Autosomal genome-wide analysis of diversity, adaptation and morphological traits in African indigenous sheep

By

Abulgasim Mustafa Ahbara

Thesis submitted to the University of Nottingham for the award of the degree of Doctor of Philosophy

July 2019

Declaration

I hereby declare that this thesis has not been previously submitted for the award of any degree either at the University of Nottingham or any other. The research documented in this thesis was done by me, except where references to other work have been duly acknowledged.

Abulgasim Ahbara

July 2019

Dedication

To my parents, wife (Nadra), son (Mustafa) and daughters (Shada, Rafah and Jenan), brother and sisters, and future generations.

Abstract

Since their domestication in the Near East approximately 10,000 years ago, sheep have adapted to a wide spectrum of geographic, ecosystems and management regimes due to their easy handling, manageable body size and better adaptability to variable biotic and abiotic stress factors such as feed shortages and extreme climatic conditions. African indigenous sheep, as are global sheep, originated in the Near East. They arrived, in the first instance, in North Africa *via* the Isthmus of Suez by the seventh millennium before present. These sheep were of thin-tail type and their dispersion southwards into the wider eastern Africa followed possibly the Nile river valley and the Red Sea coastline. The second wave brought fat-tail sheep into North and Northeast Africa *via* two entry points, the Isthmus of Suez and the Horn of Africa across the strait of Bab-el-Mandeb, respectively. The fat-rump sheep are a recent introduction and represent the third wave of arrival and dispersal of the species into eastern Africa. Consequently, the phenotypic variation within and among these sheep populations could be explained by differences in their ancestral origins, migratory history and local natural and human-driven selection.

Sheep represents one of the most economically important livestock species in Africa, playing an important role to resource-poor farmers by providing a wide range of products and services (e.g. meat, milk, skin, hair, and manure for cash, security, gifts and religious rituals), and a form of savings and investments. However, their diversity level is likely shrinking and urgent action is required to conserve the indigenous breeds. The dependency of African agro-pastoral societies on sheep products likely co-evolved with the wide range of environmental adaptability and productivity traits present today in African indigenous sheep following natural and human-mediated selection. This work aimed to provide a comprehensive overview of genome-wide diversity, population structure and signatures of adaptive divergence of African indigenous sheep populations through the analysis of Illumina Ovine SNP50 Genotyping BeadChip and whole-genome resequencing data.

In **Chapter 2**, we investigate the autosomal genome-wide profiles of 11 Ethiopian indigenous sheep populations using the Illumina Ovine 50K SNP BeadChip assay. Sheep from the Caribbean, Europe, Middle East, China, and western, northern and southern Africa were included to address globally, the genetic variation and history of Ethiopian sheep. Population relationship and structure analysis separated Ethiopian indigenous fat-tail sheep from their North African and Middle Eastern counterparts. It indicates two main genetic backgrounds and supports two distinct genetic histories for African fat-tail sheep. Within Ethiopian sheep, our

results show that the short fat-tail sheep do not represent a monophyletic group. Four genetic backgrounds occur in Ethiopian indigenous sheep but at different proportions among the fat-rump and the long fat-tail sheep from western and southern Ethiopia. The Ethiopian fat-rump sheep share a genetic background with Sudanese thin-tail sheep. Genome-wide selection signature analysis identified putative candidate regions spanning genes influencing growth traits and fat deposition (*NPR2, HINT2, SPAG8, INSR*), development of limbs and skeletal structure, and tail formation (*ALX4, HOXB13, BMP4*), the occurrence and morphology of horns (*RXFP2*) and response to heat stress (*DNAJC18*). These findings suggest that the Ethiopian fat-tail sheep represent a uniquely admixed but distinct genepool that presents an important resource for understanding the genetic control of skeletal growth, fat metabolism and associated physiological processes.

In **Chapter 3**, we generated and analyzed whole-genome data from 60 long fat-tail, 32 short fat-tail and 38 fat-rump sheep from Ethiopia and Libya (~30x coverage) as well as 20 thin-tail sheep from Sudan (~10x coverage). Overall, 34.8 million high-quality single nucleotide polymorphisms were identified and used to assess within and among population-level genome diversity. The results from this chapter were in close agreement with those reported in **Chapter 2**. The overall results are in line with the archeological history of African sheep relating to their proposed entry points into, and subsequent dispersal across the continent. The Libyan sheep population displays a high level of genetic variation combined with lower inbreeding values and effective population sizes (N_e), a consequence, most likely of the random mating breeding applied by the owners. Further genome-wide analysis including thin-tail sheep from Ethiopia including thin and fat-tail sheep from West, North and South Africa will provide a continent-wide comprehensive overview of the genome diversity of African indigenous sheep.

In **Chapter 4**, using the whole-genome sequences, we identify candidate genome regions and genes associated with adaptation to environmental challenges and tail morphology in African sheep representing the different introduction events of the species on the continent. Based on the population relationship and structure analysis (**Chapter 3**), the study populations were initially clustered into four groups: Ethiopian fat-rump (ET_G1), Ethiopian long fat-tail (ET_G2), Libyan long fat-tail (LB) and Sudanese thin tail (SD) sheep. These groups represent sheep reared at sea level (< 50 m above sea level (asl)) (LB), in a desert environment (SD) (< 1000 m asl) and at higher altitudes (> 1300 m asl) (ET_G1 and ET_G2). Following morphological characterization of the caudal vertebrae, these populations were further

categorized into two groups: long- (ET_G2 and SD) and short-tailed (ET_G1 and LB) sheep populations. Candidate regions associated with environmental adaptation and tail morphology were identified with three selection scan indexes (*ZHp, ZF_{ST}* and *XP-EHH*), using 34 million autosomal SNPs. Many candidate genes under selection were identified including; the salt-sensitivity related gene (*PLEKHA7*) in Libyan sheep, the hypoxia associated gene (*EGLN1*) in Ethiopian sheep, the *HOXB13* gene contrasting long- and short-tailed sheep and *ALOX12* a candidate gene associated with fat deposition. These most likely are candidates encoding the unique adaptations of African sheep to diverse environments and tail phenotypes.

In conclusion, our results have provided novel insights into sheep genomic adaptations to extreme environments and they illustrate the impact that environmental challenges may have had on the tail morphology of African sheep. They offer a valuable repository of new information for future research on indigenous sheep breeding in response to current and future climatic challenges. Although the results augment archeological findings relating to sheep dispersal into and across the African continent, future genome-wide analysis including thin-tail sheep from Ethiopia as well as thin- and fat-tail sheep from western, northern and southern Africa will provide a comprehensive continent-wide insight on the genome diversity of African sheep. Furthermore, functional genomic and association studies, supported with precise phenotypic recording on larger samples sizes, will be necessary to identify the causative mutations in the candidate genes identified in this study and using genome-editing techniques further validate their functional significance.

To the best of our knowledge, this is the first study that has assessed the genome profiles and dynamics of African indigenous sheep using both SNP Chip genotypes and whole-genome resequencing in the same populations. Additionally, this may be the first study to comprehensively identify such a repertoire of adaptive candidate genes in African indigenous sheep. This study, therefore, represents a major milestone in supporting breeding programs aimed at sustainable conservation of African sheep genetic resources.

Acknowledgement

First, praise be to Allah, the Most Beneficent, the Most Merciful, who blessed me with strength and patience to undertake this work to completion.

I would like to express my sincere gratitude to my supervisor Professor Olivier Hanotte for his endless support during the course of my study and related research, for his patience, motivation, and immense knowledge. His guidance helped me immensely during the time of research and writing of the thesis. I could not have imagined having a better advisor and mentor for my PhD study. My sincere appreciation also goes to the Libyan Ministry of Higher Education and Scientific Research and University of Misurata for funding this PhD work and making my dream a reality. Great thanks to the Agricultural Research Centre (Misurata station), Libya for their kind support.

I am grateful to several people and institutions, whose advice and assistance contributed to the accomplishment of this thesis. Special thanks go to my co-supervisor's Dr Sara Goodacre for her valuable comments and Dr Joram Mwacharo for his great support during all the stages of my thesis. Great thanks to Libyan Agricultural Research Centre (Misurata station) and faculty of Agreculture (Misurata University) Libya, for collecting samples from Libya, especially Dr Mukhtar Agoub and Dr Alateirsh. Many thanks are due to the sheep sampling team in Ethiopia, especially Dr Ayelle Abeba, Dr Adebabay Kebede and Mr Agraw Amane. Big thanks to Drs Salvatore Mastrangelo, Fabio Pilla and Elena Ciani for their support in genotyping our samples. Special thanks to Professor Hassan Hussein Musa who provided Sudanese sheep whole-genome sequences. Great thanks go to my brother Abdullah Ahbara and Mr Ryad Rmaida for their great effort in processing the caudal vertebrae of Libyan and Sudanese sheep. My thanks and deep gratitude also go to the Roslin Institute in particular, Drs Emily Clark and Christelle Robert, who performed whole genome sequencing of the samples.

I wish to acknowledge the support of my colleagues at The University of Nottingham, School of Life Sciences in Professor Hanotte's lab; thank you for what you have done to me. It was a pleasure to work with you.

Publications

Ahbara, A., Bahbahani, H., Almathen, F., Al Abri, M., Agoub, M. O., Abeba, A., et al. (2019). Genome-wide variation, candidate regions and genes associated with fat deposition and tail morphology in Ethiopian indigenous sheep. *Front. Genet.* 9, 699.

Mastrangelo, S., Bahbahani, H., Moioli, B., Ahbara, A., Al Abri, M., Almathen, F., et al. (2019). Novel and known signals of selection for fat deposition in domestic sheep breeds from Africa and Eurasia. *PloS ONE* 14, e0209632.

Mastrangelo, S., Moioli, B., Ahbara, A., Latairish, S., Portolano, B., Pilla, F., et al. (2018). Genome-wide scan of fat-tail sheep identifies signals of selection for fat deposition and adaptation. *Anim. Prod. Sci.* 59, 835-848.

Selected conferences

Ahbara, A., Bahbahani, H., Mastrangelo, S., Pilla, F., Ciani, E., Hanotte, O., et al. (2017). Genome-wide scans in Ethiopian indigenous sheep populations reveal putative candidate genes associated with fat deposition and tail morphotype. In "1st World Conference on Sheep", 20 to 22, November, Beijing, China.

Ahbara, A., Mwacharo, J., Bahbahani, H., Almathen, F., Al Abri, M., Mastrangelo, S., et al. (2017). Genomic diversity and population structure analysis reveal few genetic differences among Ethiopian indigenous sheep populations. In "36th International Society for Animal Genetics Conference (ISAG)", 16 to 21, July, Dublin, Ireland.

Ahbara, A., Musa, H., Clark, E., Robert, C., Watson, M., Abeba, A., et al. (2019). Indigenous African sheep genomes reveal insights on fat-tail deposition and morphology. In "37th International Society for Animal Genetics Conference (ISAG)", 16 to 21, July, Lleida, Spain.

Table of contents

List of figures

List of tables

Chapter 1

Opening scene

Evidence of sheep domestication

Zoo-archaeology perspective

The process of sheep (*Ovis aries*) domestication, was almost certainly conducted by settled crop agriculturists (Ryder, 1984; Zeder and Smith, 2009). After the domestication of dogs, which is thought to be the first animal to have been domesticated, sheep and goats appear to have been the next based on Neolithic archaeological evidence (Clutton-Brock, 1995; Zohary et al., 1998).

