04	1.00E+00
		GO:0044406	adhesion of symbiont to host	9.59E-04	1.00E+00
	95	**GO:0018101**	**protein citrullination**	**1.96E-06**	**2.34E-02**
		GO:0000052	citrulline metabolic process	8.91E-05	5.31E-01
		GO:0019240	citrulline biosynthetic process	9.12E-05	3.63E-01
		GO:0071322	cellular response to carbohydrate stimulus	2.45E-04	7.30E-01
		GO:0018195	peptidyl-arginine modification	6.26E-04	1.00E+00

RAK	*99*	GO:1902224	ketone body metabolic process	3.27E-04	1.00E+00
		GO:0046950	cellular ketone body metabolic process	3.27E-04	1.00E+00
		GO:2001022	positive regulation of response to DNA damage stimulus	5.89E-04	1.00E+00
	95	**GO:0018101**	**protein citrullination**	**4.08E-08**	**4.85E-04**
		GO:0019240	**citrulline biosynthetic process**	**2.10E-06**	**1.25E-02**
		GO:0018195	peptidyl-arginine modification	1.60E-05	6.32E-02
		GO:0000052	**citrulline metabolic process**	**1.60E-05**	**4.74E-02**
		GO:0044260	cellular macromolecule metabolic process	2.50E-04	5.95E-01
		GO:0031929	TOR signaling	3.38E-04	6.69E-01
		GO:0044237	cellular metabolic process	3.77E-04	6.40E-01
		GO:0043170	macromolecule metabolic process	5.26E-04	7.81E-01
		GO:0006955	immune response	5.51E-04	7.28E-01
		GO:0007220	Notch receptor processing	6.94E-04	8.25E-01
		GO:0010468	regulation of gene expression	8.50E-04	9.19E-01
CFT	*99*	**GO:0018101**	**protein citrullination**	**1.64E-08**	**1.95E-04**
		GO:0019240	**citrulline biosynthetic process**	**8.57E-07**	**5.10E-03**
		GO:0018195	**peptidyl-arginine modification**	**6.60E-06**	**2.62E-02**
		GO:0000052	**citrulline metabolic process**	**6.60E-06**	**1.96E-02**
		GO:0045109	intermediate filament organization	1.10E-04	2.61E-01
		GO:0071680	response to indole-3-methanol	2.07E-04	4.10E-01
		GO:0071681	cellular response to indole-3-methanol	2.07E-04	3.52E-01
		GO:0006266	DNA ligation	2.07E-04	3.08E-01
		GO:0045103	intermediate filament-based process	3.65E-04	4.82E-01
		GO:0045104	intermediate filament cytoskeleton organization	3.65E-04	4.34E-01
		GO:0086073	bundle of His cell-Purkinje myocyte adhesion involved in cell communication	4.05E-04	4.38E-01
		GO:0014070	response to organic cyclic compound	5.77E-04	5.72E-01
		GO:0097305	response to alcohol	6.08E-04	5.57E-01
		GO:0072364	regulation of cellular ketone metabolic process by regulation of transcription from RNA polymerase II promoter	6.95E-04	5.90E-01

		GO:0086042	cardiac muscle cell-cardiac muscle cell adhesion	6.95E-04	5.51E-01
		GO:0051987	positive regulation of attachment of spindle microtubules to kinetochore	7.78E-04	5.78E-01
		GO:0036413	histone H3-R26 citrullination	7.78E-04	5.44E-01
		GO:0036414	histone citrullination	7.78E-04	5.14E-01
		GO:2001199	negative regulation of dendritic cell differentiation	7.78E-04	4.87E-01
		GO:1901624	negative regulation of lymphocyte chemotaxis	7.78E-04	4.63E-01
	95	GO:0018101	protein citrullination	5.85E-06	6.99E-02
		GO:0070192	chromosome organization involved in meiosis	6.30E-05	3.77E-01
		GO:0018195	peptidyl-arginine modification	1.63E-04	6.48E-01
		GO:0019240	citrulline biosynthetic process	2.60E-04	7.76E-01
		GO:0090280	positive regulation of calcium ion import	2.60E-04	6.21E-01
		GO:0022038	corpus callosum development	3.03E-04	6.03E-01
		GO:0090279	regulation of calcium ion import	4.06E-04	6.93E-01
		GO:0006307	DNA dealkylation involved in DNA repair	8.44E-04	1.00E+00
		GO:0051216	cartilage development	9.32E-04	1.00E+00
SOA	*99*	GO:0010165	response to X-ray	3.62E-04	1.00E+00
		GO:0007185	transmembrane receptor protein tyrosine phosphatase signaling pathway	9.17E-04	1.00E+00
	95	**GO:0001580**	**detection of chemical stimulus involved in sensory perception of bitter taste**	**1.79E-06**	**2.14E-02**
		GO:0050912	detection of chemical stimulus involved in sensory perception of taste	4.58E-05	2.74E-01
		GO:0060193	positive regulation of lipase activity	2.32E-04	9.24E-01
		GO:2000269	regulation of fibroblast apoptotic process	2.82E-04	8.42E-01
		GO:0010518	positive regulation of phospholipase activity	4.61E-04	1.00E+00
		GO:1902106	negative regulation of leukocyte differentiation	7.09E-04	1.00E+00
		GO:1902751	positive regulation of cell cycle G2/M phase transition	7.89E-04	1.00E+00
		GO:0010971	positive regulation of G2/M transition of mitotic cell cycle	7.89E-04	1.00E+00
		GO:0060191	regulation of lipase activity	8.82E-04	1.00E+00

SPW	*99*	GO:0070997	neuron death	8.04E-04	1.00E+00
	95	GO:0036149	phosphatidylinositol acyl-chain remodeling	4.18E-05	4.99E-01
		GO:1902031	regulation of NADP metabolic process	1.45E-04	8.65E-01
		GO:0036152	phosphatidylethanolamine acyl-chain remodeling	5.33E-04	1.00E+00
		GO:0030901	midbrain development	7.65E-04	1.00E+00
		GO:0072329	monocarboxylic acid catabolic process	7.93E-04	1.00E+00
SBF	*99*	**GO:0018101**	**protein citrullination**	**2.38E-06**	**2.82E-02**
		GO:0000052	**citrulline metabolic process**	**5.14E-06**	**3.05E-02**
		GO:0019240	citrulline biosynthetic process	3.12E-05	1.24E-01
		GO:0018195	peptidyl-arginine modification	1.38E-04	4.10E-01
		GO:0017038	protein import	2.70E-04	6.41E-01
		GO:0034504	protein localization to nucleus	4.11E-04	8.14E-01
		GO:0003149	membranous septum morphogenesis	7.02E-04	1.00E+00
	95	GO:0018101	protein citrullination	4.20E-06	5.02E-02
		GO:0000052	citrulline metabolic process	1.12E-04	6.70E-01
		GO:0019240	citrulline biosynthetic process	1.90E-04	7.54E-01
		GO:0046599	regulation of centriole replication	7.39E-04	1.00E+00

Table S 3-3 The genome-wide ancestry assignment inferred by PCAdmix expressed as percentage. In each of the four reference population combinations comprising a representative of the Sardinian mouflon (MSar2), of the Corsican mouflon (MHun) and domestic sheep (CAS, ASM, LAC and SAB_p) the percentage of genome assigned with posterior probability >0.95 is displayed for each population analysed. For each section the cumulative percentage of regions assigned below the same threshold are also displayed. For population abbreviations see Table 1-1.

	Reference populations															
	MSar2	***MHun***	***CAS***	***<95***	***MSar2***	***MHun***	***ASM***	***<95***	***MSar2***	***MHun***	***LAC***	***<95***	***MSar2***	***MHun***	***SAB_p***	***<95***
Mouflon																
MSar1	**26.6**	**18.0**	12.7	42.7	**25.7**	**18.6**	11.3	44.4	**26.4**	**19.7**	10.7	43.2	**26.3**	**25.1**	8.5	40.0
MSar3	***15.9***	***10.1***	*33.3*	*40.7*	***16.1***	***10.2***	*32.4*	*41.2*	***16.8***	***10.8***	*31.1*	*41.3*	***17.2***	***14.8***	*25.5*	*42.5*
MSpa	**19.0**	**43.6**	4.2	33.3	**18.2**	**44.8**	4.2	32.9	**16.0**	**49.2**	3.6	31.2	**14.1**	**55.7**	3.9	26.3
Sheep																
ALT	0.9	0.8	**90.1**	8.2	1.0	1.1	**88.9**	9.0	1.4	1.8	**84.7**	12.1	2.8	6.6	**65.5**	25.2
CHI	0.6	0.4	**93.9**	5.0	0.8	0.5	**92.9**	5.9	1.0	1.4	**89.1**	8.5	3.1	6.4	**67.4**	23.1
CHU	0.9	1.0	**91.5**	6.7	1.1	1.0	**88.5**	9.4	1.6	1.8	**84.7**	11.9	3.1	7.3	**63.5**	26.1
COM	0.9	0.9	**90.5**	7.7	1.6	1.0	**88.3**	9.2	1.4	1.7	**85.0**	11.9	3.0	6.5	**65.5**	25.0
CTF	0.6	0.4	**94.7**	4.4	0.6	0.6	**93.4**	5.4	0.7	1.1	**91.5**	6.7	2.1	7.0	**71.2**	19.7
RAK	0.8	0.6	**93.3**	5.4	1.3	0.8	**91.2**	6.7	1.1	1.3	**90.2**	7.5	2.2	6.5	**70.2**	21.1
SAR	1.1	0.7	**91.2**	7.0	1.0	0.9	**90.2**	8.0	1.4	1.5	**87.0**	10.1	2.4	5.4	**73.4**	18.7
SBF	1.1	0.9	**89.5**	8.5	1.1	1.3	**88.1**	9.4	1.3	2.4	**83.9**	12.5	2.8	8.7	**61.9**	26.6
SOA	1.8	1.8	**88.0**	8.5	2.1	2.5	**84.6**	10.9	2.3	3.1	**82.3**	12.4	3.2	9.1	**65.0**	22.7
SPW	1.0	1.0	**89.4**	8.6	1.3	1.2	**87.8**	9.7	1.2	2.1	**84.8**	12.0	3.2	8.6	**62.1**	26.0
VBN	0.8	0.7	**92.0**	6.4	1.1	0.9	**90.9**	7.1	1.2	1.5	**88.4**	8.9	2.8	7.4	**66.5**	23.4

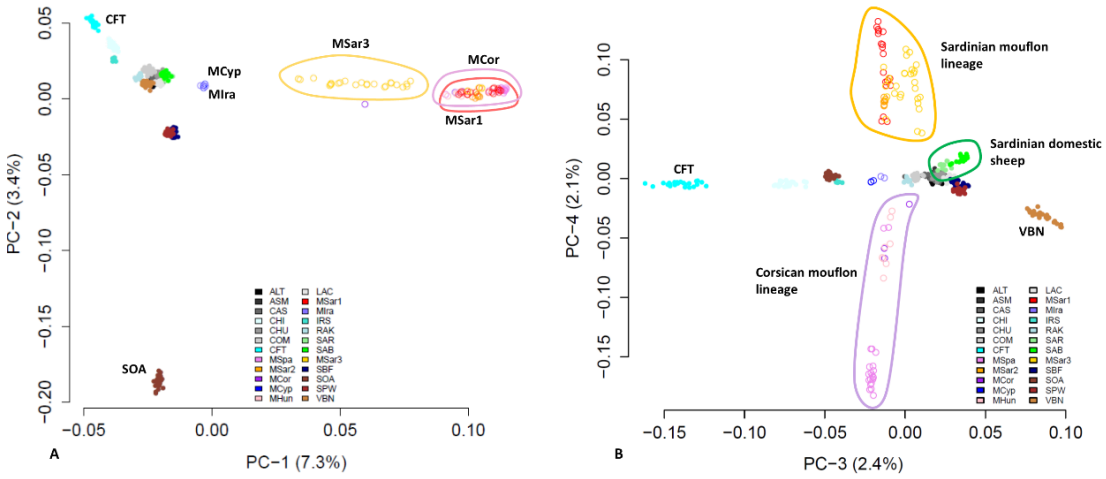


Figure S 3-1 Principal component analysis of the dataset. The first four components are displayed. Mouflon individuals are identified by empty circles whereas domestic sheep individuals are identified by solid circles. For population abbreviations see Table 1-1.

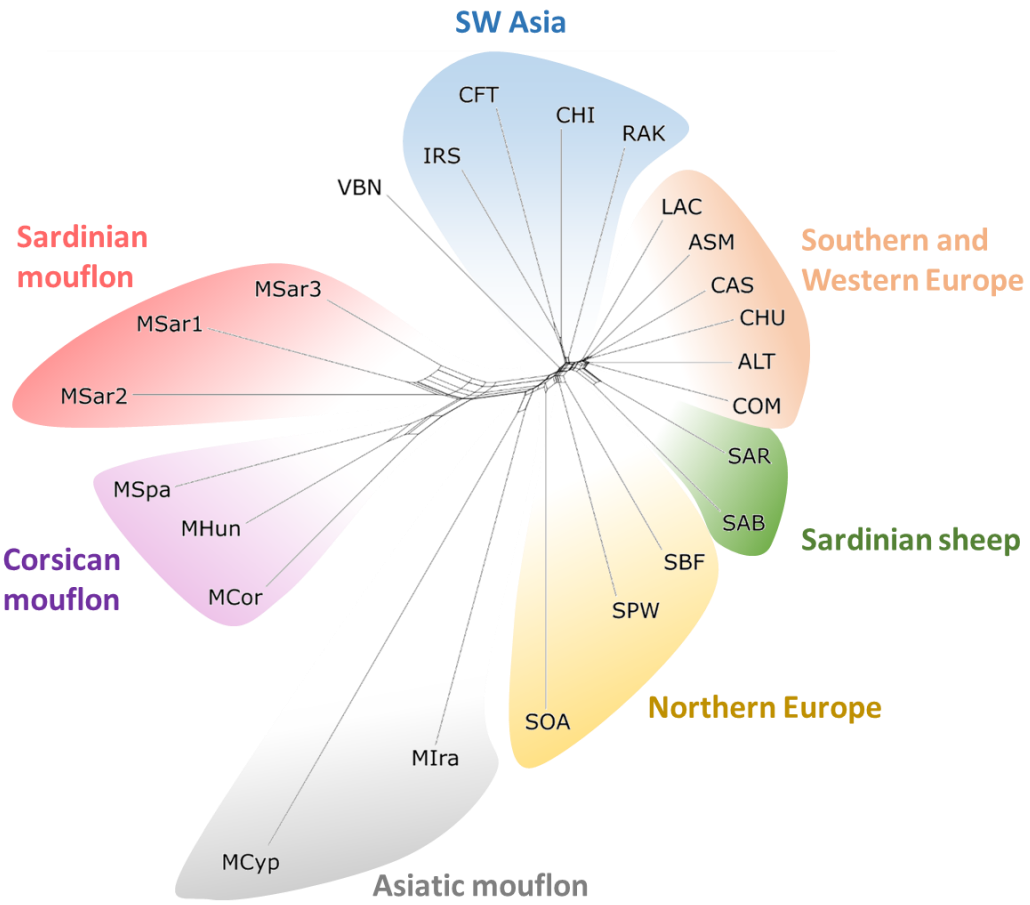


Figure S 3-2 Neighbour-Net from Splitstree4 (Huson and Bryant, 2006), based on pairwise Reynolds' distances. The coloured shades indicate the geographical origin of the main clusters identified. For population abbreviations see Table 1-1.

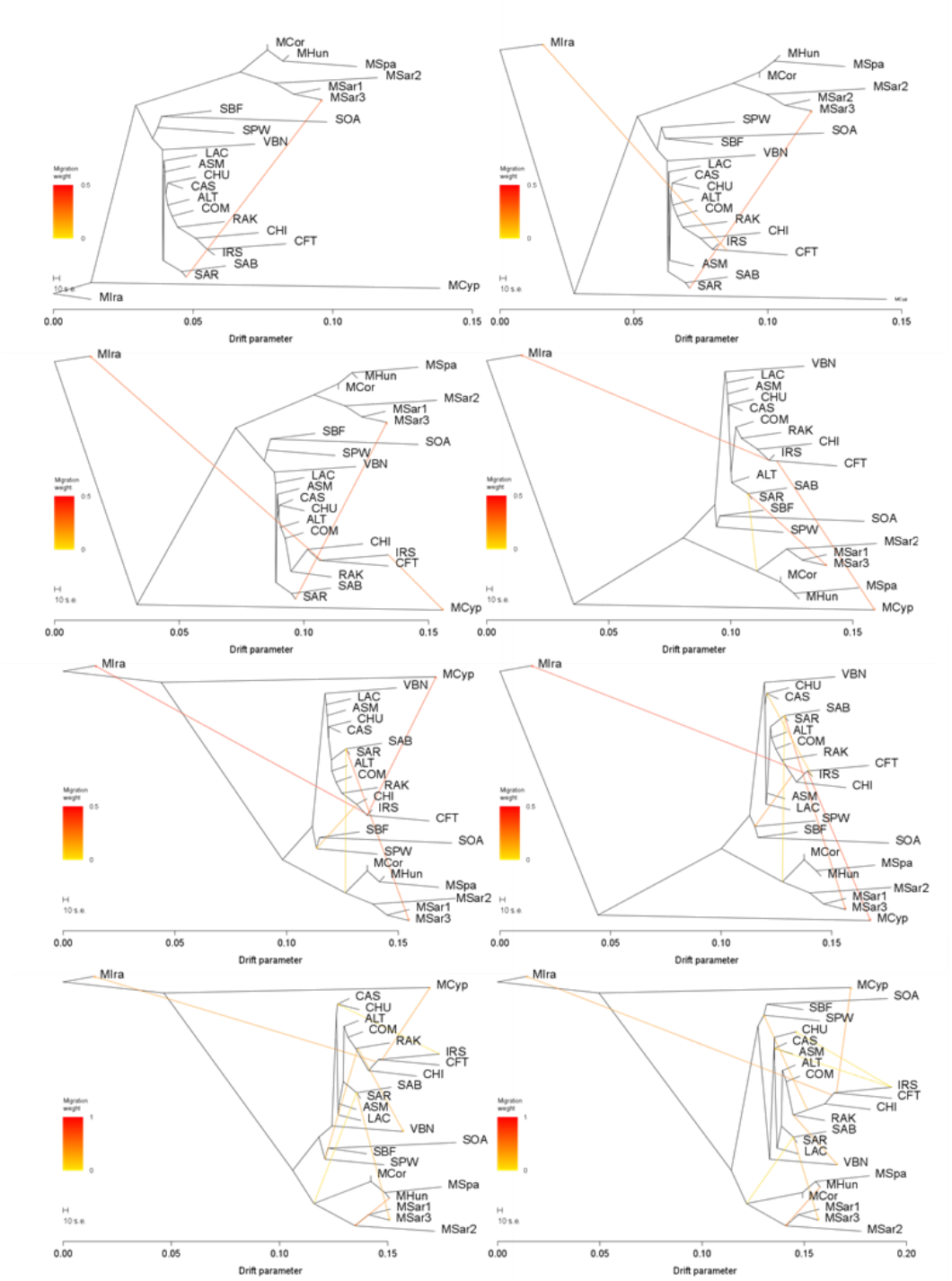
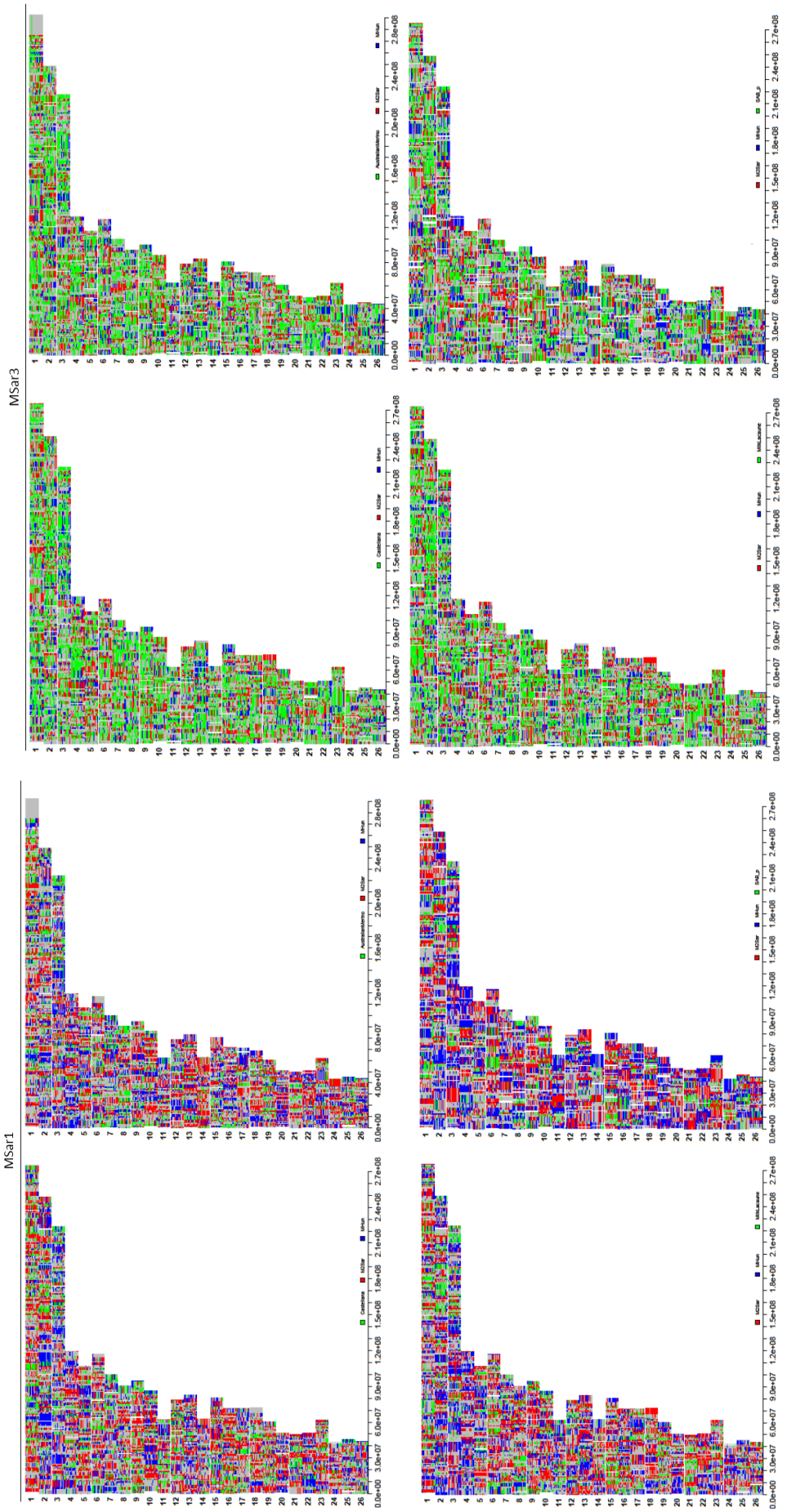
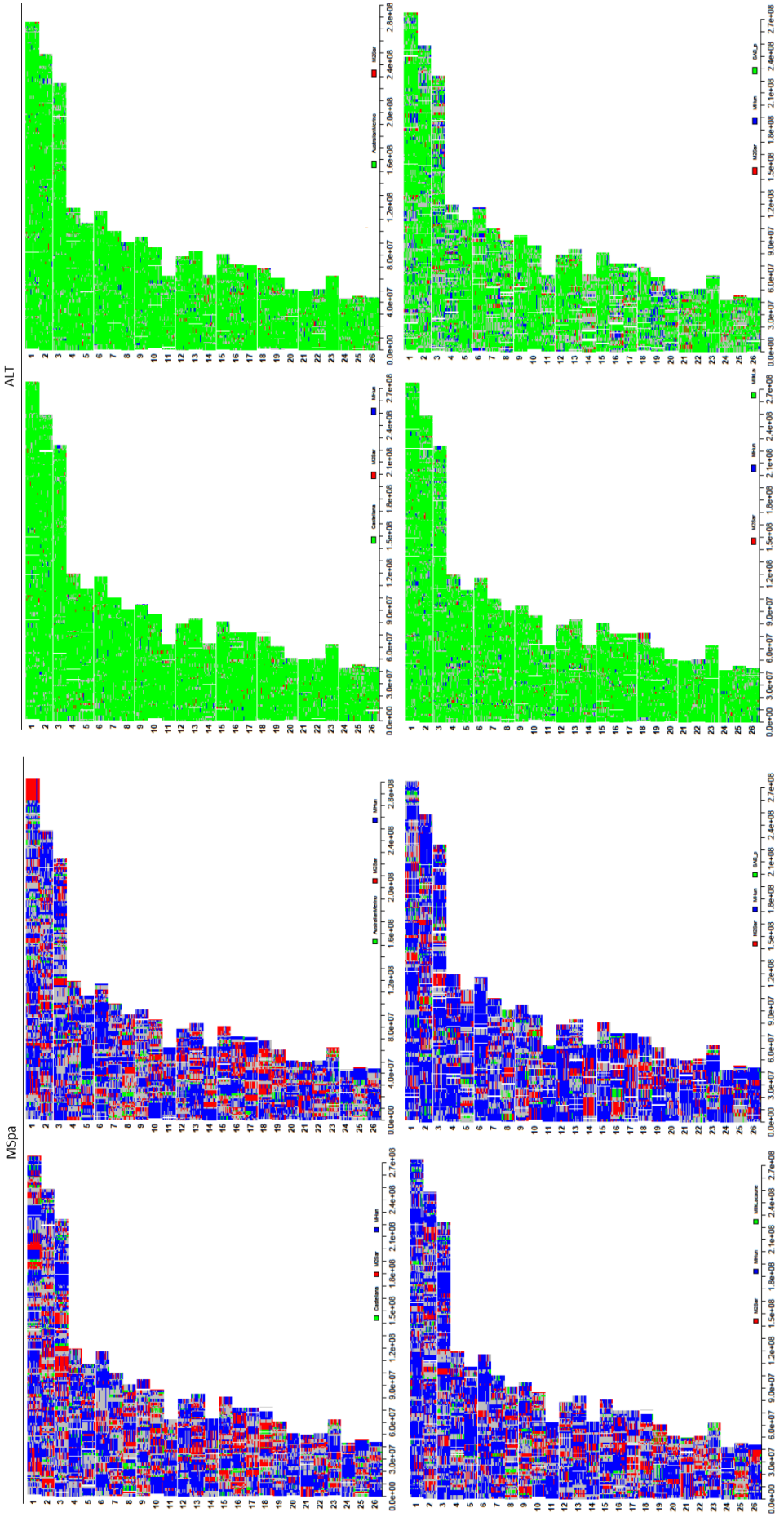
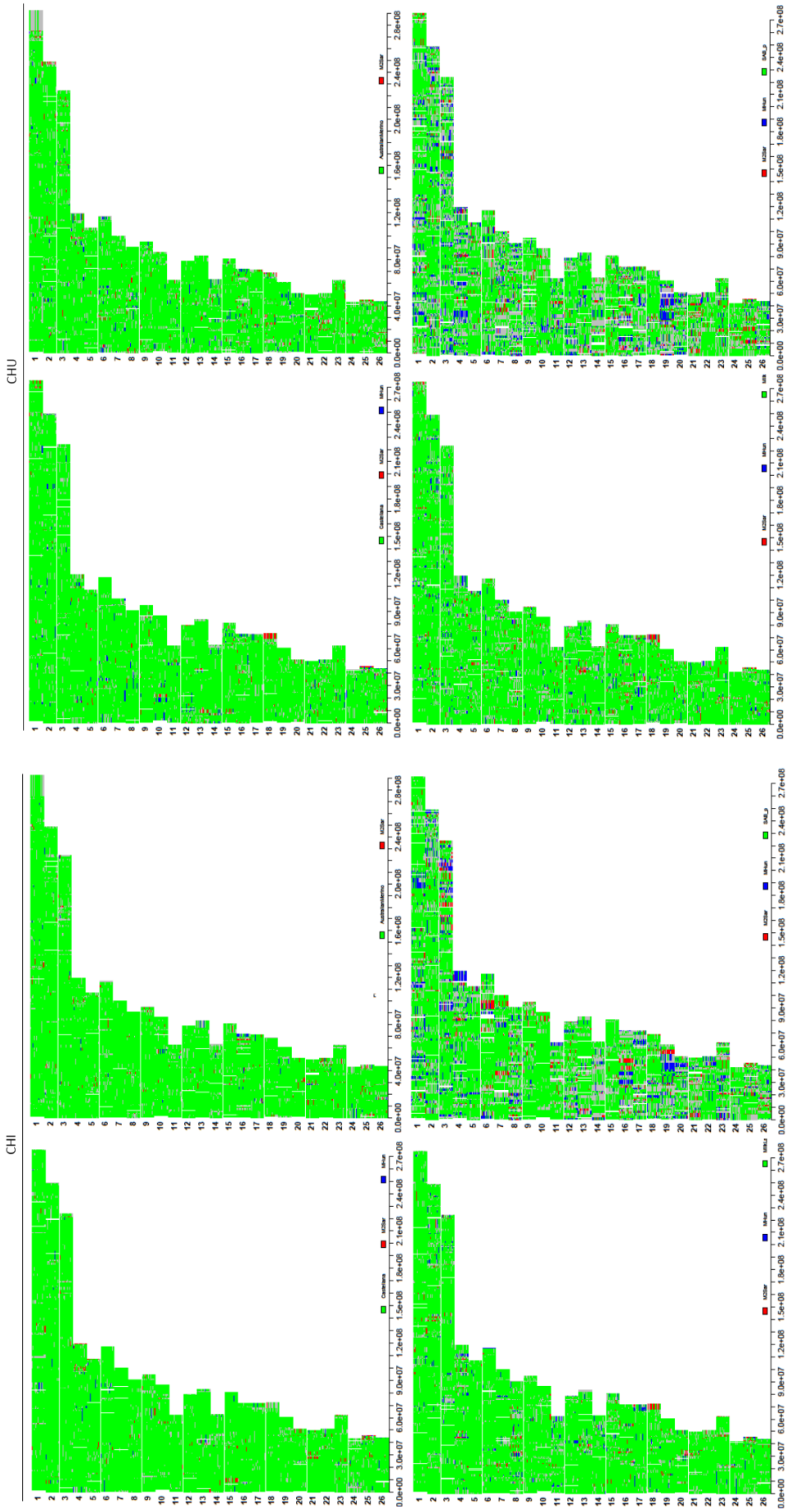


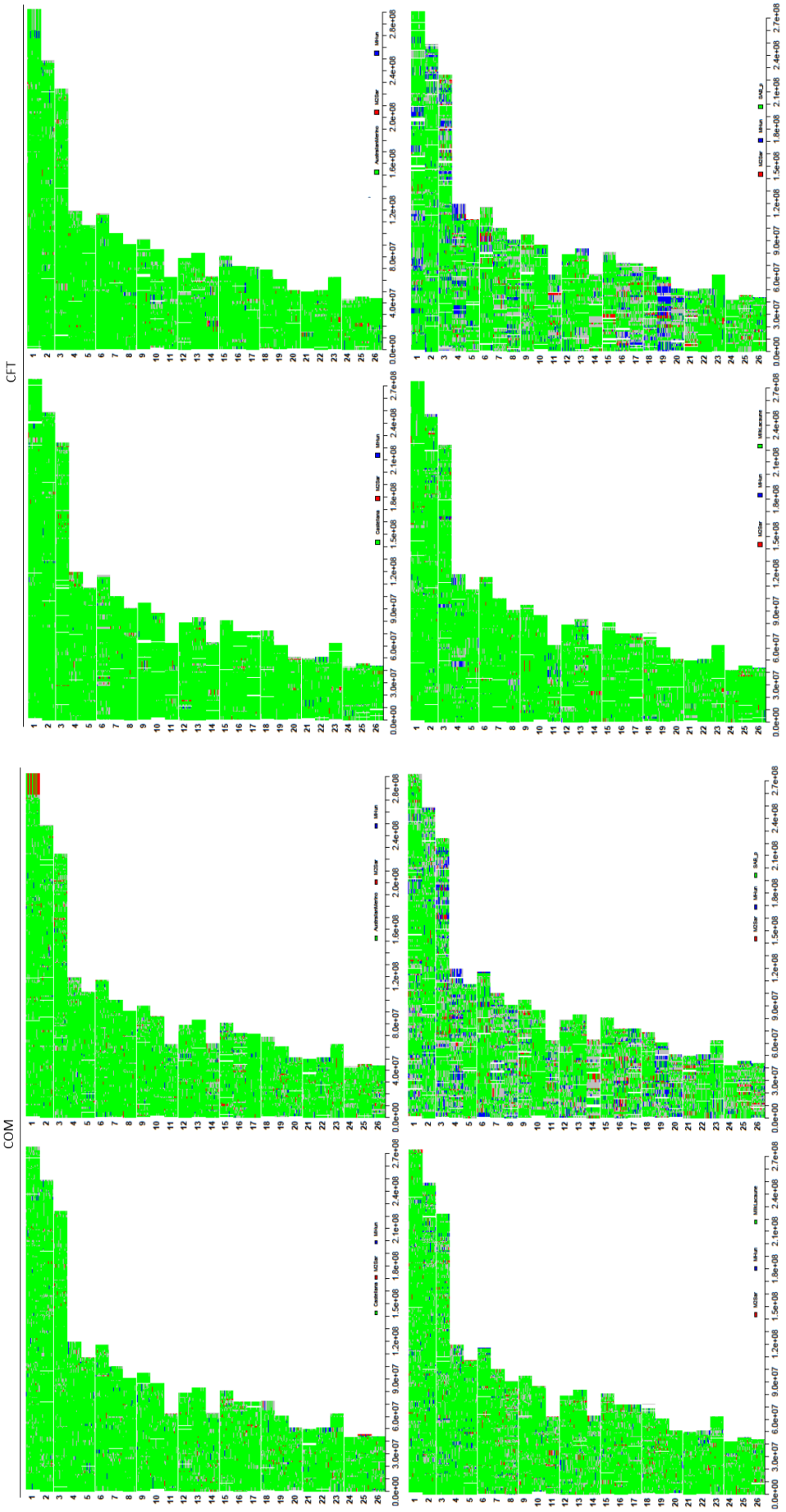
Figure S 3-3 . Phylogenetic network of the inferred relationships between mouflon and domestic sheep inferred by Treemix. Eight migration edges were computed.