Based on osteological remains from Swiss Neolithic lake dwellings, it was speculated that domestic sheep (*Ovis aries*) could have descended from both the mouflon (*O. musimon*) and Urial (*O. vignei*) ancestors (Ewart et al., 1913). Later, it was concluded that the Asiatic Mouflon (*O. orientalis*) was the most likely ancestor of domestic sheep rather than the Urial (Nadler et al., 1973; Zeder, 2008). Based on the criteria of reduction in body size, domestication of goats and sheep might have occurred around 10,000 to 9,500 before present (BP) in the southern Levant within geographic areas where crops had been domesticated (Zeder, 2008). The so-called demographic archaeological approach, which is based on the structure and dynamics of animal populations rather than body size (Chamberlain, 2009), suggests that the domestication of sheep occurred about 10,500 to 11,000 BP in a region spanning the northern Zagros mountains and south-eastern Anatolia in the Fertile Crescent (**Figure 1.1**) (Zeder, 2008; Oner et al., 2011).

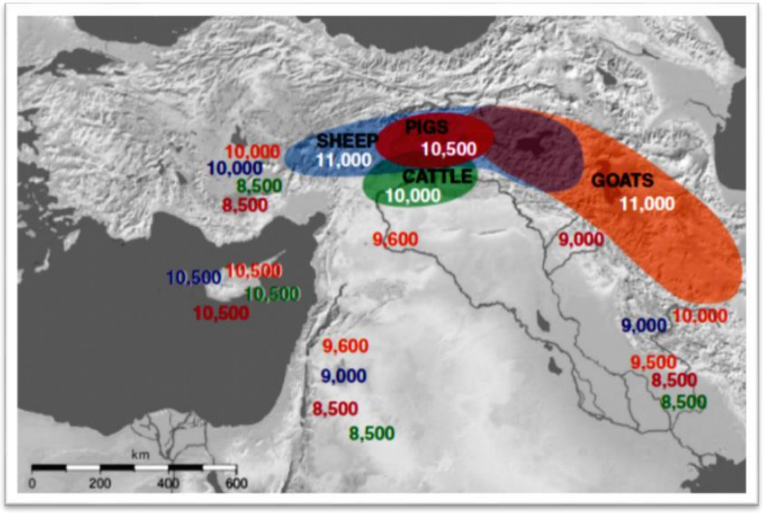


Figure 1.1 The origin and dispersal of domestic livestock species in the Fertile Crescent (Shaded areas = domestication sites, white numbers = domestication date and coloured numbers outside shaded area = approximate date when the domesticates first appeared in a region. Adapted from Zeder (2008).

Molecular perspective

The application of mitochondrial (mtDNA) and nuclear DNA sequencing to modern breeds has opened new windows for understanding domestication processes and its dating (Bruford et al., 2003). In this regard, to investigate the origin of domestic sheep (*O. aries*), the control region of mtDNA of wild mouflon (*O. musimon, O. orientalis*), urial (*O. vignei*), argali (*O. ammon*), bighorn (*O. canadensis*), and domestic sheep from Asia, Europe and New Zealand were sequenced (Hiendleder et al., 2002). The results indicate two clear separate lineages; one includes the Asiatic (*O. orientalis*) and European (*O. musimon*) mouflon with domestic sheep, while the other wild species are organized in two additional well-separated lineages. The results suggested that the domestication of sheep may have involved more than one mouflon species (Gaouar et al., 2015; Hiendleder et al., 2002). Using the same mtDNA lineage approach, several domestication events have been implicated in explaining modern-day diversity of Turkish sheep breeds (Pedrosa et al., 2005).

Using endogenous retroviruses as genetic markers, Chessa et al. (2009) suggested two different dispersal and migratory events of sheep from the centre(s) of domestication in the Fertile Crescent. The first episode is represented today by some "European mouflon" populations (Mediterranean Islands) and North-West Europe "primitive" breeds, while the second wave includes a large majority of modern-day breeds with morphological diversity corresponding to their productivity traits. Large scale genomics analysis also inferred two migratory waves of sheep across eastern Eurasia (Lv et al., 2015). In addition, in several studies, mitochondrial DNA analysis has revealed several maternal lineages (e.g. A, B, C, D, E), each with distinct geographic distribution patterns, suggesting different maternal ancestors and perhaps distinct events of sheep domestication (Tapio et al., 2006; Singh et al., 2013).

Dispersal of domestic sheep into and across Africa

Zoo-archaeology perspective

It seems that the movement of people with their animals, for environmental, commercial and political reasons, played a major role in the dispersal of domestic sheep in the Mediterranean Basin, including along the shores of North Africa, as well as across eastern Africa. African

sheep, like other domestic sheep, originate from the Asiatic mouflon (*O. orientalis*). Given that there are no wild sheep in Africa, the continent can be excluded as a centre of domestication (Gifford-Gonzalez and Hanotte, 2011; Muigai and Hanotte, 2013).

African sheep are thought to be of Near-Eastern origin (Epstein, 1954; Epstein, 1971; Ryder, 1984). According to Marshall and Weissbrod (2011), the first introduction of sheep and goats to Africa occurred during the early eighth millennium calibrated (cal) BP from Southwestern Asia. The earliest sheep in Africa were thin-tailed and hairy and were introduced to East Africa through North Africa (Marshal, 2000). Following their arrival in Egypt, archaeological evidence indicates that, together with goats, the thin-tailed sheep dispersed southwards into Sudan and Ethiopia following the Nile river basin (Chaix and Grant, 1987; Clutton-Brock, 2000).

The second wave of sheep introduction to Africa involved fat-tailed sheep that entered North Africa *via* the Isthmus of Suez and East Africa *via* the straits of Bab-el-Mandeb (Ryder, 1984). Fat-rump sheep entered East Africa much later, though their origin remains unknown (Epstein, 1954; Epstein, 1971; Ryder, 1984). In addition, it is thought that the Mediterranean maritime trading routes may have facilitated the dispersal of sheep along the coastal regions of northern Africa, and towards the western and southern parts of the continent (Ryder, 1983; Muigai and Hanotte, 2013). However, the movement of sheep towards the southern part of the continent remains largely speculative. They may have accompanied Bantu speaking populations from West Africa to South Africa through central Africa, following a western rather than the more traditional postulated eastern routes from East to South Africa (Newman, 1995). In general, the possible routes of introduction and spread of sheep pastoralism into Africa are presented in **Figure 1.2**.

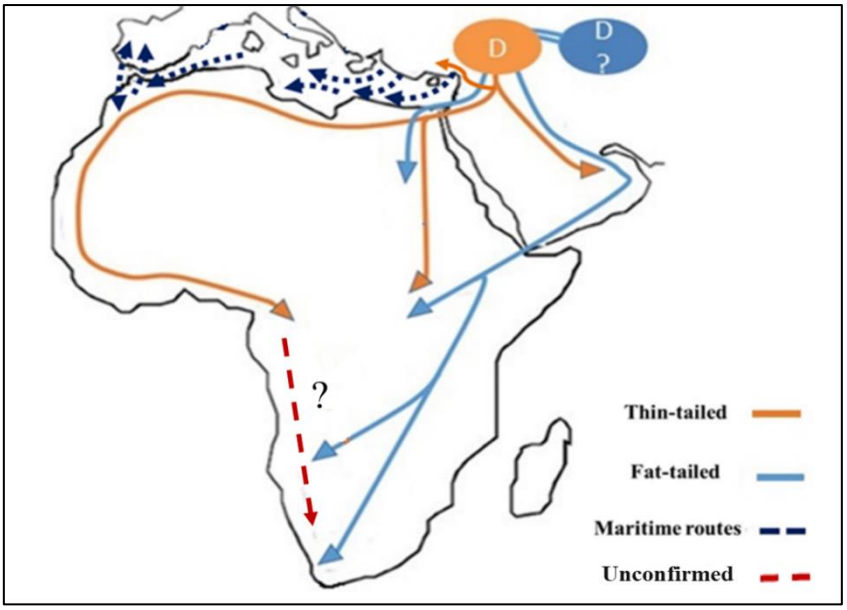


Figure 1.2 Possible routes of introduction and spread of sheep pastoralism into Africa (Gifford-Gonzalez and Hanotte, 2011; Muigai and Hanotte, 2013).

Molecular perspectives

From the molecular point of view, the spread of domestic sheep into and across northern Africa has been studied through the analysis of mtDNA and *Y*-chromosomal marker polymorphisms. MtDNA studies indicate that they probably descended from the same domestication centres as European and Asiatic sheep, as they share a common maternal ancestry.

Two mtDNA haplogroups have been identified on the African continent. Gornas et al. (2011) indicate the presence predominantly of the B haplogroup in the Sudanese sheep populations they studied in addition to haplogroup A which was detected in 10% of the samples. MtDNA was also used to analyse the maternal genetic diversity in Egyptian (Barki, Ossimi and Rahmani) and Italian (Sarda, Laticauda and Italian Muflon) sheep breeds. The results indicate that although the two groups of sheep were separately clustered, the B haplogroup was the most common followed by haplogroup C, while E was the least prevalent (Othman et al., 2014; Othman et al., 2015).

On the other hand, full mitochondrial genome analysis indicated that the West African Djallonke and Sahelian sheep breeds represent a Eurasian origin, while other West African sheep breeds are a genetic admixture of these two and the European breeds (Brahi et al., 2015). Kdidi et al. (2015a) investigated *Y*-chromosome haplotype variation in Tunisian sheep breeds and evaluated their relationship with Middle Eastern, African and European breeds. Five haplotypes (H4, H5, H6, H8, and H12) were found in Tunisian breeds indicating considerable genetic variation on the *Y* chromosome.

The mtDNA and *Y*-chromosome diversity, however, seems to be threatened by massive crossbreeding in many regions across North Africa. For instance, by investigating 30 autosomal microsatellite markers in the Rembi and Taadmit Algerian indigenous breeds, Gaouar et al., (2015) found the two breeds to have likely lost most of their autosomal genetic identity following uncontrolled admixture with the Ouled-Djellal breed. Likewise, using random amplified polymorphic DNA-polymerase chain reaction (RAPD-PCR) markers, a high level of gene flow was observed between Tunisian Barbarine and western thin-tail breeds, threatening the purity of both breeds (El Hentati et al., 2012; Wiedemar and Drögemüller, 2015). Kdidi et al. (2015b) who studied Barbarine sheep along with other Tunisian breeds (Western Thin Tail, Thibar and Sicilo Sarde), indicated that mtDNA haplotype diversity underlines the genetic relation of this breed with the Middle Eastern and European breeds.

Phenotypic diversity and spatial distribution of the studied sheep populations

Ethiopian indigenous sheep

Ethiopia is an eastern Africa state located in the Horn of Africa. It is considered to be one of the main entry points of sheep into the continent from the main centres of domestication. Sheep represents one of the most economically important livestock species in Ethiopia, playing an important role for resource-poor farmers by providing a wide range of products and services (e.g. meat, milk, skin, hair, and manure for cash, security, gifts and religious rituals), and as a source of saving and investment (Assefa et al., 2015). Ethiopia hosts a large number of distinct breeds or ecotypes of sheep that are characterized by different morphologies (Rege et al., 2007). The ecological, climatological, and human ethnic and cultural diversity of Ethiopia is reflected in its large sheep population (25.5 million heads), among which 99.8% are local breeds (Leta and Mesele, 2014), and the tail phenotype is one of the most distinctive features for their identification.

Three tail types are observed in Ethiopian sheep, fat-tailed, thin-tailed and fat-rumped. The fat-tail sheep exhibit great variation in shape and sizes with some having short tails, and other long ones. Based on the combination of tail type and length, Ethiopian sheep have been categorised into four broad breed groups: short fat-tailed, long fat-tailed, thin-tailed and fat-rumped (Gebremichael, 2008; Gizaw, 2009). The short fat-tailed population mainly inhabit the sub-alpine regions; long fat-tailed are dispersed in mid- to high-altitude environments; fat-rump are predominantly found in the dry lowland towards the East and thin-tail sheep occupy the western lowlands bordering Sudan (**Figure 1.3**; Gizaw et al., 2007).

Fibre type is another major distinguishing trait of Ethiopian indigenous sheep. They are categorised into two breed groups: course-wool sheep, which occupy the cold and sub-alpine environments, and the short-hair sheep found in warmer regions. Accordingly, Ethiopia indigenous sheep have been categorized in to 14 populations or sheep types, that are named following their geographic location and/or the ethnic communities keeping them (**Figure 1.3)**. These types have been further classified into nine breeds (**Table 1.1**; Gizaw, 2009).

Table 1.1 Main tail groups, sheep breeds and types in Ethiopia (Gizaw, 2009)

Mainn tail groups	Breeds	Sheep types	Tail type/ length	Fibre type
Short	Simien	Simien	Fatty and short	Wool/fleece
fat-tail	Short fat-tailed	Sekota, Farta, Tikur, Wollo, Menz	Fatty and short	Wool/fleece
	Washera	Washera	Fatty and short	Short hair
Long	Horro	Horro	Fatty and long	Short hair
fat-tail	Arsi-Bale	Arsi-Bale, Adilo	Fatty and long	Short hair
	Bonga	Bonga	Fatty and long	Short hair
Fat-rump	Afar	Afar	Fat rump/ fat tail	Short hair
	Blackhead Somali	Black head Somali	Fat rump/tinny tail	Short hair
Thin tail	Gumz	Gumz	Thin and long	Short hair

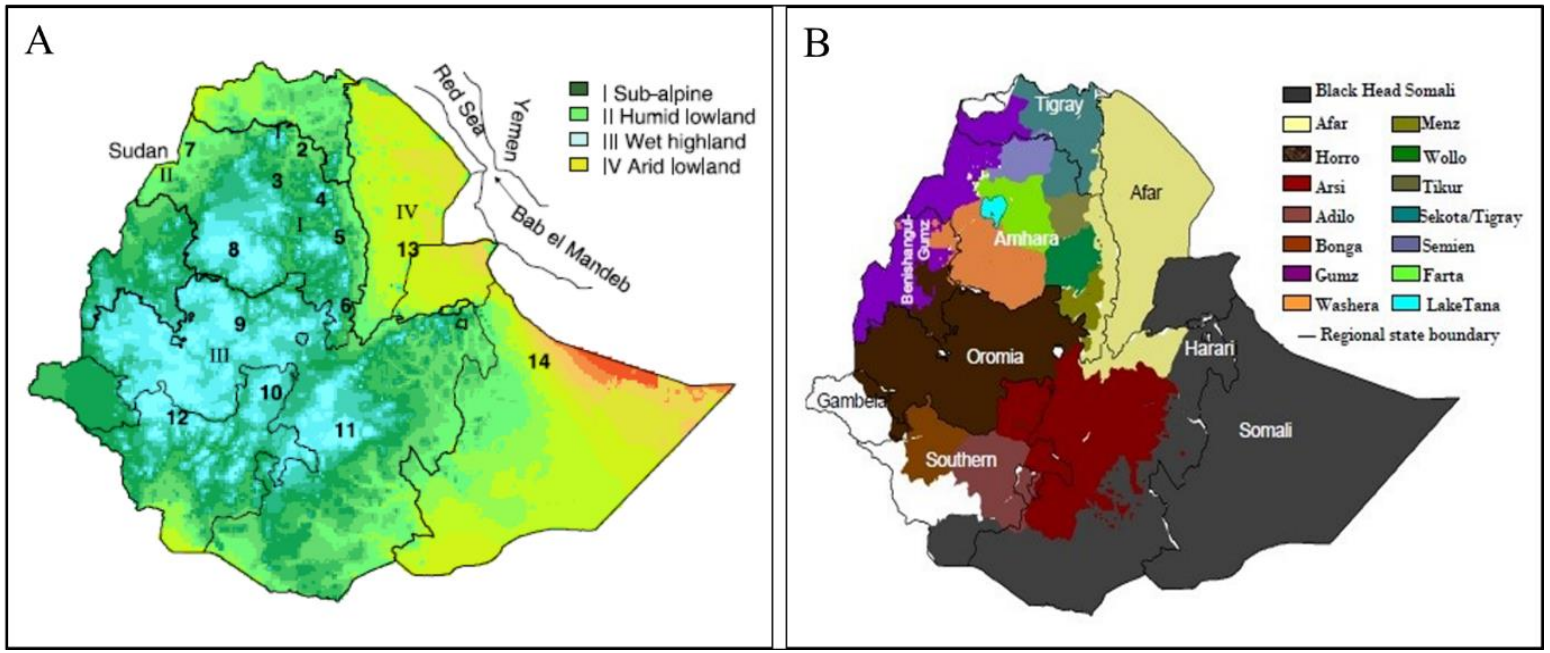


Figure 1.3 Geographic distribution of sheep breeds in Ethiopia. A. Ecological zones: I. Cool to very cold sub-moist/dry alpine mountains and plateaus, low vegetation, average elevation 3008 m above sea level (asl), average annual rainfall 1102 mm, average daily maximum temperature 22.1^{o}C, average daily minimum temperature 7.6^{o}C. II. Hot sub-humid lowland plain, high vegetation, average elevation 637 m a.s.l., average annual rainfall 894 mm, average maximum daily temperature 37.7^{o}C, average daily minimum temperature 20.1^{o}C. III. Tepid to cool wet highlands, very high vegetation, average elevation 2091 m a.s.l., average annual rainfall 1437 mm rain, average daily maximum temperature 24.8^{o}C, average daily minimum temperature 10.1^{o}C. IV. Hot arid lowland plain, very low vegetation, average elevation 894 m a.s.l., average annual rainfall 404.5 mm, average daily maximum temperature 33.2^{o}C, average daily minimum temperature 17.4^{o}C. B. The spatial distribution of Ethiopian sheep breeds according to the predominant ecotype. Source: (Gizaw et al., 2007; Gizaw, 2009).

Libyan indigenous sheep

Libya is a North African country that borders an area of 1,759,500 km^2 along the shores of the Mediterranean Sea, between 19° and 33° North latitude, with around 6.3 million people (Daw et al., 2015). Except for the coastal regions, the country is an uninhabited sand and stone wilderness which is part of the Sahara Desert. In general, except Jebel Nufusah and Jebel Akhder in the North and Tibesti mountains in the south, the topographic feature of Libya is flat and dominated by low steppes and tablelands below 762 metres above sea level. Due to its geographic location, Libya is characterised by low precipitation ranging between 100 to 350 millimetres per annum along the narrow coastal strip (approximately 150 km wide to the south). The land receiving more than 200 mm annual rainfall is considered dry farming area, mainly for barley or wheat production, while that which receives less than 200 mm is classified as rangeland, fit for livestock production (Jansen, 1988).

Small ruminants (sheep and goats) play a major role in the socio-economy of Libya providing an important source of meat, wool and milk. The demand for meat is high. Al-Mabruk and Alimon (2015) estimated that meat consumption was about 38 kg per capita in the year 2000 and about 500,000 tonnes of meat, fish and poultry, are imported annually. Therefore, sheep are mainly raised for meat production as they represent the main source of red meat, and despite the harsh environment, they efficiently utilise available feed sources to provide high quality meat that is preferred by most of the citizens (Al-Mabruk and Alimon, 2015). Estimates by Abdulkarim (2015) indicate that a 26% increase in sheep population was observed between 1990 and 2004, followed by a reduction in population size from 7,500,000 to 3,988,000 heads in 2007. It seems that this reduction relates to massive over-slaughtering and a severe shortage of vegetation in the country following suboptimal rainfall.

Sheep production systems in Libya depend on grazing and they may be classified into four types (Al-Mabruk (2013). The most valuable and common production system is the open grazing where animals are raised on native pastures moving over long distances between locations due to feeding and water scarcity. The intensive system is where sheep graze and are nutritionally supplemented with crop residues, mainly barley and wheat that are cultivated under irrigation. The mixed system is applied where grazing area is limited for large numbers of sheep, especially in summer and early spring, since animals are reared in small flock sizes of 20 to 40 heads under the guidance of employed experienced shepherds. Some Bedouin

communities have traditionally practiced the nomadic (pastoral) production system, found in some regions of the country, across generations.

The Barbary fat-tailed sheep are the most dominant breed in Libya due to its adaptation to the local harsh environment. It represents about 95% of the Libyan sheep population with around 6.5 million head and its bred for meat and wool production (Abdulkarim, 2015). Archaeological evidence indicates that the first appearance of domestic ovi-caprine, including possibly the wild Barbary sheep (*Ammotragus lervia*), was in Haua Fteah in north-eastern Libya about 6,800 Years BP (Klein and Scott, 1986). Marshall and Weissbrod (2011) argue that ovine coprolites suggest the penning of wild Barbary sheep at Una Afuda Uan Tabu and Fozzigiaren in Acacus Mountain in southern Libya during the 9^{th} to 10^{th} millennia Cal BP. However, the current contribution of this species to the domestic Barbary sheep gene pool has not yet been established (Ann Horsburgh and Rhines, 2010).

Phenotypically, the Libyan Barbary sheep are medium-sized and sexually dimorphic; adult males weigh between 40 to 60 kg and adult females between 35 to 50 kg. The height at withers of their white-coloured body ranges from 65 to 75 cm. According to their facial colouration, Libyan Barbary sheep have black, brown, white and spotted faces (Abdulkarim, 2015). The main external features of Libyan Barbary sheep are shown in **Figure 1.4**. In contrast to the other colours observed in Libyan Barbary sheep, white-faced sheep have not been observed in the neighbouring Tunisia (Ben Salem et al., 2011; Othman et al., 2015), which may indicate that these populations could have a different genetic history, or possible different recent admixture compared to Libyan breed.

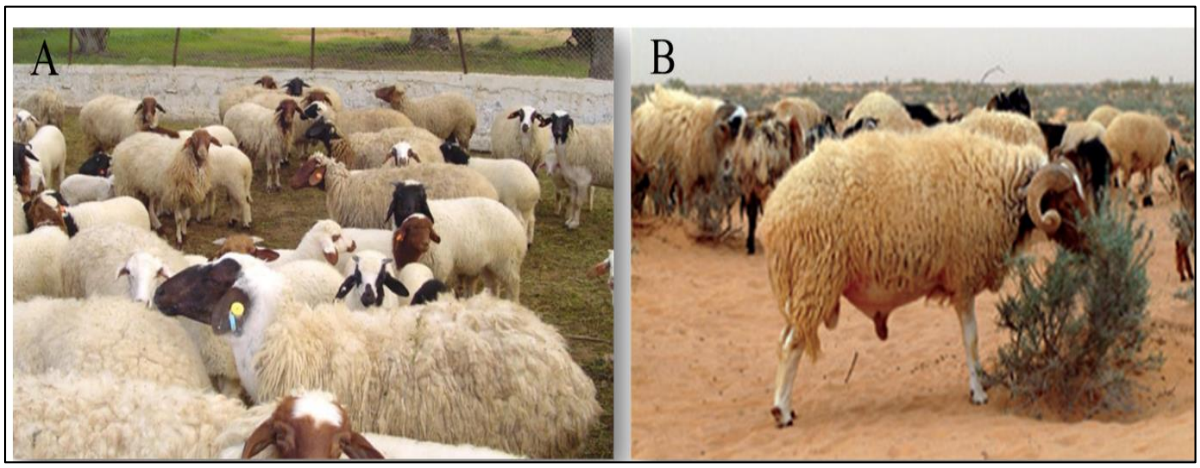


Figure 1.4 Libyan Barbary sheep. A. A different head coloured Libyan Barbary ewes. B. Libyan Barbary sheep reared in natural rangeland. Source: Abdulkarim (2015)

Sudanese indigenous sheep

Sudan is the largest country in Africa, occupying an area of about 2.5 million km2. It also has one of the largest livestock populations in the continent. In 2004, sheep represented about 36% of the estimated 140 million heads of livestock (cattle, sheep, goat, and camel) in the country (Gornas et al., 2011). Subsequently, sheep play an important role in food security and the sustenance of rural households. Furthermore, sheep also play a fundamental socio-economic role in the country and are a valuable strategic resource for local and export trade. They are also of major cultural importance due to their use in traditional rites and celebrations (Alvarez et al., 2009).