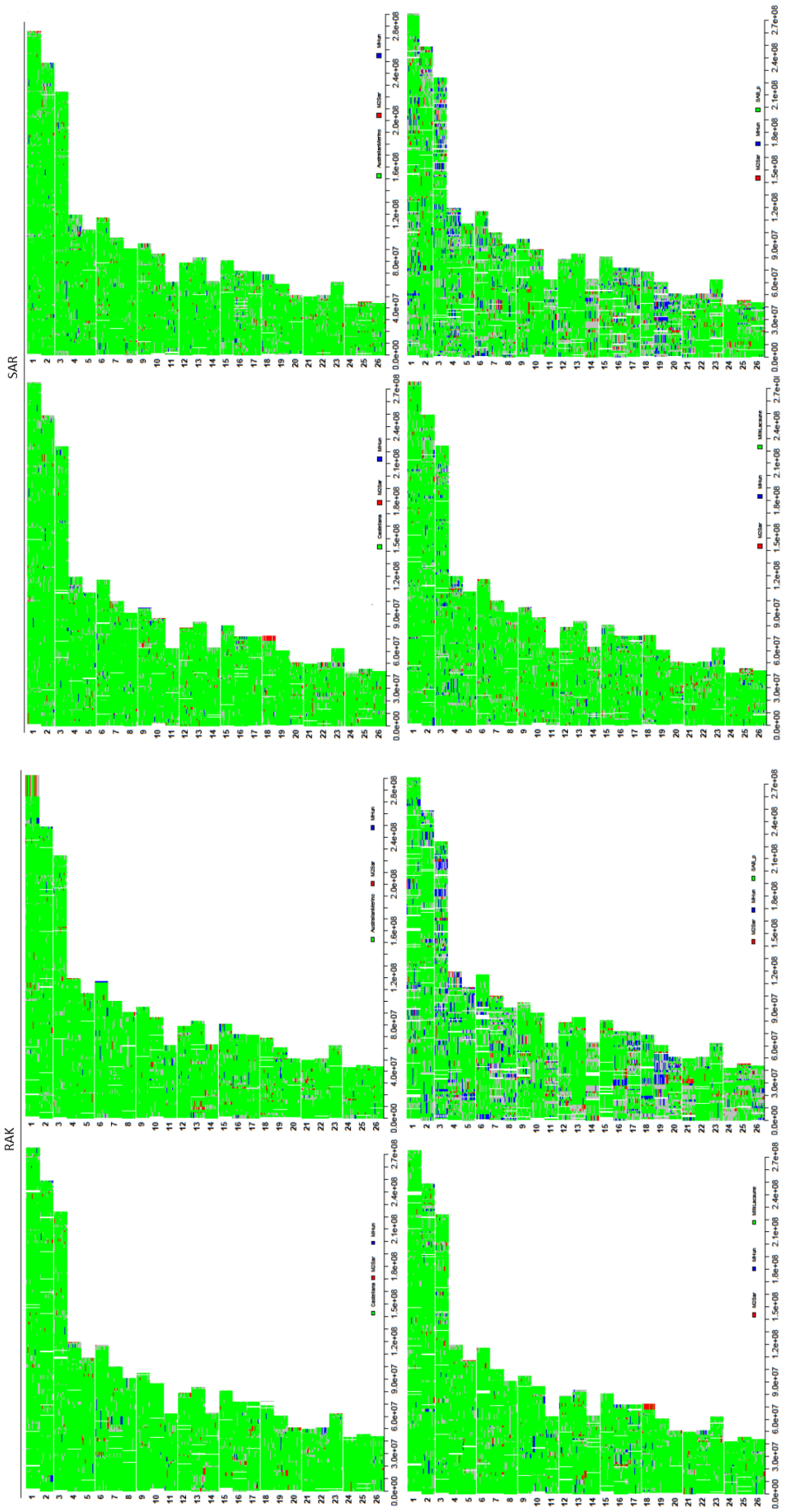
Figure S 3-4 (from page 88 to 94) Graphical representation of the local ancestry analysis results. Each group of four graphs represents the results obtained for a single population (the population abbreviation is displayed on top-of each group). M2Sar (red) and MHun (blue) were always used as references for the Sardinian and Corsican mouflon lineage, respectively. For each population analysed four domestic sheep (green) reference populations were used: CAS (top-left), ASM (top-right), LAC (bottom-left) and SAB_p (bottom-right). Genomic regions assigned by PCAdmix with posterior probability <0.95 are displayed in grey. The horizontal axis represents the chromosome position in bp. The numbers in the vertical axis identifies the chromosome. Each line within a chromosome, represents a haploid individual.

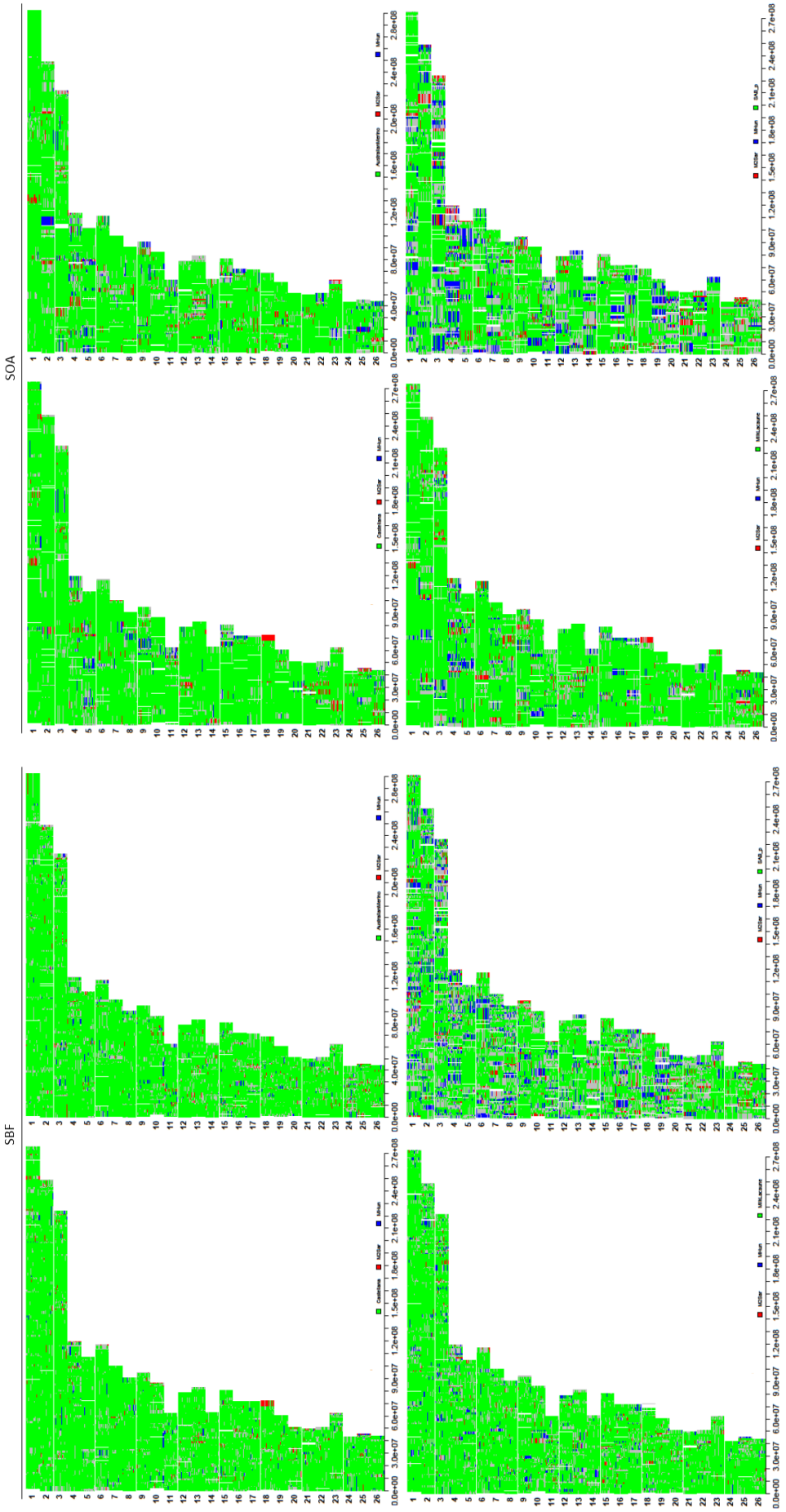

MSar3
MSar1


ALT
MSpa


CHU
CHI


CFT
COM


SAR
RAK


SOA
SBF

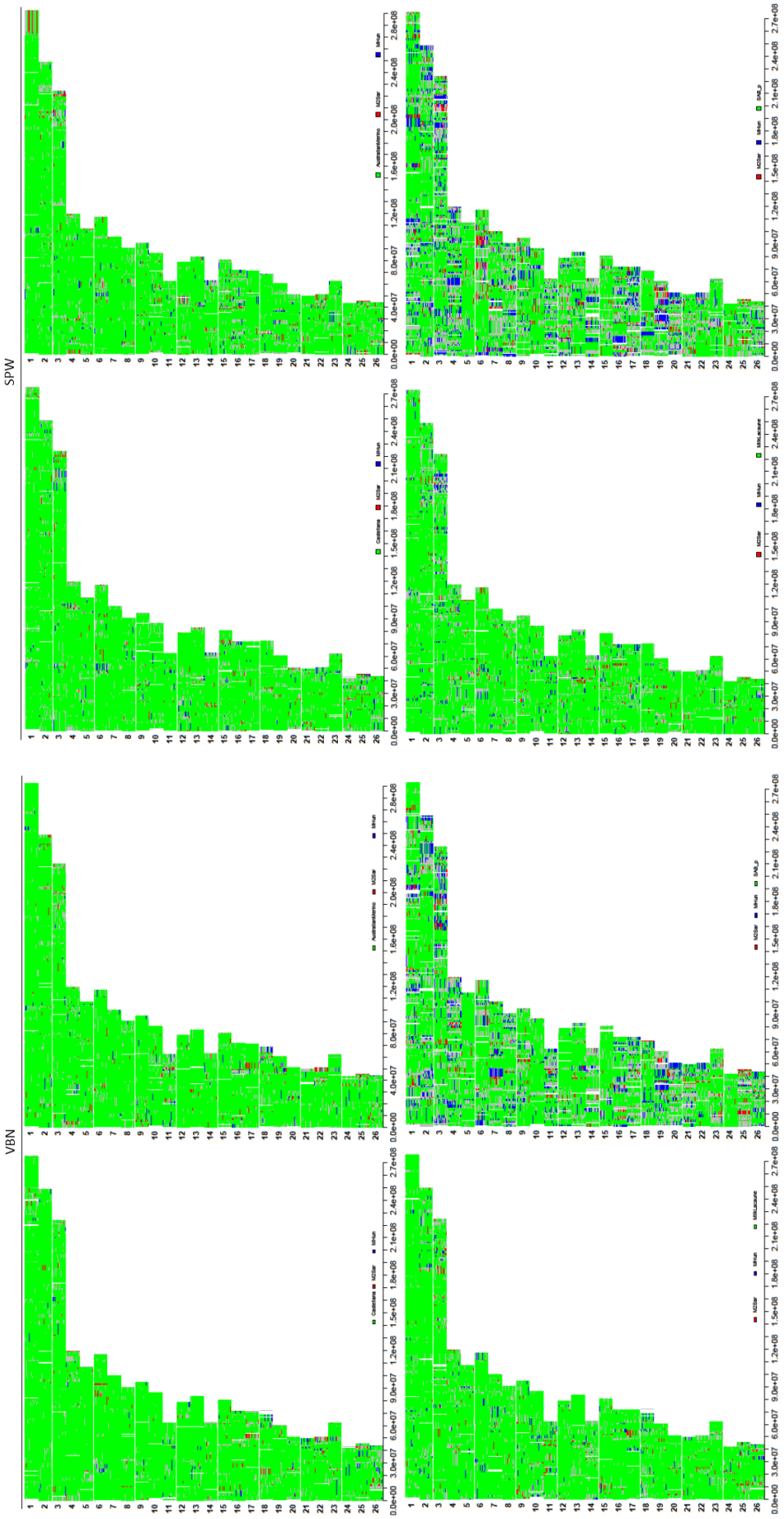
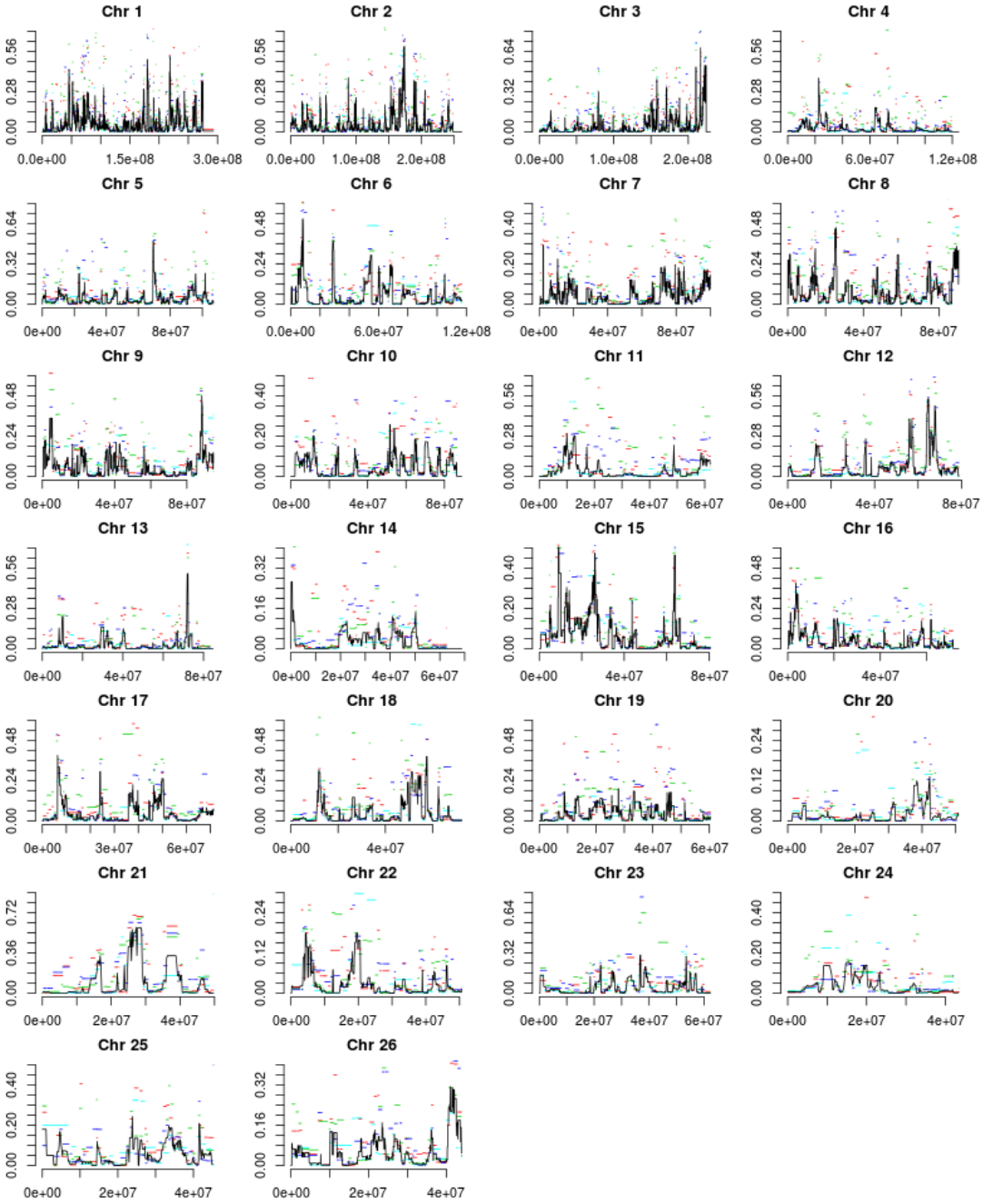
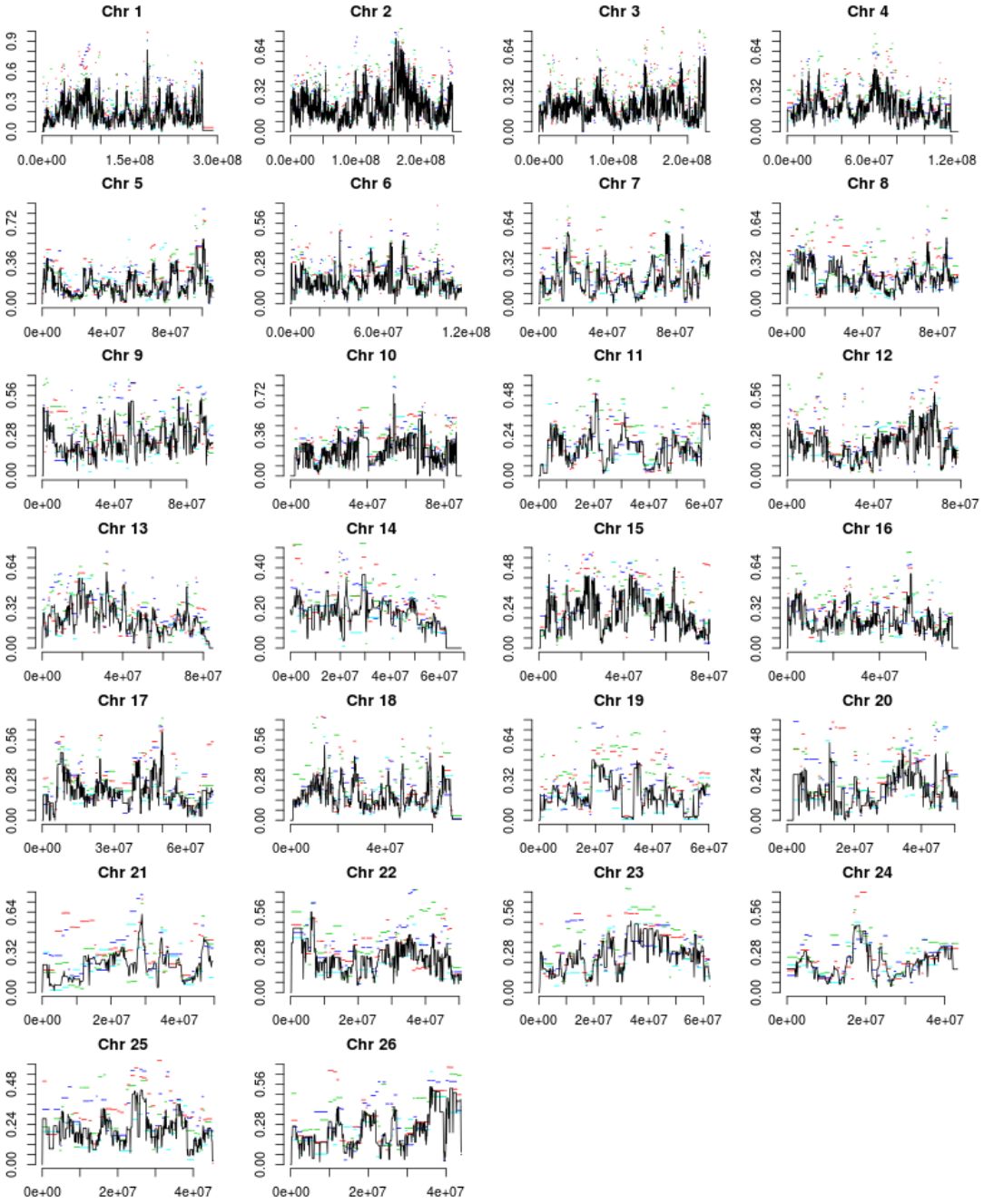

SPW
VBN

Figure S 3-5 (from page 95 to 108) Consistently Introgressed Windows of Interest (CIWI). In each of the following figures the results obtained for a single population are displayed (abbreviation in top-left corner of each page). In each page the CIWI obtained for each of the 26 autosomes (Chr) are displayed. The horizontal axis represents the chromosome position expressed in bp. The vertical axis represent the CIWI score. Within each graph, the segments represent the A-scores obtained using the four reference combinations (each reference set is identified by a different colour). The CIWI score is represented by the black line. For mouflon populations the CIWIs represent concordance of sheep ancestry, whereas for sheep populations they indicate mouflon ancestry.