Based on their geographic distribution and phenotypes, Sudanese sheep are classified into five basic and three fused populations. The five basic populations are Desert, Nilotic, Arid equatorial, Arid upland and West African. The fused populations are the crosses of Desert and Nilotic (Nilo-desert), the crosses of Desert and Arid upland and crosses of Nilotic and Arid Equatorial (McLeroy, 1961). As indicated by Abualazayium (2004), Sudanese sheep descended from two original breeds: the Ovilongeps, the ancestors of Nilotic sheep, and the Oviplatre, the ancestors of Desert sheep of North Sudan. Numerically, only three populations are important; the Desert population which comprises 65% of the total sheep population, the Nilo-desert and the Nilotic populations, which represent 18% and 12% of the total population, respectively (Abualazayium, 2004).

Sudan Desert sheep are strictly confined to the semi-arid climatic zone. Their homeland is roughly bound in the South by latitude 12° N, although this southern boundary has recently retreated further south due to the southward advance of the Sahara Desert. The western border is marked by the range of rocky hills from Jebel Marra in the South to the Zaghawa plateau in the North. To the East, the range in distribution extends to the Red Sea hills. To the North, it fades away with an undulating border in the Nubian desert (Mufarrih, 1991).

As reported by Mufarrih (1991), this area experiences intense solar radiation from March to the end of June and has a mild, moist temperature from July to the end of October. During winter, November to February, the temperature is mild during the day and cold at night. The coastal ranges of the Red Sea receive winter showers which stimulate green grazing which varies with the rainfall amount. In the main desert sheep land, the rainfall varies from 75 mm in the far North to 400 mm in the South.

The Desert sheep comprise seven sub-types, namely Kabashi, Hammari, Meidob, Beja, Butana, Gezira and Watish. Kabashi is raised in the northern and eastern parts of North Kordofan State while Hammari occupies the western part of Kordofan State with different grades of crosses between these two tribal subtypes in central Kordofan (Ali et al., 2014). The predominant colours of Kabashi are brown, light brown (Ashgur) and spotted black or red and white (Abrug), while the main colour for Hammari is red. There are several similarities between the localities and management of these subtypes, although Kabashi migration routes are longer, averaging more than 643.6 km one way. These sheep are raised mainly under extensive nomadic conditions depending on natural grazing and fattening is rarely practiced (Abusuwar et al., 2012).

Animals managed under rangelands face many constraints including nutritional limitations, high ambient temperatures, scarcity of feed and water. These constraints have a huge effect on the reproduction and production performance of sheep in semi-arid areas of Sudan (Ebrahiem et al., 2014). During these seasons, some flock management practices adopted include night grazing which is a common practice in the region (Mufarrih, 1991). In addition to grazing during the day, night grazing is essential especially in the dry season, when available forage is low and of low quality (Bayer et al., 1987). As reported by King (1983), night grazing helps to reduce the effect of heat stress on animals. Sudan Desert sheep are generally described as long-legged. In the northern ranges where scarcity of feeds and water imposes walking long distances, the sheep have developed longer legs and light bodies. The sheep of the southern regions (such as Hammari), have shorter legs and heavy bodies. In these regions, range grazing and drinking water are abundant, grazing distances are short and the seasonal migration range is comparatively shorter (Mufarrih, 1991). Figure 1.5 shows the distinctive features and appearance of Kabashi and Hammari sheep.

Figure 1.5 The main phenotypic characteristics of Kabashi and Hammari desert sheep. Source; (https://www.slideshare.net/nahidfawi5/hamari-sheep).

Characterization of genetic diversity

Livestock play an important role in humankind as sources of meat, milk, eggs, fibre and other products and services. According to FAO (2011), there are around 8000 livestock breeds and strains representing domestic animal genetic resources (AnGR). These are of importance to food and agriculture. Since their domestication, livestock have dispersed in different habitats because of human migrations and developed through natural and artificial selection (Hanotte and Jianlin, 2006). Hence, they are considered an important part of global biodiversity.

This diversity is likely shrinking and urgent action is needed to conserve these breeds. A possible cause of this reduction relates to the emergence and spread of commercial breeds, such as Holstein-Friesian, that has been introduced globally due to its high milk production (FAO, 2011). Replacement or even cross-breeding of such highly productive breeds with local ones has resulted in growing concern on the likelihood of erosion of genetic diversity (Groeneveld et al., 2010). A case in point in sheep is two breeds of Algerian local sheep (Rembi and Taâdmit) that seem to have lost almost all of their genetic identity following intensive cross-breeding with the high meat producing Ouled-Djellal breed (Gaouar et al., 2015). Thus, under this situation, focusing efforts towards genetic and phenotypic characterisation of livestock breeds is recommended. Therefore, a plan of action needs to be adopted to genetically characterize native breeds, especially in developing countries, to guide decision-making in prioritizing breeds for conservation.

Until recently, researchers in developing countries have concentrated their work on phenotypic characterization of breeds, with relatively few studies conducted on molecular characterization. The use of the later in combination with statistical analysis could open new windows towards setting conservation priorities (Hanotte and Jianlin, 2006). In the early 1990's, awareness was raised by FAO for the need to monitor and consider molecular approaches to characterize livestock genetic resources (FAO, 2011).

Through cytochrome b (Cyt-b) sequences analysis, the divergence of the two most common bovine and ovine lineages (A and B) has been estimated to be around 30 million years BP, which is considerably earlier than proposed previously (Hiendleder et al., 1998). Geographically, sheep mitochondrial A and B haplogroups are mainly present in Asia, while B is mostly found in Europe (Meadows et al., 2007). Haplogroup C has been reported in different areas such as Portugal, Turkey, the Caucasus and China (Tapio et al., 2006). Other haplogroups are less common. Haplogroup D, possibly related to haplogroup A, has been found

in Roumanian Karachai and Caucasian sheep, while haplogroup E, intermediate between A and C, is present in some Turkish animals (Groeneveld et al., 2010). Despite the comparable diversity in mtDNA haplogroups, it seems that sheep haplogroups are less distinct than the taurine–zebu mtDNA ones (Bruford et al., 2003). Although it is an informative marker, the use of mtDNA in domestication studies is limited due to the poor prediction of the overall diversity of the genome because it acts as a single locus (Bruford et al., 2003). Furthermore, mtDNA is maternally inherited, hence it does not consider the paternal contribution to gene flow that is also critical in the evolution of livestock species (MacHugh et al., 1997). Alternatively, *Y* chromosomal diversity can be used as a marker for male introgression (FAO, 2011). For instance, recent paternal gene flow between sheep breeds from distinct geographic regions in North Eurasia was revealed through the analysis of two SNPs (A-oY1 and G-oY2) on the *Y* chromosome (Zhang et al., 2014). However, less diversity has been observed in sheep paternal lineages and very few markers are available for paternal analysis. Previous *Y* chromosomal analyses in sheep revealed 18 haplotypes and two main lineages with distinct geographic distribution patterns (Meadows et al., 2006; Ferencakovic et al., 2013; Kdidi et al., 2015a).

So far, autosomal microsatellites have been widely used in genetic characterization and diversity studies of sheep as they provide a high range of variability (FAO, 2011). They allow estimates to be made of several parameters including expected heterozygosity and allelic richness and may reveal effects of genetic isolation, inbreeding, population bottlenecks, introgression and subdivision. Using microsatellites to assess genetic structure and variability of Pramenka (Adriatic and western Dinaric) sheep populations revealed that, among the 12 breeds studied, three were under the risk of inbreeding (Salamon et al., 2014). Despite large population sizes, recommendations were made to conserve three endangered Italian sheep breeds because of high levels of heterozygote deficiency at microsatellite loci (Bozzi et al., 2009). Leroy et al. (2015) reported that ancient genetic introgression episodes have occurred in French sheep breeds following the use of Merino sheep in improving wool production. In Tunisia, low genetic diversity between sheep breeds and deficiency of heterozygotes were observed (Kdidi et al., 2015b). However, despite their high informativeness and wide use in diversity studies, microsatellite markers are labour intensive in genotyping and scoring (Ajmone-Marsan et al., 2014).

The recent development of high-density single nucleotide polymorphisms (SNP) chips have made possible the assessment and characterisation of levels of sheep breed diversity genome-wide. Depending on the species considered, panels with 50,000 to 1 million SNPs are now

available. These have replaced microsatellites in paternity testing and genetic diversity studies as well as in genome-wide association analysis, due to their robustness, low cost and automated allele calling (Ajmone-Marsan et al., 2014). In this regard, SNPs display higher levels of variation in sheep perhaps due to larger effective population sizes compared to goats and cattle (Kijas et al., 2012). Genome-wide SNP variation could be used as an effective tool to reveal population structure in different sheep populations. A study by Kijas et al. (2009) using Illumina OvineSNP50 BeadChip shows that sheep breeds cluster following their geographic origins, indicating that SNPs can effectively identify population structure even when a small number of animals represent a breed. Similarly, the same chip differentiated Suffolk sheep from Rambouillet, Columbia, Polypay, and Targhee. This was not surprising, as the Rambouillet has contributed to the genetic make-up of the later three breeds (Zhang et al., 2013).

One drawback of SNP arrays is ascertainment bias, a consequence of the restricted number of populations that were initially selected to design the arrays, which can result in inaccurate estimates of genetic diversity (FAO, 2011).

Signatures of selection and adaptation in domestic sheep

Detecting genes associated with signatures of selection can provide valuable insights in understanding how evolution has shaped the genome and how the phenotypic traits in modern livestock have been shaped by domestication. Detection of the footprints left by differential selection (natural and human) assumes the absence of the effects of population expansion, subdivision and bottlenecks (Gouveia et al., 2014). However, footprints of selection can be mimicked by population demographic factors, which may lead to assessment bias (Kim and Stephan, 2002; Stephan, 2010). Identifying regions under selection depends on the approach used. Methods for detecting selection sweeps can be classified according to the test they are based on (Gouveia et al., 2014). First, selection footprints can be estimated as a ratio between non-synonymous (dN) and synonymous (dS) substitutions, where $dN/dS > 1$ and $dN/dS < 1$ are interpreted as positive and negative selection signals respectively. The ratio is not significantly different from one ($dN/dS = 1$) under the neutral evolution hypothesis (Yang and Nielsen, 1998).

The other method is based on frequency spectrum, which refers to the difference between the average number of nucleotides and the number of polymorphic sites along a DNA sequence. This difference when standardized is called the Tajima's D test. Tajima's D is expected to be

zero (D = 0) under neutrality, less than zero (D < 0) indicates recent selective sweep because of the surplus of rare variants surrounding the selected locus due to reduced heterozygosity. Under long-term balancing selection, Tajima's D is expected to be greater than zero (D > 0) because of increased heterozygosity around the selected site (Tajima, 1989; Charlesworth, 2006). To distinguish between positive and negative selection, Fay and Wu (2000) proposed a test similar to Tajima's D but, instead of the number of segregating sites along the DNA sequence, it takes in account the homozygosity of the derived alleles.

The third approach includes several statistical tests that rely on patterns of linkage disequilibrium (LD). It includes the long-range haplotype (LRH), extended haplotype homozygosity (EHH), the relative EHH (REHH), the integrated haplotype score (*iHS*), the cross population EHH (*XP-EHH*) and the LD decay tests (Sabeti et al., 2002; Gouveia et al., 2014). Sabeti et al. (2002) defined the Extended Haplotype Homozygosity (EHH) as "the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent for the entire interval from the core loci to a point x". It may be recognized when positive selection causes a rapid increase in the frequency of a favorable allele within a short time period, and consequently, leads to the generation of an unusually long haplotype.