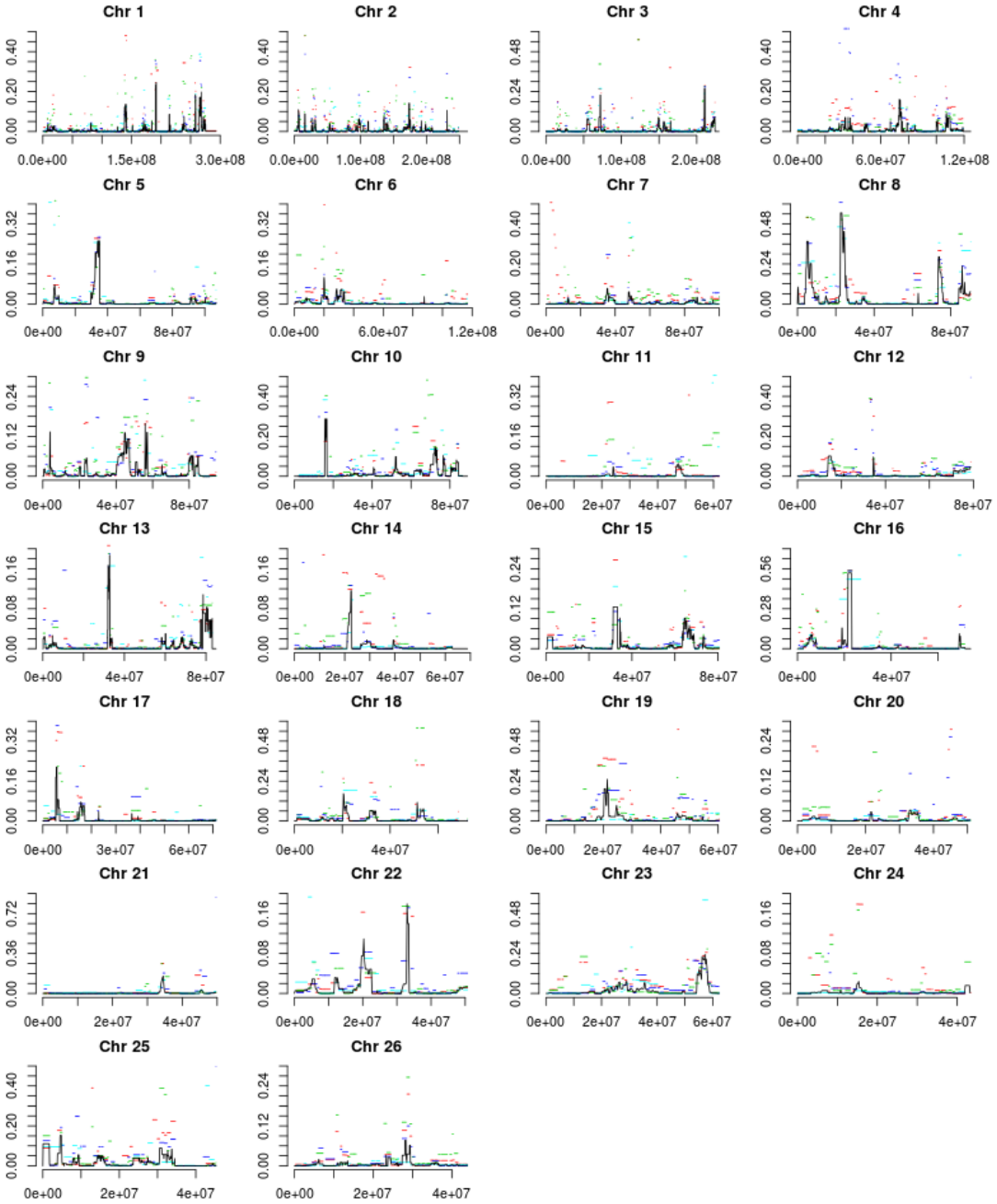
MSar1



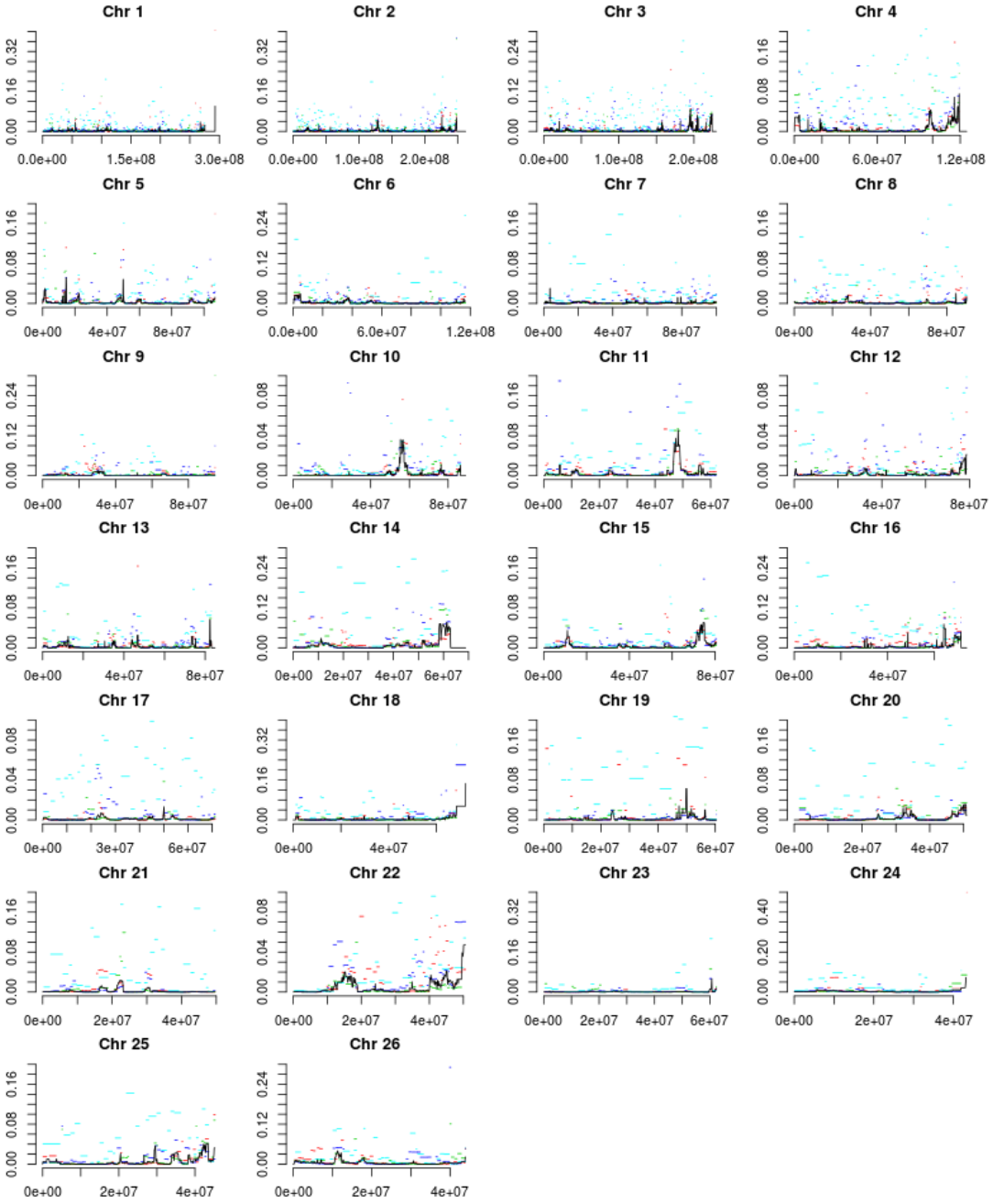
MSar3


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

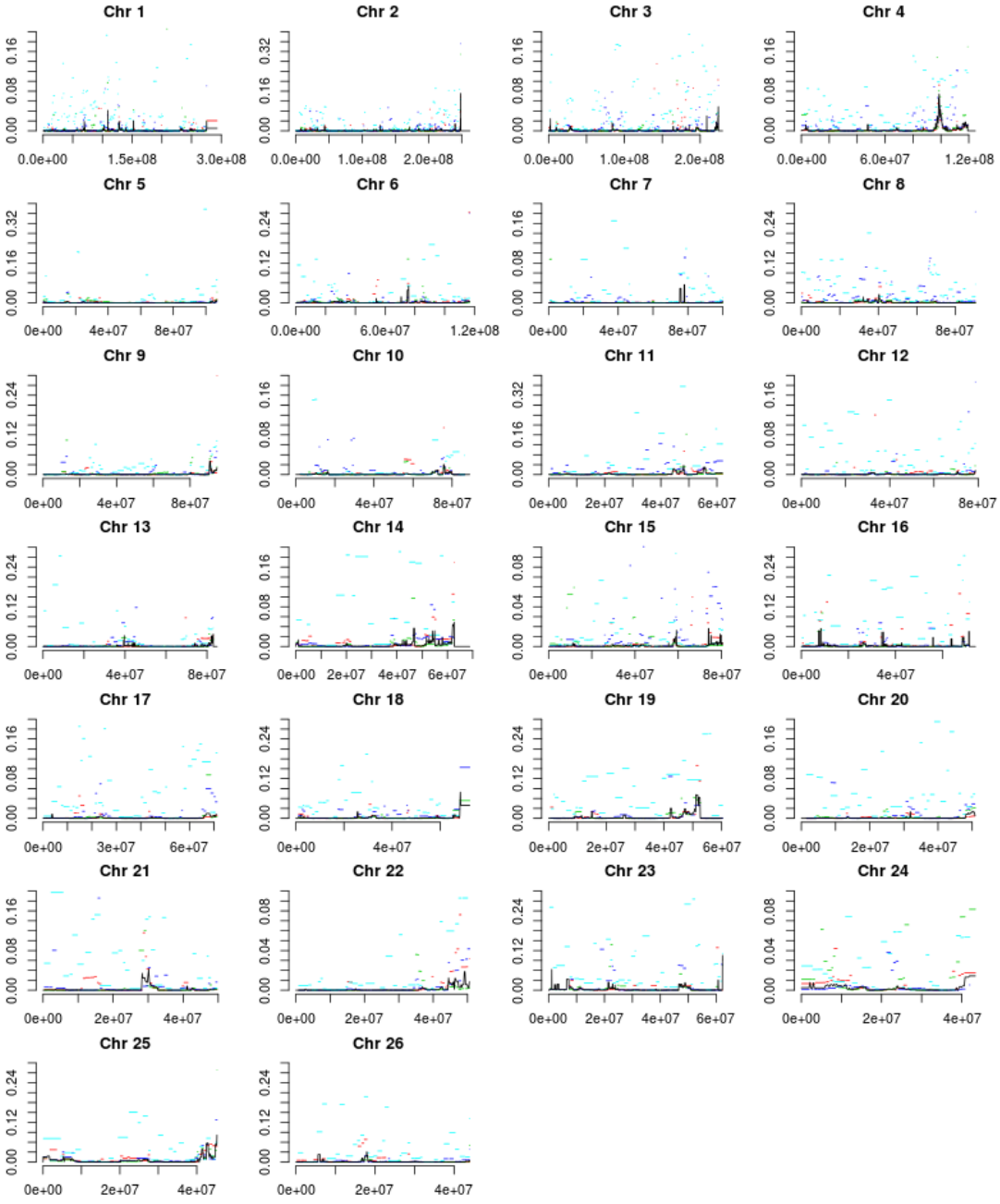
MSpa


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

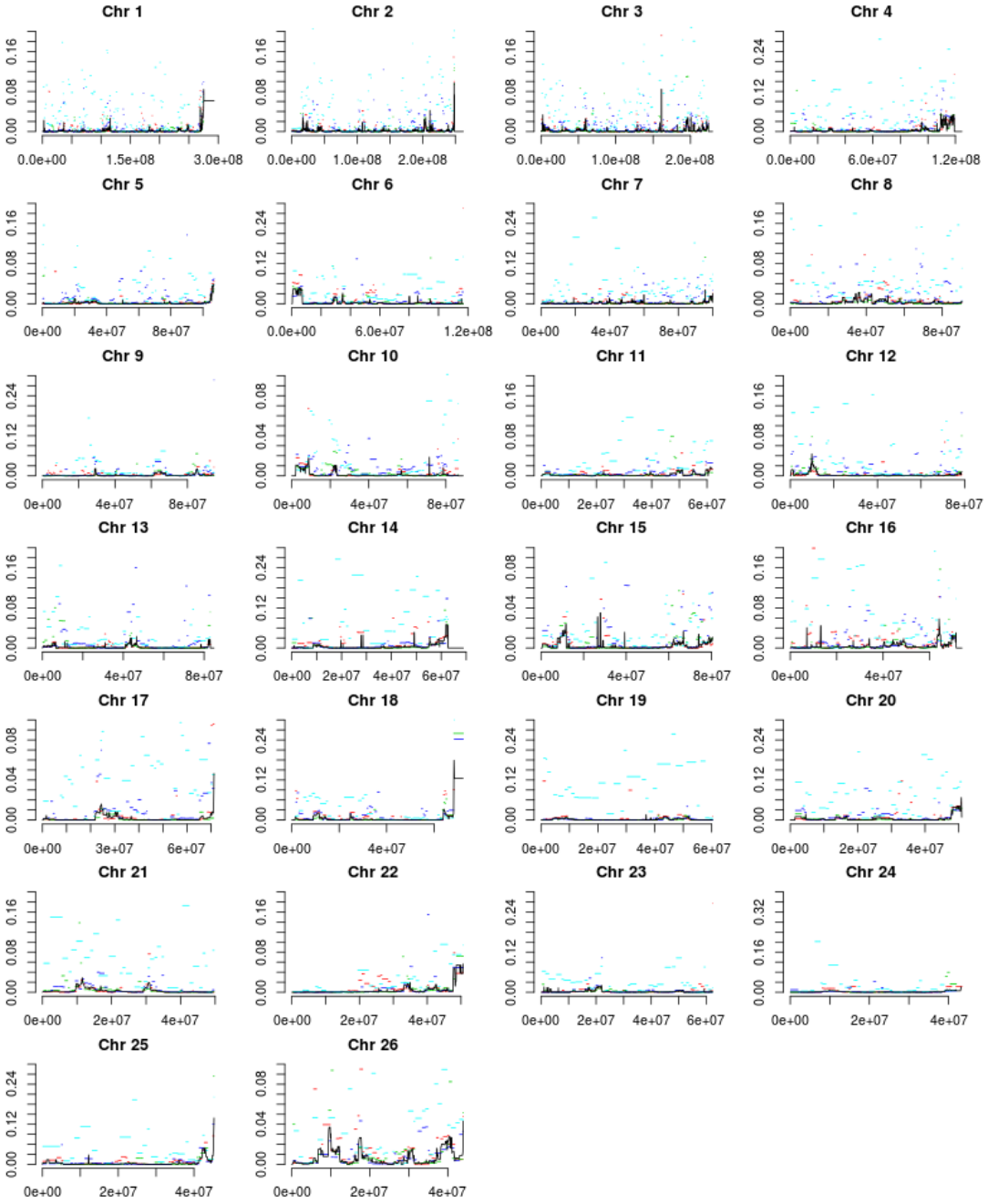
ALT


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

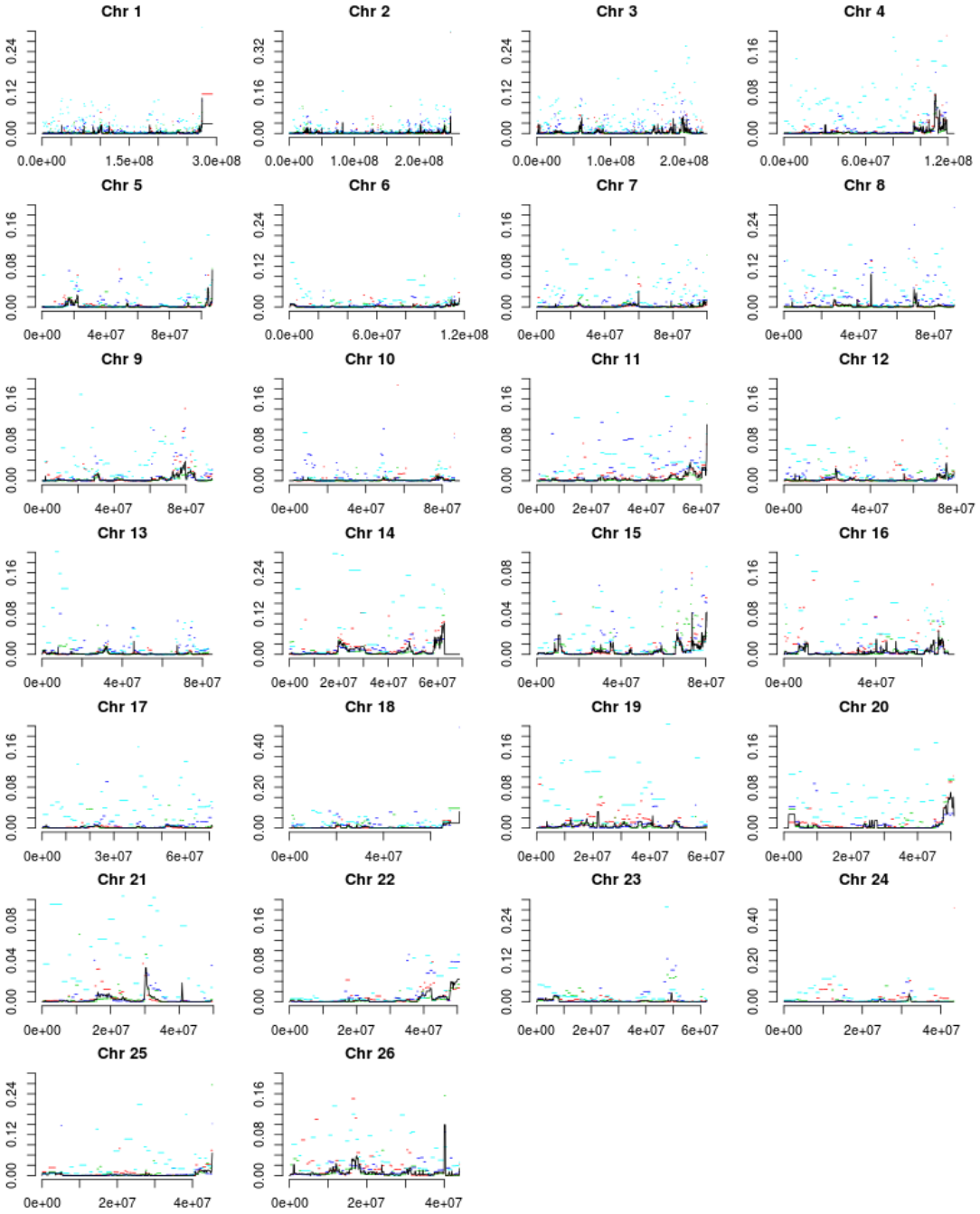
CHI


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

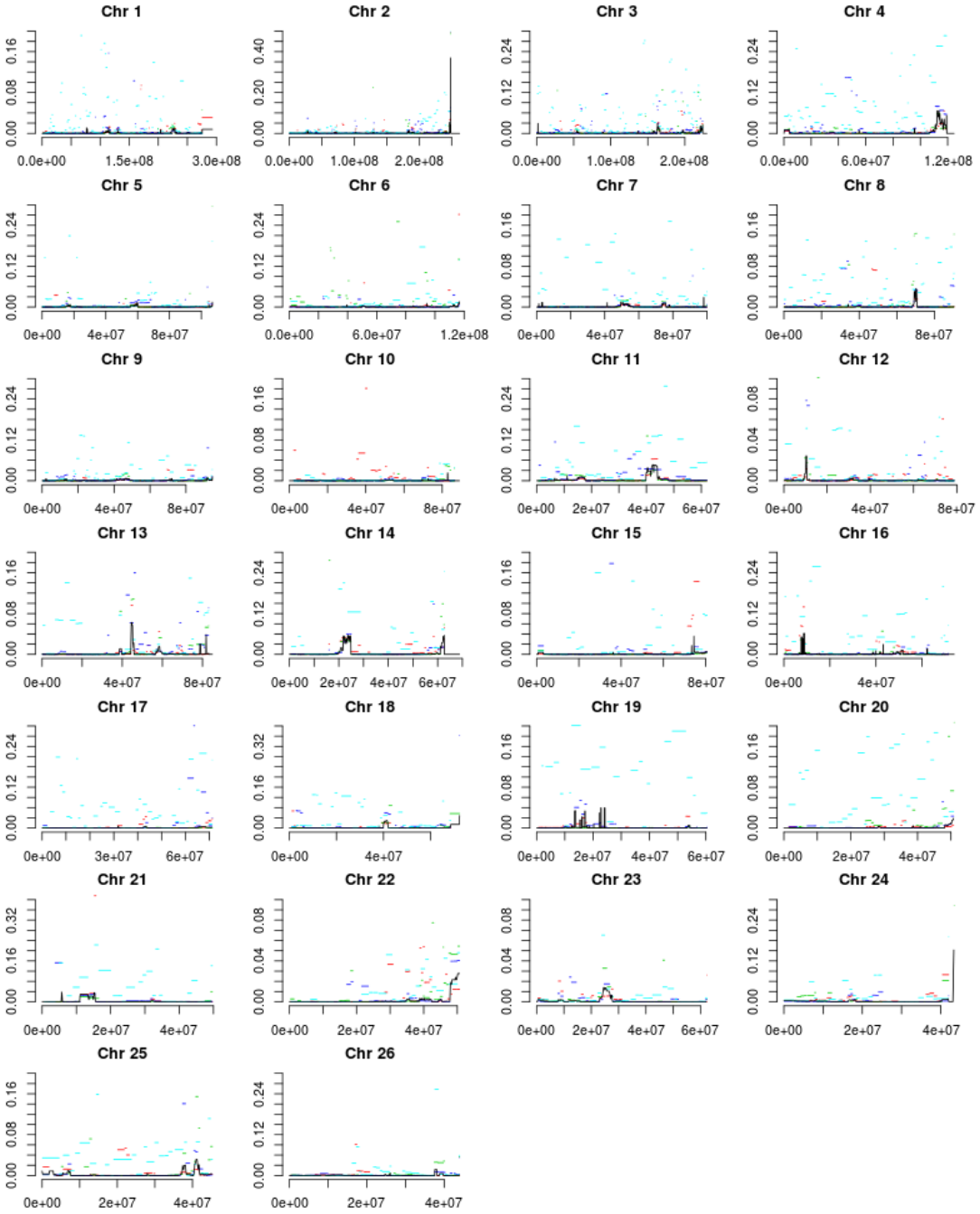
CHU


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

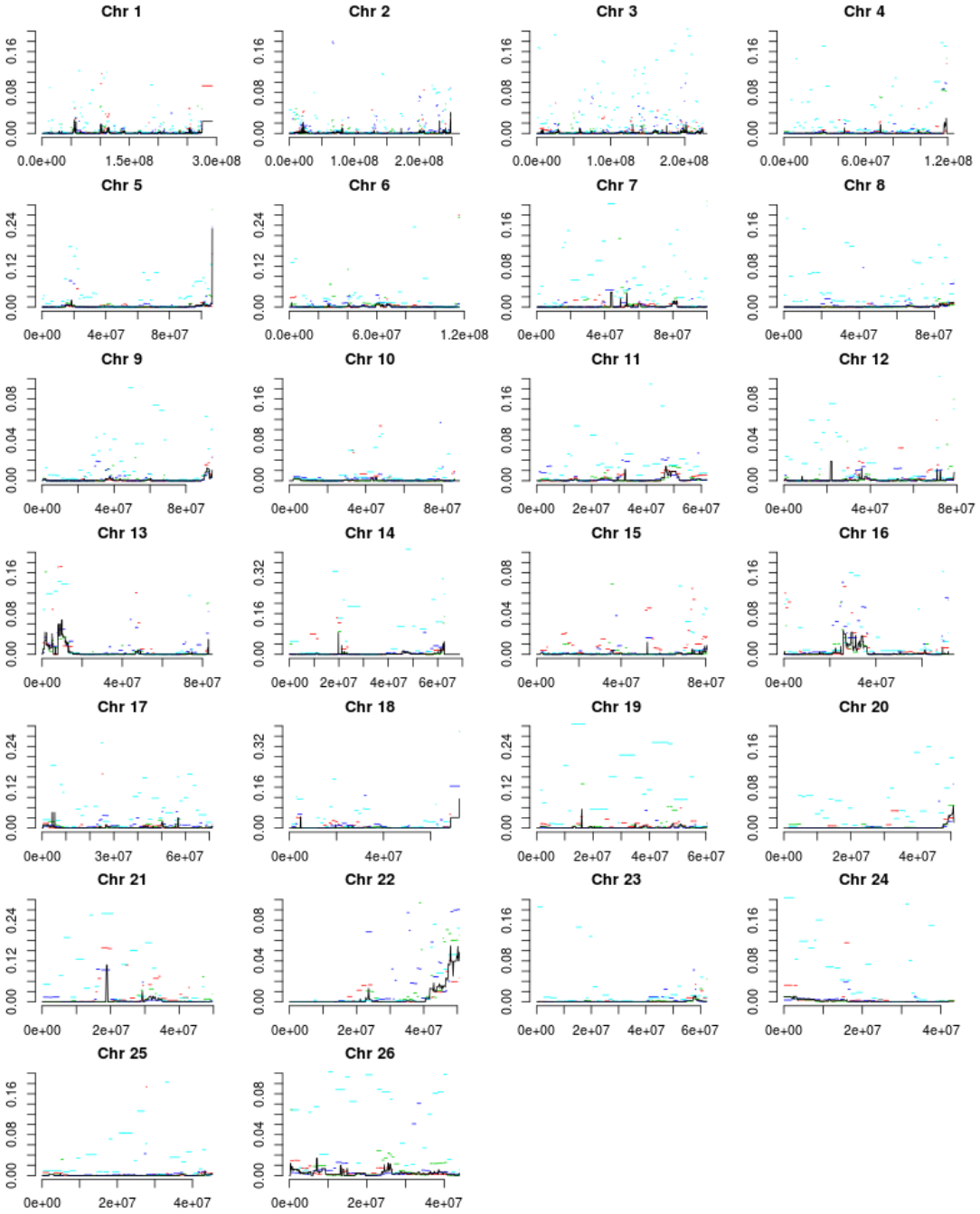
COM


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

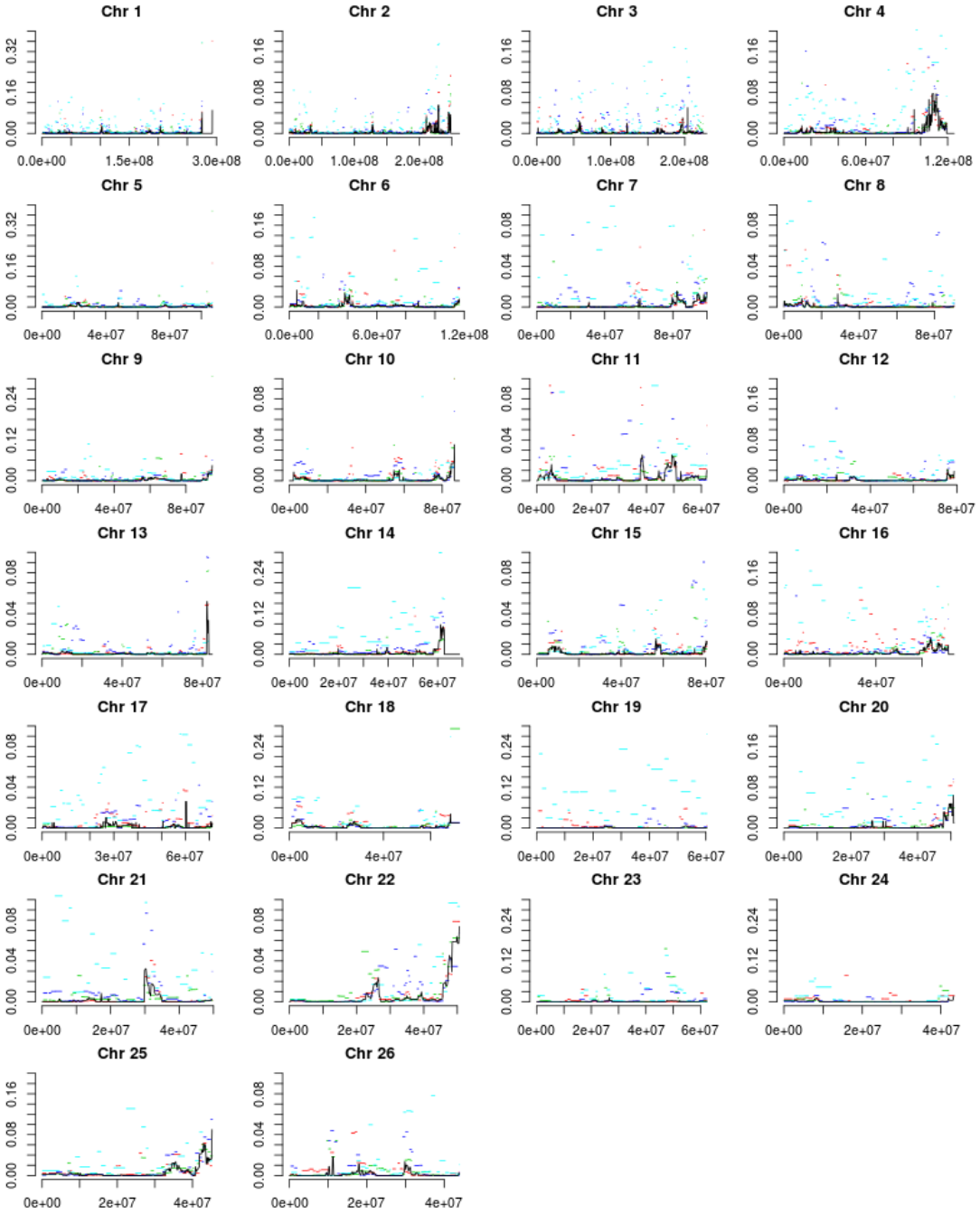
CFT


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

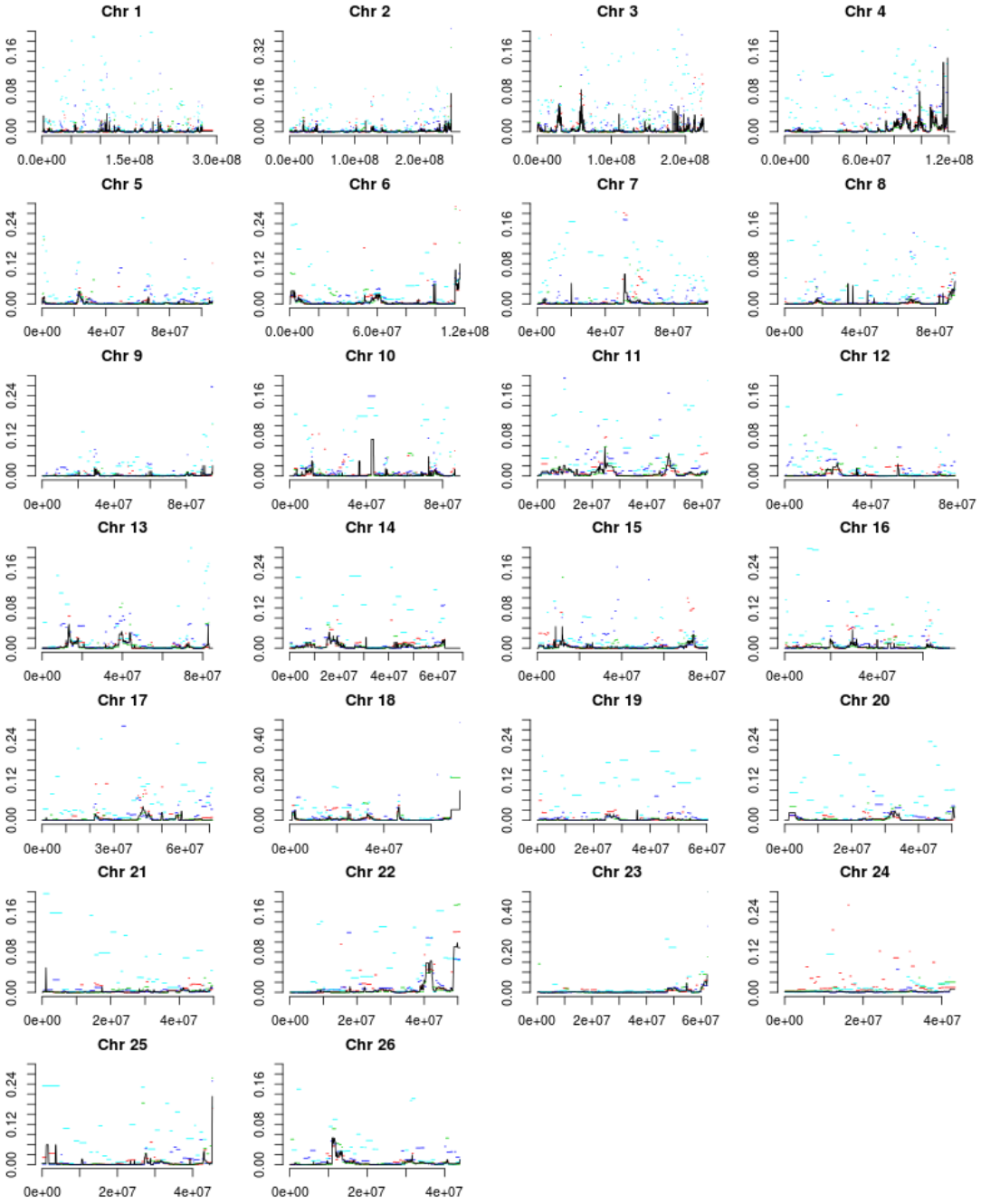
RAK


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

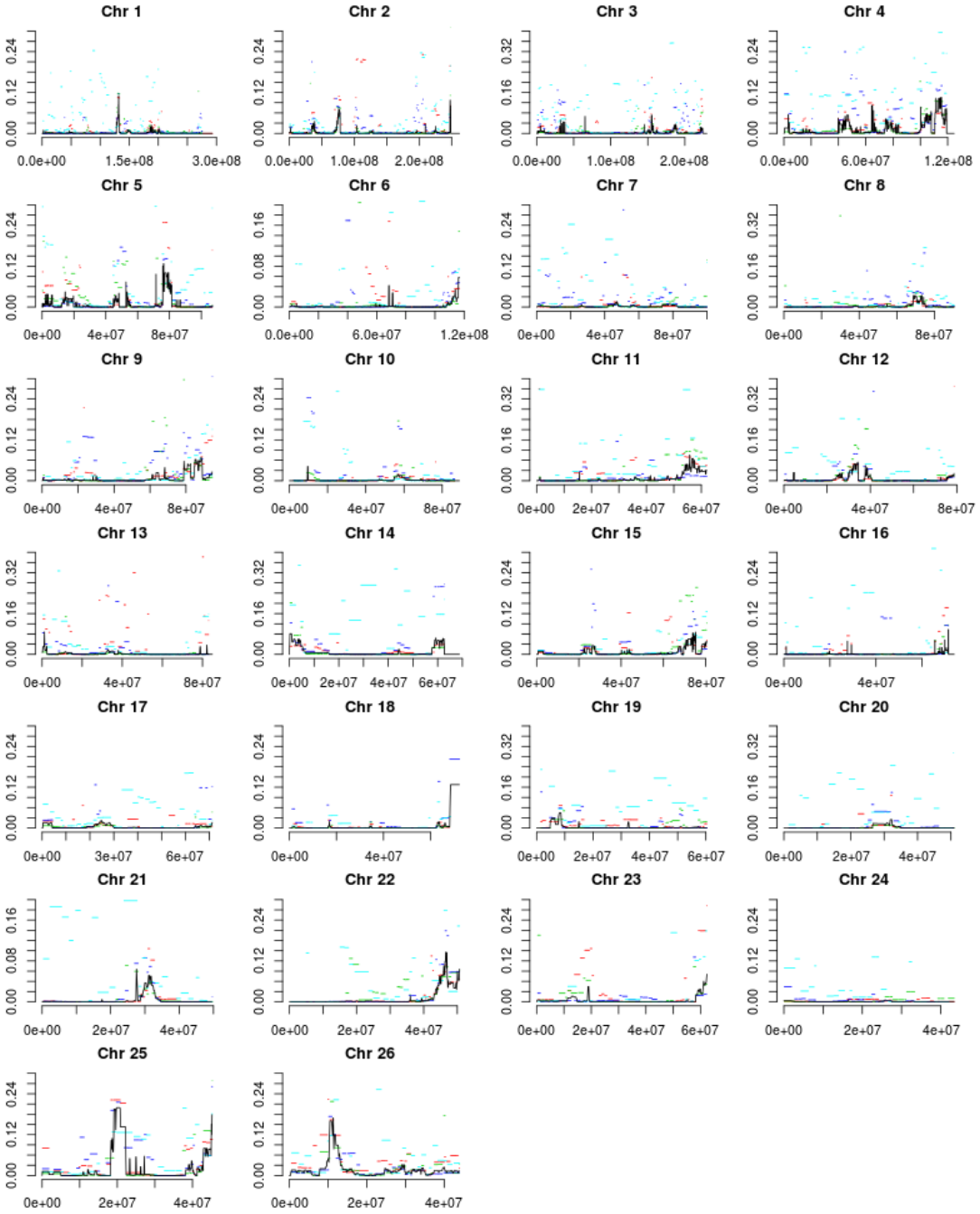
SAR


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

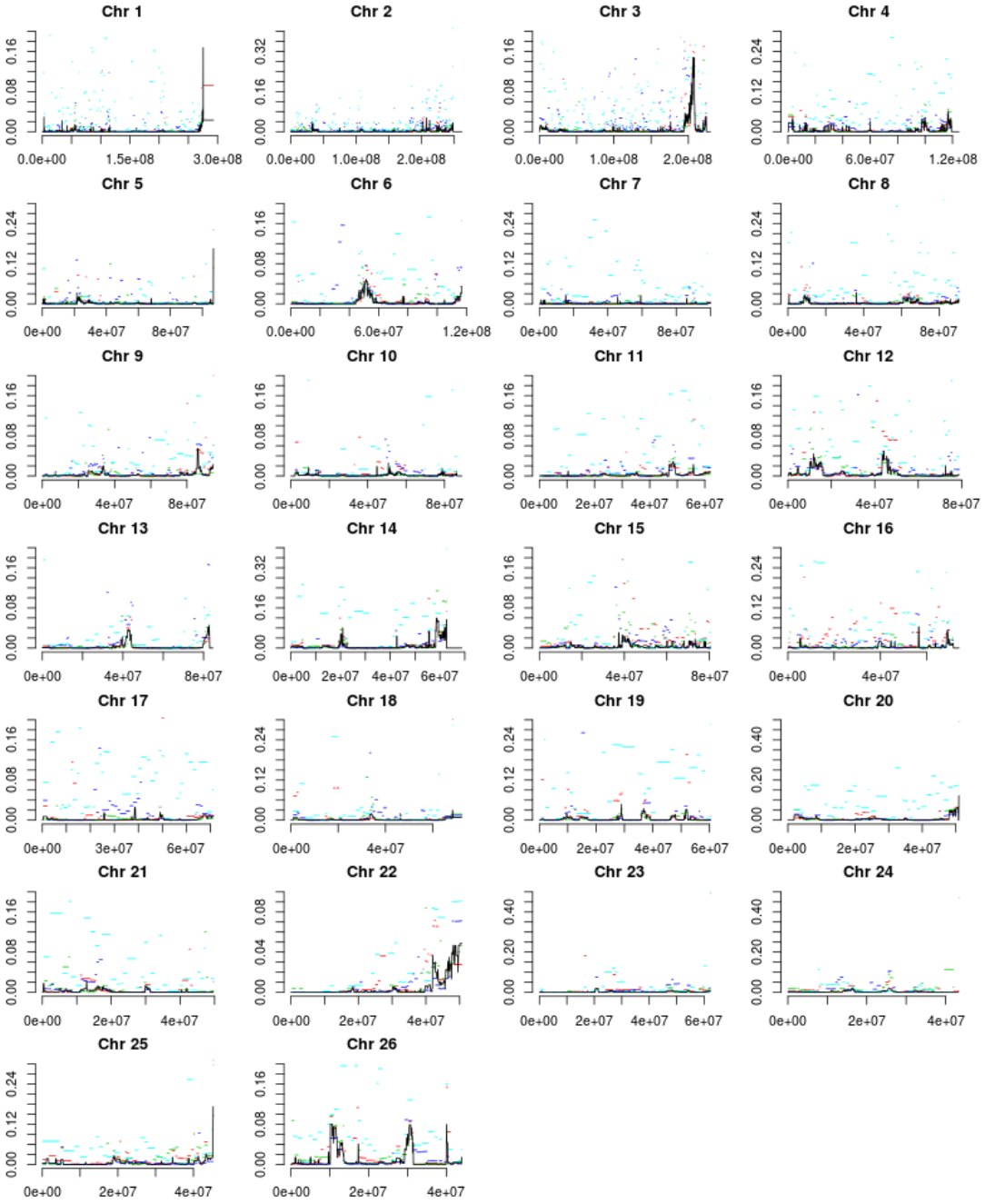
SBF


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

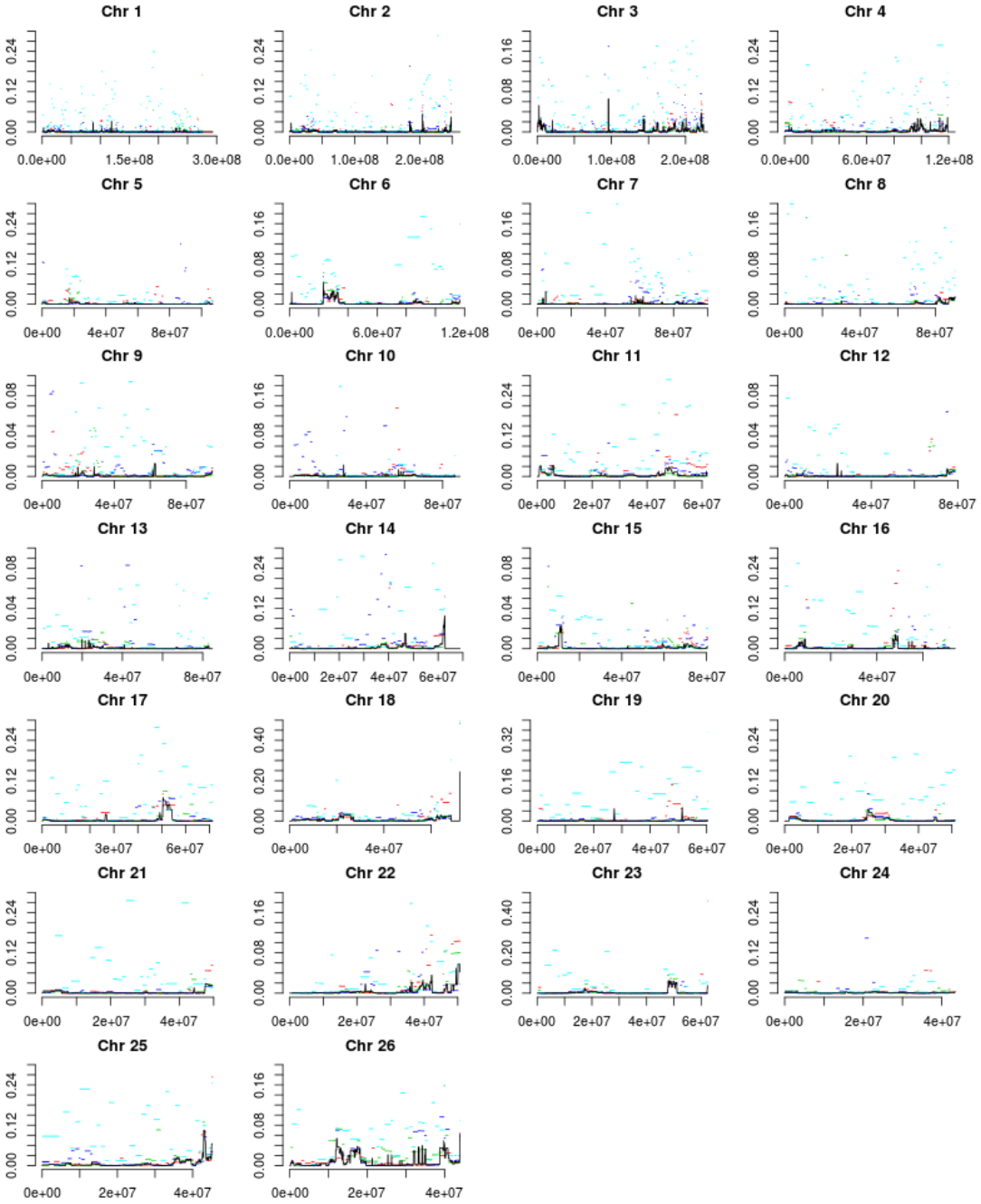
SOA


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

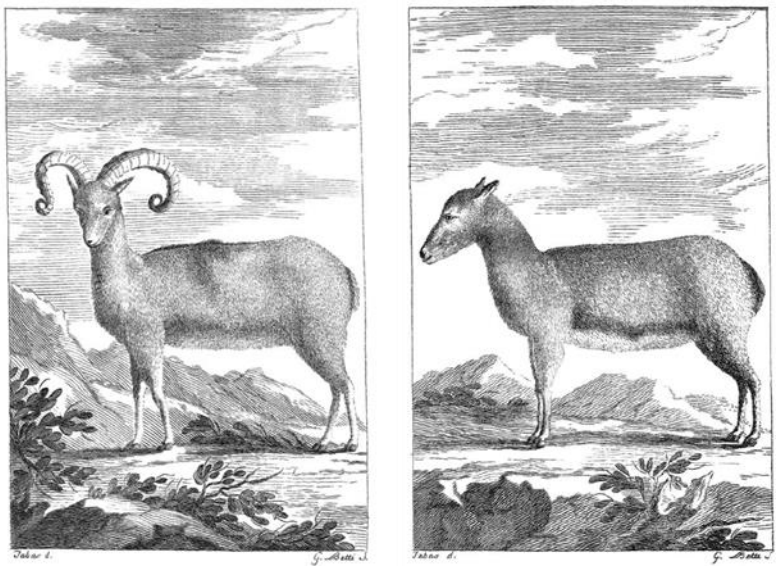
SPW


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

VBN


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
Chr 25
Chr 26

Chapter four



Sardinian mouflon, engraving (Cetti, 1774)

4 Chapter four - Selection signatures in feral and domestic Sardinian sheep

Mario Barbato[1*], Pablo Orozco-terWengel[1], and Michael W. Bruford[1]

[1] *School of Biosciences, Cardiff University, Cardiff CF10 3AX, Wales, UK.*

** Lead author contribution: developed the analytical and informatics methods, performed all the analysis and wrote the first draft, modified the manuscript based on co-author comments.*

Keywords: Selection signatures, local adaptation, domestication, sheep, European mouflon, SNP array.

4.1 Abstract

Since domestication, both artificial and natural selection have shaped the genomes of domestic species towards economically important traits and adaptation to the anthropogenic domain. In this study genome-wide single nucleotide polymorphism (SNP) data were investigated to identify selection signatures and local adaptation in sympatrically distributed feral and domestic sheep populations in Sardinia. Selection signatures were studied using Cross Population Extended Haplotype Homozygosity (XP-EHH) while multiple logistic regression methods were applied to identify local adaptation. A novel locus-specific empirical p-value assessment approach was developed to define a confidence threshold for XP-EHH results and windows of selection were defined by including neighbouring positive results. By using this approach, we identified 475 windows of selection harbouring 693 and 180 genes under putative selection in sheep and mouflon, respectively. In sheep, genes of known agricultural value such as: milk production (e.g., CCL2 and PLCE1) and fertility (e.g., NKD1) were found. In addition, sheep signatures for pigmentation (SLC24A4 and SLC24A5) were also identified for the first time. Genes involved with body size (PLAG1 and CHCHD7) that could be related to mating advantage, were identified in mouflon. The investigation of local adaptation did not produce any significant results. Power and sample-size inference analysis identified the number of samples available to be insufficient to reliably identify local adaptation signals in our dataset. By investigating domestic and feral populations with

genome-wide data and applying novel methods of data filtering, we identified regions under strong selection harbouring promising candidate genes in both feral and domestic sheep that can help a better understanding of the effects of domestication at the genome level. Further, landscape genomics can be computationally intense and we suggest the use of a statistical power inference approach as a preliminary analysis to assess the power for the given dataset.

4.2 Introduction

The advent of modern genotyping technologies has allowed cheaper and faster ways to investigate genome-wide data for large numbers of samples. Among these, SNP (Single Nucleotide Polymorphism) array technology is one of the preferred tools to simultaneously obtain information for a large number of loci (tens to hundreds of thousands) in a relatively cheap and quick way. SNP arrays have been developed for the major domestic species, primarily for quantitative trait investigation and genome-wide association studies, but also for phylogeny, phylogeography and demographic inference (Kijas et al., 2012; McTavish et al., 2013; Orozco-terWengel et al., 2015).

However, a central application for SNP arrays remains the identification of selection signatures, defined as regions of the genome where functionally important sequence variants are present and consequently are preserved or promoted by either natural or artificial selection (positive and/or directional selection). The occurrence of selection translates in the genome by the increase in frequency of the beneficial polymorphism and of the linkage block surrounding it, due to hitchhiking effect (Vitti et al., 2013). The detection of selection signatures applied to humans has allowed the identification of recent positive selection (e.g., lactase persistence) (Bersaglieri et al., 2004), of beneficial traits for productivity in domestic breeds (e.g., meat quality) (Wang et al., 2015) and traits involved with mating success in wild species (e.g., sexual weaponry) (Johnston et al., 2011).

Since the latest technological advances in next-generation sequencing have facilitated the transition from genetics to genomics, new analytical tools to detect selection at genome level have also been developed (Vitti et al., 2013). Among

them, the extended haplotype homozygosity approaches (EHH, Sabeti et al., 2007) have been successfully used to identify selection in a variety of species, first in humans (Pickrell et al., 2009; Racimo et al., 2015) and soon after in cattle and sheep (Moradi et al., 2012; Flori et al., 2014). EHH is defined as the probability that two randomly chosen chromosomes carrying the same allele at a focal SNP are identical by descent over a given map distance surrounding it (Gautier and Vitalis, 2012). In particular, EHH methods in combination with SNP arrays data have been widely used to investigate selection signatures in domestic populations (Moradi et al., 2012; Orozco-terWengel et al., 2015; Kim et al., 2015).

The use of SNP arrays involves an inherent degree of ascertainment bias caused by the SNP discovery process in which a small number of individuals from selected populations are used as discovery panels (Albrechtsen et al., 2010) (see 1.3.1.2). This issue inflates along with the increasing phylogenetic distance between the breed/species analysed and the panel of breeds used during SNP discovery (McTavish and Hillis, 2015). For this reason, the use of the commercially available SNP arrays in species different from those used in the SNP discovery panel (e.g., in wild populations) requires the application of methodological approaches able to reduce the effects of ascertainment bias (Kijas et al., 2012; McTavish and Hillis, 2015).

Local adaptation is the response to differential selective pressures among populations and environments, acting on genetically controlled fitness differences among individuals (Savolainen et al., 2013). The availability of tens of thousands markers has fuelled the development of informatic tools capable of assessing local adaptation to environmental variables for large datasets (Frichot et al., 2013; Guillot et al., 2014; Stucki et al., 2014). Such methods invariably attempt to correlate either genotype or allelic frequencies with environmental data and assess the probability of the presence of that genetic signal given the environment, and are generally referred as part of the 'Landscape genomics' field (Rellstab et al., 2015) (see 1.4.4). The application of such an analytical framework can help to selectively identify genome-wide patterns of selection due to specific drivers (e.g., altitude, solar radiation) in contrast with candidate-gene-based methods (e.g., QTL), where *a priori* information on gene function is needed

(Mackay, 2004), or with population genomics studies, where genome-wide patterns of selection can be identified, but are difficult to link to a specific selective force (Joost et al., 2007). Landscape genomics can fill this gap, complementing and supporting the results obtained by such methods (Joost et al., 2007).

An interesting case study for analysis of selection is the autochthonous European mouflon (*Ovis aries musimon*) population (Figure 4-1) living on the Mediterranean island of Sardinia. Archeozoological and genetic data describe two waves of sheep domestication, dated ~11,000 YA and ~8000 YA, respectively (Chessa et al., 2009), with the first domesticated sheep thought to have shown 'primitive' features typical of the wild ancestor (e.g., presence of hair, horns), and the second wave of domesticated animals showing 'modern' characteristics typical of many present-day breeds (e.g., presence of wool, polled). European mouflon are described as a remnant of the first wave of sheep domestication that subsequently became feral (Chessa et al., 2009). Once present exclusively on the Mediterranean islands of Sardinia and Corsica where they were introduced ~7000-6000 YA (Vigne, 2011), at the end of the 18th century European mouflon populations were established in mainland Europe and more recently elsewhere (e.g., Hawaii) as game and park animals (Santiago-Moreno et al., 2004). Historical records describe the range of the Sardinian mouflon population as covering the whole island with very large numbers (Cetti, 1774). However, in the last century anthropogenic pressure (e.g., hunting, habitat loss) led the mouflon to disappear from the majority of Sardinia except the "Gennargentu" massif in the east-central part of the island (Biagini, 1948). This population can be considered as one of the most ancient autochthonous representatives of the first wave of domestication in Europe.

An autochthonous sheep (*Ovis aries aries*) breed known as the Sarda (Figure 4-1) occurs across Sardinia, currently comprising almost four million individuals. It is the main breed farmed in the island. Despite this number, the Sarda is almost exclusively farmed with low technological support, often using primitive shepherding, especially in the Central mountainous area of Sardinia where traditional farming dominates daily life. Results from chapter 3 show that despite

Sardinian mouflon and Sarda sheep residing in sympatry for thousands of years, no obvious signs of ancient or recent admixture are present in the mouflon.

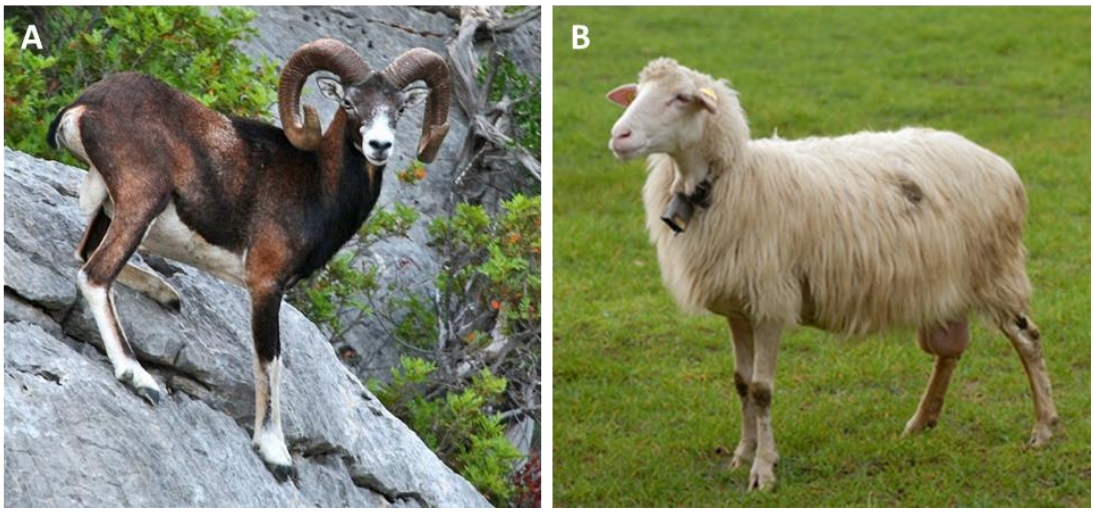


Figure 4-1 Sardinian mouflon (A) and Sarda sheep (B).

Applying both a haplotype-based and a landscape genomics approach using SNP array data, I analysed Sardinian Mouflon and Sarda Sheep populations to identify signatures of selection that reflect adaptation to agriculture and/or to the local environment. Here, i) post-analysis pipeline was developed to establish a locus specific significance distribution for population pairwise statistics such as Cross Population Extended Haplotype Homozygosity (XP-EHH; Sabeti et al., 2007), ii) the effect of sampling on landscape genomic approaches was assessed and iii) a protocol was developed to establish the statistical power of proposed studies.

4.3 Material and Methods

4.3.1 Samples

The Sarda samples analysed were those described as SAR in chapter 3, while the mouflon samples (MSar) were a combination of populations MSar1 and MSar2 from the same chapter. Both Sarda sheep and mouflon sampled for this work were collected in an area of ~450 km^2 in the east-central mountainous part of Sardinia (Figure 4-2).

4.3.1.1 Data filtering

Genomic data for the two populations were obtained as described in chapter 3, with the 26 autosomes being retained for analysis. SNP loci with minimum allele

frequencies (MAF) < 0.05 or unknown positioning were removed. To reduce the impact of this ascertainment bias (Albrechtsen et al., 2010), the dataset was filtered for loci in strong linkage disequilibrium (LD > 0.5) (Kijas et al., 2012; Iacolina et al., 2015) (see 1.3.1.3 and 3.3).

4.3.2 Identification of loci under selection

Genomic regions under positive selection were inferred using XP-EHH, since it is known to perform well with small sample sizes (Pickrell et al., 2009), and does not require ancestral allele information, which is not available for *Ovis*. XP-EHH was performed separately for each chromosome in selscan v1.0.0 (Szpiech and Hernandez, 2014) using default parameters except that the maximum gap allowed between loci that was set to 400,000 bp, appropriate for the estimated OvineSNP50 SNP density (Orozco-terWengel et al., 2015). As the software uses phased data, phasing was performed using fastPHASE v1.4 (Scheet and Stephens, 2006). Default parameters were used in fastPHASE, except that we allowed for the incorporation of subpopulation labels, as this has been shown to significantly improve imputation accuracy (Hayes et al., 2012). Genetic position is used to perform EHH related calculations, however, a complete linkage map for the sheep genome is currently not available, consequently an approximation was used to transform loci physical position to linkage position according to the relation Mb ≈ cM (Kijas et al., 2012; Rothammer et al., 2013). Selscan does not analyse monomorphic loci, consequently these were removed from both population prior analyses. The computed XP-EHH raw scores (rXP-EHH) were then normalized by z-score (sXP-EHH).

4.3.2.1 Empirical p-value obtained by permutation

A permutation approach was developed to determine the significance of the sXP-EHH results. XP-EHH compares haplotype lengths between two populations for local variation in recombination rates (Vitti et al., 2013). The underlying assumption is that the two populations differ in terms of haplotype length and variation. Hence, I formulated a null-hypothesis (H_0) on the assumption that the two populations do not differ, and that neighbouring SNPs are not independent. To test H_0 I developed a pipeline that performed permutations on the two populations by i) pooling their individuals, ii) randomly resampling from this

pool without substitution to produce two new populations with the same size as the original. Then, iii) for each permutation generated, the filtering steps and rXP-EHH values were calculated using the same parameters as in the original populations. A total of 10,000 unique permutations were generated and analysed. Among the permutations produced, populations with identical individual composition could in principle be generated; however, given the large number of replicates performed, we ignored the effect of inflated values toward one specific population composition. Then, for step iv) I calculated *p-values* for each XP-EHH result obtained using the original data following Davison and Hinkley (1997):

$$(1) \qquad pval = (1 + r)/(1 + n),$$

where *r* is the number of permutations that produced an rXPEHH score greater than or equal to that calculated for the actual data and *n* is the total number of permutations obtained for that specific locus. Lastly, in step v) p-values were used to identify loci under selection as those with p-value <0.05. All the computations related to generating the empirical p-values were performed using software written in C++.

4.3.3 Locus and Gene selection

XP-EHH scans the genome for stretches of conserved haplotypes (then compared to a reference population). Consequently, I chose to define windows under selection (Wsel) as those genomic regions with significant neighbouring loci. When isolated loci were found, both downstream and upstream regions flanking the focal SNP were considered and the downstream and upstream sides of each Wsel were extended with flanking regions. To determine each flanking region size, the halved distance between a SNP under selection and the closest non-significant SNP was used. If the focal SNP was the first or last within the chromosome, the flanking region size was set to 27kb, that is half of the median gap among SNPs in the ovine 50k SNP array. Windows of selection were determined via VBA custom-made scripts.

Genes lying within these defined windows were obtained by matching each window against the latest gene annotation for the Ovine genome. The SNP

distribution in the array is relatively homogenous; however the distance among neighbouring loci can vary from <10kb to >150kb (Ovine SNP50 BeadChip leaflet). If the distance between adjacent significant loci is large, very large windows (in terms of bp) represented by a small number of SNPs can be generated. Such uninformative windows are likely to inflate the identification of false positives by artificially overextending the selection sweep range. Given the homogenous SNP density in the array, window size and number of SNPs is expected to be linearly correlated. Hence, a linear regression was used to fit the correlation between window size and number of SNPs, and the windows in the extreme 5% of the lower tail distribution of the residuals were removed. Computation was performed using the R package.

Analyses were also performed to identify putative gene ontology terms (GO) significantly represented among the selected genes. The analysis was performed using the software GOrilla (Eden et al., 2009) and a set of 11,089 genes, related to the SNPs present in the 50k SNP chip as in Kijas and colleagues (2012) was used as background. All the genes obtained for both sheep and mouflon were aggregated according to biological function using PANTHER (Thomas et al., 2003). No annotation reference for *Ovis aries* was available through the software at the time this work is written and therefore the *Bos taurus* annotation reference was used.

4.3.4 Landscape genomics

A total of 20 environmental variables were analysed with the genomic data (Table S 4-2). A collection of 18 climatic variables and a digital elevation model (DEM) relative to the Central part of Sardinia were obtained from online resources (www.worldClim.org and www.sardegnageoportale.it, respectively; Table S 4-2). Additionally, the slope of the area under analysis was inferred from the DEM by using the terrain analyses toolbox within QGIS (QGIS Development Team 2015, QGIS Geographic Information System. Open Source Geospatial Foundation Project). The same software was used to extract environmental variable values coinciding with the sampling locations of the datasets through the 'Point Sampling Tool' extension (http://hub.qgis.org/projects/pointsampling-tool).

In order to investigate signals of local adaptation in Sarda sheep and Sardinian mouflon we used the software Samβada (Stucki et al., 2014). Samβada uses logistic regression to model the probability of presence of each genotype of a polymorphic marker given the environmental conditions at the sampling locations (Joost et al., 2007). Whilst all SAR individuals were georeferenced, among MSar, only 17 mouflons were georeferenced and could be used for the landscape genomics analysis (Figure 4-2). The sheep and mouflon datasets comprising spatial data were analysed by using both a univariate and bivariate model. When the former was applied, each model involving a genotype and an environmental variable was compared with a constant model, in which the probability of presence of the genotype is the same at each location in the landscape and is equal to its frequency in the dataset. When the bivariate model was used, pairs of environmental variables contributed to model the presence of each genotype.

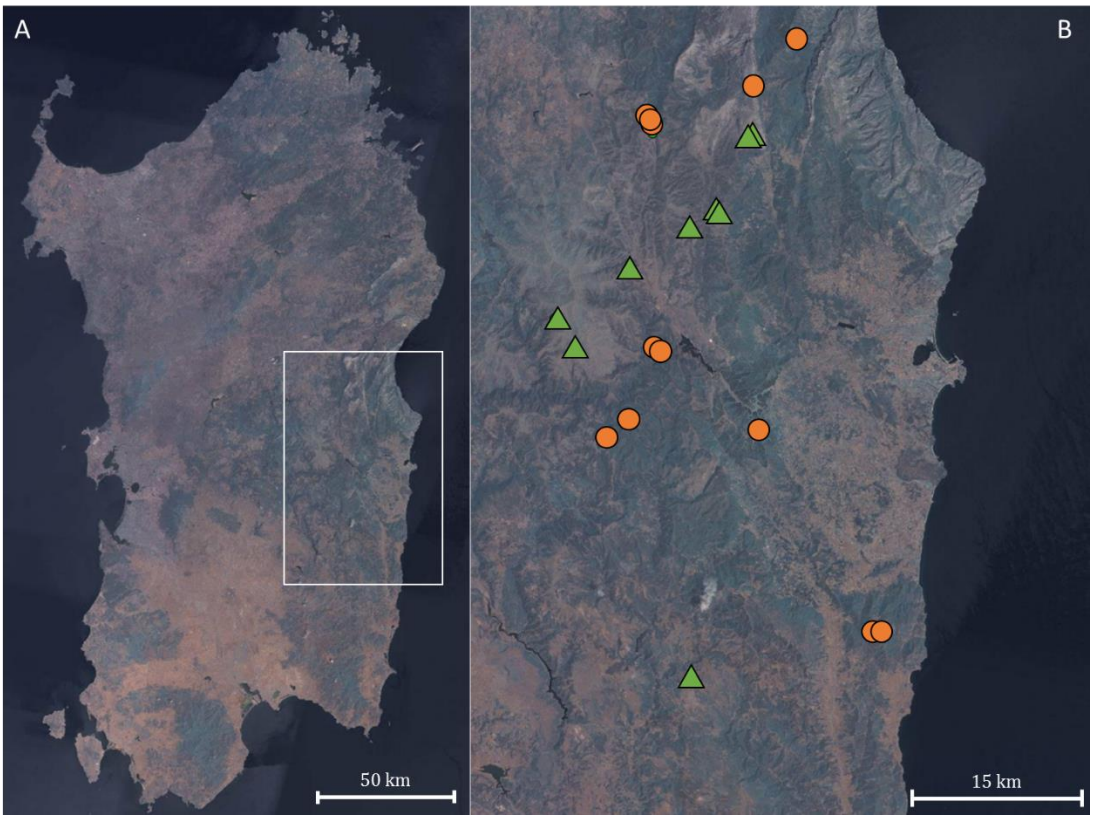


Figure 4-2 Map of mouflon and sheep sampling locations for georeferenced data. A) Overview of Sardinia, mouflon and sheep samples were collected from within the area in the white box. B) Detail of the sampling area, orange circles and green triangles refer to mouflon and sheep samples, respectively.

When the univariate model was tested, all 20 environmental variables were used. When the bivariate model was applied, a test for collinearity was necessary to remove highly correlated variables. The presence of strong collinearity among predictor variables can inflate the analysis towards I type error (Stucki et al., 2014). A pairwise analysis was performed to remove linearly correlated environmental variables using Pearson's correlation index (r). Multicollinearity among variables has been described as acceptable for correlations <0.9 (Stucki et al., 2014). Hence, pairs of variables were removed iteratively until the maximum number of variables retained had a total correlation factor less than 0.9. For both univariate and bivariate analyses a Bonferroni corrected p-value of 0.01 was used to define the significance of each model.

4.3.4.1 Power and sample size analyses

The R function 'powerLogisticCon' from the 'powerMediation' package (Qiu, 2015) was used to calculate the statistical power for a simple logistic regression with a continuous predictor (Hsieh et al., 1998) to detect a signal of local adaptation.

The logistic regression model can be expressed in terms of probability (p) as:

$$(2)\ p = e^{\beta_0+\beta_1 X}/\left(1+e^{\beta_0+\beta_1 X}\right)$$

where the B_0 and B_1 are the regression coefficients for the logistic model and X is the continuous predictor.

For a continuous independent variable the odds ratio (OR) can be defined as:

$$(3)\ OR = e^{\beta_1}$$

And applied to the following:

$$(4)\ n = \left(Z_{1-\alpha/2} + Z_{power}\right)^2/[p_1(1-p_1)[\log(OR)]^2]$$

where p_1 is computed from Equation (2) when X represents the mean of the continuous predictor, n is the sample size required and Z_u is the u-th percentile of the standard normal distribution (Hsieh et al., 1998).

In order to define the sample size required to be able to identify signals of selection without expanding the environmental sampling we used Equation (4)

as implemented in the function 'SSizeLogisticCon' from the same R package. This function calculates the sample size required for a simple logistic regression with continuous predictor in order to have a specific statistical power.

To perform the sampling size analysis, models with statistical power <0.01 were excluded. Sample sizes were estimated for the remaining models using different power categories (for description of categories see Figure 4-5). For each category the number of individuals needed for the analysis to reflect different degrees of statistical power was estimated.

4.4 Results

After pruning for minimum allele frequency and linkage disequilibrium, 36,961 autosomal loci remained for analysis of 27 mouflon and 10 domestic sheep.

4.4.1 XPEHH

XP-EHH was used to identify SNPs showing signals of positive selection in the dataset. Pruning for monomorphic loci led to different locus pools in the two populations; only common loci could be used by the software, leaving 29,961 genome-wide autosomal polymorphic loci. Also, the software excludes a SNP from the results if the EHH decay calculated for that locus reaches the end of a sequence before reaching the EHH cut-off value (Szpiech and Hernandez, 2014). Of the 29,961 loci submitted to analysis, 988 failed to produce a score, leaving 28,973 successfully generated rXP-EHH scores, corresponding to ~55% of the loci originally present in the SNP chip. When compared to the full SNP set, the difference in median distance between neighbouring SNPs was ~15kb (57kb in our dataset against 43kb with the full set from the Ovine SNP array) (Figure S 4-1).

Each permutation of the original populations was analysed using the same approach, and the distribution of the rXP-EHH scores for each SNP was used to define a locus specific p-value (Equation 1). A total of 2,922 loci accounted for the extreme 5% of the standardised XP-EHH scores for both mouflon and sheep (data not shown). After filtering for p-values >0.05, ~45% of the selected loci were removed, leaving 1,324 and 242 putative loci under selection in sheep and mouflon, respectively (Table 4-1, Figure S 4-2).

Table 4-1 Number of loci (top 5%, p-value ≤ 0.05) identified by XP-EHH for each chromosome (Chr).

Chr	*Sheep*	*Mouflon*
1	157	11
2	103	6
3	110	38
4	77	13
5	66	12
6	74	15
7	65	10
8	60	7
9	52	16
10	48	3
11	33	8
12	48	15
13	42	7
14	31	10
15	46	6
16	45	7
17	41	4
18	39	6
19	35	3
20	31	7
21	24	5
22	30	6
23	33	9
24	18	5
25	29	11
26	27	2
TOTAL	*1364*	*242*

The distribution of the loci identified was different in terms of homogeneity, with large genomic regions with many loci identified in sheep (e.g., a region spanning 10-20Mb on Chromosome 1, Figure S 4-2), whereas in mouflon a more sparse an evenly distributed representation of loci under selection was seen (Figure S 4-2). Windows of selection were defined to identify genes but those with a large 'size/number of SNPs' ratio were removed as being poorly informative and prone to produce false positives; 18 and eight windows were removed from the sheep

and mouflon data set, respectively (Figure S 4-3). After pruning for outliers, 475 Wsel were identified, 331 and 144 describing regions under selection for sheep and mouflon, respectively. The Wsel size distribution between sheep and mouflon showed windows as large as 32 SNPs in sheep and a maximum of six in mouflon (Figure S 4-4).

4.4.1.1 Genes under selection

By crosschecking each Wsel with the latest sheep genome annotation (as in OAR v3.1) a total of 693 and 180 genes were identified for sheep and mouflon, respectively. The complete list of the genes found with chromosome and position information is found in Table S 4-1. In the following section those genes that previous research documented to have a recognizable role in the species biology will be discussed.

4.4.1.1.1 Genes under selection in Sarda sheep

Two genes involved in milk production were identified (Table 4-2): CCL2 that is involved with milk yield in sheep (Cecchinato et al., 2014) and PLCE1 that has been related to protein quantity (Wang et al., 2015; Cecchinato et al., 2014). Two genes involved in morphological trait variation were identified: SLC24A4 and SLC24A5 involved in cattle melanogenesis (Rothammer et al., 2013). Six genes related to reproduction and development were identified on four different chromosomes. FOXG1, RELN and MAGI1 are involved in neural stem development and nervous system development (Kijas et al., 2012; Gautier et al., 2009; Cerri et al., 2012), whereas RERE and PRKCH are related to embryonic growth and reproductive development (Randhawa et al., 2014; Chomwisarutkun et al., 2012), lastly NKD1 to fertility traits (Rothammer et al., 2013). MANEA, a gene involved in protein metabolism, was identified on the chromosome 8 (Buzanskas et al., 2014). APBA2, related to behavioural traits in mammals was also identified (Bertolini et al., 2015; Frantz et al., 2015).

4.4.1.1.2 Genes under selection in Sardinian mouflon

Four genes identified as being under selection in mouflon were already described by previous research to have a role in both cattle and sheep. The neighbouring genes RPS20-PLAG1-CHCHD7 related to body size were found, on chromosome

9 (Randhawa et al., 2014; Frantz et al., 2015). Further, RGS1 and SLC38A2 were identified (located in chromosome 12 and 3, respectively) and have been previously related to cell-mediated immunity and small molecule transport, respectively (Gautier et al., 2009).

Table 4-2 Main genes and functions under selection in sheep and mouflon. The chromosome (Chr) and position of each gene are provided.

Gene function	*Chr*	*Position*	*Gene symbol*	*Reference*
Sheep				
Milk production				
milk yield	11	15.51-15.51	CCL2	Cecchinato et al., 2014
total protein weight in milk	22	14.97-14.97	PLCE1	Cecchinato et al., 2014
Morphology				
melanogenesis	7	59.22-59.24	SLC24A5	Rothammer et al., 2013
melanogenesis	18	55.95-56.14	SLC24A4	Rothammer et al., 2013
Reproduction and development				
neural development	4	44.67-45.20	RELN	Rothammer et al., 2014
developmental processes	7	70.36-70.72	PRKCH	Rothammer et al., 2013
embryonic growth reproductive development	12	42.85-43.16	RERE	Randhawa et al., 2014
fertility	14	18.45-18.55	NKD1	Rothammer et al., 2014
neural stem development and differentiation	18	38.11-38.11	FOXG1	Kijas et al., 2012
nervous system	19	35.64-35.64	MAGI1	Gautier et al., 2009
Metabolic processes				
metabolism of proteins	8	41.36-41.36	MANEA	Buzanskas et al., 2014
Behaviour				
behavioural traits	18	27.09-27.09	APBA2	Frantz et al., 2015
Cell functions				
cell cycle	1	17.82-17.82	CDC20	Gautier et al., 2009
cell function	3	196.15-196.15	PIK3C2G	Flori et al. 2009
cellular processes	7	51.44-51.44	NEDD4	Flori et al., 2014
Mouflon				
Morphology				
stature	9	36.66-36.11	RPS20	Frantz et al., 2015
stature	9	36.16-36.21	PLAG1	Randhawa et al., 2014
stature	9	36.21-36.22	CHCHD7	Randhawa et al., 2014
Immune response				
cell-mediated immunity	12	10.79-10.80	RGS1	Gautier et al., 2009
Cell function				
transport of 2-(methylamino)isobutyric acid	3	139.94-139.95	SLC38A2	Gautier et al., 2009

4.4.1.1.3 GO analysis

Analysis of GO term enrichment was performed on both the sheep and mouflon genes sets identified. However, even without correcting for multiple testing not a single GO term was identified as being enriched in the dataset.

Genes identified for both sheep and mouflon were aggregated by biological function (Figure 4-3). The distribution of the aggregated data was mostly comparable between sheep and mouflon with the metabolic and cellular processes showing as the most represented classes in both sub-species.

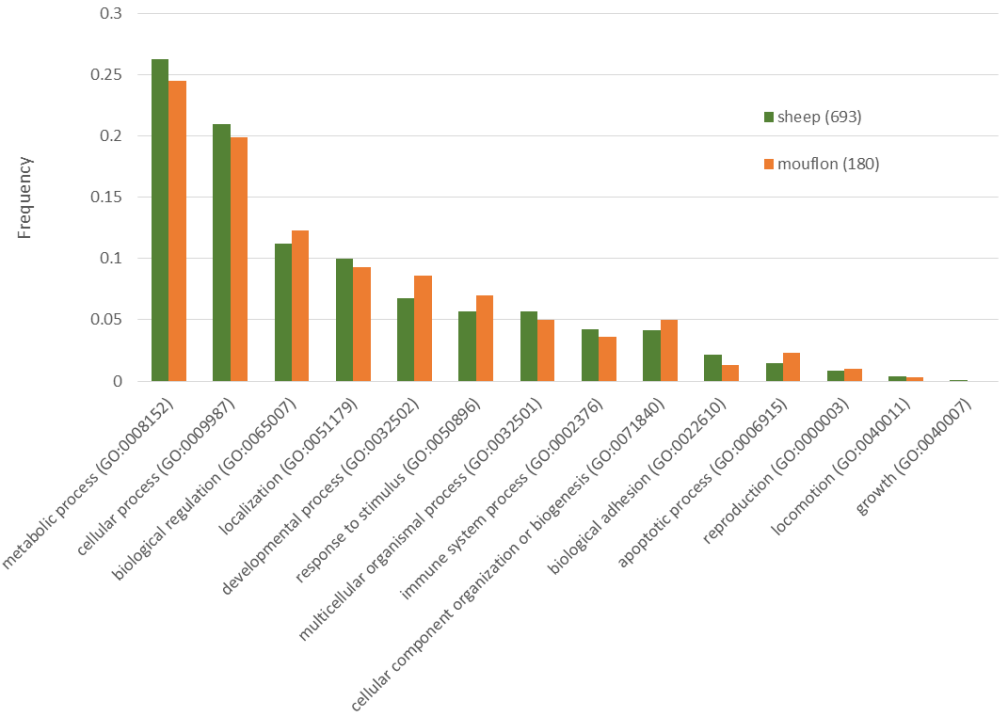


Figure 4-3 Genes identified as being under selection in sheep and mouflon aggregated for biological function. Biological functions are sorted from most to least represented according to sheep results. On the horizontal axis the biological categories and related GO code are shown. The number of genes identified for sheep and mouflon are indicated in brackets in the inset legend.

4.4.2 Landscape genomics

4.4.2.1 Analysis

Neither the univariate nor the bivariate analysis produced any significant model after Bonferroni correction for both the mouflon and sheep dataset. Consequently, in this section I describe the results obtained from the sole univariate model as the approach with less stringent parameterisation. Samβada computed a total of 1,483,249 and 1,651,948 models for the mouflon and sheep

dataset, respectively. To understand the statistical power distribution for these models a logistical power analysis was performed on both datasets.

4.4.2.2 Power

The distribution of statistical power showed that for both the sheep and mouflon populations, 95-96% of the models had a power lower than 0.1, with a median of 0.006 for both species (Table 4-3, Figure 4-4). The other ~5% of mouflon models computed were evenly allocated across the rest of the distribution, with 0.3% of models having a statistical power larger than 0.9 (Table 4-3). In contrast we found the sheep models possessing the highest statistical power were located between 0.8 and 0.9 (Table 4-3).

Table 4-3 Number of models present for each class of statistical power.

Power	Mouflon	Sheep
0.1	1,412,178 (95.2%)	1,585,065 (96%)
0.2	25,475 (1.7%)	30,741 (1.9%)
0.3	10,541 (0.7%)	13,669 (0.8%)
0.4	7,706 (0.5%)	4,680 (0.3%)
0.5	6,699 (0.5%)	10,549 (0.6%)
0.6	3,665 (0.2%)	419 (0%)
0.7	4,774 (0.3%)	0 (0%)
0.8	4,231 (0.3%)	815 (0%)
0.9	2,893 (0.2%)	6,010 (0.4%)
1	5,087 (0.3%)	0 (0%)
TOTAL	*1,483,249*	*1,651,948*

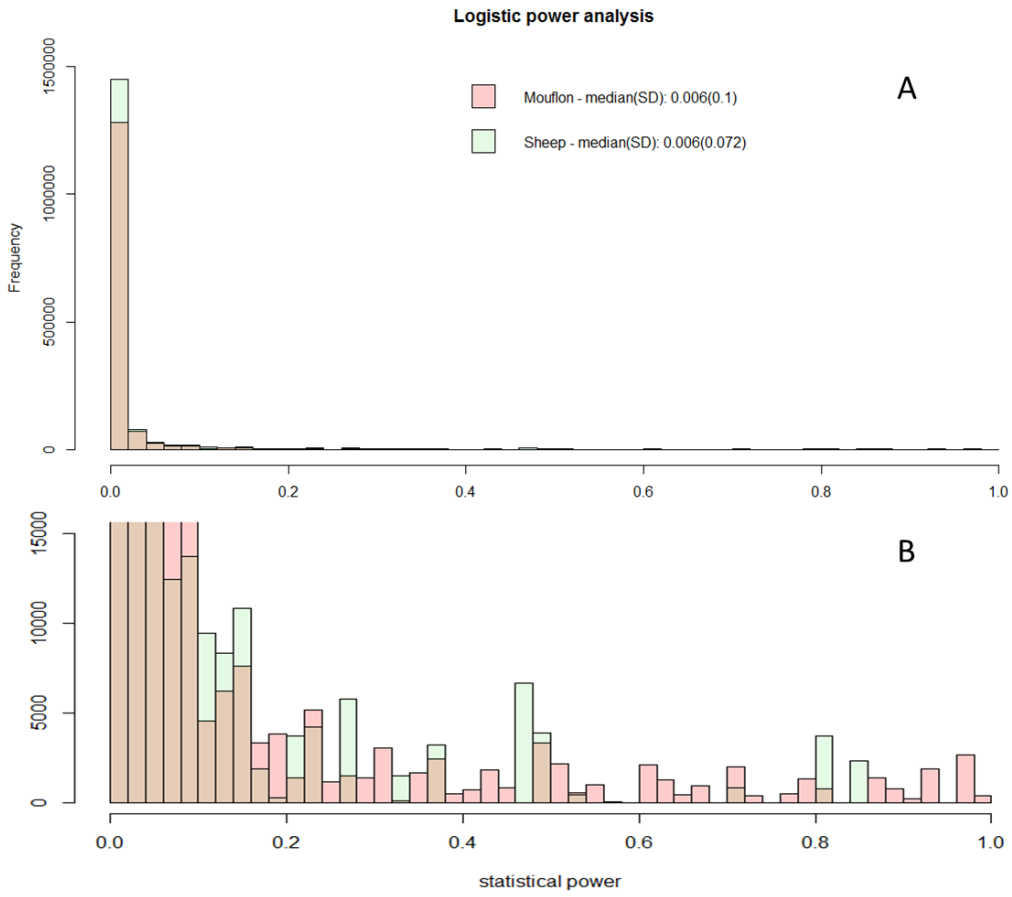


Figure 4-4 Distribution of the statistical power calculated for the simple logistic regression with continuous predictor applied to Samβada results for sheep and mouflon. A) Full scale histogram. B) Magnification of the low frequency values.

4.4.2.3 Sample size

Given the sampling on the environmental variables, we investigated which sample size would have been necessary to adequately improve the statistical power of the logistic regression analysis. For this purpose we divided the frequency range in bins of 0.2 from 0 to 1 and attempted to estimate the expected sample size needed for a given statistical power on the basis of our dataset. Instead of following this approach we could have estimated the sample size under a certain statistical power for the whole distribution of statistical models used, however, as the statistical models for our dataset were not evenly distributed across the whole range of statistical power, we opted for the use of the bins instead (Figure 4-4).