Based on EHH, Voight et al. (2006) proposed the *iHS* statistic, which compares the observed decay of EHH with genomic distance for the ancestral allele to that of the derived allele for a core SNP in a population. However, this approach lacks the robustness to identify a selective sweep when the selected allele is at a high frequency or near complete fixation (Tang et al., 2007; Gautier and Naves, 2011). Later, Tang et al. (2007) developed a new statistic,*Rsb*, which compares the observed decay of EHH for a specific SNP (EHHS), instead of an allele, with the genomic distance between two populations. EHHS is estimated by averaging the EHH for the two alternative alleles weighted by the square of their allele frequencies.

The fixation index (F_{ST}; Wright, 1951), estimates the degree of genetic differentiation between two populations based on the correlation between gametes chosen at random from within the same sub-population relative to the entire population. In other words, it indicates the reduction in heterozygosity within sub-populations relative to the total population (Holsinger and Weir, 2009). Under the condition of differential selection, two populations may favor alternative alleles at the loci under selection. Subsequently, the heterozygosity of the two populations will be reduced and, therefore, an increase in their genetic differentiation at these loci (high F_{ST}) will be observed. This algorithm can be estimated using the following equation:

$$F_{ST} = \frac{H_T - H_S}{H_T} \quad \textbf{(1.1)}$$

Where F_{ST} is the fixation index, H_T is the heterozygosity of the total population and H_S is the average heterozygosity of the two sub-populations. Fixation index, however, suffers from limitations likely to direct the search for genes under selection towards false positives. This most likely happens when allele frequency differences are being compared between populations with divergent effective population sizes (*N*e). As a consequence, populations with divergent *Ne* will exhibit raised homozygosity due to co-ancestry and/or genetic drift, rather than elevated homozygosity resulting from positive recent selection (Kijas, 2014).

Moreover, the fixation index (F_{ST}) ignores hidden ancestral relationships between subsets of breeds that are hierarchically structured. This means, recently developed breeds from the same ancestral population are likely to have a large number of highly correlated allele frequencies when compared to a genetically divergent breed. To address these drawbacks, Fariello et al. (2013) developed a haplotype-based statistic termed *hapFLK* that accounts for hierarchical population structure. This algorithm estimates differences in haplotype frequency between populations after accounting for the relationship between breeds. The *hapFLK* approach can be applied to un-phased genotypic data and does not require information describing the ancestral allele at each locus, meaning it is well suited to SNP Chip datasets (Kijas, 2014).

As positive selection reduces genomic diversity (increases homozygosity) at specific regions, such regions are likely to be informative in revealing genome-wide selective sweeps. Therefore, the *Hp* selective sweep analysis was developed to detect genomic signatures of selection in different livestock species using full genome sequences. It has been applied in chickens (Rubin et al., 2010), pigs (Rubin et al., 2012), cattle (Liao et al., 2013) and sheep (Zhi et al., 2018). This statistic considers pooled heterozygosity (*Hp*) across the genome to detect regions with low heterozygosity based on adequate intervals of pre-defined windows.

Importance of identifying genome-wide signatures of positive selection in indigenous African sheep

The process of domestication and selection has been shaping the genetic pool of sheep breeds resulting in various traits of ecological and commercial importance including environmental adaptation, appearance, and production traits (Randhawa et al., 2016; Taye et al., 2017). Therefore, determining the genes and genomic regions influenced by selection is crucial in understanding the biological mechanisms underlying phenotypic differences between sheep

breeds that have been selected for different purposes and under different environmental conditions (Rothammer et al., 2013).

The availability of genomic data (SNPs and genome sequences) and the comparison of genomic patterns of SNP variation between divergent breeds has successfully opened the window towards identifying putative genomic regions and genes associated with differences in selection pressure in sheep (Kijas et al., 2012; Moradi et al., 2012; Lv et al., 2014; Moioli et al., 2015; Zhi et al., 2018). Among which, are several investigations concerning fat-tail morphology, the most distinguishing feature of sheep breeds. It reflects how the ovine gene pool has been shaped since domestication up to their final destination of dispersion and afterward as an adaptive mechanism. It can be argued that elevated fat tail weight was not driven by natural selection alone since none of the wild ancestors of domestic sheep occupying similar tropical conditions have fat tails (Nejati-Javaremi et al., 2007). This indicates the importance of artificial selection driven by cultural differences between communities in shaping sheep tails.

In Ethiopia, variations in tail length and morphology have been associated with ecological, climatological, ethnic, and human cultural diversity, which are reflected in its large number of sheep populations (Gizaw et al., 2007; Leta and Mesele, 2014; Edea et al., 2017). Tail morphology, including fat quantity and tail length, acts as a good indicator of meat production. Positive phenotypic covariance exists between body and fat weight (Zamiri and Izadifard, 1997), while increased carcass quality with decreased fat tail weight was reported in the progeny of the fat-tailed Baluchi and Mehraban sheep crossed with Targhee and Corriedal (Farid, 1991). The length of the sheep tail may be associated with vertebral length and number. This may alternatively reflect higher carcass weights and meat quantity. For instance, a positive association was observed between thoracolumbar vertebrae number and carcass length and weight in Kazakh sheep. It supports the economic importance of vertebral numbers in lamb production. (Zhang et al., 2017).

The genomes of African indigenous sheep have been subjected mainly to natural selection driven by tropical and sub-tropical climates, diseases and parasites. These sheep can be considered of interest in understanding the genetic history of indigenous sheep in the continent and the genomic mechanisms underlying their adaptation to different ecosystems. The later remains poorly investigated. Furthermore, despite their low productivity, compared to their commercial counterparts, indigenous sheep are the only available choice to contribute to food

security for millions of people under restricted feed and water resources and high ecto- and endo-parasite and disease challenges.

Identifying candidate genomic regions and genes displaying signatures of positive selection, associated with African indigenous sheep adaptations is a vital first step to conserve and utilize these genetic resources. This will help in establishing breeding programs to select sheep for breeding improvement while conserving their valuable adaptations. Consequently, these programs can contribute to reducing poverty in developing countries.

Here, it is assumed that African sheep genomes may have been shaped by their translocation and introduction events into the continent. The varied environmental conditions of the current heartland of each group and population of sheep may also have played a role in shaping their genome profiles. This may have affected their genome diversity and evolution under different ecosystems.

This thesis aims to provide a comprehensive view of the genetic diversity, population structure and adaptation of African indigenous sheep populations through the analysis of Illumina Ovine SNP50 BeadChip and whole-genome resequencing data. Specifically, the history and current level of diversity of African indigenous sheep will be assessed. Possible signatures of adaptation to local environments in northern and eastern Africa will be assessed through a comparative analysis of sheep populations from Ethiopia, Sudan and Libya. The study also aimed to identify candidate regions controlling variations in tail morphology in sheep.

References

Abdulkarim, A. (2015). Small ruminant contribution in meat production in Libya. *Egyptian Journal of Sheep and Goats Sciences*, 10(1), pp. 1-6.

Abualazayium, M. (2004). Animal Wealth and Animal Production in Sudan. Khartoum University Press, Khartoum, Sudan.

Abusuwar, A. O., Ahmed, E. O., and Blal, A. (2012). Effect of feeding treated groundnut hulls with molasses on performance of Desert Sheep during Late Summer in Arid Rangelands of Western Sudan. *Int. J. of Theoretical & Applied Sciences* 4, 122-127.

Ajmone-Marsan, P., Colli, L., Han, J. L., Achilli, A., Lancioni, H., Joost, S., et al. (2014). The characterization of goat genetic diversity: Towards a genomic approach. *Small Ruminant Res.* 121, 58-72.

Ali, M., Abdella, M., Elimam, H., Sulieman, M., El-Hag, A., Neama, F., et al. (2014). Pre-weaning Body Measurements and Performance of Desert Sheep (Tribal Subtypes Hamari and Kabashi) Lambs of Kordofan Region, Sudan. *Mal. J. Anim. Sci.* 17, 35-45.

Al-Mabruk, R. M. (2013). Nutritional Part of Sheep and Goat in Libya. *Int. J. Adv. Biol. Biomed. Res.* 1, 112-133.

Al-Mabruk, R., and Alimon, A. (2015). Feed resources for small ruminants in Libya: A review. *Mal. J. Anim. Sci.* 18, 1-21.

Alvarez, I., Traore, A., Tamboura, H., Kaboré, A., Royo, L., Fernández, I., et al. (2009). Microsatellite analysis characterizes Burkina Faso as a genetic contact zone between Sahelian and Djallonké sheep. *Anim. Biotechnol.* 20, 47-57.

Ann Horsburgh, K., and Rhines, A. (2010). Genetic characterization of an archaeological sheep assemblage from South Africa's Western Cape. *J. Archaeol. Sci.* 37, 2906-2910.

Assefa, F., Tigistu, T., and Lambebo, A. (2015). Assessment of the production system, major constraints and opportunities of sheep production in Doyogena Woreda, Kembata Tambaro Zone, Southern Ethiopia. *J. Biol. Agric. Healthc.* 5, 37–41.

Bayer, W., Suleiman, H., Kaufmann, R. v., and Waters-Bayer, A. (1987). Resource use and strategies for development of pastoral systems in subhumid West Africa-The case of Nigeria. *Q. J. Int. Agric.* 26, 58-71.

Ben Salem, H., Lassoued, N., and Rekik, M. (2011). Merits of the fat-tailed Barbarine sheep raised in different production systems in Tunisia: digestive, productive and reproductive characteristics. *Trop. Anim. Health. Prod.* 43, 1357-70.

Bozzi, R., Degl'Innocenti, P., Diaz, P. R., Nardi, L., Crovetti, A., Sargentini, C., et al. (2009). Genetic characterization and breed assignment in five Italian sheep breeds using microsatellite markers. *Small Ruminant Res.* 85, 50-57.

Brahi, O., Xiang, H., Chen, X., Farougou, S., and Zhao, X. (2015). Mitogenome revealed multiple postdomestication genetic mixtures of West African sheep. *J. Anim. Breed. Genet.* 132, 399-405.

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

Chaix, L., and Grant, A. (1987). A study of a prehistoric population of sheep (Ovis aries L.) from Kerma (Sudan). Archaeozoological and archaeological implications. *Archaeozoologia* 1, 77-92.

Chamberlain, A. (2009). Archaeological demography. *Hum. Biol.* 81, 275-287.

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2(4): e64.

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

Clutton-Brock, J. (1995). Origins of the dog: domestication and early history. In "The domestic dog: Its evolution, behaviour and interactions with people" (J. Serpell, ed.), pp. 7-20. Cambridge University, Kambridge, UK.

Clutton-Brock, J. (2000). Cattle, sheep, and goats south of the Sahara: an archaeozoological perspective. In "The origins and development of African livestock: Archaeology, genetics, linguistics and ethnography" (R. M. Blench and K. C. MacDonald, eds.), pp. 30-37. Routledge, London, UK.

Daw, M. A., El-Bouzedi, A., and Dau, A. A. (2015). Libyan armed conflict 2011: mortality, injury and population displacement. *Afr. J. of Emergency Medicine* 5, 101-107.

Ebrahiem, M., Turki, I., Haroun, H., Bushara, I., and Mekki, D. (2014). The effect of natural pastures grazing conditions on skin\leather quality of Sudan desert sheep. *Glob. J. Anim. Sci. Res.* 2(3): 299-303.

Edea, Z., Dessie, T., Dadi, H., Do, K.-T., and Kim, K.-S. (2017). Genetic Diversity and Population Structure of Ethiopian Sheep Populations Revealed by High-Density SNP Markers. *Front. Genet.* 8, 218.

El Hentati, H., Hamouda, M. B., and Chriki, A. (2012). Genetic differentiation and gene flow between the Tunisian ovine breeds Barbarine and Western thin tail using random amplified polymorphic DNA-polymerase chain reaction (RAPD-PCR) analysis. *Afr. J. Biotechnol.* 11, 16291-16296.