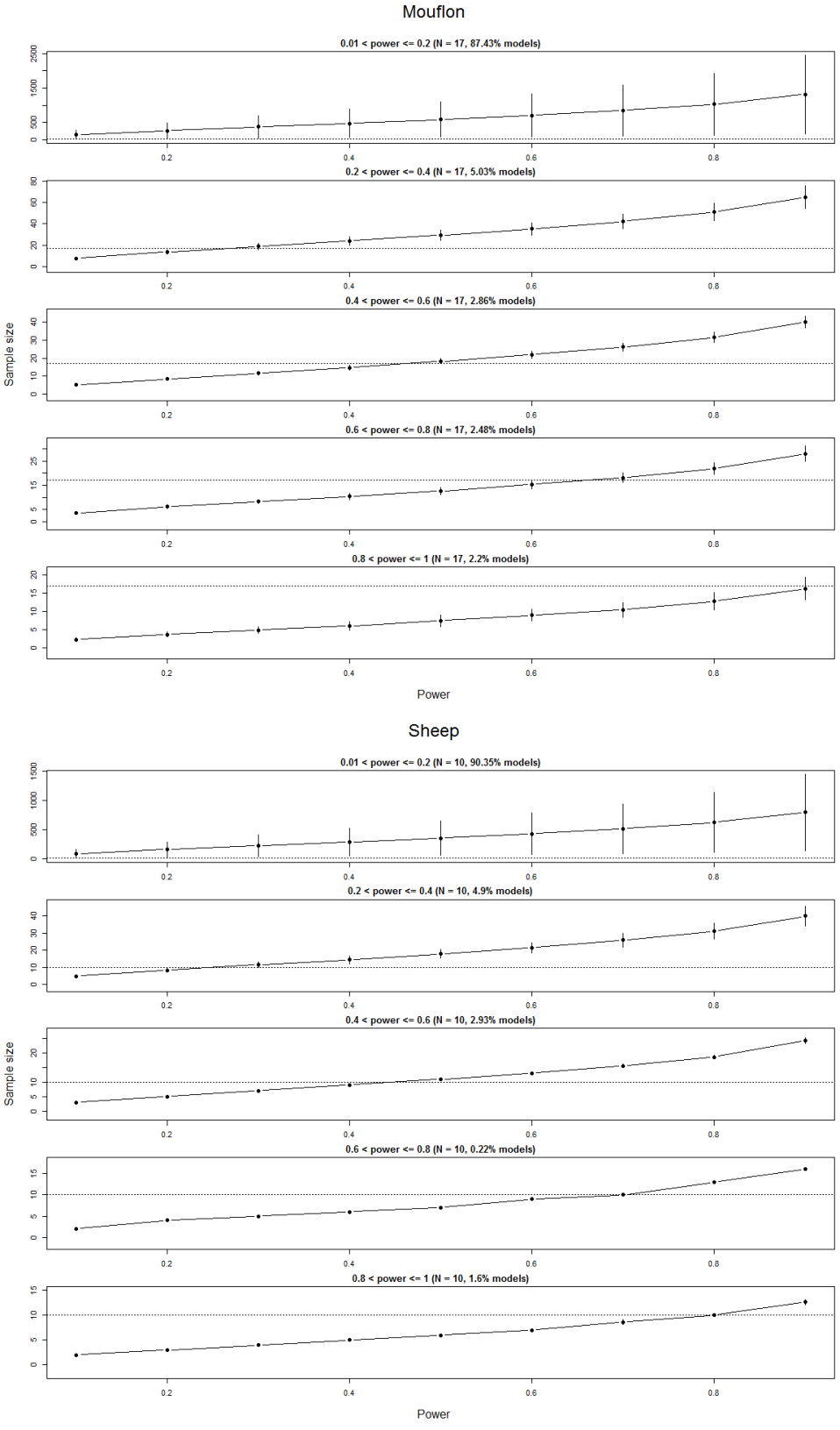


Figure 4-5 Sample size inference for logistic regression with continuous variable data on the mouflon (top) and sheep (bottom). The vertical lines on each data point represent standard deviation values. On top of each graph the power class represented is indicated. In brackets is the actual sample size (N, also represented within the graph by a dashed line) and the proportion of models represented by that particular power class.

The results described here are those pertaining the first bin of the statistical power distribution (0-0.2), as these represent the bulk of the models in our dataset, approximately 87% for mouflon and 90% for sheep (Figure 4-4). The mouflon sample size necessary to achieve a statistical power of 0.1 in the first bin was 131 individuals, while to achieve 0.9 a total of 1,156 individuals were needed (Figure 4-5). Similarly, the sheep sample sizes needed for these thresholds are 82 and 719 individuals, respectively (Figure 4-5). For the remaining bins, the higher the statistical power, the lower the sample size needed (Figure 4-4, Figure 4-5).

4.5 Discussion

4.5.1 XPEHH

Selection signatures were inferred using XP-EHH via pairwise analyses between MSar and SAR of the extent of EHH surrounding each locus in the dataset (Sabeti et al., 2007). After pruning for extreme values and significance threshold, ~5% of the loci submitted to analysis were identified as within a selective sweep. Similarly, Rothammer et al. (2013) found a total of ~3% of loci under selection and Orozco-terWengel et al. (2015) recorded a total of ~7% while investigating cattle breeds using XP-EHH analyses.

The combination of pre-analysis filtering and success in assigning XP-EHH scores gave a reduced representation across the genome, with information-free gaps spanning several Mb (Figure S 4-1, Figure S 4-2). The distribution gap of SNPs contributing to the XPEHH analysis and the full SNP set in the Ovine SNP array were not perfectly comparable (Figure S 4-1). However, their medians differed by only 15kb. This result suggests a non-homogenous distribution of gaps along the chromosomes, with the overall spacing of the original SNP set being mostly retained but with an increased occurrence of large and extremely large gaps (Figure S 4-1, Figure S 4-2).

Is important to remember that analytical tools such as EHH methods were developed for high density datasets (e.g., 3M SNPs in the Human HapMap project; see Sabeti et al., 2007), and are therefore able to identify the occurrence of narrow selection sweeps with high accuracy (provided the frequency of the core allele meets the requirement of the analytical method applied). The presence of

larger gaps such as in the 50k SNP chips poses the challenge of less detailed identification of a selective SNP. A consequence is that narrower selective sweeps might go undetected, if the focal point of selection lies within a large gap and the surrounding loci do not sufficiently represent the selection event, or could be aggregated into a single inferred locus if selection happened close to an available SNP. While the latter problem can be countered by allowing a window of selection surrounding a selected locus to be set according to the linkage disequilibrium decay, as done for this and other studies (Rothammer et al., 2013; Orozco-terWengel et al., 2015), dealing with the former imposes more articulated solutions such as the use of composite metrics (e.g., F_{ST} and XPEHH as in Randhawa et al., 2014) or the use of windows of selection as in this work (Rothammer et al., 2013). However, while haplotype based methods are ideal to deal with ascertainment data and XP-EHH has been shown to produce meaningful results even with small samples size, allele frequency based methods such as F_{ST} are highly affected by both ascertainment bias and sample size. Our dataset suffers from both effects due to the absence of the European mouflon from the original SNP discovery panel (Kijas et al., 2012) and the small sample size of the domestic population (N = 10) making XP-EHH the least biased approach for the data.

4.5.1.1 Windows of selection

Neighbouring loci under selection were used to define windows of selection in the genome. The number of loci identified in domestic sheep was 5-fold greater than those identified in mouflon (Table 4-1); however, the number of windows identified in sheep was only twice as large (349/152). This observation indicates larger genomic regions under selection identified in sheep (with many SNPs included), while smaller windows (with less SNPs included) were found in mouflon (Figure S 4-2, Figure S 4-3).

The presence of larger windows of selection in sheep than mouflon is compatible with the timing and magnitude of selection in the genome. A selection sweep is expected to have a larger effect (detected by the XP-EHH score) upon fixation; however both recombination and mutation break the linkage disequilibrium around the locus under selection over time, *de facto* reducing the haplotype size

(Sabeti et al., 2002). Consequently, large windows can also be related to recent or ongoing selection, as in the case for sheep, whereas shorter windows could be due to older fixed variants with the surrounding selection sweep size reduced by evolutionary processes.

The investigation of known genes within the windows of selection identified in sheep revealed genes involved in milk production, morphology, reproduction, metabolism and behaviour (Table 4-2). All these biological functions can be related to features typical of the modern sheep breeds.

4.5.1.1.1 Sheep genes under selection

A gene related to increased milk yield (CCL2) was identified in the Sarda sheep, an interesting result given that this is a breed highly selected for milk production (Lancioni et al., 2013). The vast majority of the sheep milk produced in Italy comes from this particular breed and is mostly made into pecorino cheese. A gene related with total protein weight in milk was also identified (PLCE1). Protein content in milk is a major index for milk quality, but is especially important for cheese production, and depends on the protein intake of the lactating ewe (Walker et al., 2004). More efficient protein metabolism can increase the total protein mass in milk. Additionally, among the other selected genes we identified was endo-alpha mannosidase (MANEA) previously identified to be involved in protein metabolism (Buzanskas et al., 2014).

Two genes in the Solute Carrier family (SLC) related to melanogenesis were identified: SLC24A4 and SLC24A5. Pigmentation is a complex trait and hundreds of genes have been identified to be responsible for one or more pigmentation traits in mammals by influencing either the production or distribution of pheomelanin and eumelanin (Cieslak et al., 2011; Zhang et al., 2015). Importantly, similar phenotypes can be the result of different combinations of genes involved (Cieslak et al., 2011). SLC24A4 and SLC24A5 have been previously linked to eye, hair and skin lightening in humans (Hudson et al., 2014; Sabeti et al., 2007; Sturm, 2009; Sulem et al., 2007). The Sarda sheep phenotype is characterised by a white coat and very light skin, while the mouflon phenotype, as in the majority of wild sheep, possesses darker coat and skin (Figure 4-1), is

therefore possible/likely that, similar to humans, these two SLC genes contribute to Sarda light coloured phenotype.

Six genes from different chromosomes were found to be related to reproduction and development. In the wild the reproductive cycle is strictly seasonal and ruled by photoperiod with one lambing period in spring that allows milk production optimised to pasture availability (Lincoln, 1998). In domestic sheep optimisation of reproductive output requires up to two gestations per year, and for lambs to grow at a faster rate for quick sale. The Sarda is a highly prolific breed, with twins occurring at 10-40%, in stark contrast to both Corsican and Sardinian mouflon (Garel et al., 2005). It this context our finding of genes involved in fertility (NKD1, RERE) and development (RELN, PRKCH, FOXG1 and MAGI1) could be explained by artificial selection for reproductive traits in the Sarda. In this regard, the protein metabolism related gene previously discussed (MANEA) could also play a role as previous research suggested MANEA to be involved with increased birth weight in cattle (Buzanskas et al., 2014). Additionally, PIK3C2G was also identified, and belongs to an intracellular signalling pathway primarily involved with the cell cycle (PI3K/AKG), which recent research suggests is related to the recruitment of primordial follicles (Bromfield et al., 2015) in oogenesis.

Selection for behavioural traits is of paramount importance in the domestication process (Clutton-Brock, 1999). In our analyses a gene related to the locomotory behaviour in response to stimulus was found (APBA2). The relaxation of natural selection, could cause a modification of behaviours towards less energy demanding strategies (Lindqvist and Jensen, 2009) such a reduced flight distance for sheep compared to mouflon.

4.5.1.1.2 Mouflon genes under selection

Genes under selection in mouflon were also identified. Among these, the biological function of four has been previously identified (Table 4-2). The neighbouring genes RPS20-PLAG1-CHCHD7 were identified, which previous research has related to stature in cattle (Karim et al., 2011), pig (Frantz et al., 2015), horses (Petersen et al., 2013) and humans (Weedon and Frayling, 2008). The cross-species relevance of these genes seems to be suggested since mouflon are on average taller than Sarda sheep (height at withers 75 and 70 cm,

respectively). Body size has a substantial influence on reproductive success in ungulates (Grignolio et al., 2008); however, during domestication the imposition of controlled mating by herders has rendered this feature less relevant in domestic sheep (Clutton-Brock, 1999).

4.5.1.1.3 Summary on selection in sheep and mouflon

The XP-EHH approach uses both populations as reciprocal references for each other. However, if selection is acting in the same genomic region in both populations with equal intensity, the resulting XP-EHH score is expected to become 0, with one population masking the selection of the other. So, while this methodology potentially provides a perfect reference to investigate features involved with the second wave of sheep domestication compared to the first, only strongly/heavily unbalanced signals of selection can be detected with this approach, whereas coincident or parallel selection of similar intensity in both species (e.g., environmentally mediated selection) could be masked.

The results concerning genes under selection in Sarda sheep seem to identify features linked to those that specifically characterise the breed. Two main classes of genes were found to be directly or indirectly related to milk production/quality and fertility/development of lambs: two agronomical traits for which Sarda sheep are known to excel, with ~200 litres of milk produced per year and ~1.3 lambs per birth (see also 1.2.1). In mouflon, no strong and obvious signals of selection related to previously described genes were found, except for morphological traits coherent with mouflon anatomy when compared with sheep.

Importantly, several of the genes identified here were previously described in other studies, but mostly concerning cattle. The coherency of the results obtained here with the background knowledge on the species investigated suggests them as potential targets for selection across multiple mammalian lineages as also observed in other studies (Kijas et al., 2012).

4.5.1.2 Comments on identification of windows of selection and the methodological framework

In several apparently large sheep regions of selection (e.g., in chromosomes 2, 10 and 15, Figure S 4-2) no gene with known function was identified. However, several coincided with regions previously recognised to harbour SNPs under selection in sheep. That is the case of a region in the second chromosome spanning 55-56Mb and another one spanning 25-31Mb (residing in the second and tenth chromosome, respectively). These regions, among others, were recorded by Kijas et al. (2012) while identifying regions under selection in a dataset of 74 domestic breeds by means of global F_{ST} (Nicholson et al., 2002). Within these windows Kijas et al., (2012) identified NPR2 and RXFP2, the first being involved with skeletal development and the second being consistently linked to the polled phenotype in ungulates. The latter was expected to be identified in our dataset considering that while mouflon is usually horned, Sarda sheep is almost exclusively polled (Figure 4-1). Indeed, the use of a less stringent size for the flanking regions extending from a Wsel, allowed the identification of RXFP2 among the genes under selection (data not shown). This also supports the negative effect of larger gaps between the loci analysed on detecting selection signatures as previously discussed. A more relaxed parameterisation could, however, have resulted in an inflated number of type I errors, given the sparsity of our dataset and the unfeasibility of using composite methods to generate a consensus result on positive signals. In the case of RXFP2, the nearest locus was not excluded due to pre- or post-analysis filtering, and was instead removed by the internal routine of the XP-EHH implementation. The application of analytical frameworks developed for dense datasets to non-ideal scenarios, implies a certain decrease in the signal detection accuracy and precision. Indeed, sparse datasets can promote both type I and II error. Using methods with different sensitivity can help reduce/minimise false negative results (e.g., iHS can detect shallower signals of selection), but different methods imply assumptions that do not always fit with the available data, as in this study. However, statistical methods can be applied post-analysis to quantify the strength of the positive results against an alternative hypothesis.

In this study locus-specific empirical p-values were computed to assess the confidence of each XP-EHH result and help to compensate for the lack of comparative methods to assess the robustness of the selection signals. In previous research, empirical p-values based on a sliding-window approach were computed for large genomic regions (e.g., ~200k) and used to define confidence intervals for both XP-EHH and iHS (Pickrell et al., 2009; Simonson et al., 2010; Cheeseman et al., 2012). The latter methodology has the advantage of being applicable to any statistic where a distribution of results is produced. However, such an approach requires a high-density dataset to guarantee accurate empirical p-values in sufficiently narrow genomic regions (e.g., 1 SNP/10kb as in Pickrell et al., 2009). Instead, the allele-specific empirical p-values presented here can be applied to any data density; although it is method specific.

More extreme top score thresholds (e.g., extreme 1%) have previously been applied to identified top SNPs and reduce the chance of false positives (e.g., Simonson et al., 2010). However, although such a strategy can identify the most significant loci under selection, it also reduces the chances of recognising extended regions under selection when a sparse dataset is used. Instead, a three-fold approach was applied to recognise genomic regions under selection:

i) We considered the extreme 5% of all the available scores.
ii) We retained the XP-EHH results with sufficient statistical support via empirical p-value, and:
iii) We used the window-based approach to identify regions of selection.

This pipeline allowed the identification of several genomic region with strong and coherent signals of selection and to identify genes consistent with sheep domestication and demographic history in the Sarda but also in the European mouflon.

4.5.2 Landscape genomics

In this work an exploratory analysis of selection signatures driven by environmental variables in mouflon and sheep was performed. However, no genotype at any locus could be related to any of the investigated environmental variables with significant statistical support. The absence of positive results could be due to a genuine lack of correlation among the genotypes and the

environmental variables analysed. However, it must also be stated that the analysis could only be performed on a reduced subset of geo-referenced individuals (10 sheep and 17 mouflon), and sample size has a major effect on several aspect of the analysis. Firstly, larger sample size allows a more accurate description of the marker frequency, whereas small sample sizes can wrongly approximate the true value. Secondly, more samples can provide a stronger habitat association or can cover it more efficiently, effectively providing a better sampling of the environmental variables, and consequently increasing the chances of finding a correlation if it exists.

The statistical power of the analysis and the necessary sample sizes required to improve it, given the dataset available, were therefore estimated. The majority of the models obtained had little or no power to detect any signal; however, 5,087 models with power >0.9 and representing all 20 the environmental variables used in this work could also be found. Yet, no significant correlation between genotype and environmental variable was detected, suggesting that no statistically significant correlation exists, at least for this sub-set of models. When the sample size needed to improve the overall statistical power of the analysis was inferred, hundreds of individuals were found to be necessary in order to have the statistical power to pick-up a correlation, if it exists, given the same sampling of environmental variables.

Local adaptation to environmental variables can be subtle and difficult to detect (Rellstab et al., 2015) and accurately sampling the diversity of the landscape is fundamental to identify positive results (Rellstab et al., 2015). Consequently, several sampling strategies have been identified to maximise the chance of recording signals of adaptation if present (Manel et al., 2010; Lotterhos and Whitlock, 2015). However, especially in the case of wild and elusive animals, refined sampling strategies might be not possible, with the dataset being limited by logistical constraints, as for the Sardinian mouflon here. As a consequence, the majority of the landscape variable distributions were highly discontinuous and their range poorly represented (e.g., only three isothermality values were sampled for both sheep and mouflon, Figure S 4-6).

The reduced sampling explored here with landscape genomics tools is highly unlikely to have been sufficient to identify signals of local adaptation. The power analyses performed showed that although a small percentage of the results provided true negative results, the vast majority (~95%) of the negative results were due to a lack of statistical power as a consequence of the reduced sampling, and could not have identified a putative signal even if it existed. Consequently, no inference can be confidently made on the effect of local adaptation on the Sardinian population investigated.

However, I have suggested a methodological framework to assess the capability of the logistic regression-based method given a specific dataset (power and sample size analyses), as an investigatory and diagnostic procedure in landscape genomics analyses. Firstly, to assess the nature of negative results as we did in this work. Secondly, as an exploratory tool to evaluate the quality of the sampling at hand in terms of signal discovery power. Indeed, landscape genomic studies can, and often do, include large sampling (e.g., >800 individuals in Stucky et al., 2014). These numbers, even when the latest informatic tools are applied (Stucki et al., 2014), translate into extremely large computational time, according to the number of variables considered in the analysis. However, large numbers of individuals may not necessarily reflect sufficient sampling of their environment. It is the suggested, therefore, that several randomised sub-samples of a dataset could be analysed in parallel, their power assessed, and appropriate considerations drawn according to the results.

4.5.3 A note on sampling: *it is easier for a camel to pass through the eye of a needle than to sample mouflons or sheep in Sardinia*

The research project for this thesis was designed to have an explicit focus on the local adaptation of feral and domestic animals using the Sardinian mouflon and sheep as models. As mentioned in 1.4.4, landscape sampling is a function of the sampling coverage and a fundamental step in the experimental design. My expectation was to be able to sample Sarda sheep across the whole island of Sardinia. Such coverage, along with the high heterogeneity of the Sardinian

landscape, would have provided a promising picture of locally adaptive features in Sarda sheep. Such data could have been then compared with the local adaptation results obtained from mouflon at different spatial scales, and the convergence of traits potentially been used to draw general conclusion on local adaptation strategies for ovines on the island. In the following sections I will discuss the main factors that hindered the collection of both mouflon and sheep samples.

4.5.3.1 Mouflon samples

Among the samples collected and then analysed in this thesis are two Sardinian mouflon populations. Once heavily hunted (Cetti, 1774; Biagini, 1948), the Sardinian Mouflon has recently been granted the status of endangered and is protected from hunting in the whole region (see 1.1.4.2). Consequently, while the mouflon samples collected in Hungary (chapter 3) were collected from hunted animals, this sampling strategy could not be carried out in Sardinia.

The alternative available options were therefore reduced to:

1) Non-invasive sampling.
2) Biopsy darts.
3) Trapping.

The use of non-invasive sampling (e.g., hairs, feathers or faeces) is feasible when few microsatellites or small portions of the mtDNA are foreseen to be analysed (Taberlet and Luikart, 1999). The use of SNP genotyping arrays however requires good quality DNA, difficult to obtain from samples other than peripheral blood or tissue, hence, making non-invasive sampling unsuitable. Biopsy dart sampling would have provided the needed samples. Unfortunately, to approach mouflons at the distance required to ensure a clean shot (<10 m) is extremely difficult due to mouflon's vigilant behaviour and the harsh mountainous terrain. Although a long distance shot could be attempted (<70m) the inaccuracy of such a shot could put the animal at risk if anything different from a large well-muscled area is hit (Karesh et al., 1987). Trapping strategies involving cages and baits can be extremely effective when a desirable bait can be provided, e.g., salt is recognised to be very efficient for several species of the genus *Capra* (Perez et al., 1997).

However, mouflons are extremely suspicious and their nutritional needs are well met by the habitat, making trapping with baits unfeasible (Marco Apollonio, *personal communication*).

A successful method normally employed by the Sardinian forestry department to collect animals for translocation is represented by the use of hundred meter-long nets, set at the edges of forested areas, and several beaters to push flocks towards the nets. This approach is the safest for the animals, but requires a large deployment of manpower, and the rate of success is rather low as the mouflon often change direction when they sense the presence of the nets (Marco Apollonio, *personal communication*). When mouflon are captured, a health check is performed by Regional Animal health authorities operating in the Gennargentu area, and blood samples are opportunistically taken. The mouflon material collected for this thesis project was provided by the institute, along with biopsies taken from poached mouflon carcasses found by the forestry department. The samples are therefore temporally heterogeneous, with the oldest sample collected in 2001 and the most recent in 2013.

4.5.3.2 Sheep samples

Despite been confined in enclosures, the collection of sheep samples proved to be more difficult and by far less successful even than the mouflon sampling. No technical difficulty was encountered apart from the necessity of having veterinary sampling the peripheral blood. Yet, the bureaucracy involved in coordinating veterinarians from different districts of the island and the uncooperative nature of some Sardinian agricultural institutes proved challenging to the extent of having to reconsider the focal aim of the project. The Sarda sheep samples here used were sampled by the personnel of the same institute that provided the mouflon samples within the boundaries of their district jurisdiction.

4.6 Acknowledgements

Many thanks to Elia for useful comments and advices on the landscape genomics analyses.

4.7 Supplementary materials

Table S 4-1 Complete list of genes identified under selection in sheep (A) and mouflon (B). In brackets is the chromosome were the gene is mapped.

A) Genes under selection in Sarda sheep

GJB5 (1)	MYCL (1)	SLC4A10 (2)	BTBD11 (3)	C19orf25 (5)	PARM1 (6)	AREL1 (7)	PRPF8 (11)	XPR1 (12)	E2F4 (14)	SNORA73 (15)	SERPINA1 (18)	DRAP1 (21)	TMEM159 (24)
GJB4 (1)	MFSD2A (1)	ITGB6 (2)	PWP1 (3)	PCSK4 (5)	WDFY3 (6)	ECHDC1 (8)	TLCD2 (11)	KIAA1614 (12)	ELMO3 (14)	FAM169A (16)	SERPINA14 (18)	TSGA10IP (21)	ZP2 (24)
U6 (1)	CAP1 (1)	PLA2R1 (2)	SLCO1C1 (3)	PCYOX1L (5)	C4orf36 (6)	U6atac (8)	oar-mir-22 (11)	STX6 (12)	LRRC29 (14)	NSA2 (16)	SERPINA12 (18)	SART1 (21)	ANKS4B (24)
GJA4 (1)	PPT1 (1)	SNORA21 (2)	PIK3C2G (3)	IL17B (5)	KLHL8 (6)	SLC35F1 (8)	WDR81 (11)	MR1 (12)	LRRC36 (14)	GFM2 (16)	SERPINA5 (18)	EIF1AD (21)	OTOA (24)
SMIM12 (1)	SNORA79 (1)	CD302 (2)	MAGI2 (4)	oar-mir-143 (5)	HSD17B11 (6)	FAM162B (8)	SERPINF2 (11)	PRG4 (12)	TPPP3 (14)	IPO11 (16)	SERPINA4 (18)	BANF1 (21)	HS3ST2 (24)
DLGAP3 (1)	RLF (1)	BAZ2B (2)	GSAP (4)	CSNK1A1 (5)	HSD17B13 (6)	KPNA5 (8)	SERPINF1 (11)	TPR (12)	ZDHHC1 (14)	SMIM15 (16)	TGFBR2 (19)	CST6 (21)	HS3ST4 (24)
SFPQ (1)	SMAP2 (1)	CCDC146 (4)	ABHGEF37 (5)	ZBTB49 (6)	SCML4 (8)	SMYD4 (11)	C1orf27 (12)	FAM65A (14)	NDUFAF2 (16)	GADD1 (19)	CATSPER1 (21)	ACTN2 (25)	
ZMYM4 (1)	ZFP69 (1)	PKP4 (2)	FGL2 (4)	SLC6A7 (5)	LYAR (6)	SOBP (8)	RPA1 (11)	PARD3 (13)	CTCF (14)	ERCC8 (16)	SNORA11 (19)	GAL3ST3 (21)	Y_RNA (25)
KIAA0319L (1)	EXOS (1)	CCDC148 (2)	FAM185A (4)	CAMK2 (5)	TMEM128 (6)	PDSS2 (8)	RTN4RL1 (11)	NRP1 (13)	RLTPR (14)	ELOVL7 (16)	CTNNB (19)	SF3B2 (21)	MTR (25)
NCDN (1)	ZNF684 (1)	U12 (2)	FBXL13 (4)	ARSI (5)	OTOP1 (6)	BEND3 (8)	DPH1 (11)	ITGB1 (13)	ACD (14)	DEPDC1B (16)	ULK4 (19)	PACS1 (21)	BICC1 (25)
TFAP2E (1)	RIMS3 (1)	GALNT5 (2)	LRRC17 (4)	TCOF1 (5)	KCNN2 (7)	C6orf203 (8)	OVCA2 (11)	GPR158 (13)	PARD6A (14)	PART1_2 (16)	MAGI1 (19)	MRGPRF (21)	TMEM26 (25)
PSMB2 (1)	NFYC (1)	GPD2 (2)	ARMC10 (4)	CD74 (5)	TRIM36 (7)	GPR63 (8)	OPTN (13)	ENKD1 (14)	PART1_1 (16)	C3orf67 (19)	TPCN2 (21)	C10orf107 (25)	
C1orf216 (1)	oar-mir-30c (1)	NR4A2 (2)	NAPEPLD (4)	NDST1 (5)	PGGT1B (7)	MANEA (8)	SMG6 (11)	MCM10 (13)	C16orf86 (14)	PDE4D (16)	FAM3D (19)	CTTN (21)	ARID5B (25)
CLSPN (1)	KCNQ4 (1)	U1 (2)	PMPCB (4)	SYNPO (5)	CCDC112 (7)	NHSL1 (8)	USP43 (11)	PHYH (13)	GFOD2 (14)	RAB3C (16)	FAM107A (19)	SHANK2 (21)	DDX21 (25)
AGO4 (1)	CITED4 (1)	GALNT13 (2)	DNAJC2 (4)	MYO23 (5)	FEM1C (7)	RAB32 (8)	DHRS7C (11)	SEPHS1 (13)	RANBP10 (14)	C5orf42 (16)	ACOX2 (19)	PLCE1 (22)	KIAA1279 (25)
AGO1 (1)	CTPS1 (1)	AR16IP6 (2)	P5MC2 (4)	RBM22 (5)	THOC3 (7)	ESR1 (8)	GLP-2R (11)	FAM107B (13)	TSNAXIP1 (14)	NIPBL (16)	KCTD6 (19)	NOC3L (22)	HKDC1 (25)
AGO3 (1)	SLFNL1 (1)	PRPF40A (2)	SLC26A5 (4)	DCTN4 (5)	CPLX2 (7)	ARID1B (8)	RCVRN (11)	CDNF (13)	CENPT (14)	MAP9 (17)	ABHD6 (19)	TBC1D12 (22)	HK1 (25)
TEKT2 (1)	SCMH1 (1)	FMNL2 (2)	RELN (4)	SMIM3 (5)	HRH2 (7)	LDHAL6B (8)	GAS7 (11)	HSPA14 (13)	NUTF2 (14)	SFT2D1 (17)	DNASE1L3 (19)	HELLS (22)	MICU1 (25)
ADPRHL2 (1)	FOXO6 (1)	LYPD6 (2)	TMEM168 (4)	ZNF300 (5)	AGGF1 (7)	TMEM242 (8)	MRPL45 (11)	SNORD22 (13)	EDC4 (14)	PDGFC (17)	CDCP1 (19)	SORBS1 (22)	MCU (25)
COL8A2 (1)	EDN2 (1)	LYPD6B (2)	NUDCD3 (4)	TNIP1 (5)	SLC12A6 (7)	ZDHHC14 (8)	GPR179 (11)	SUV39H2 (13)	PSKH1 (14)	ZNF84 (17)	CLEC3B (19)	ALDH18A1 (22)	OIT3 (25)
TRAPPC3 (1)	HIVEP3 (1)	KIF5C (2)	CAMK2B (4)	ANXA6 (5)	EMC4 (7)	KCNQ5 (9)	SOCS7 (11)	DCLRE1C (13)	PSMB10 (14)	ZNF140 (17)	EXOSC7 (19)	TCTN3 (22)	PLA2G12B (25)
MAP7D1 (1)	GUCA2B (1)	ORC4 (2)	ING3 (4)	CCDC69 (5)	EMC7 (7)	RIMS1 (9)	TLK2 (11)	MEIG1 (13)	LCAT (14)	PXMP2 (17)	ZDHHC3 (19)	ENTPD1 (22)	MYO2L (25)
THRAP3 (1)	GUCA2A (1)	ACVR2A (2)	CPED1 (4)	FAT2 (5)	CHRM5 (7)	oar-mir-30a (9)	MRC2 (11)	OLAH (13)	SLC12A4 (14)	POLE (17)	TMEM42 (19)	PIK3AP1 (22)	SEC24C (25)
SH3D21 (1)	FOXI3 (1)	NBAS (3)	WNT16 (4)	SPARC (5)	UBR1 (7)	B3GAT2 (9)	TANC2 (11)	ACBD7 (13)	DPEP3 (14)	P2RX2 (17)	GHRL (19)	LCOR (22)	FUT11 (25)
EVA1B (1)	RIMKLA (1)	DDX1 (3)	FAM3C (4)	ATOX1 (5)	GNG2 (7)	SMAP1 (9)	SLC41A1 (12)	ARMC4 (13)	DPEP2 (14)	LRCOL1 (17)	SEC13 (19)	C10orf12 (22)	CHCHD1 (25)
STK40 (1)	ZMYND12 (1)	FAM49A (3)	PTPR21 (4)	G3BP1 (5)	C14orf166 (7)	ADGRB3 (9)	PM20D1 (12)	NETO2 (14)	ZNF550 (14)	FBRSL1 (17)	ATP2B2 (19)	SLIT1 (22)	ZSWIM8 (25)
LSM10 (1)	PPCS (1)	RAD51AP2 (3)	AASS (4)	GLRA1 (5)	NID2 (7)	LY6K (9)	SLC26A9 (12)	PDLIM1 (14)	ZNF549 (14)	NOC4L (17)	KHDRBS2 (20)	SNORD88 (22)	CAMK2G (25)
OSCP1 (1)	CCDC30 (1)	VSNL1 (3)	FEZF1 (4)	CCNG1 (5)	RPS27L (7)	PKHD1L1 (9)	PSEN2 (12)	ABCC12 (14)	POU2AF1 (15)	DDX51 (17)	TNFRSF21 (20)	ATRNL1 (22)	ADK (25)
MRPS15 (1)	ST3GAL3 (1)	SWC6 (3)	CADPS2 (4)	HMMR (5)	LACTB (7)	NUDCD3 (9)	AHCTF1 (12)	ABCC11 (14)	BTG4 (15)	EP400 (17)	CD2AP (20)	FGFR2 (22)	OPN4 (25)
CSF3R (1)	IPO13 (1)	GEN1 (3)	TAS2R16 (4)	TMEM167A (5)	TPM1 (7)	SNORA40 (10)	SCCPDH (12)	ZNF423 (14)	C11orf88 (15)	SNORA49 (17)	ADGRF2 (20)	ATE1 (22)	LDB3 (25)
GRIK3 (1)	DPH2 (1)	MSGN1 (3)	SLC13A1 (4)	snoU109 (5)	ANXA2 (7)	CCNA1 (10)	PLD5 (12)	ADCY7 (14)	LAYN (15)	PUS1 (17)	OPN5 (20)	ZNRD1-AS1_2 (22)	FECB (25)
ZC3H12A (1)	ATP6V0B (1)	KCNS3 (3)	IQUB (4)	XRCC4 (5)	ALDH1A2 (7)	SPG20 (10)	XCL1 (12)	BRD7 (14)	FDXACB1 (15)	ULK1 (17)	PTCHD4 (20)	NSMCE4A (22)	CYP4V2 (26)
MEAF6 (1)	B4GALT2 (1)	FOSL2 (3)	U7 (4)	SPATA9 (5)	MYZAP (7)	NBEA (10)	F5 (12)	NKD1 (14)	C11orf1 (15)	MMP17 (17)	C6orf141 (20)	TACC2 (22)	KLKB1 (26)
C1orf109 (1)	KLF17 (1)	PPP1CB (3)	ASB15 (4)	RHOBTB3 (5)	NEDD4 (7)	RFC3 (10)	SELP (12)	MMP2 (14)	CRYAB (15)	SFSWAP (17)	RHAG (20)	BTBD16 (22)	F11 (26)
CDCA8 (1)	DMAP1 (1)	SPDYA (3)	LMOD2 (4)	LNPEP (5)	PRTG (7)	UBL3 (10)	METTL18 (12)	LPCAT2 (14)	C11orf52 (15)	ADGRD1 (17)	PGK2 (20)	PLEKHA1 (22)	TRIML2 (26)
MTF1 (1)	ERI3 (1)	RPS26 (3)	U3 (4)	CAMK2D (6)	SLC12A1 (7)	CCL1 (11)	C1orf112 (12)	SLC6A2 (14)	DIXDC1 (15)	RGMA (18)	CRISP1 (20)	C10orf120 (22)	TRIML1 (26)
INPP5B (1)	APP (1)	YPEL5 (3)	HYAL4 (4)	PPP3CA (6)	CTXN2 (7)	CCL8 (11)	SCYL3 (12)	NUP93 (14)	PIH1D2 (15)	ZO1 (18)	PKHD1 (20)	CUZD1 (22)	MSR1 (26)
SF3A3 (1)	7SK (1)	CCDC85A (3)	TMEM229A (4)	DHX15 (6)	MYEF2 (7)	CCL11 (11)	KIFAP3 (12)	SLC12A3 (14)	C11orf57 (15)	FOXG1 (18)	IL-17 (20)	DSEL (23)	KAT6A (26)
FHL3 (1)	NCAM2 (1)	ATXN7L3B (3)	C19orf26 (5)	SOD3 (6)	SLC24A5 (7)	U4 (11)	PTCHD2 (12)	HERPUD1 (14)	TIMMBB (15)	NUBPL (18)	ZNF311 (20)	CDH19 (23)	AP3M2 (26)
UTP11L (1)	TMPRSS15 (1)	KCNC2 (3)	ATP5D (5)	CCDC149 (6)	SAMD4A (7)	NOS2 (11)	UBIAD1 (12)	NLRC5 (14)	IL18 (15)	HTR1D (18)	Metazoa_SRP (20)	CELF4 (23)	PLAT (26)
RRAGC (1)	CHODL (1)	5S_rRNA (3)	MIDN (5)	LGI2 (6)	SLC38A6 (7)	VPS53 (11)	MTOR (12)	CES3 (14)	TEX12 (15)	FBLN5 (18)	FAT3 (21)	KIAA1328 (23)	IKBKETA (26)
GJA9 (1)	C21orf91 (1)	NUAK1 (3)	CIRBP (5)	SEPSECS (6)	PRKCH (7)	FAM101B (11)	ANGPTL7 (12)	CES4A (14)	BCO2 (15)	TRIP11 (18)	DLG2 (21)	CCDC178 (23)	POLB (26)
RHBDL2 (1)	BTG3 (1)	CKAP4 (3)	C19orf24 (5)	ARAP2 (6)	HIF1A (7)	C17orf97 (11)	EXOSC10 (12)	CBFB (14)	PLET1 (15)	ATXN3 (18)	SNX32 (21)	GAREM (23)	ERBA BETA1 (26)
AKIRIN1 (1)	CXADR (1)	TCP11L2 (3)	EFNA2 (5)	DTHD1 (6)	SNAPC1 (7)	RPH3AL (11)	SRM (12)	C16orf70 (14)	TTC12 (15)	CPSF2 (18)	CFL1 (21)	LPIN2 (23)	
NDUFS5 (1)	oar-let-7c (1)	POLR3B (3)	MUM1 (5)	ERVW-1 (6)	SYT16 (7)	DOC2B (11)	MASP2 (12)	B3GNT9 (14)	ANKK1 (15)	SLC24A4 (18)	MUS81 (21)	MYOM1 (23)	
U2 (1)	oar-mir-99a (1)	RFX4 (3)	NDUFS7 (5)	TMPRSS11A (6)	FUT8 (7)	CRK (11)	APITD1 (12)	TRADD (14)	DRD2 (15)	RIN3 (18)	EFEMP2 (21)	MRC13 (23)	
NT5C1A (1)	TLE1 (2)	RIC8B (3)	GAMT (5)	TMPRSS11F (6)	DPF3 (7)	MYO1C (11)	PGD (12)	HSF4 (14)	TMPRSS5 (15)	LGMN (18)	CTSW (21)	MYL12B (23)	
HPCAL4 (1)	FAP (2)	TMEM263 (3)	DAZAP1 (5)	MUC7 (6)	SYNDIG1L (7)	INPP5K (11)	KIF1B (12)	NOL3 (14)	HTR3B (15)	TMEM251 (18)	FIBP (21)	VAPA (23)	
PPIE (1)	GCG (2)	MTERF2 (3)	RPS15 (5)	MOB1B (6)	NPC2 (7)	SLC43A2 (11)	RERE (12)	KIAA0895L (14)	NXPE4 (15)	BTBD7 (18)	FOSL1 (21)	WDR7 (23)	
TRIT1 (1)	DPP4 (2)	CRY1 (3)	APC2 (5)	BTC (6)	LTBP2 (7)	SCARF1 (11)	SLC45A1 (12)	EXOC3L1 (14)	NXPE2 (15)	COX8C (18)	C11orf68 (21)	GSPT1 (24)	