El-Barasi, Y. M. M., and Saaed, M. W. B. (2013). Threats to plant diversity in the North Eastern part of Libya (El-Jabal El-Akahdar and Marmarica Plateau). *J. Environ. Sci.* 2, 41-58.

Epstein, H. (1954). The fat-tailed sheep of East Africa. *East Afr. J. Agr.* 20, 109-117.

Epstein, H. (1971). "The origin of the domestic animals of Africa" Africana publishing corporation, New York.

Ewart, J. C., Highland, R., and Translations, A. S. o. S. (1913). "Domestic Sheep and Their Wild Ancestors: I. Sheep of the Mouflon and Urial Types," University of California, California, USA.

FAO (2011). "Molecular Genetic Characterization of Animal Genetic Resources. FAO Animal Production and Health Guidelines.

Farid, A. (1991). Slaughter and carcass characteristics of three fat-tailed sheep breeds and their crosses with Corriedale and Targhee rams. *Small Ruminant Res.* 5, 255-271.

Fariello, M. I., Boitard, S., Naya, H., San Cristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193, 929-941.

Fay, J. C., and Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. Genetics 155, 1405-1413.

Ferencakovic, M., Curik, I., Perez-Pardal, L., Royo, L. J., Cubric-Curik, V., Fernandez, I., et al. (2013). Mitochondrial DNA and Y-chromosome diversity in East Adriatic sheep. *Anim. Genet.* 44, 184-92.

Gaouar, S. B., Da Silva, A., Ciani, E., Kdidi, S., Aouissat, M., Dhimi, L., et al. (2015). Admixture and local breed marginalization threaten Algerian sheep diversity. *PLoS ONE* 10, e0122667.

Gautier, M., and Naves, M. (2011). Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.* 20, 3128-3143.

Gebremichael, S. (2008). "Sheep resources of Ethiopia: genetic diversity and breeding strategy," Wageningen University, The Netherlands.

Gifford-Gonzalez, D., and Hanotte, O. (2011). Domesticating animals in Africa: implications of genetic and archaeological findings. *J. of World Prehistory* 24, 1-23.

Gizaw, S. (2009). Sheep Breeds of Ethiopia: a guide for identification and utilization. *Technical Bulletin*.

Gizaw, S., Van Arendonk, J. A., Komen, H., Windig, J., and Hanotte, O. (2007). Population structure, genetic variation and morphological diversity in indigenous sheep of Ethiopia. *Anim. Genet.* 38, 621-628.

Gornas, N., Weimann, C., El Hussien, A., and Erhardt, G. (2011). Genetic characterization of local Sudanese sheep breeds using DNA markers. *Small Ruminant Res.* 95, 27-33.

Gouveia, J. J. d. S., Silva, M. V. G. B. d., Paiva, S. R., and Oliveira, S. M. P. d. (2014). Identification of selection signatures in livestock species. *Genet. Mol. Biol.* 37, 330-342.

Groeneveld, L., Lenstra, J., Eding, H., Toro, M., Scherf, B., Pilling, D., et al. (2010). Genetic diversity in farm animals–a review. *Anim. Genet.* 41, 6-31.

Hanotte, O., and Jianlin, H. (2006). Genetic characterization of livestock populations and its use in conservation decision-making. The Role of Biotechnology in Exploring and Protecting Agricultural Genetic Resources. Food and Agriculture Organization of the United Nations, Rome, 89-96.

Hiendleder, S., Lewalski, H., Wassmuth, R., and Janke, A. (1998). The complete mitochondrial DNA sequence of the domestic sheep (Ovis aries) and comparison with the other major ovine haplotype. *J. Mol. Evol.* 47, 441-448.

Holsinger, K. E., and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F ST. *Nat. Rev. Genet.* 10, 639.

Jansen, H. C. (1988). Range development in northern Libya. *Rangelands* 4, 178-182.

Kdidi, S., Calvo, J., González-Calvo, L., Sassi, M. B., Khorchani, T., and Yahyaoui, M. (2015b). Genetic relationship and admixture in four Tunisian sheep breeds revealed by microsatellite markers. *Small Ruminant Res.* 131, 64-69.

Kdidi, S., Yahyaoui, M. H., Garcia-Manrique, B., Sarto, P., SASSI, M. B., Khorchani, T., et al. (2015a). Y chromosome haplotype characterization of Tunisian sheep breeds. *Turk. J. Vet. Anim. Sci.* 39, 333-337.

Kijas, J. W. (2014). Haplotype-based analysis of selective sweeps in sheep. *Genome* 57, 433-437.

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M., et al. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 10, e1001258.

Kijas, J. W., Townley, D., Dalrymple, B. P., Heaton, M. P., Maddox, J. F., McGrath, A., et al. (2009). A genome-wide survey of SNP variation reveals the genetic structure of sheep breeds. *PloS ONE* 4, e4668.

Kim, Y., and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160, 765-777.

King, J. M. (1983). Livestock water needs in pastoral Africa in relation to climate and forage. ILCA. Addis Ababa, Ethiopia.

Klein, R. G., and Scott, K. (1986). Re-analysis of faunal assemblages from the Haua Fteah and other Late Quaternary archaeological sites in Cyrenaican Libya. *J. Archaeol. Sci.* 13, 515-542.

Leroy, G., Danchin-Burge, C., Palhière, I., SanCristobal, M., Nédélec, Y., Verrier, E., et al. (2015). How do introgression events shape the partitioning of diversity among breeds: a case study in sheep. *Genet. Sel. Evol.* 47, 48.

Leta, S., and Mesele, F. (2014). Spatial analysis of cattle and shoat population in Ethiopia: growth trend, distribution and market access. *Springerplus* 3, 310.

Liao, X., Peng, F., Forni, S., McLaren, D., Plastow, G., and Stothard, P. (2013). Whole-genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. *Genome* 56, 592-598.

Lv, F. H., Peng, W. F., Yang, J., Zhao, Y. X., Li, W. R., Liu, M. J., et al. (2015). Mitogenomic Meta-Analysis Identifies Two Phases of Migration in the History of Eastern Eurasian Sheep. *Mol. Biol. Evol.* 32, 2515-2533.

Lv, F.-H., Agha, S., Kantanen, J., Colli, L., Stucki, S., Kijas, J. W., et al. (2014). Adaptations to climate-mediated selective pressures in sheep. *Mol. Biol. Evol.* 31, 3324-3343.

MacHugh, D. E., Shriver, M. D., Loftus, R. T., Cunningham, P., and Bradley, D. G. (1997). Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (Bos taurus and Bos indicus). *Genetics* 146, 1071-86.

Marshal, F. (2000). The origins and spread of domestic animals in East Africa. In "The Origins and Development of African Livestock: Archaeology, Genetics, Linguistics and Ethnography" (R. M. Blench and K. C. MacDonald, eds.), pp. 191-218. University College London Press, London.

Marshall, F., and Weissbrod, L. (2011). Domestication processes and morphological change: through the lens of the donkey and African pastoralism. *Curr. Anthropol.* 52, S397-S413.

McLeroy, G. B. (1961). The sheep of the Sudan. 2. Ecotypes and tribal breeds. *Sudan J. Vet. Sci. Anim. Husb.* 2, 1151.

Meadows, J. R., 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, 1371-1379.

Meadows, J. R., Hanotte, O., Drögemüller, C., Calvo, J., Godfrey, R., Coltman, D., et al. (2006). Globally dispersed Y chromosomal haplotypes in wild and domestic sheep. *Anim. Genet.* 37, 444-453.

Melinda A. Zeder, and Bruce D. Smith (2009). A Conversation on Agricultural Origins: Talking Past Each Other in a Crowded Room. *Curr. Anthropol.* 50, 681-690.

Moioli, B., Pilla, F., and Ciani, E. (2015). Signatures of selection identify loci associated with fat tail in sheep. *J. Anim. Sci.* 93, 4660–4669.

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

Mufarrih, M. E. (1991). Sudan desert sheep: Their origin, ecology and production potential. *World Anim. Rev.* 66, 23-31.

Muigai, A. W. T., and Hanotte, O. (2013). Then origin of African sheep: Archaeological and genetics perspectives. *Afr. Archaeol. Rev.* 30, 39–50.

Nadler, C., Hoffmann, R., and Woolf, A. (1973). G-band patterns as chromosomal markers, and the interpretation of chromosomal evolution in wild sheep (Ovis). *Experientia* 29, 117-119.

Nejati-Javaremi, A., Izadi, F., Rahmati, G., and Moradi, M. (2007). Selection in fat-tailed sheep based on two traits of fat-tail and body weight versus single-trait total body weight. *Int. J Agri. Biol.* 9, 645–648.

Newman, J. L. (1995). The Peopling of Africa: A Geographic Interpretation. Dexter, MI: Yale University Press.

Oner, Y., Calvo, J. H., and Elmaci, C. (2011). Y chromosomal characterization of Turkish native sheep breeds. *Livest. Sci.* 136, 277-280.

Othman, O. E., Balabel, E. A., and Abdel-Samad, M. F. (2014). Mitochondrial DNA diversity in five Egyptian sheep breeds. *Glob. Veterin.* 12, 369-375.

Othman, O. E., Pariset, L., Balabel, E. A., and Marioti, M. (2015). Genetic characterization of Egyptian and Italian sheep breeds using mitochondrial DNA. J. Genet. Eng. *Biotechnol.* 13, 79-86.

Randhawa, I. A., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2016). A meta-assembly of selection signatures in cattle. *PLoS ONE* 11, e0153013.

Rege, J., Ayalew, W., Getahun, E., Hanotte, O., and Dessie, T. (2007). Domestic Animal Genetic Resources Information System (DAGRIS). Addis Ababa, Ethiopia: International Livestock Research Institute.

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 genom.* 14, 908.

Rubin, C.-J., Megens, H.-J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., et al. (2012). Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences* 109, 19529-19536.

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R., Sherwood, E., Webster, M. T., et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587–591.

Ryder, M. L. (1983). Sheep and Man. London: Gerald Duckworth & Co. Ltd.

Ryder, M. L. (1984). Sheep. In "Evolution of Domestic Animals" (I. L. Mason, ed.), pp. 63-85. Longman, London.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837.

Salamon, D., Gutierrez-Gil, B., Arranz, J., Barreta, J., Batinic, V., and Dzidic, A. (2014). Genetic diversity and differentiation of 12 eastern Adriatic and western Dinaric native sheep breeds using microsatellites. *Animal* 8, 200-207.

Singh, S., Kumar, S., Jr., Kolte, A. P., and Kumar, S. (2013). Extensive variation and sub-structuring in lineage A mtDNA in Indian sheep: genetic evidence for domestication of sheep in India. *PLoS ONE* 8, e77858.

Stephan, W. (2010). Genetic hitchhiking versus background selection: the controversy and its implications. *Phil. Trans. R. Soc. B: Biol. Sci.* 365, 1245-1253.

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

Tang, K., Thornton, K. R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5, e171.

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.

Taye, M., Lee, W., Jeon, S., Yoon, J., Dessie, T., Hanotte, O., et al. (2017). Exploring evidence of positive selection signatures in cattle breeds selected for different traits. *Mamm. Genome* 28, 528-541.

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.

Wiedemar, N., and Drögemüller, C. (2015). A 1.8-kb insertion in the 3′-UTR of RXFP2 is associated with polledness in sheep. *Anim. Genet.* 46, 457–461.

Wright, S. (1951). The genetical structure of populations. *Ann. Genet.* 15, 323-354.

Yang, Z., and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46, 409-418.

Zamiri, M., and Izadifard, J. (1997). Relationships of fat-tail weight with fat-tail measurements and carcass characteristics of Mehraban and Ghezel rams. *Small Ruminant Res.* 26, 261-266.

Zeder, M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci. USA* 105, 11597-604.