B) Genes under selection in mouflon.

FGGY (1)	MRPL35 (3)	EPDR1 (4)	C14orf37 (7)	TMEM68 (9)	CD7 (11)	RASAL2 (12)	SLC25A1 (17)	ALG8 (21)
HOOK1 (1)	REEP1 (3)	AMPH (4)	TTLL5 (7)	TGS1 (9)	CSNK1D (11)	DENND1B (12)	HIRA (17)	THRSP (21)
DNASE2B (1)	SNORA68 (3)	CHCHD3 (4)	TMEM63C (7)	LYN (9)	SLC16A3 (11)	LHX9 (12)	MRPL40 (17)	OR8D4 (21)
GAP43 (1)	DYSF (3)	MRPS33 (4)	SMPDL3A (8)	RPS20 (9)	GRB2 (11)	HAO1 (13)	C22orf39 (17)	OR6T1 (21)
CCDC50 (1)	MYPT1 (3)	AGK (4)	FABP7 (8)	snoU54 (9)	SLC25A19 (11)	TGIF2 (13)	UFD1L (17)	OR10G6 (21)
PIK3CB (1)	CEP83 (3)	KIAA1147 (4)	PKIB (8)	U1 (9)	MIF4GD (11)	TGIF2-C20orf24 (13)	CDC45 (17)	PDCD4 (22)
UBE2R2 (2)	TMCC3 (3)	CSNK1G3 (5)	MAP3K7 (8)	MOS (9)	MRPS7 (11)	SLA2 (13)	CERS3 (18)	SHOC2 (22)
DCAF12 (2)	ATF7 (3)	SNCAIP (5)	TBX18 (8)	PLAG1 (9)	GGA3 (11)	NDRG3 (13)	SYNPR (19)	ADRA2A (22)
UBAP2 (2)	NPFF (3)	BBS7 (6)	MAP3K5 (8)	CHCHD7 (9)	NUP85 (11)	ZNFX1 (13)	TDP2 (20)	TECTB (22)
CPO (2)	TARBP2 (3)	CCNA2 (6)	NSMCE2 (9)	SDR16C5 (9)	TMEM104 (11)	ZNFX1-AS1_1 (13)	KIAA0319 (20)	ADAM12 (22)
EHMT1 (3)	MAP3K12 (3)	EXOSC9 (6)	KIAA0196 (9)	SNORA42 (9)	NAT9 (11)	ZNFX1-AS1_2 (13)	ALDH5A1 (20)	NOL4 (23)
MBOAT2 (3)	uc_338 (3)	SYNPO2 (6)	SQLE (9)	TNFRSF11B (9)	SLC9A3R1 (11)	SNORD12 (13)	GPLD1 (20)	PSTPIP2 (23)
C2orf43 (3)	PRR13 (3)	SEC24D (6)	ZNF572 (9)	MMP16 (9)	RGS13 (12)	KCNB1 (13)	MRS2 (20)	AUTS2 (24)
U6 (3)	AMHR2 (3)	GABRA2 (6)	5S_rRNA (9)	KLF12 (10)	RGS1 (12)	CDH11 (14)	DCDC2 (20)	RYR2 (25)
POLR1A (3)	SLC38A2 (3)	ATP10D (6)	SPIDR (9)	GDPD1 (11)	USH2A (12)	DPY19L3 (14)	NRSN1 (20)	FAM13C (25)
SNORA76 (3)	BICD1 (3)	CORIN (6)	PRKDC (9)	MYO1D (11)	ACBD3 (12)	ANKRD27 (14)	EDN1 (20)	NRBF2 (25)
PTCD3 (3)	ETNK1 (3)	TBCA (7)	ST18 (9)	NF1 (11)	SNORA70 (12)	NUDT19 (14)	ECI2 (20)	JMJD1C (25)
SNORD94 (3)	SEMA3E (4)	NR2E3 (7)	RGS20 (9)	EVI2B (11)	PER3 (12)	ANO3 (15)	C6orf201 (20)	REEP3 (25)
IMMT (3)	NME8 (4)	MYO9A (7)	TCEA1 (9)	OMG (11)	CAMTA1 (12)	SPDL1 (16)	FAM217A (20)	DNAJC12 (25)
SNORA51 (3)	SFRP4 (4)	WDR72 (7)	LYPLA1 (9)	UTS2R (11)	PAPPA2 (12)	DGCR14 (17)	PRPF4B (20)	SORBS2 (26)

Table S 4-2 Environmental variables and abbreviations used in this study.

Abbreviation	***Environmental variable***
NUORO_alti	*Altitude*
NUORO_slop	*Slope*
bio2 _16	*Mean Diurnal Range (Mean of monthly (max temp - min temp))*
bio3 _16	*Isothermality (BIO2/BIO7) (* 100)*
bio4 _16	*Temperature Seasonality (standard deviation *100)*
bio5 _16	*Max Temperature of Warmest Month*
bio6 _16	*Min Temperature of Coldest Month*
bio7 _16	*Temperature Annual Range (BIO5-BIO6)*
bio8 _16	*Mean Temperature of Wettest Quarter*
bio9 _16	*Mean Temperature of Driest Quarter*
bio10 _16	*Mean Temperature of Warmest Quarter*
bio11 _16	*Mean Temperature of Coldest Quarter*
bio12 _16	*Annual Precipitation*
bio13 _16	*Precipitation of Wettest Month*
bio14 _16	*Precipitation of Driest Month*
bio15 _16	*Precipitation Seasonality (Coefficient of Variation)*
bio16 _16	*Precipitation of Wettest Quarter*
bio17 _16	*Precipitation of Driest Quarter*
bio18 _16	*Precipitation of Warmest Quarter*
bio19 _16	*Precipitation of Coldest Quarter*

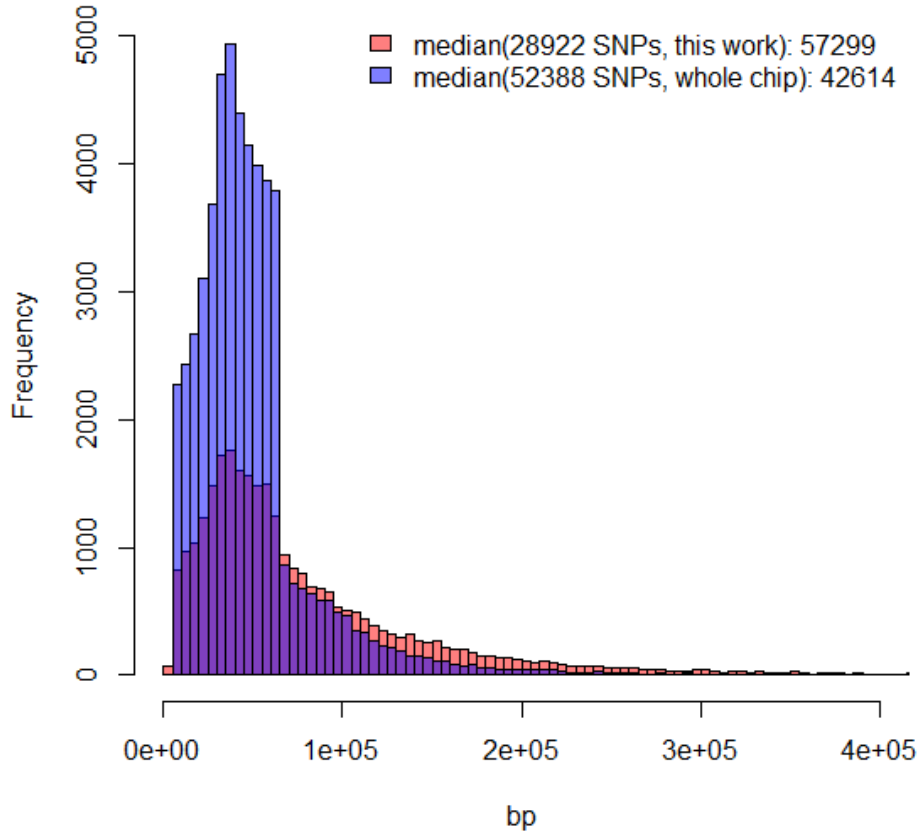
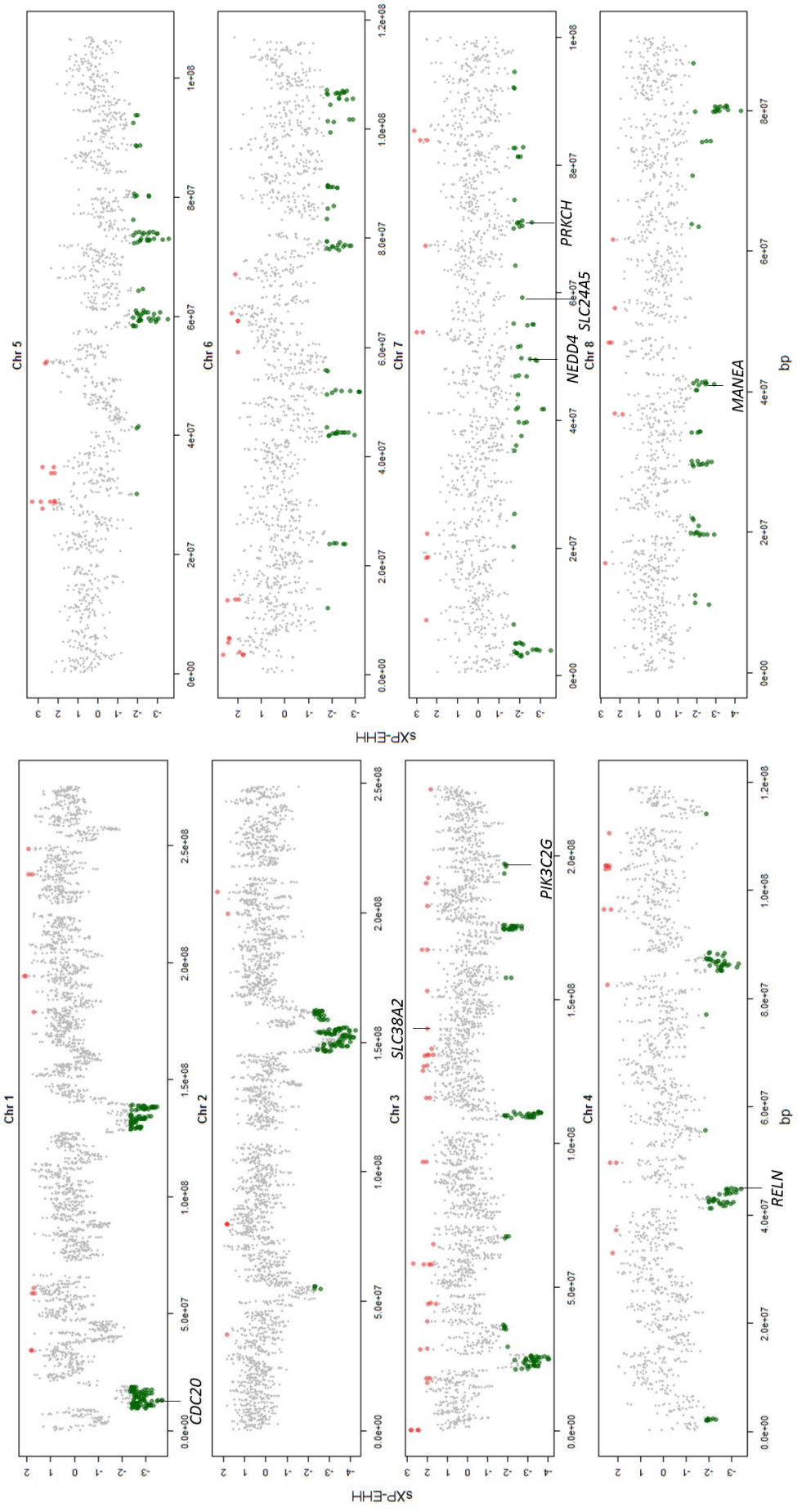
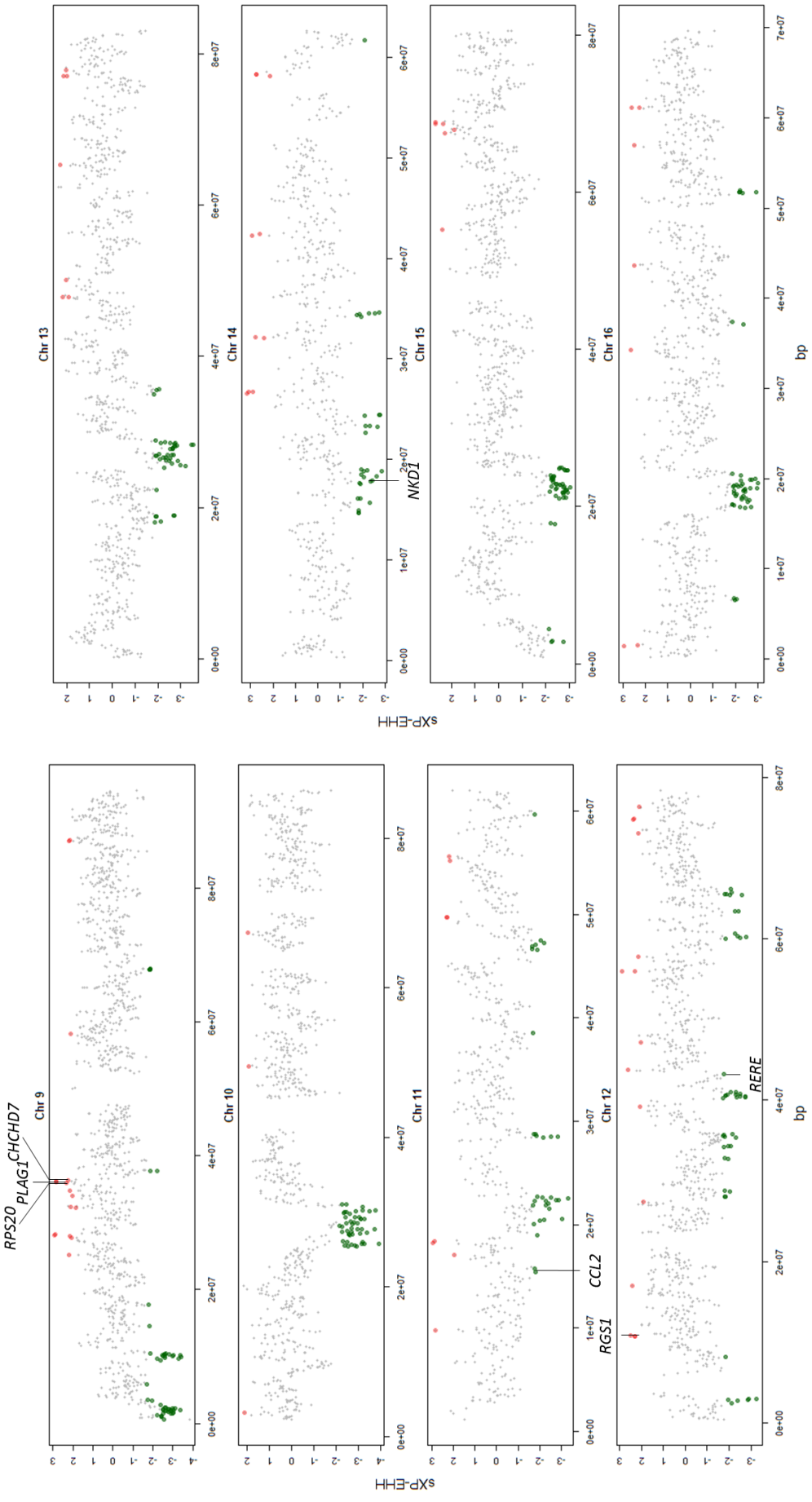
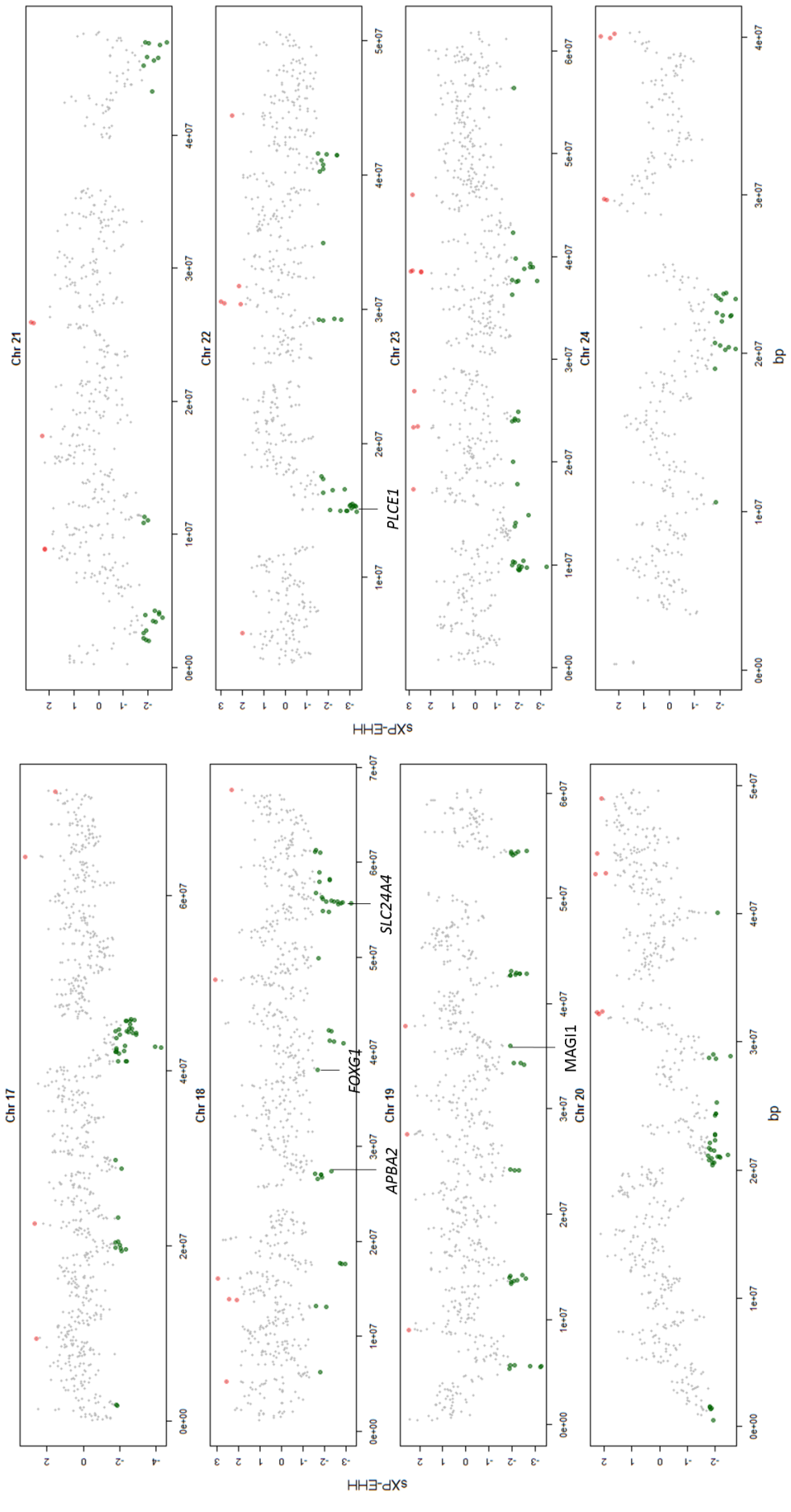


Figure S 4-1 Comparison between the distributions of the distances between post-analysis neighbouring SNPs represented in the selection analysis results (red) and in the full set of autosomal markers in the Ovine SNP chip (blue). The histogram is truncated at 400,000 bp for readability.


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
CDC20
SLC38A2
PIK3C2G
RELN
NEDD4
SLC24A5
PRKCH
MANEA
bp
bp
sXP-EHH
sXP-EHH


Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
RPS20
PLAG1
CHCHD7
NKD1
CCL2
RGS1
RERE
XP-EHH
bp


Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23
Chr 24
FOXG1
SLC24A4
APBA2
MAGI1
PLCE1
bp
bp
xXP-EHH
xXP-EHH

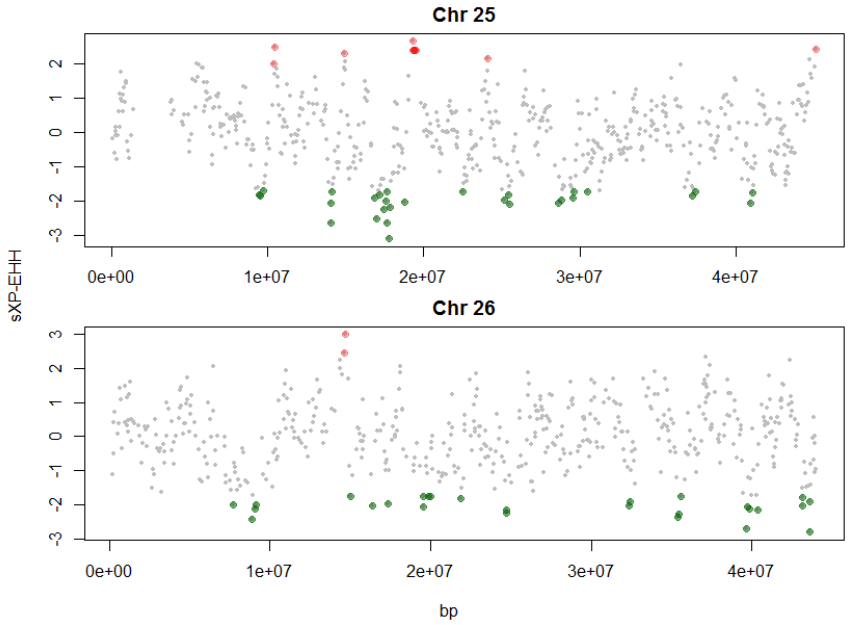


Figure S 4-2 Standardised XP-EHH results for 26 autosomes. On top of each graph is the chromosome (Chr) number. The standardised XP-EHH score value (sXPEHH) for each locus analysed according to its position in the chromosome (bp) is represented by a grey dot. The extreme 5% of all the loci for each chromosome with p-value <0.05 are encircled in green and red (bottom and top part of the figure, respectively), representing SNPs under putative selection for sheep and mouflon, respectively. The genes from table 2 are reported in these figures in the corresponding chromosome and physical position.

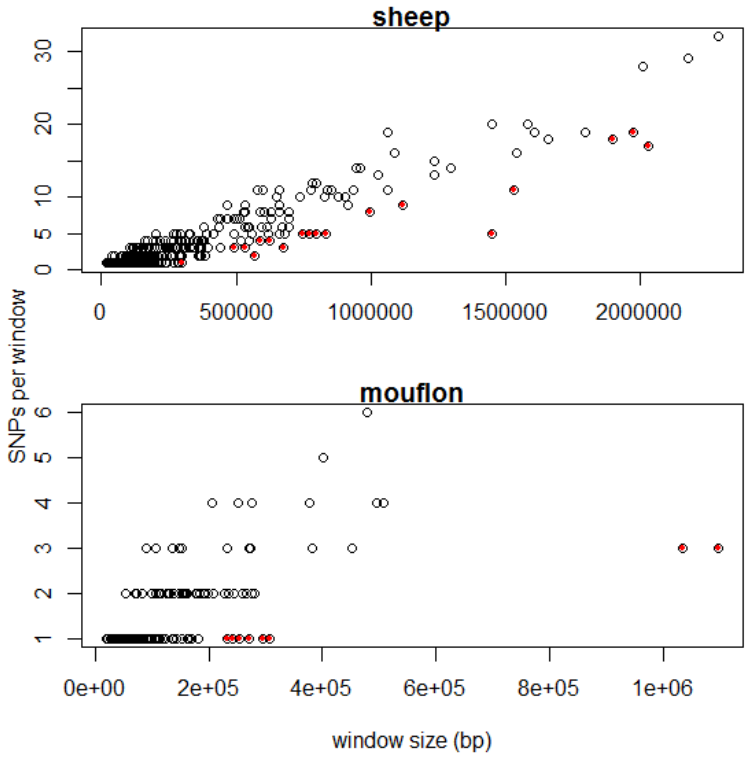


Figure S 4-3 Correlation between window size and number of SNPs for each of the windows identified as under selection in sheep (top) and mouflon (bottom). Linear regression was performed for both sheep and mouflon data and those windows represented in the 5% lower tail of the residual distribution were removed as outliers (filled red circles).

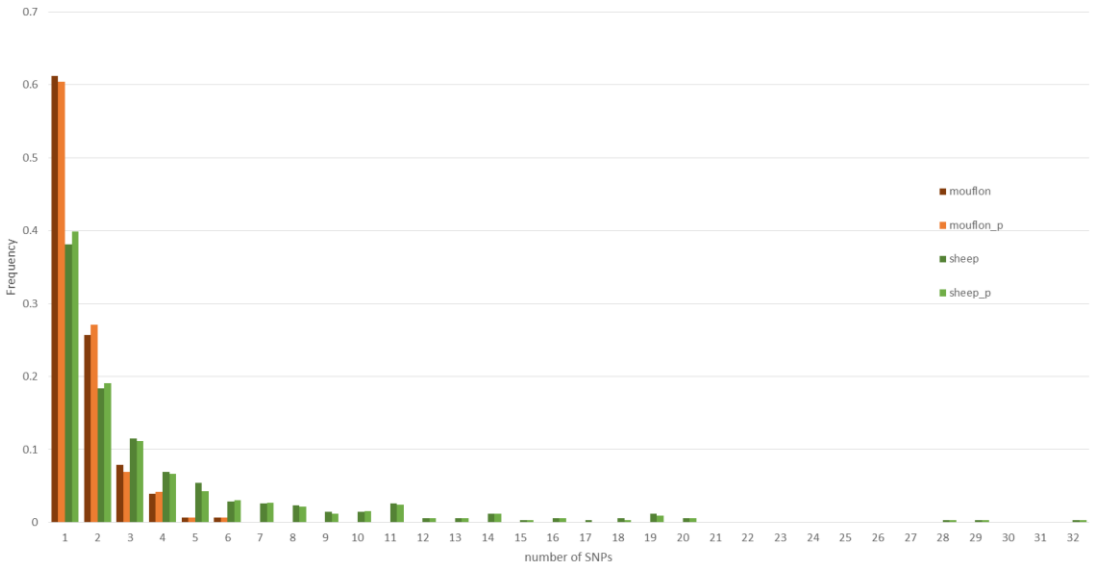


Figure S 4-4 Window size distribution (according to the number of loci) for both sheep and mouflon. The windows size frequency is shown before and after pruning for outliers. Pruned data are identified in the legend with the extension '_p'.

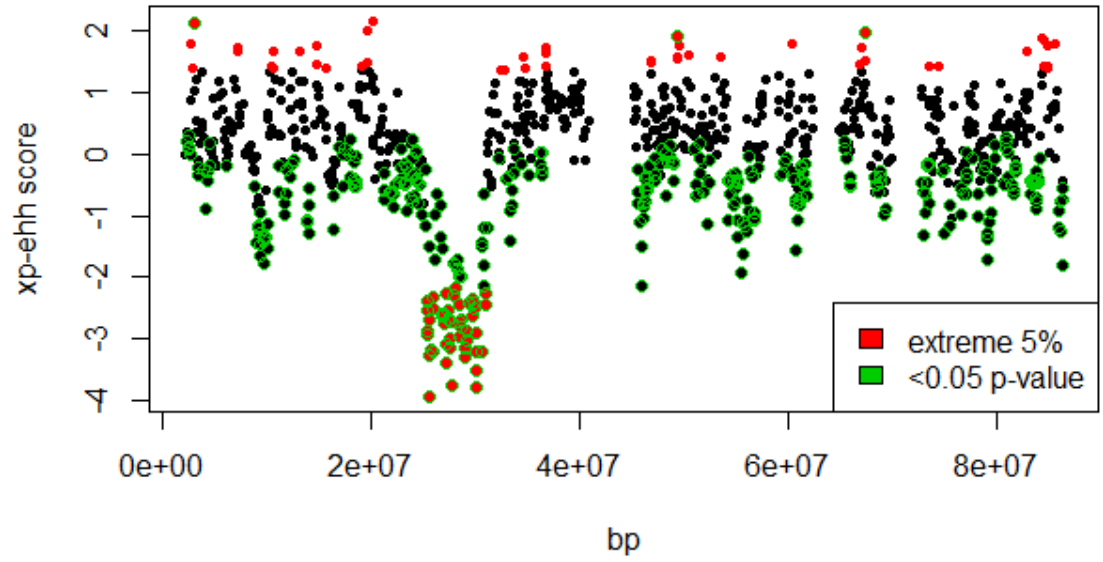


Figure S 4-5 Effect of top scores and p-value filtering on the 10th chromosome. In black are the 956 sXP-EHH scores obtained for chromosome 10. The loci used to identify the windows of selection are those both filled in red (extreme 5% of the chromosome –wide distribution) and encircled in green (p-value < 0.05).