Zhang, L., Mousel, M. R., Wu, X., Michal, J. J., Zhou, X., Ding, B., et al. (2013). Genome-wide genetic diversity and differentially selected regions among Suffolk, Rambouillet, Columbia, Polypay, and Targhee sheep. *PloS ONE* 8, p.e65942.

Zhang, M., Peng, W. F., Yang, G. L., Lv, F. H., Liu, M. J., Li, W. R., et al. (2014). Y chromosome haplotype diversity of domestic sheep (Ovis aries) in northern Eurasia. *Anim. Genet.* 45, 903-907.

Zhang, Z., Sun, Y., Du, W., He, S., Liu, M., and Tian, C. (2017). Effects of vertebral number variations on carcass traits and genotyping of Vertnin candidate gene in Kazakh sheep. Asian-Australas. *J. Anim. Sci.* 30, 1234-1238.

Zhi, D., Da, L., Liu, M., Cheng, C., Zhang, Wang, Y., et al. (2018). Whole genome sequencing of hulunbuir short-tailed sheep for identifying candidate genes related to the short-tail phenotype. *Gene Genom. Genet.* 8, 377-383.

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.

Chapter 2

Genome-Wide Variation, Candidate Regions and Genes Associated with Fat Deposition and Tail Morphology in Ethiopian Indigenous Sheep[1]

[1]The results presented in this chapter have been published:

Ahbara A., et al. (2019) Genome-Wide Variation, Candidate Regions and Genes Associated With Fat Deposition and Tail Morphology in Ethiopian Indigenous Sheep. Front. Genet. 9:699.

Abstract

Variations in body weight and in the distribution of fat are associated with feed availability, thermoregulation, and energy reserve. Ethiopia is characterized by distinct agro-ecological and human ethnic farmer diversity of ancient origin, which have impacted the variation of its indigenous livestock. Here, we investigate autosomal genome-wide profiles of 11 Ethiopian indigenous sheep populations using the Illumina Ovine 50K SNP BeadChip assay. Sheep from the Caribbean, Europe, Middle East, China, and western, northern and southern Africa were included to address globally, the genetic variation and history of Ethiopian populations. Population relationship and structure analysis separated Ethiopian indigenous fat-tail sheep from their North African and Middle Eastern counterparts. It indicates two main genetic backgrounds and supports two distinct genetic histories for African fat-tail sheep. Within Ethiopian sheep, our results show that the short fat-tail sheep do not represent a monophyletic group. Four genetic backgrounds are present in Ethiopian indigenous sheep but at different proportions among the fat-rump and the long fat-tail sheep from western and southern Ethiopia. The Ethiopian fat-rump sheep share a genetic background with Sudanese thin-tail sheep. Genome-wide selection signature analysis identified eight putative candidate regions spanning genes influencing growth traits and fat deposition (*NPR2, HINT2, SPAG8, INSR*), development of limbs and skeleton, and tail formation (*ALX4*, *HOXB13*, *BMP4*), embryonic development of tendons, bones and cartilages (*EYA2*, *SULF2*), regulation of body temperature (*TRPM8*), body weight and height variation (*DIS3L2*), control of lipogenesis and intracellular transport of long-chain fatty acids (*FABP3*), the occurrence and morphology of horns (*RXFP2*), and response to heat stress (*DNAJC18*). Our findings suggest that Ethiopian fat-tail sheep represent a uniquely admixed but distinct genepool that presents an important resource for understanding the genetic control of skeletal growth, fat metabolism and associated physiological processes.

Introduction

African indigenous sheep originated in the Near East. They arrived, in the first instance, in North Africa *via* the Isthmus of Suez by the seventh millennium BP (Marshall, 2000). These sheep were of a thin-tail type and their dispersion southwards into East Africa followed possibly the Nile river valley and the Red Sea coastline (Blench and MacDonald, 2006; Gifford-Gonzalez and Hanotte, 2011). The second wave brought fat-tail sheep into North and Northeast Africa *via* two entry points, the Isthmus of Suez and the Horn of Africa across the straits of Bab-el-Mandeb, respectively. The fat-rump sheep are a recent introduction and represent the third wave of arrival and dispersal of the species into eastern Africa (Epstein, 1971; Ryder, 1983; Marshall, 2000).

Sheep fulfill important socio-cultural and economic roles in the Horn of Africa. In Ethiopia, sheep provide a wide range of products, including meat, milk, skin, hair, and manure, and are a form of savings and investment (Assefa et al., 2015). Ethiopia hosts many indigenous breeds of sheep, with currently 14 recognized populations/breeds, which are defined based on their geographic location and/or the ethnic communities managing them (Gizaw, 2008). Based on structure analysis, Edea et al. (2017) showed that the five Ethiopian indigenous sheep populations they analyzed clustered together based on their geographic distribution and tail phenotypes.

Fat depots act as an energy reserve that allows sheep to survive extreme environments and conditions such as prolonged droughts, cold, and food scarcity (Atti et al., 2004; Nejati-Javaremi et al., 2007; Moradi et al., 2012). Based on the combination of tail type and length, Ethiopian indigenous sheep can be classified as short fat-tail, long fat-tail, thin-tail, and fat-rump sheep. The short fat-tail inhabit sub-alpine mountainous regions, the long fat-tail predominate in mid- to high-altitude environments and the fat-rump sheep occur in semi-arid and arid environments (Gizaw et al., 2007). These populations are considered to be adapted to their production environments and they represent an important model species for investigating and enhancing our knowledge on the genome profiles of environmental adaptation, tail morphology, and fat localization.

Different approaches, that contrast groups of fat- and thin-tail sheep, have been used to identify candidate regions and genes associated with tail formation and morphotypes. Moradi et al. (2012) identified three regions on chromosomes 5, 7 and X associated with tail fat deposition

in Iranian breeds. Using two fat-tail (Laticauda and Cyprus fat-tail) and 13 Italian thin-tail breeds, Moioli et al. (2015) identified *BMP2* and *VRTN* as the likely candidate genes explaining fat-tail phenotype in the studied populations/breeds. Zhu et al. (2016) detected several copy number variations intersecting genes (*PPARA, RXRA,* and *KLF11*) associated with fat deposition in three Chinese native sheep (Large-tail Han, Altay, and Tibetan). Several candidate genes with possible links to fat-tail development, i.e., *HOXA11, BMP2, PPP1CC, SP3, SP9, WDR92, PROKR1,* and *ETAA1,* were identified using genome scans that contrasted fat- and thin-tail Chinese sheep (Yuan et al., 2017). Whole-genome sequencing of extremely short-tail Chinese sheep revealed the *T* gene as the best possible candidate, among other nine genes, influencing tail size, following its association with vertebral development (Zhi et al., 2018). There is, so far, no information on the genetic basis of variation in tail fat distribution and size in African indigenous sheep.

In this study, using the Ovine 50K SNP BeadChip genotypes, we investigated the (i) genetic relationships and structure within and between Ethiopian indigenous sheep of different fat-tail morphotypes alongside other sheep populations and breeds from the Caribbean, European, Middle East, China and Africa, and (ii) candidate genome regions and genes associated with tail morphology, fat deposition and possible eco-climatic adaptation in African indigenous sheep. For the later, 11 Ethiopian indigenous sheep of different fat-tail morphotypes and two populations of thin-tail sheep from Sudan were analyzed.

Materials and methods

DNA Samples and SNP Genotyping

The sampling strategy targeted breeds of indigenous sheep from different geographic regions in Ethiopia (**Figure 2.1** and **Table 2.1).** Geographic positioning system (GPS) coordinates were recorded for all the populations. We used altitude to determine the agro-eco-climatic zones of the geographic locations where the sheep were sampled. The sampling procedure targeted indigenous sheep in the grazing area where each is reared. A maximum of three samples (one male per flock) were collected from small-sized flocks (< 10 heads). All efforts were made to include populations representing the different tail phenotypes found in Ethiopia. Twenty DNA samples from two thin-tail sheep, Hammari and Kabashi, were obtained from Sudan. Genomic DNA was extracted from 146 ear tissue samples, collected from 11 Ethiopian indigenous sheep populations, using the NucleoSpin® Tissue Kit (www.mn-net.com) following the manufactures protocol. All 166 genomic DNA samples were genotyped using the Ovine 50K SNP BeadChip assay. The assay includes 54,240 SNPs composed of 52,413 autosomal, 1449 X-chromosome and 378 mitochondrial SNPs, respectively.

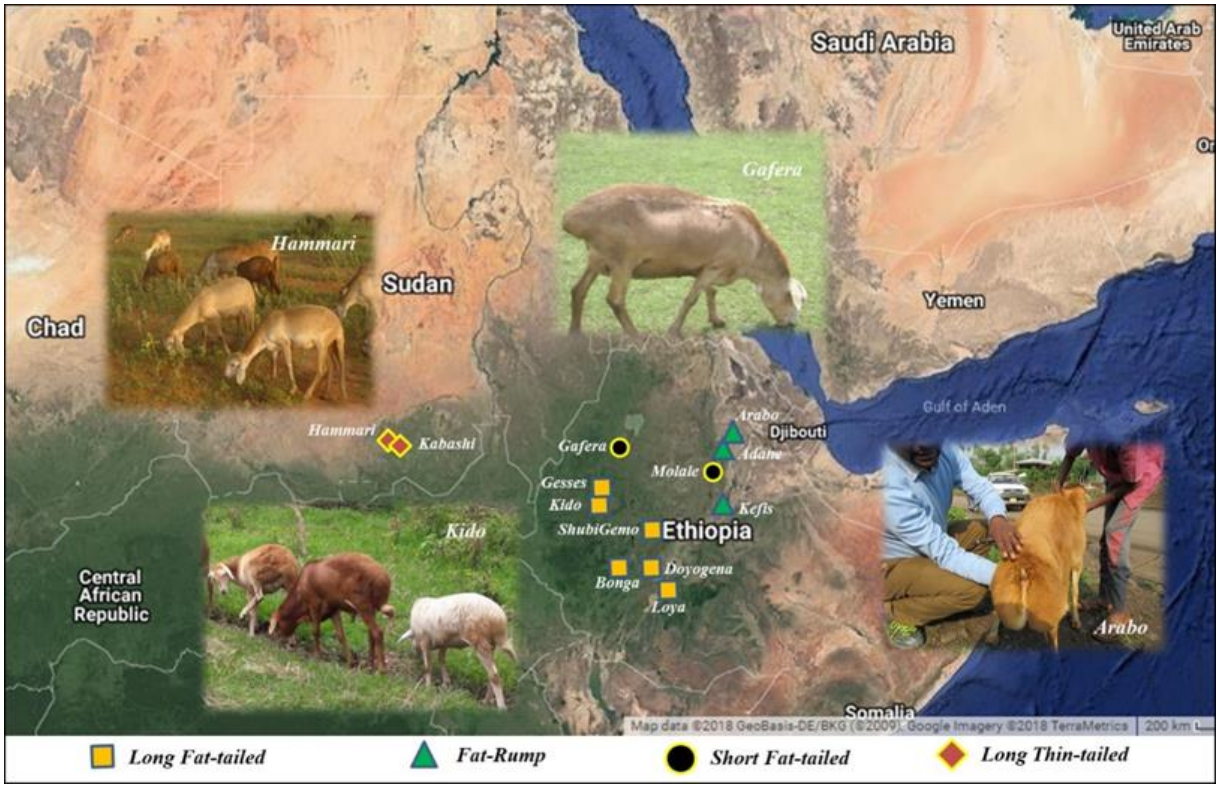


Figure 2.1 The locations where the Ethiopian and Sudanese sheep populations were sampled.

Table 2.1 Description of the populations that were sampled for this study.