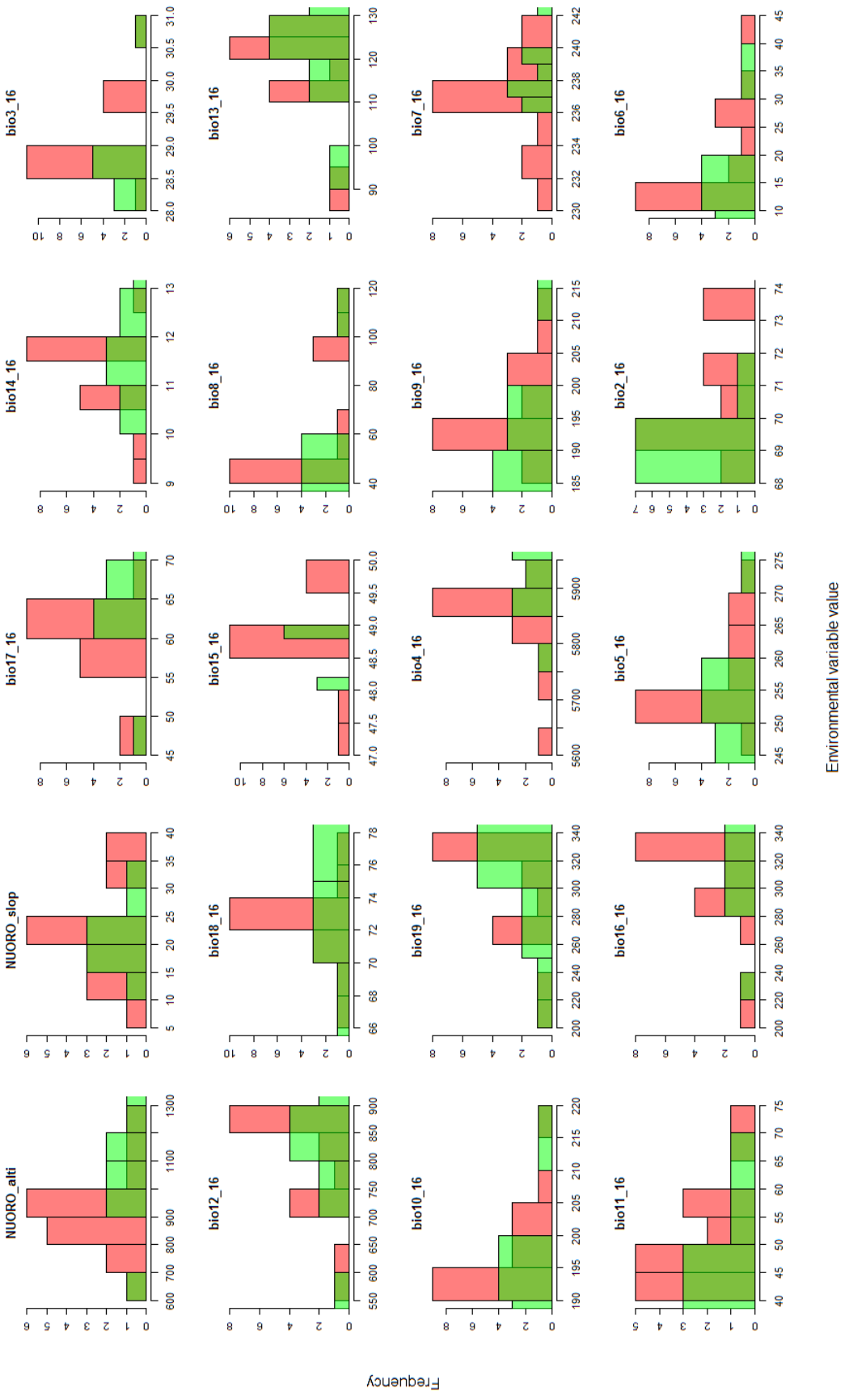
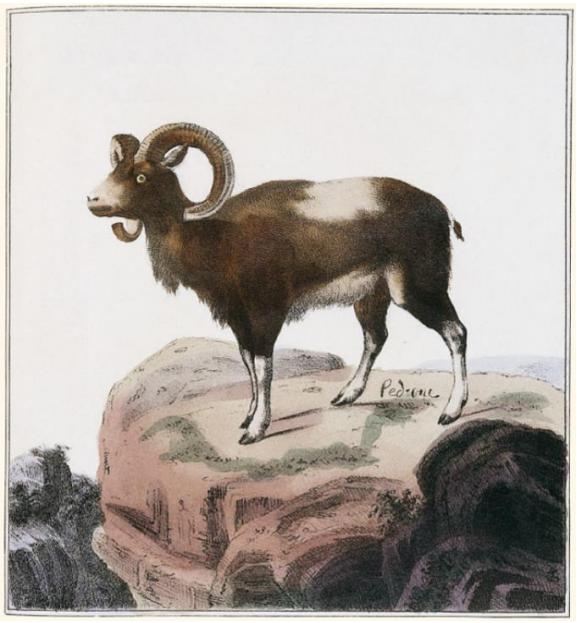


Figure S 4-6 Distribution of the 20 environmental variables sampled by using 17 mouflons (red) and 10 sheep (green). See Table S 4-2 for the abbreviations.

Chapter five



Mouflon, lithography (Baldassarre, 1841)

5 Chapter five – General discussion

5.1 Aims

This thesis aimed to use genomic tools to better understand the effects of natural and artificial selection in sheep by studying SNP array data of feral and domestic populations, mainly in Sardinia. Admixture between mouflon and domestic sheep, signatures of selection and local adaptation, were investigated. Additionally, the utility of commercially available mid-density SNP array data to address these questions in non-model species was evaluated.

5.1.1 Completion of aims

In **chapter 2** the development of *SNeP* (Barbato et al., 2015) was described**.** This computer program provides a computationally efficient alternative to using a combination of pre-existing software and custom scripts (Tenesa et al., 2007; Flury et al., 2010; Corbin et al., 2012) for the estimation of *Ne* from LD. *SNeP* was used at different stages in this thesis to assess *Ne* (**chapter 3**) and estimate the decay in LD (**chapter 4**) in sheep populations. Recently, other researchers have started to use and publish analyses using *SNeP* to investigate *Ne* trends in different species (Orozco-terWengel et al., 2015; Makina et al., 2015; Kim et al., 2015)**.**

In **chapter 3** we investigated the presence, intensity and adaptive significance of crossbreeding between domestic sheep and mouflon populations by studying the largest collection of European mouflon samples to date, comprising sites from the Mediterranean Islands, mainland Europe and Middle-East along with sheep samples overlapping the same regions of past and present mouflon populations. The analyses provided no evidence of diffuse sheep introgression in either the Sardinian or Corsican mouflons. Signals of recent introgression were instead found in a single Sardinian population. However, this population was sampled within an enclosure where crossbred mouflon were known to have been reared. Signature of mouflon introgression in domestic sheep were identified. Additionally, thanks to a novel approach developed to aid the investigation of local ancestry data (Consistently Introgressed Windows of Interest, CIWI), admixture signals of mouflon ancestry in sheep were related to biological

functions involved with innate immunity processes in modern sheep breeds and bitter taste recognition in two breeds known for their broad dietary choices.

In **chapter 4** signals of positive selection and local adaptation in feral and domestic sheep were investigated. By using the cross-population extended haplotype homozygosity (XP-EHH) along with a novel locus-specific empirical p-value inference, signatures of selection were identified in both the Sardinian mouflon and Sarda sheep breed. Traits such as fertility, pigmentation and behaviour were identified in Sarda sheep to be under selection. These traits are common among most of the modern sheep breeds, with the first two features heavily characterising Sarda sheep. Genes involved with stature - probably related to mating success - were found in mouflon. Signals of local adaptation related to environmental variables were investigated in both subspecies. However, no results were obtained, which was likely due to the sample-size analysed, as assessed by post-analysis tests.

5.2 Conclusions

5.2.1 Methodological solutions for low- to mid-density SNP array data

In this thesis an informatic tool (*SNeP*, **chapter 2**, Barbato et al., 2015) and two novel methodological approaches (the Consistently Introgressed Windows of Interest or 'CIWI': **chapters 3** and the allele-specific XP-EHH empirical p-value: **chapter 4**) were developed.

5.2.1.1 Towards SNeP v2.0

SNeP is a computer program implemented to fill a gap in the informatic tools availability for LD-based demography inference. It explicitly aims to become a user-friendly tool in the hands of those population geneticists interested in the recent demographic history for species where large SNP datasets are available. By using *SNeP*, hundreds of thousands SNPs can be efficiently analysed in minutes (Barbato et al., 2015), a graphical user interface makes it extremely easy to use and the output files generated are easy to interpret and to transfer into a variety of post-analysis tools.

Since its publication *SNeP* has been updated to version 1.1 (http://sourceforge.net/projects/snepnetrends/). Thanks to the feedback received from *SNeP* users, several minor improvements have been implemented and few minor bugs identified and fixed. Other improvements, mostly focused on computational optimisation and additional methods for parameter estimation will be available with *SNeP* v2.0, which will be provided in the first half of 2016.

Among the implemented in the latest version is the possibility for the user to 'thin' the dataset to match a user-defined maximum number of loci. The seed used to remove loci can be randomly generated (default) or set by the user. This option enables the analysis of extremely large datasets (millions of SNPs) in a reduced amount of time, while the statistics provided can be used to define the best trade-off between computational time and variance in results. Moreover, multiple analyses can be performed on the same dataset using different seeds and therefore changing the pool of markers analysed each time, consequently allowing the user to define the best thinning parameters that provide consistent results across the analyses.

The theory and application of *Ne* estimation using LD is progressing, with regular improvements in the methods available (e.g., (Saura et al., 2015)). *SNeP* takes advantage of this interest and tries to facilitate the application of *Ne* inference for large datasets. Additionally, while *SNeP*'s primary task is to provide *Ne* estimates, it can be used to investigate the extent of LD at the same time, since the relationship of LD with distance is provided in the standard output. Understanding the distribution of LD is important because it underlies all forms of genetic mapping and is, in part, determined by population history and demography (Amaral et al., 2008; Tenesa et al., 2007; Kijas et al., 2014).

5.2.1.2 CIWIs and XP-EHH empirical p-values to the rescue of sparse data

In this study we used CIWIs to identify genomic regions of concordant ancestry in chromosome painting studies (**chapter 3**) and the XP-EHH empirical p-value approach to generate allele-specific confidence intervals for XP-EHH results (**chapter 4**). Both address analytical barriers by applying investigation methods developed for high-density data to sparser datasets. Importantly, the genotyping of non-model species by means of SNP arrays translates in a reduced call rate that

linearly and inversely correlates with the increasing divergence between the species for which the array was developed and the species to which it was applied (Miller et al., 2012a), where more loci than usual are lost to quality pruning (Kijas et al., 2012) and implementation/method specific issues (**chapter 4**). By applying these methodological frameworks, biologically meaningful and consistent signals could be identified despite the reduced number of loci available for analysis (~55% of the original dataset). Although developed to analyse the dataset used in this thesis, the applicability of the two approaches can be extended to other SNP based investigations. Therefore, a possible application of the methodological frameworks described here and as done in this thesis, is the combined use of SNP arrays data of both the model species used for the chip development and their wild/feral counterpart (e.g., caprine SNP array used both on goat and bezoar) where the chance of confounding signals is higher (**chapter 3**). Additionally, even in the case of an 'ideal' dataset (i.e., the breed analysed have been used in SNP ascertainment), the CIWI approach can help in identifying the most concordant – and possibly relevant – signals of local ancestry across populations, independently of the method used to define the chromosomes local ancestry.

5.2.2 Admixture between feral and domestic sheep

Interbreeding between mouflon and domestic sheep is thought to have happened for more than 2,000 years (Cetti, 1774), was occasionally encouraged in more recent times (Tomiczek and Türcke, 2003) and a recent genetic investigation identified some levels of admixture in a mouflon population sampled in Sardinia (Lorenzini et al., 2011). More generally, hybridisation of wild populations with domestic counterparts can have several – and sometime contrasting – effects: the loss of wild genetic integrity, outbreeding depression, loss of adaptive features for wild living (e.g., adaptation to the environment, coat colour and mimetic), pathogen susceptibility and disease transmission, adaptation to captivity, hybrid vigour, increased growth rates and larger litter size (Goedbloed et al., 2013; Randi, 2008; Fang et al., 2009; Puigcerver et al., 2014; Levin et al., 1996; Rhymer and Simberloff, 1996; Kidd et al., 2009; McDowell, 2002; Hedrick, 2013).

Technologies such as the SNP array genotyping have enabled the genome-wide detection of hybridisation, allowing the detection of several generations of back-crossing (vonHoldt et al., 2013) otherwise barely detectable with fast evolving markers such as microsatellites (Oliveira et al., 2015). This approach has been used to investigate domestics and their wild relatives (e.g., Iacolina et al., 2015). To our knowledge, the study presented here is the first to address the occurrence of crossbreeding among feral and domestic sheep using genome-wide data.

Here we could find no obvious signals of sheep introgression into mouflon (**chapter 3**). Our results cannot exclude the occasional occurrence of hybridization, but we hypothesize that either species management (e.g., the protected status of the Sardinian population or selective hunting in the mainland populations), a large historical population size (historical records from the 1800 describing flocks comprising hundreds of individuals) or the lower fitness of hybrids in the wild might have reduced/diluted its effect at population level (**chapter 3**).

5.2.2.1 Hybrids in an autochthonous population: the case of MSar3

An enclosed Sardinian population showed clear signals of recent admixture (MSar3, **chapter 3**). Unfortunately, it was not possible to retrieve detailed information on the configuration of this enclosure and whether the admixture was accidental or planned. In recent decades, several reintroduction efforts have been endorsed by the Sardinian Regional government to re-establish mouflon presence in several districts of Sardinia and new guidelines have been established to manage reintroductions, taking into account ecological, behavioural and genetic considerations (Apollonio et al., 2005). Our results suggest that admixture with domestic sheep is not threatening the Sardinian mouflon, however, the presence of an extremely admixed cluster – although kept in an enclosure – which was initially identified to us as being purebred, should generate serious concern. Until recently, the widely practiced reintroduction procedure relied on using animals kept in enclosures where they were allowed to reproduce. This practice raises some issues in terms of 1) low genetic diversity given by the reduced number of founders 2) a chance for rare crossbred animals

to generate a cluster of hybrids that can then become founders for reintroduced populations.

Additionally, it has been reported that the recognisable phenotypic features of sheep introgression in mouflon (e.g., woolly fleece, white patches), tend to disappear within two generations of backcrossing with purebred mouflon (Lauvergne et al., 1977) complicating the visual identification of hybrids. However, in the case of MSar3, the Admixture results (**chapter 3**) described a 21-50% sheep component among the individuals, values compatible with first and second generation crossbred and backcrosses, making it likely that at least the first generation of crossbred animals showed visible traits of hybridisation with domestics. Indeed, the latest guidelines for mouflon reintroduction in Sardinia suggest to avoid the use of animals kept in enclosures as founders, while recommending instead to capture individuals from the historical range (i.e., the Gennargentu massif) (Apollonio et al., 2005). However, this approach is extremely resource intensive (see 4.5.3). A cheaper strategy would be to screen the animals within the enclosures with the SNP array or few microsatellite markers in order to identify the level of crossbreeding (see Lorenzini et al., 2011) and determine the most suitable founders accordingly. The use of SNP data has been also suggested to aid the identification of wolf *x* dog hybrids (vonHoldt et al., 2013). In this case a set of 100 diagnostic loci able to maximise wolf/dog ancestry assignment was identified among those provided in the canine SNP array.

A similar genetic test panel could be defined for mouflon and sheep and applied to long-term screening for the Sardinian mouflon population, which - along with the small population on Corsica – represents the last of the extant autochthonous European mouflon populations, and whose conservation should be prioritised. In addition mouflon is regarded as a flagship species in Sardinia and therefore requires protection by the Sardinian government.

5.2.3 Selection in feral and domestic sheep

In **chapters 3** and **4** insights into the evolutionary history of mouflon and sheep have been provided with the identification of regions of ancient adaptive

introgression in modern sheep breeds. The signals we identified using the CIWI approach were related to genes known to trigger the Neutrophil Extracellular Traps (NETs; Wang et al., 2009), an innate immunity process recently identified to be involved with response to mastitis in dairy sheep (Pisanu et al., 2015). Innate immunity provides the front-line of defence against infection, and is constantly shaped by selection to face the ever-changing threat of rapidly evolving pathogens (Webb et al., 2015). Evidence of adaptive introgression related to innate immunity has been previously recorded in humans (Mendez et al., 2012), with ~54% of the Eurasian population carrying a Neanderthal derived version of the innate immune gene *STAT2*. In this context the results obtained in **chapter 3** might suggest a multifaceted evolutionary plasticity of nonspecific immunity, characterised by genomic regions of high adaptive value due to ancient adaptive introgression that do not change or change at a slower rate compared to the majority of the regions involved in innate immune processes (Webb et al., 2015). However, further studies are needed to investigate this hypothesis.

When selection between feral and domestic sheep was investigated using genome-wide scans for sweeps (**chapter 4**), we identified several genes under selection in sheep that were previously identified in other domestic species including cattle and pigs and also in humans (Sabeti et al., 2007; Rothammer et al., 2013; Frantz et al., 2015), suggesting potential targets for selection across multiple mammalian lineages.

5.2.4 Future perspective

5.2.4.1 More markers, more samples, or improved sampling?

The speed at which the next-generation sequencing throughput is improving has been estimated to be doubling every five months (Davey et al., 2011). Is therefore likely that whole genome sequencing will soon become the standard choice for genome research.

In the context of this thesis, genome wide sequence data would have removed ascertainment bias, allowing meaningful comparisons among distant clades. For both the chromosome painting and selection scan analysis the increased data

density would have allowed a more precise identification of the windows of ancestry and regions of selection respectively, allowing a less stringent parameterisation applied to the investigation methods. Additionally, ascertainment-free data would have improved inference for the *O. orientalis* populations: the Iranian mouflon and the Cypriot mouflon (**chapter 3**). However, these two populations were heavily underrepresented in the dataset with only two (Iranian mouflon) and three (Cypriot mouflon) individuals sampled, and EHH-based methods selection analysis require population data to correctly infer the haplotypes and infer selection sweeps (Browning and Browning, 2011; Szpiech and Hernandez, 2014). However, at least in the short term and for large numbers of samples, reduced-representation methods such as SNP arrays will remain a valuable choice in terms of cost and speed of analysis. Consequently, analytical frameworks such as those developed in this thesis can be valuable research tools in extending the applicability of SNP array data, e.g., to non-model species.

In **chapter 4** the occurrence of local adaptation in both feral and domestic sheep from Sardinia was investigated. Our analyses did not identify any signal of local adaptation. However, over 90% of the negative results obtained could be attributed to the very low statistical power of the analysis. To significantly increase the detection power of positive signals - given the same environmental variable (EV) sampling – more than one thousand individuals would be required.

In local adaptation analyses the sampling of the EVs is of paramount importance (see 1.4.4, 4.5.3). Such data is a function of sample size, more and/or better geographically distributed samples are able to represent a wider range of values for a given EV. Consequently, this would improve the statistical power, which increases along with the number of different genotypes associated to different environment (Rellstab et al., 2015).

As mentioned in section 4.5.3, sampling Sarda sheep across the whole island would improve our statistical power. However, substantial efforts are necessary to successfully complete such sampling, as the bureaucratic difficulties involved can be – as described – overwhelming

Appendix

Mouflon, watercolours (Ferreira, 2015)

6 Appendix

6.1 Appendix A1 - The first mitogenome of the Cyprus mouflon (*Ovis gmelini ophion*): new insights into the phylogeny of the genus Ovis

Daria Sanna, Mario Barbato, Eleftherios Hadjisterkotis, Piero Cossu, Luca Decandia, Sandro Trova, Monica Pirastru, Giovanni Giuseppe Leoni, Salvatore Naitana, Paolo Francalacci, Bruno Masala, Laura Manca, Paolo Mereu

Published: (2015), PLoS ONE 10(12): e0144257. doi: 10.1371/journal.pone.0144257

Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144257

6.2 Appendix A2 - Revisiting demographic processes in cattle with genome-wide population genetic analysis

Pablo Orozco-terWengel, Mario Barbato, Ezequiel Nicolazzi, Filippo Biscarini, Marco Milanesi, Wyn Davies, Don Williams, Alessandra Stella, Paolo Ajmone-Marsan and Michael W. Bruford

Published: (2015) *Frontiers in Genetics*, 6:191. doi: 10.3389/fgene.2015.00191

Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4451420/

7 Bibliography

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–73. doi:10.1038/nature09534.

Abraham, G., and Inouye, M. (2014). Fast principal component analysis of large-scale genome-wide data. *PLoS One* 9, 1–5. doi:10.1371/journal.pone.0093766.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res.* 19, 711–722. doi:10.1101/gr.086652.108.19.

Albrechtsen, A., Nielsen, F. C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* 27, 2534–2547. doi:10.1093/molbev/msq148.

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

Allendorf, F. W., Hohenlohe, P. a, and Luikart, G. (2010). Genomics and the future of conservation genetics. *Nat. Rev. Genet.* 11, 697–709. doi:10.1038/nrg2844.

Amaral, A. J., Megens, H.-J., Crooijmans, R. P. M. A., Heuven, H. C. M., and Groenen, M. A. M. (2008). Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* 179, 569–79. doi:10.1534/genetics.107.084277.

Apollonio, M., Luccarini, S., Giustini, D., Scandura, M., and Ghiandai, F. (2005). Carta delle vocazioni faunistiche della sardegna, sottoprogetto 3 (Studio relativo agli ungulati).

Aulchenko, Y. S., Ripke, S., Isaacs, A., and van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23, 1294–6. doi:10.1093/bioinformatics/btm108.

Barbato, M., Orozco-terWengel, P., Tapio, M., and Bruford, M. W. (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front. Genet.* 6, 1–6. doi:10.3389/fgene.2015.00109.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., and Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74, 1111–1120. doi:10.1086/421051.

Bertolini, F., Galimberti, G., Calò, D. G., Schiavo, G., Matassino, D., and Fontanesi, L. (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, 346–56. doi:10.1111/jbg.12155.

Biagini, E. (1948). La caccia al Muflone in Sardegna. *Le Vie d'Italia*, 937–942.

Björnerfeldt, S., Hailer, F., Nord, M., and Vilà, C. (2008). Assortative mating and fragmentation within dog breeds. *BMC Evol. Biol.* 8, 28. doi:10.1186/1471-2148-8-28.

Blanquart, F., Kaltz, O., Nuismer, S. L., and Gandon, S. (2013). A practical guide to measuring local adaptation. *Ecol. Lett.* 16, 1195–1205. doi:10.1111/ele.12150.

Bon, R., Cugnasse, J. M., Dubray, D., Gilbert, P., Houard, T., and Rigaud, P. (1991). Le Mouflon de Corse. *Reveud Ecol.*, 67–110.

Botta, S. ed. (1841). *Cenni sulla Sardegna*. Torino.

Brinkmann, V., Reichard, U., Goosmann, C., Fauler, B., Uhlemann, Y., Weiss, D. S., Weinrauch, Y.,

and Zychlinsky, A. (2004). Neutrophil extracellular traps kill bacteria. *Science* 303, 1532–1535. doi:10.1126/science.1092385.

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G., and Bustamante, C. D. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.* 84, 343–64. doi:10.3378/027.084.0401.

Bromfield, J. J., Santos, J. E. P., Block, J., Williams, R. S., and Sheldon, I. M. (2015). PHYSIOLOGY AND ENDOCRINOLOGY SYMPOSIUM: Uterine infection: Linking infection and innate immunity with infertility in the high-producing dairy cow. *J. Anim. Sci.* 93, 2021. doi:10.2527/jas.2014-8496.

Browning, S. R., and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 12, 703–14. doi:10.1038/nrg3054.

Bruford, M. W., Bradley, D. G., and Luikart, G. (2003). DNA markers reveal the complexity of livestock domestication. *Nat. Rev. Genet.* 4, 900–10. doi:10.1038/nrg1203.

Bruford, M. W., Ginja, C., Hoffmann, I., Joost, S., Orozco-terWengel, P., Alberto, F. J., Amaral, A. J., Barbato, M., Biscarini, F., Colli, L., et al. (2015). Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. *Front. Genet.* 6, 314. doi:10.3389/fgene.2015.00314.

Bruford, M. W., and Townsend, S. J. (2006). "Mitchondrial DNA Diversity in Modern Sheep: Implications for Domestication," in *Documenting Domestication: New Genetic and Archaeological Paradigms*, eds. M. A. Zeder, D. G. Bradley, E. Emshwiller, and B. D. Smith (London: Univeristy of California Press, Ltd.), 307–317.

Brust, V., and Guenther, A. (2015). Domestication effects on behavioural traits and learning performance: comparing wild cavies to guinea pigs. *Anim. Cogn.* 18, 99–109. doi:10.1007/s10071-014-0781-9.

Burren, A., Signer-Hasler, H., Neuditschko, M., Tetens, J., Kijas, J. W., Drögemüller, C., and Flury, C. (2014). Fine-scale population structure analysis of seven local Swiss sheep breeds using genome-wide SNP data. *Anim. Genet. Resour. génétiques Anim. genéticos Anim.* 55, 67–76. doi:10.1017/S2078633614000253.

Buzanskas, M. E., Grossi, D. a, Ventura, R. V, Schenkel, F. S., Sargolzaei, M., Meirelles, S. L. C., Mokry, F. B., Higa, R. H., Mudadu, M. a, da Silva, M. V. G. B., et al. (2014). Genome-wide association for growth traits in Canchim beef cattle. *PLoS One* 9, e94802. doi:10.1371/journal.pone.0094802.

Calvo, J. H., Alvarez-Rodriguez, J., Marcos-Carcavilla, A., Serrano, M., and Sanz, A. (2011). Genetic diversity in the Churra tensina and Churra lebrijana endangered Spanish sheep breeds and relationship with other Churra group breeds and Spanish mouflon. *Small Rumin. Res.* 95, 34–39. doi:10.1016/j.smallrumres.2010.09.003.

Carnevali, L., Pedrotti, L., Riga, F., and Toso, S. (2009). Banca Dati Ungulati: Status, distribuzione, consistenza, gestione e prelievo venatorio delle popolazioni di Ungulati in Italia. Rapporto 2001-2005. doi:10.1007/s13398-014-0173-7.2.

Cecchinato, A., Ribeca, C., Chessa, S., Cipolat-Gotet, C., Maretto, F., Casellas, J., and Bittante, G. (2014). Candidate gene association analysis for milk yield, composition, urea nitrogen and somatic cell scores in Brown Swiss cows. *Animal* 8, 1062–1070. doi:10.1017/S1751731114001098.

Cerri, R. L. a, Thompson, I. M., Kim, I. H., Ealy, a D., Hansen, P. J., Staples, C. R., Li, J. L., Santos, J. E. P., and Thatcher, W. W. (2012). Effects of lactation and pregnancy on gene expression of endometrium of Holstein cows at day 17 of the estrous cycle or pregnancy. *J. Dairy Sci.* 95, 5657–75. doi:10.3168/jds.2011-5114.

Cetti, F. (1774). *Storia Naturale di Sardegna*. ILISSO.

Charlesworth, B., Nordborg, M., and Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* 70, 155–74. doi:10.1111/j.1469-1809.1943.tb02321.x.

Cheeseman, I. H., Miller, B. A., Nair, S., Nkhoma, S., Tan, A., Tan, J. C., Al Saai, S., Phyo, A. P., Moo, C. L., Lwin, K. M., et al. (2012). A major genome region underlying artemisinin resistance in malaria. *Science* 336, 79–82. doi:10.1126/science.1215966.

Chen, J., Kallman, T., Ma, X., Gyllenstrand, N., Zaina, G., Morgante, M., Bousquet, J., Eckert, A., Wegrzyn, J., Neale, D., et al. (2012). Disentangling the Roles of History and Local Selection in Shaping Clinal Variation of Allele Frequencies and Gene Expression in Norway Spruce (Picea abies). *Genetics* 191, 865–881. doi:10.1534/genetics.112.140749.

Chessa, B., Pereira, F., Arnaud, F., Amorim, A., Goyache, F., Mainland, I., Kao, R. R., Pemberton, J. M., Beraldi, D., Stear, M. J., et al. (2009). Revealing the history of sheep domestication using retrovirus integrations. *Science* 324, 532–6. doi:10.1126/science.1170587.

Chomwisarutkun, K., Murani, E., Ponsuksili, S., and Wimmers, K. (2012). Gene expression analysis of mammary tissue during fetal bud formation and growth in two pig breeds--indications of prenatal initiation of postnatal phenotypic differences. *BMC Dev. Biol.* 12, 13. doi:10.1186/1471-213X-12-13.

Ciani, E., Crepaldi, P., Nicoloso, L., Lasagna, E., Sarti, F. M., Moioli, B., Napolitano, F., Carta, A., Usai, G., D'Andrea, M., et al. (2013). Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. *Anim. Genet.* doi:10.1111/age.12106.

Ciani, E., Lasagna, E., D'Andrea, M., Alloggio, I., Marroni, F., Ceccobelli, S., Delgado Bermejo, J. V., Sarti, F. M., Kijas, J. W., Lenstra, J. A., et al. (2015). Merino and Merino-derived sheep breeds: a genome-wide intercontinental study. *Genet. Sel. Evol.* 47, 64. doi:10.1186/s12711-015-0139-z.

Cieslak, M., Reissmann, M., Hofreiter, M., and Ludwig, A. (2011). Colours of domestication. *Biol. Rev. Camb. Philos. Soc.* 86, 885–99. doi:10.1111/j.1469-185X.2011.00177.x.

Clutton-Brock, J. (1999). *A Natural History of Domesticated Mammals*. Second Edi. Cambridge: Cambridge University Press.

Clutton-Brock, J. (1992). The process of domestication. *Mamm. Rev.* 22, 79–85. doi:10.1111/j.1365-2907.1992.tb00122.x.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185, 1411–23. doi:10.1534/genetics.110.114819.

Corbin, L. J., Blott, S. C., Swinburne, J. E., Vaudin, M., Bishop, S. C., and Woolliams, J. A. (2010). Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Anim. Genet.* 41 Suppl 2, 8–15. doi:10.1111/j.1365-2052.2010.02092.x.

Corbin, L. J., Liu, A. Y. H., Bishop, S. C., and Woolliams, J. A. (2012). Estimation of historical effective population size using linkage disequilibria with marker data. *J. Anim. Breed. Genet.* 129, 257–70. doi:10.1111/j.1439-0388.2012.01003.x.

Crow, J. F., and Kimura, M. (1970). *An introduction to population genetics theory*. New York: Harper and Row.

Davey, J. W., Hohenlohe, P. a, Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing.

Nat. Rev. Genet. 12, 499–510. doi:10.1038/nrg3012.

Davison, A. C., and Hinkley, D. V (1997). *Bootstrap methods and their application.* Cambridge, United Kingdom: Cambridge University Press.

Demirci, S., Koban Baştanlar, E., Dağtaş, N. D., Pişkin, E., Engin, A., Özer, F., Yüncü, E., Doğan, Ş. A., and Togan, İ. (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, e81952. doi:10.1371/journal.pone.0081952.

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., and Ovenden, J. R. (2014). NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Mol. Ecol. Resour.* 14, 209–214. doi:10.1111/1755-0998.12157.

Eckert, A. J., Bower, A. D., GonzÁlez-MartÍnez, S. C., Wegrzyn, J. L., Coop, G., and Neale, D. B. (2010). Back to nature: Ecological genomics of loblolly pine (Pinus taeda, Pinaceae). *Mol. Ecol.* 19, 3789–3805. doi:10.1111/j.1365-294X.2010.04698.x.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10, 48. doi:10.1186/1471-2105-10-48.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. a., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, 1–10. doi:10.1371/journal.pone.0019379.

England, P. R., Cornuet, J.-M., Berthier, P., Tallmon, D. a., and Luikart, G. (2006). Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conserv. Genet.* 7, 303–308. doi:10.1007/s10592-005-9103-8.

Fang, M., Larson, G., Ribeiro, H. S., Li, N., and Andersson, L. (2009). Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genet.* 5, e1000341. doi:10.1371/journal.pgen.1000341.

Favreau, A., Baumont, R., Ferreira, G., Dumont, B., and Ginane, C. (2010). Do sheep use umami and bitter tastes as cues of post-ingestive consequences when selecting their diet? *Appl. Anim. Behav. Sci.* 125, 115–123. doi:10.1016/j.applanim.2010.04.007.

Fay, J. C., and Wu, C.-I. (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics* 155, 1405–1413.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.* 31, 1275–91. doi:10.1093/molbev/msu077.

Flori, L., Thevenon, S., Dayo, G. K., Senou, M., Sylla, S., Berthier, D., Moazami-Goudarzi, K., and Gautier, M. (2014). Adaptive admixture in the West African bovine hybrid zone: Insight from the Borgou population. *Mol. Ecol.* 23, 3241–3257. doi:10.1111/mec.12816.

Flury, C., Tapio, M., Sonstegard, T., Drögemüller, C., Leeb, T., Simianer, H., Hanotte, O., and Rieder, S. (2010). Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *J. Anim. Breed. Genet.* 127, 339–47. doi:10.1111/j.1439-0388.2010.00862.x.

Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H.-J., Cagan, A., Bosse, M., Paudel, Y., Crooijmans, R. P. M. A., Larson, G., and Groenen, M. A. M. (2015). Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat. Genet.* 47, 1141–1148. doi:10.1038/ng.3394.

Frascaroli, E., Schrag, T. a., and Melchinger, A. E. (2013). Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals

ascertainment bias for a subset of SNPs. *Theor. Appl. Genet.* 126, 133–141. doi:10.1007/s00122-012-1968-6.

Frichot, E., and François, O. (2015). LEA : An R package for landscape and ecological association studies. *Methods Ecol. Evol.* 6, 925–929. doi:10.1111/2041-210X.12382.

Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* 30, 1687–99. doi:10.1093/molbev/mst063.

Frisina, M. R., and Frisina, R. M. (2013). Phenotype evaluation of free-ranging European mouflon. *Taprobanica* 05, 157–162.

Garel, M., Cugnasse, J. M., Gaillard, J.-M., Loison, A., Gibert, P., Douvre, P., and Dubray, D. (2005). Reproductive output of female mouflon (Ovis gmelini musimon × Ovis sp.): a comparative analysis. *J. Zool.* 266, 65–71. doi:10.1017/S0952836905006667.

Gautier, M., Flori, L., Riebler, A., Jaffrézic, F., Laloé, D., Gut, I., Moazami-Goudarzi, K., and Foulley, J.-L. (2009). A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics* 10, 550. doi:10.1186/1471-2164-10-550.

Gautier, M., and Vitalis, R. (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.

Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., Green, R. D., Hamernik, D. L., Kappes, S. M., Lien, S., et al. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324, 528–32. doi:10.1126/science.1167936.

Goedbloed, D. J., van Hooft, P., Megens, H.-J., Langenbeck, K., Lutz, W., Crooijmans, R. P. M. A., van Wieren, S. E., Ydenberg, R. C., and Prins, H. H. T. (2013). Reintroductions and genetic introgression from domestic pigs have shaped the genetic population structure of Northwest European wild boar. *BMC Genet.* 14, 43. doi:10.1186/1471-2156-14-43.

Gratten, J., Pilkington, J. G., Brown, E. a, Clutton-Brock, T. H., Pemberton, J. M., and Slate, J. (2012). Selection and microevolution of coat pattern are cryptic in a wild population of sheep. *Mol. Ecol.* 21, 2977–90. doi:10.1111/j.1365-294X.2012.05536.x.

Grignolio, S., Madau, R., Pipia, A., Apollonio, M., Luchetti, S., and Ciuti, S. (2008). Influence of sex, season, temperature and reproductive status on daily activity patterns in Sardinian mouflon (Ovis orientalis musimon). *Behaviour* 145, 1723–1745. doi:10.1163/156853908786279628.

Guerrini, M., Forcina, G., Panayides, P., Lorenzini, R., Garel, M., Anayiotos, P., Kassinis, N., and Barbanera, F. (2015). Molecular DNA identity of the mouflon of Cyprus (*Ovis orientalis ophion* , Bovidae): Near Eastern origin and divergence from Western Mediterranean conspecific populations. *Syst. Biodivers.*, 1–12. doi:10.1080/14772000.2015.1046409.

Guillot, G., Vitalis, R., Rouzic, A. le, and Gautier, M. (2014). Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. *Spat. Stat.* 8, 145–155. doi:10.1016/j.spasta.2013.08.001.

Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* 195, 205–20. doi:10.1534/genetics.113.152462.

Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 8, 299–309.

Hao, K., Di, X., and Cawley, S. (2007). LdCompare: rapid computation of single- and multiple-

marker r2 and genetic coverage. *Bioinformatics* 23, 252–4. doi:10.1093/bioinformatics/btl574.

Harrison, R. G., and Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *J. Hered.* 105 Suppl , 795–809. doi:10.1093/jhered/esu033.

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

Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W., and van der Werf, J. H. J. (2012). Accuracy of genotype imputation in sheep breeds. *Anim. Genet.* 43, 72–80. doi:10.1111/j.1365-2052.2011.02208.x.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13, 635–43. doi:10.1101/gr.387103.

Hedrick, P. W. (2013). Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* 22, 4606–4618. doi:10.1111/mec.12415.

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, a, Maes, G. E., Diopere, E., Carvalho, G. R., et al. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol. Ecol. Resour.* 11 Suppl 1, 123–36. doi:10.1111/j.1755-0998.2010.02943.x.

Hernández-Orduño, G., Torres-Acosta, J. F. J., Sandoval-Castro, C. a., Capetillo-Leal, C. M., Aguilar-Caballero, a. J., and Alonso-Díaz, M. a. (2015). A tannin-blocking agent does not modify the preference of sheep towards tannin-containing plants. *Physiol. Behav.* 145, 106–111. doi:10.1016/j.physbeh.2015.04.006.

Hess, S. C., Kawakami, B., Okita, D., and Medeiros, K. (2006). A preliminary assessment of mouflon abundance at the Kahuku unit of Hawai'i Volcanoes National Park.

Hiendleder, S., Kaupe, B., Wassmuth, R., and Janke, A. (2002). Molecular analysis of wild and domestic sheep questions current nomenclature and provides evidence for domestication from two different subspecies. *Proc. Biol. Sci.* 269, 893–904. doi:10.1098/rspb.2002.1975.

Hiendleder, S., Mainz, K., Plante, Y., and Lewalski, H. (1998). Analysis of mitochondrial DNA indicates that domestic sheep are derived from two different ancestral maternal sources: no evidence for contributions from urial and argali sheep. *J. Hered.* 89, 113–20.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.* 38, 209–216. doi:10.1017/S0016672300020553.

Hill, W. G., and Robertson, A. (1968). Linkage Disequilibrium in Finite Populations. *Theor. Appl. Genet.* 38, 226–231. doi:10.1007/BF01245622.

Hsieh, F. Y., Bloch, D. A., and Larsen, M. D. (1998). A Simple Method of Sample Size Calculation for Linear and Logistic Regression. *Stat. Med.* 17, 1623–1634.

Hudson, N. J., Porto Neto, L. R., Kijas, J. W., McWilliam, S., Taft, R. J., and Reverter, A. (2014). Information compression exploits patterns of genome composition to discriminate populations and highlight regions of evolutionary interest. *BMC Bioinformatics* 15, 66. doi:10.1186/1471-2105-15-66.

Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18, 486–487. doi:10.1016/S0168-9525(02)02722-1.

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi:10.1093/molbev/msj030.

Iacolina, L., Scandura, M., Goedbloed, D. J., Alexandri, P., Crooijmans, R. P. M. a, Larson, G.,

Archibald, A. L., Apollonio, M., Schook, L. B., Groenen, M. a M., et al. (2015). Genomic diversity and differentiation of a managed island wild boar population. *Heredity (Edinb).*, 1–8. doi:10.1038/hdy.2015.70.

Johnston, S. E., McEwan, J. C., Pickering, N. K., Kijas, J. W., Beraldi, D., Pilkington, J. G., Pemberton, J. M., and Slate, J. (2011). Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Mol. Ecol.* 20, 2555–66. doi:10.1111/j.1365-294X.2011.05076.x.

Johnston, S. E., Slate, J., and Pemberton, J. M. (2015). A genomic region containing RNF212 is associated with sexually-dimorphic recombination rate variation in wild Soay sheep (Ovis aries). Author Summary .

Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., and Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol. Ecol.* 16, 3955–69. doi:10.1111/j.1365-294X.2007.03442.x.

Karesh, W. B., Smith, F., and Frazier-Taylor, H. (1987). A Remote Method for Obtaining Skin Biopsy Samples. *Conserv. Biol.* 1, 261–262. doi:10.1111/j.1523-1739.1987.tb00041.x.

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J. A. C., Baurain, D., Cambisano, N., Davis, S. R., Farnir, F., Grisart, B., et al. (2011). Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nat. Genet.* 43, 405–13. doi:10.1038/ng.814.

Kidd, A. G., Bowman, J., Lesbarrères, D., and Schulte-Hostedde, A. I. (2009). Hybridization between escaped domestic and wild American mink (Neovison vison). *Mol. Ecol.* 18, 1175–86. doi:10.1111/j.1365-294X.2009.04100.x.

Kijas, J. W., Lenstra, J. A., Hayes, B. J., Boitard, S., Porto Neto, L. R., San Cristobal, M., Servin, B., McCulloch, R., Whan, V., Gietzen, K., et al. (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., Porto-Neto, L., Dominik, S., Reverter, A., Bunch, R., McCulloch, R., Hayes, B. J., Brauning, R., and McEwan, J. (2014). Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. *Anim. Genet.* 45, 754–7. doi:10.1111/age.12197.

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

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Jpn. J. Genet.* 66, 367–386. doi:10.1266/jjg.66.367.

Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Ann. Eugenetics* 12, 172–175.

Kwok, P.-Y., and Chen, X. (2003). Detection of single nucleotide polymorphisms. *Curr. Issues Mol. Biol.* 5, 43–60.

Lalueza-Fox, C., Castresana, J., Sampietro, L., Marquès-Bonet, T., Alcover, J. A., and Bertranpetit, J. (2005). Molecular dating of caprines using ancient DNA sequences of Myotragus balearicus, an extinct endemic Balearic mammal. *BMC Evol. Biol.* 5, 70. doi:10.1186/1471-2148-5-70.

Lancioni, H., Di Lorenzo, P., Ceccobelli, S., Perego, U. a, Miglio, A., Landi, V., Antognoni, M. T., Sarti, F. M., Lasagna, E., and Achilli, A. (2013). Phylogenetic relationships of three Italian merino-derived sheep breeds evaluated through a complete mitogenome analysis. *PLoS One* 8, e73712. doi:10.1371/journal.pone.0073712.

Lauvergne, J.-J., Denis, B., and Théret, M. (1977). Hybridation entre un Mouflon de Corse (Ovis

ammon musimon Schreber, 1872) et des brebis de divers génotypes: gènes pour la coloration pigmentaire. *Genet. Sel. Evol.* 9, 151. doi:10.1186/1297-9686-9-2-151.

Lawson Handley, L.-J., Byrne, K., Santucci, F., Townsend, S. J., Taylor, M., Bruford, M. W., and Hewitt, G. M. (2007). Genetic structure of European sheep breeds. *Heredity (Edinb).* 99, 620–31. doi:10.1038/sj.hdy.6801039.

Le Razze Ovine e Caprine in Italia (2002). Roma.

Lee, R. J., and Cohen, N. a. (2014). Taste receptors in innate immunity. *Cell. Mol. Life Sci.* 72, 217–236. doi:10.1007/s00018-014-1736-7.

Levin, D. A., Francisco-Ortega, J., and Jansen, R. K. (1996). Hybridization and the Extinction of Rare Plant Species. *Conserv. Biol.* 10, 10–16. doi:10.1046/j.1523-1739.1996.10010010.x.

Li, B., and Kimmel, M. (2013). Factors influencing ascertainment bias of microsatellite allele sizes: Impact on estimates of mutation rates. *Genetics* 195, 563–572. doi:10.1534/genetics.113.154161.

Li, P., Li, M., Lindberg, M. R., Kennett, M. J., Xiong, N., and Wang, Y. (2010). PAD4 is essential for antibacterial innate immunity mediated by neutrophil extracellular traps. *J. Exp. Med.* 207, 1853–1862. doi:10.1084/jem.20100239.

Lincoln, G. a (1998). Reproductive seasonality and maturation throughout the complete life-cycle in the mouflon ram (Ovis musimon). *Anim. Reprod. Sci.* 53, 87–105. doi:10.1016/S0378-4320(98)00129-8.

Lindgren, G., Backström, N., Swinburne, J., Hellborg, L., Einarsson, A., Sandberg, K., Cothran, G., Vilà, C., Binns, M. M., and Ellegren, H. (2004). Limited number of patrilines in horse domestication. *Nat. Genet.* 36, 335–6. doi:10.1038/ng1326.

Lindqvist, C., and Jensen, P. (2009). Domestication and stress effects on contrafreeloading and spatial learning performance in red jungle fowl (Gallus gallus) and White Leghorn layers. *Behav. Processes* 81, 80–4. doi:10.1016/j.beproc.2009.02.005.

Lorenzini, R., Cabras, P., Fanelli, R., and Carboni, G. L. (2011). Wildlife molecular forensics: identification of the Sardinian mouflon using STR profiling and the Bayesian assignment test. *Forensic Sci. Int. Genet.* 5, 345–9. doi:10.1016/j.fsigen.2011.01.012.

Lotterhos, K. E., and Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.* 24, n/a–n/a. doi:10.1111/mec.13100.

Mackay, T. F. C. (2004). The genetic architecture of quantitative traits: lessons from Drosophila. *Curr. Opin. Genet. Dev.* 14, 253–7. doi:10.1016/j.gde.2004.04.003.

Magri, D., Fineschi, S., Bellarosa, R., Buonamici, A., Sebastiani, F., Schirone, B., Simeone, M. C., and Vendramin, G. G. (2007). The distribution of Quercus suber chloroplast haplotypes matches the palaeogeographical history of the western Mediterranean. *Mol. Ecol.* 16, 5259–66. doi:10.1111/j.1365-294X.2007.03587.x.

Makina, S. O., Taylor, J. F., van Marle-Köster, E., Muchadeyi, F. C., Makgahlela, M. L., MacNeil, M. D., and Maiwashe, A. (2015). Extent of Linkage Disequilibrium and Effective Population Size in Four South African Sanga Cattle Breeds. *Front. Genet.* 6. doi:10.3389/fgene.2015.00337.

Manel, S., Gaggiotti, O. E., and Waples, R. S. (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol. Evol.* 20, 136–42. doi:10.1016/j.tree.2004.12.004.

Manel, S., Joost, S., Epperson, B. K., Holderegger, R., Storfer, A., Rosenberg, M. S., Scribner, K. T., Bonin, A., and Fortin, M.-J. (2010). Perspectives on the use of landscape genetics to detect

genetic adaptive variation in the field. *Mol. Ecol.* 19, 3760–72. doi:10.1111/j.1365-294X.2010.04717.x.

Mbole-Kariuki, M. N., Sonstegard, T., Orth, A., Thumbi, S. M., Bronsvoort, B. M. D. C., Kiara, H., Toye, P., Conradie, I., Jennings, A., Coetzer, K., et al. (2014). Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya. *Heredity (Edinb).* 113, 297–305. doi:10.1038/hdy.2014.31.

McDowell, N. (2002). Stream of escaped farm fish raises fears for wild salmon. *Nature* 416, 571–571. doi:10.1038/416571a.

McEvoy, B. P., Powell, J. E., Goddard, M. E., and Visscher, P. M. (2011). Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res.* 21, 821–9. doi:10.1101/gr.119636.110.

McTavish, E. J., Decker, J. E., Schnabel, R. D., Taylor, J. F., and Hillis, D. M. (2013). New World cattle show ancestry from multiple independent domestication events. *Proc. Natl. Acad. Sci. U. S. A.* 110, E1398–406. doi:10.1073/pnas.1303367110.

McTavish, E. J., and Hillis, D. M. (2015). How do SNP ascertainment schemes and population demographics affect inferences about population history? *BMC Genomics* 16. doi:10.1186/s12864-015-1469-5.

Meadows, J. R. S., Cemal, I., Karaca, O., Gootwine, E., and Kijas, J. W. (2007). Five ovine mitochondrial lineages identified from sheep breeds of the near East. *Genetics* 175, 1371–9. doi:10.1534/genetics.106.068353.

Meadows, J. R. S., Hiendleder, S., and Kijas, J. W. (2011). Haplogroup relationships between domestic and wild sheep resolved using a mitogenome panel. *Heredity (Edinb).* 106, 700–6. doi:10.1038/hdy.2010.122.

Meadows, J. R. S., Li, K., Kantanen, J., Tapio, M., Sipos, W., Pardeshi, V. C., Gupta, V. S., Calvo, J. H., Whan, V., Norris, B., et al. (2005). Mitochondrial sequence reveals high levels of gene flow between breeds of domestic sheep from Asia and Europe. *J. Hered.* 96, 494–501. doi:10.1093/jhered/esi100.

Mendez, F. L., Watkins, J. C., and Hammer, M. F. (2012). A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am. J. Hum. Genet.* 91, 265–74. doi:10.1016/j.ajhg.2012.06.015.

Miller, J. M., Kijas, J. W., Heaton, M. P., McEwan, J. C., and Coltman, D. W. (2012a). Consistent divergence times and allele sharing measured from cross-species application of SNP chips developed for three domestic species. *Mol. Ecol. Resour.* 12, 1145–1150. doi:10.1111/1755-0998.12017.

Miller, J. M., Poissant, J., Hogg, J. T., and Coltman, D. W. (2012b). Genomic consequences of genetic rescue in an insular population of bighorn sheep (Ovis canadensis). *Mol. Ecol.* doi:10.1111/j.1365-294X.2011.05427.x.

Miller, J. M., Poissant, J., Kijas, J. W., and Coltman, D. W. (2011). A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Mol. Ecol. Resour.* 11, 314–322. doi:10.1111/j.1755-0998.2010.02918.x.

Moioli, B., Napolitano, F., Orrù, L., and Catillo, G. (2010). Analysis of the genetic diversity between Gentile di Puglia, Sopravissana and Sarda sheep breeds using microsatellite markers. *Ital. J. Anim. Sci.* 5, 73–78.

Moradi, M. H., Nejati-Javaremi, A., Moradi-Shahrbabak, M., Dodds, K. G., and McEwan, J. C. (2012). Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC Genet.* 13, 10. doi:10.1186/1471-2156-13-10.

Morelli, L., Useli, A., Sanna, D., Barbato, M., Contu, D., Pala, M., Cancedda, M., and Francalacci, P. (2014). Mitochondrial DNA lineages of Italian Giara and Sarcidano horses. *Genet. Mol. Res.* 13, 8241–8257. doi:http://dx.doi.org/10.4238/2014.October.20.1.

Naderi, S., Rezaei, H.-R., Pompanon, F., Blum, M. G. B., Negrini, R., et al. (2008). The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proc. Natl. Acad. Sci. U. S. A.* 105, 17659–64. doi:10.1073/pnas.0804782105.

Nadler, C. F., Hoffmann, R. S., and A, W. (1973). G-Band Patterns as Chromosomal Markers, and the interpretation of Chromosomal Evolution in Wild Sheep (Ovis). *Experientia* 29, 117–119.

Nicholson, G., Smith, A. V., Jonsson, F., Gustafsson, O., Stefansson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* 64, 695–715. doi:10.1111/1467-9868.00357.

Nicolai, P. (1904). Alpinismo sardo. *Tour. Club Ital.*

Nicolazzi, E. L., Biffani, S., Biscarini, F., Orozco Ter Wengel, P., Caprera, A., Nazzicari, N., and Stella, A. (2015). Software solutions for the livestock genomics SNP array revolution. *Anim. Genet.* 46, 343–53. doi:10.1111/age.12295.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. D., and Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* 8, 857–868. doi:10.1038/nrg2187.

Nielsen, R., Hubisz, M. J., and Clark, A. G. (2004). Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168, 2373–2382. doi:10.1534/genetics.104.031039.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi:10.1038/nrg2986.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., et al. (2008). Genes mirror geography within Europe. *Nature* 456, 98–101. doi:10.1038/nature07566.

Ohta, T., and Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* 68, 571–580.

Oliveira, R., Randi, E., Mattucci, F., Kurushima, J. D., Lyons, L. A., and Alves, P. C. (2015). Toward a genome-wide approach for detecting hybrids: informative SNPs to detect introgression between domestic cats and European wildcats (Felis silvestris). *Heredity (Edinb).* 115, 195–205. doi:10.1038/hdy.2015.25.

Orozco-terWengel, P., Barbato, M., Nicolazzi, E. L., Biscarini, F., Milanesi, M., Davies, W., Williams, D., Stella, A., Ajmone-Marsan, P., and Bruford, M. W. (2015). Revisiting demographic processes in cattle with genome-wide population genetic analysis. *Front. Genet.* 6, 1–15. doi:10.3389/fgene.2015.00191.

Orozco-Terwengel, P., and Bruford, M. W. (2014). Mixed signals from hybrid genomes. *Mol. Ecol.* 23, 3941–3943. doi:10.1111/mec.12863.

Orozco-terWengel, P., Corander, J., and Schlötterer, C. (2011). Genealogical lineage sorting leads to significant, but incorrect Bayesian multilocus inference of population structure. *Mol. Ecol.* 20, 1108–21. doi:10.1111/j.1365-294X.2010.04990.x.

Pasaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics* 25, i213–21. doi:10.1093/bioinformatics/btp197.

Pedrosa, S., Uzun, M., Arranz, J.-J., Gutiérrez-Gil, B., San Primitivo, F., and Bayón, Y. (2005). Evidence of three maternal lineages in Near Eastern sheep supporting multiple domestication events. *Proc. Biol. Sci.* 272, 2211–7. doi:10.1098/rspb.2005.3204.

Perez, J. M., Granados, J. E., Ruiz-Martinez, I., and Chirosa, M. (1997). Capturing Spanish ibexes with corral traps. *Wildl. Soc. Bullettin* 25, 89–92.

Petersen, J. L., Mickelson, J. R., Rendahl, A. K., Valberg, S. J., Andersson, L. S., Axelsson, J., et al. (2013). Genome-Wide Analysis Reveals Selection for Important Traits in Domestic Horse Breeds. *PLoS Genet.* 9. doi:10.1371/journal.pgen.1003211.

Petit, E., Aulagnier, S., Vaiman, D., Bouissou, C., and Crouau-roy, B. (1997). Microsatellite Variation in an Introduced Mouflon Population. *J. Hered.* 88, 517–520.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19, 826–37. doi:10.1101/gr.087577.108.

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* 8. doi:10.1371/journal.pgen.1002967.

Pipia, A., Ciuti, S., Grignolio, S., Luchetti, S., Madau, R., and Apollonio, M. (2009). Effect of predation risk on grouping pattern and whistling behaviour in a wild mouflon Ovis aries population. *Acta Theriol. (Warsz).* 54, 77–86.

Piras, M., Casu, S., Salaris, S., Usai, M. G., and Carta, A. (2009). The Pecora Nera di Arbus: a new sheep breed in Sardinia, Italy. *Anim. Genet. Resour. Inf.* 45, 91. doi:10.1017/S1014233909990393.

Pisanu, S., Cubeddu, T., Pagnozzi, D., Rocca, S., Cacciotto, C., Alberti, A., Marogna, G., Uzzau, S., and Addis, M. F. (2015). Neutrophil extracellular traps in sheep mastitis. *Vet. Res.* 46, 59. doi:10.1186/s13567-015-0196-x.

Poplin, F. (1979). Origine du Mouflone de Corse dans une nouvelle perspective paléontologique: par marronnage. *Ann. Genet. Sel. anim* 11, 133–143.

Pritchard, J. K., and Di Rienzo, A. (2010). Adaptation – not by sweeps alone. *Nat. Rev. Genet.* 11, 665–667. doi:10.1038/nrg2880.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155, 945–959.

Pugach, I., and Stoneking, M. (2015). Genome-wide insights into the genetic history of human populations. *Investig. Genet.* 6, 1–20. doi:10.1186/s13323-015-0024-0.

Puigcerver, M., Sanchez-Donoso, I., Vilà, C., Sardà-Palomera, F., García-Galea, E., and Rodríguez-Teijeiro, J. D. (2014). Decreased fitness of restocked hybrid quails prevents fast admixture with wild European quails. *Biol. Conserv.* 171, 74–81. doi:10.1016/j.biocon.2014.01.010.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–75. doi:10.1086/519795.

Qiu, W. (2015). powerMediation: Power/Sample Size Calculation for Mediation Analysis.

Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* 16, 359–371. doi:10.1038/nrg3936.

Randhawa, I. A., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2014). Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet.* 15, 34. doi:10.1186/1471-2156-15-34.

Randi, E. (2008). Detecting hybridization between wild species and their domesticated relatives. *Mol. Ecol.* 17, 285–93. doi:10.1111/j.1365-294X.2007.03417.x.

Rando, A., Di Gregorio, P., Capuano, M., Senese, C., Manca, L., Naitana, S., and Masala, B. (1996). A comparison between the β-globin gene clusters of domestic sheep (Ovis aries) and Sardinian mouflon (Ovis gmelini musimon). *Genet Sel Evol* 28, 217–222.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi:10.1111/mec.13322.

Rezaei, H.-R., Naderi, S., Chintauan-Marquier, I. C., Taberlet, P., Virk, A. T., Naghash, H.-R., Rioux, D., Kaboli, M., and Pompanon, F. (2010). Evolution and taxonomy of the wild species of the genus Ovis (Mammalia, Artiodactyla, Bovidae). *Mol. Phylogenet. Evol.* 54, 315–26. doi:10.1016/j.ympev.2009.10.037.

Rhymer, J. M., and Simberloff, D. (1996). Extinction by Hybridization and Introgression. *Annu. Rev. Ecol. Syst* 27, 83–109.

Rival, F. (2000). THE ARGALI OF THE "CAUNE DE L'ARAGO" (SOUTHERN FRANCE). PALAEOECOLOGY OF A 440,000 YEARS OLD POPULATION. in *Proceedings of the Third International Symposium on Mouflon.*, 103–113.

de Roos, a P. W., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179, 1503–12. doi:10.1534/genetics.107.084301.

Rothammer, S., Seichter, D., Förster, M., and Medugorac, I. (2013). A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. *BMC Genomics* 14, 908. doi:10.1186/1471-2164-14-908.

Rozzi, R., Palombo, M. R., and Barbieri, M. (2011). THE ARGALI (OVIS AMMON ANTIQUA) FROM THE MAGLIANA AREA (ROME). *Ital. J. Quat. Sci.* 24, 113–119.

Ryder, M. (1981). A survey of European primitive breeds of sheep. *Ann. Genet. Sel. anim* 13, 381–418. doi:10.1186/1297-9686-13-4-381.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V, Patterson, N. J., McDonald, G. J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi:10.1038/nature01140.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. a, Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–8. doi:10.1038/nature06250.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507, 354–7. doi:10.1038/nature12961.

Sanna, D., Barbato, M., Hadjisterkotis, E., Cossu, P., Decandia, L., Trova, S., Pirastru, M., Leoni, G. G., Naitana, S., Francalacci, P., et al. (2015). The First Mitogenome of the Cyprus Mouflon (Ovis gmelini ophion): New Insights into the Phylogeny of the Genus Ovis. *PLoS One* 10, e0144257. doi:10.1371/journal.pone.0144257.

Santiago-Moreno, J., Todefano-Diaz, A., Gomez-Brunet, A., and Lopez-Sebastian, A. (2004). El muflón Europeo (Ovis orientalis musimon Schreber, 1782) en España: consideraciones historícas, filogenéticas y fisiología reproductiva. *Galemys* 16, 3–20.

Saura, M., Tenesa, A., Woolliams, J. A., Fernández, A., and Villanueva, B. (2015). Evaluation of the linkage-disequilibrium method for the estimation of effective population size when

generations overlap: an empirical case. *BMC Genomics* 16, 922. doi:10.1186/s12864-015-2167-z.

Savolainen, O., Lascoux, M., and Merilä, J. (2013). Ecological genomics of local adaptation. *Nat. Rev. Genet.* 14, 807–20. doi:10.1038/nrg3522.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–44. doi:10.1086/502802.

Schraiber, J. G., and Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* 16, 727–740. doi:10.1038/nrg4005.

Serra, M. G. (2009). Effetto della tecnica di allattamento parziale sulla produzione di latte e di carne negli ovini di razza Sarda.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., et al. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science* 329, 72–75. doi:10.1126/science.1189406.

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35. doi:John Maynard Smith and John Haigh (1974). The hitch-hiking effect of a favourable gene. Genetical Research, 23, pp 23-35. doi:10.1017/S0016672300014634.

Stahlberger-Saitbekova, N., Schlapfer, J., Dolf, G., and Gaillard, C. (2001). Genetic relationships in Swiss sheep breeds based on microsatellite analysis. *Anim. Breed. Genet.* 118, 379–387.

Stucki, S., Orozco-terWengel, P., Bruford, M. W., Colli, L., Masembe, C., Negrini, R., Taberlet, P., Joost, S., and Consortium, the N. (2014). High performance computation of landscape genomic models integrating local indices of spatial association. 1–15.

Sturm, R. A. (2009). Molecular genetics of human pigmentation diversity. *Hum. Mol. Genet.* 18, R9–R17. doi:10.1093/hmg/ddp003.

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., et al. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* 39, 1443–52. doi:10.1038/ng.2007.13.

Sundqvist, A.-K., Bjornerfeldt, S., Leonard, J. A., Hailer, F., Hedhammar, A., Ellegren, H., and Vilà, C. (2005). Unequal Contribution of Sexes in the Origin of Dog Breeds. *Genetics* 172, 1121–1128. doi:10.1534/genetics.105.042358.

Sved, J. a (1971). Linkage Disequilibrium and Homozygosity of Chromosome Segments in finite Populations. *Theor. Popul. Biol.* 141, 125–141. doi:10.1016/0040-5809(71)90011-6.

Sved, J. a, Cameron, E. C., and Gilchrist, a S. (2013). Estimating Effective Population Size from Linkage Disequilibrium between Unlinked Loci: Theory and Application to Fruit Fly Outbreak Populations. *PLoS One* 8, e69078. doi:10.1371/journal.pone.0069078.

Sved, J. a, and Feldman, M. W. (1973). Correlation and probability methods for one and two loci. *Theor. Popul. Biol.* 4, 129–32.

Sved, J. a, McRae, A. F., and Visscher, P. M. (2008). Divergence between human populations estimated from linkage disequilibrium. *Am. J. Hum. Genet.* 83, 737–43. doi:10.1016/j.ajhg.2008.10.019.

Szpiech, Z. a, and Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31, 2824–7. doi:10.1093/molbev/msu211.

Taberlet, P., Coissac, E., Pansu, J., and Pompanon, F. (2011). Conservation genetics of cattle, sheep, and goats. *C. R. Biol.* 334, 247–54. doi:10.1016/j.crvi.2010.12.007.

Taberlet, P., and Luikart, G. (1999). Non-invasive genetic sampling and individual identification. *Biol. J. Linn. Soc.* 68, 41–55. doi:10.1111/j.1095-8312.1999.tb01157.x.

Taberlet, P., Valentini, A., Rezaei, H.-R., Naderi, S., Pompanon, F., Negrini, R., and Ajmone-Marsan, P. (2008). Are cattle, sheep, and goats endangered species? *Mol. Ecol.* 17, 275–84. doi:10.1111/j.1365-294X.2007.03475.x.

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

Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E. G., and Risch, N. J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am. J. Hum. Genet.* 81, 626–33. doi:10.1086/520769.

Tang, H., Peng, J., Wang, P., and Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* 28, 289–301. doi:10.1002/gepi.20064.

Tapio, M., Marzanov, N., Ozerov, M., Cinkulov, M., Gonzarenko, G., Kiselyova, T., et al. (2006). Sheep mitochondrial DNA variation in European, Caucasian, and Central Asian areas. *Mol. Biol. Evol.* 23, 1776–83. doi:10.1093/molbev/msl043.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., and Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17, 520–6. doi:10.1101/gr.6023607.

Thomas, P. D., Campbell, M. J., and Kejariwal, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome …*, 2129–2141. doi:10.1101/gr.772403.2.

Tomiczek, H., and Türcke, F. (2003). *Das Muffelwild.* Kosmos. Stuttgard.

Uimari, P., and Tapio, M. (2011). Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. *J. Anim. Sci.* 89, 609–14. doi:10.2527/jas.2010-3249.

Vigne, J.-D. (2011). The origins of animal domestication and husbandry: A major change in the history of humanity and the biosphere. *Comptes Rendus - Biol.* 334, 171–181. doi:10.1016/j.crvi.2010.12.009.

Vigne, J.-D. (1992). Zooarcheology and the biogeographical history of the mammals of Corsica and Sardinia since the last ice age. *Mamm. Rev.* 22, 87–96.

Vigne, J.-D., Carrère, I., Briois, F., and Guilaine, J. (2011). The Early Process of Mammal Domestication in the Near East. *Curr. Anthropol.* 52, S255–S271. doi:10.1086/659306.

Vilà, C., Leonard, J. a., Gotherstrom, A., Marklund, S., Sandberg, K., Liden, K., Wayne, R. K., and Ellegren, H. (2001). Widespread origins of domestic horse lineages. *Science.* 291, 474–7. doi:10.1126/science.291.5503.474.

Vilà, C., Seddon, J., and Ellegren, H. (2005). Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends Genet.* 21, 214–218. doi:10.1016/j.tig.2005.02.004.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* 47, 97–120. doi:10.1146/annurev-genet-111212-133526.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72. doi:10.1371/journal.pbio.0040072.

VonHoldt, B. M., Pollinger, J. P., Earl, D. a., Knowles, J. C., Boyko, A. R., Parker, H., Geffen, E., Pilot, M., Jedrzejewski, W., Jedrzejewska, B., et al. (2011). A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Res.* 21, 1294–1305. doi:10.1101/gr.116301.110.

vonHoldt, B. M., Pollinger, J. P., Earl, D. A., Parker, H. G., Ostrander, E. A., and Wayne, R. K. (2013). Identification of recent hybridization between gray wolves and domesticated dogs by SNP genotyping. *Mamm. Genome* 24, 80–88. doi:10.1007/s00335-012-9432-0.

Walker, G. ., Dunshea, F. ., and Doyle, P. . (2004). Effects of nutrition and management on the production and composition of milk fat and protein: a review. *Aust. J. Agric. Res.* 55, 1009. doi:10.1071/AR03173.

Wang, H., Zhang, L., Cao, J., Wu, M., Ma, X., Liu, Z., Liu, R., Zhao, F., Wei, C., and Du, L. (2015). Genome-Wide Specific Selection in Three Domestic Sheep Breeds. *PLoS One* 10, e0128688. doi:10.1371/journal.pone.0128688.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360, 1395–409. doi:10.1098/rstb.2005.1682.

Wang, Y., Li, M., Stadler, S., Correll, S., Li, P., Wang, D., Hayama, R., Leonelli, L., Han, H., Grigoryev, S. a., et al. (2009). Histone hypercitrullination mediates chromatin decondensation and neutrophil extracellular trap formation. *J. Cell Biol.* 184, 205–213. doi:10.1083/jcb.200806072.

Waples, R. S., and Do, C. (2008). LDNE: a Program for Estimating Effective Population Size From Data on Linkage Disequilibrium. *Mol. Ecol. Resour.* 8, 753–6. doi:10.1111/j.1755-0998.2007.02061.x.

Webb, A. E., Gerek, Z. N., Morgan, C. C., Walsh, T. A., Loscher, C. E., Edwards, S. V, and O'Connell, M. J. (2015). Adaptive Evolution as a Predictor of Species-Specific Innate Immune Response. *Mol. Biol. Evol.* 32, 1717–29. doi:10.1093/molbev/msv051.

Weedon, M. N., and Frayling, T. M. (2008). Reaching new heights: insights into the genetics of human stature. *Trends Genet.* 24, 595–603. doi:10.1016/j.tig.2008.09.006.

Weir, B. S., and Cockeram, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution (N. Y).* 38, 1358–1370.

Weir, B. S., and Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* 95, 477–88.

Wollstein, A., and Lao, O. (2015). Detecting individual ancestry in the human genome. *Investig. Genet.* 6, 7. doi:10.1186/s13323-015-0019-x.

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

Wright, S. (1943). Isolation by distance. *Genetics* 28, 114–138.

Zeder, M. a, Emshwiller, E., Smith, B. D., and Bradley, D. G. (2006). Documenting domestication: the intersection of genetics and archaeology. *Trends Genet.* 22, 139–55. doi:10.1016/j.tig.2006.01.007.

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.

Zhang, H., Wang, Z., Wang, S., and Li, H. (2012). Progress of genome wide association study in domestic animals. *J. Anim. Sci. Biotechnol.* 3, 26. doi:10.1186/2049-1891-3-26.

Zhang, J., Liu, F., Cao, J., and Liu, X. (2015). Skin transcriptome profiles associated with skin color in chickens. *PLoS One* 10, e0127301. doi:10.1371/journal.pone.0127301.

Zohary, D., Tchernov, E., and Horwitz, L. K. (1998). The role of unconscious selection in the domestication of sheep and goats. *J. Zool.* 245, 129–135. doi:10.1111/j.1469-7998.1998.tb00082.x.