Population	Zone	Latitude (*N*)	Longitude (*E*)	Altitude (m)	N		Tail Type	Agro-ecology
					Male	Female		
Kefis	Zone 3	9°30’	40°10’	890	2	12	Fat-Rump	Arid lowland
Adane	South Wollo	11°14’	39°50’	2450	7	5	Fat-Rump	Cool highland
Arabo	South Wollo	11°31’	36°54’	1500	3	7	Fat-Rump	Cool highland
Gafera	Agew Awi	11°31’	36°54’	2500	3	12	Short, fat tail	Wet, warm mid-highland
Molale	North Shewa	10°70’	39°39’	3068	6	9	Short, fat tail	Sub-alpine
Bonga	Keffa	7°16’	36°15’	1788	-	15	Long, fat tail	Humid mid-highland
Gesses	Metekel	10°50’	36°14’	1300	1	9	Long, fat tail	Moist lowlands
Kido	Metekel	10°71’	36°19’	1300	3	7	Long, fat tail	Moist lowland
Doyogena	Kembata	7°21’	37°47’	2324	1	14	Long, fat tail	Cool, wet highland
ShubiGemo	East Shewa	8°80’	38°51	1600	5	10	Long, fat tail	Cool, wet highland
Loya	Sidama	6°29’	38°24’	1900	3	12	Long, fat tail	Cool, wet highland
Hammari	North Kurdufan	13°09’	29°22’	620	3	7	Long, thin tail	Arid lowland
Kabashi	North Kurdufan	13°09’	29°22’	620	5	5	Long, thin tail	Arid lowland
					42	**124**		
		Total			**166**			

Ovine 50K SNP BeadChip genotypes of Caribbean, European, Middle East and Chinese, as well as western, northern and southern African sheep, respectively were obtained from the Sheep HapMap database (http://www.sheephapmap.org/ hapmap.php, (**Appendix Table 5.1**) and included in the study. The aim was to provide a global context of the genetic origins, trajectories of introduction, and dispersal of sheep into Ethiopia.

Quality Control and Genetic Diversity Analyses

The Sheep HapMap dataset were merged with the ones generated from Ethiopian and Sudanese sheep using PLINK v1.9 (Purcell et al., 2007). The data files for the final analysis were generated after pruning the merged dataset of SNPs not mapping on any autosomes, with a minor allele frequency (MAF) ≤ 0.01 and animals and markers with ≥ 10 and 5% missing genotypes, respectively. This generated a dataset with 45,102 SNPs which were further pruned, using PLINK v1.9, to be in approximate linkage equilibrium to avoid the possible influence of clusters of SNPs on population genetic relationship and structure analysis (Yuan et al., 2017). Following the later pruning, 34,088 SNPs were retained for population relatedness and structure analysis. To minimize the possible loss of informative SNPs for selection signature analysis, the data for Ethiopian and Sudanese sheep was extracted from the dataset of 45,102 autosomal SNPs that were obtained prior to linkage disequilibrium (LD) pruning.

The proportion of polymorphic SNPs (*P*n), expected (*H*e), and observed (*Ho*) heterozygosity and inbreeding coefficient (*F*) were estimated for each population and across all populations using PLINK v1.9, to evaluate the levels of genetic diversity present in Ethiopian and Sudanese sheep, respectively.

Population Genetic Analyses

Principal component analysis (PCA) was performed using PLINK v1.9 to investigate the genetic structure and relationships among the studied breeds based on genetic correlations between individuals. A graphical display of the first two principal components (PC1 and PC2) was generated using GENESIS (Buchmann and Hazelhurst, 2014). Admixture analysis implemented in ADMIXTURE v1.3 (Alexander et al., 2009) was used to investigate the underlying genetic structure and estimate the proportion of shared genome ancestry between the study populations. A 5-fold cross-validation procedure following Lawal et al. (2018), was

used to determine the optimal number of ancestral genomes (*K*) and proportions of genome ancestry that was shared among the study populations.

To further evaluate historical relationships and interactions (gene flow) within and between Ethiopian and Sudanese populations, we used the maximum likelihood tree-based approach implemented in TreeMix (Pickrell and Pritchard, 2012) and included the Soay sheep as an out-group. The number of migration events (m) varied between 1 (migration between two populations) and 15 (migrations between all the populations). The value of "m" with the highest reproducibility and consistency, among the 15 tested, and which also had the highest log-likelihood value following six replication runs of the analysis, was chosen as the most optimal.

The *f3* and *f4* tests implemented in TreeMix were also performed. The *f3*-statistics (A, B, C) were to determine if A was derived from the admixture of populations B and C; a significantly negative value of the *f3*-statistics would suggest population A is admixed. The *f4*-statistics (A, B,) (C, D) were to test the validity of hierarchical clustering patterns in four-population trees. Significant de*via*tions of the *f4*-statistics from zero for the three possible topologies of the four-population trees would provide evidence of gene flow between the populations tested. A significantly positive *Z*-score indicates gene flow between populations that are related to either A and C or B and D while a significantly negative *Z*-score indicates gene flow between populations that are related to A and D or B and C. Standard errors were estimated using blocks of 500 SNPs.

Analysis of Signatures of Selection

For this analysis, we separated 12 of the 13 Ethiopian and Sudanese populations into four genetic groups based on the population clusters revealed by PCA. The four population groups included, western (Bonga, Kido, Gesses) and southern (Loya, ShubiGemo, Doyogena) long fat-tail, and fat-rump (Kefis, Adane, Arabo) sheep from Ethiopia and thin-tail sheep (Hammari, Kabashi) from Sudan. One short fat-tail sheep (Molale) was included with the fat-rump sheep and the other (Gafera), which appeared to be genetically distinct based on the PCA, was dropped from further analysis. Equal numbers of samples were chosen at random to represent each genetic group. Three comparisons that contrasted the fat-rump (E1), western- (E2) and southern- (E3) long fat-tail sheep with the thin-tail sheep (S) from Sudan were performed. The selection signature analysis involved three approaches, $F_{ST,}$ *hapFLK* and *Rsb*.

A sliding window approach was used to perform the F_{ST} analysis using the HIERFSTAT package (Goudet, 2005) of R (R Core Team, 2012). The window size of 200 Kb was allowed to slide along the genome by a distance of 60 Kb. The window size and slide distance were determined based on LD decay analysis (**Figure 2.2**).

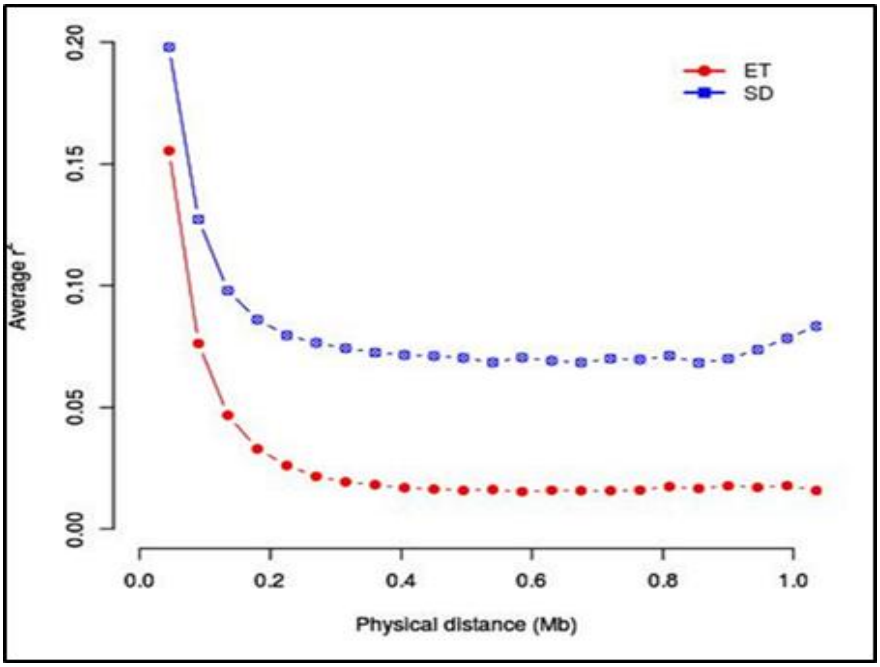


Figure 2.2 Patterns of linkage disequilibrium (LD) decay calculated within the Ethiopian (ET) and Sudanese (SD) sheep populations.

The pairwise F_{ST} (Weir and Cockerham, 1984) values for each SNP in each window and between the genetic groups being tested were estimated as follows:

$$F_{ST} = 1 - \frac{p1q1+p2q2}{2prqr} \qquad (2.1)$$

Where *p1, p2* and *q1, q2* are the frequencies of alleles "A" and "a" in the first and second group of the test populations, respectively, and *pr* and *qr* are the frequencies of alleles "A" and "a", respectively, across the tested groups (Zhi et al., 2018). The F_{ST} values were standardized into *Z*-scores as follows:

$$ZF_{ST} = \frac{F_{ST}-\mu F_{ST}}{\sigma F_{ST}} \qquad (2.2)$$

Where μF_{ST} is the overall average value of F_{ST} and σF_{ST} is the standard de*via*tion derived from all the windows tested for a given comparison. **Figure 2.3a** shows the distribution of ZF_{ST} values. We set the value of $ZF_{ST} \geq 4$ as the threshold to identify candidate genomic regions under selection.

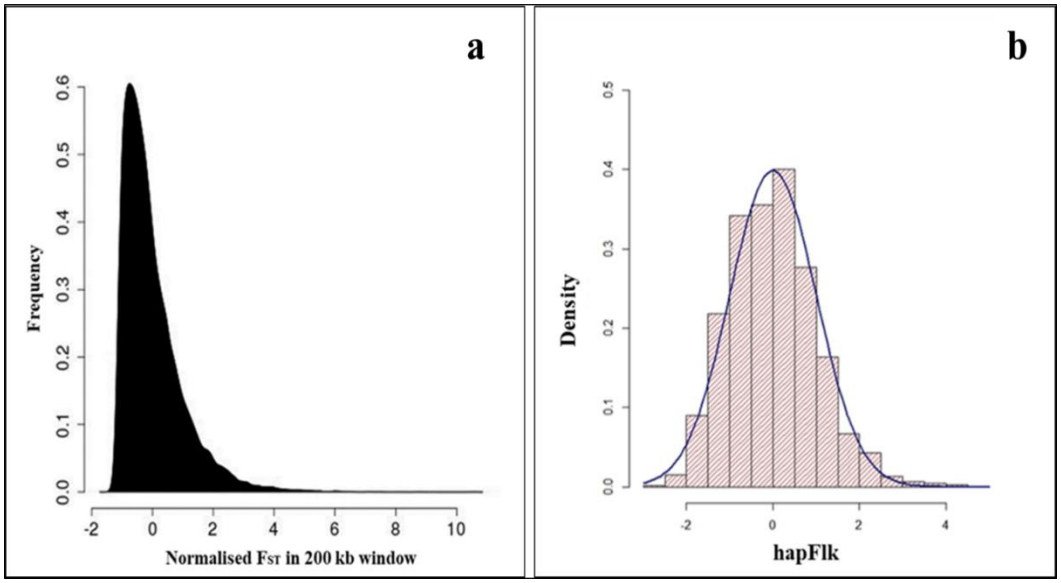


Figure 2.3 Distribution of the standardized Z-score values for (a) F_{ST} and (b) *hapFLK* for the autosomal markers.

The *hapFLK* approach was implemented with *hapFLK* package v1.2 (Fariello et al., 2013) to detect selection signatures based on differences in haplotype frequencies between groups of populations. Reynolds genetic distances were converted into kinship matrix using an R script supplied with the package. The *hapFLK* values and kinship matrix were calculated assuming 15 clusters in the fastPHASE model (-K 15). The *hapFLK* statistic was then computed as the average value across 40 expectation maximization runs to fit the LD model. The *P*-values were obtained by running a python script "Scaling_chi^2_hapFLK.py" available at (https://forge-dga.jouy. inra.fr/documents/588) which fits a chi-squared distribution to the empirical distribution. As with the F_{ST} calculations, the *hapFLK* statistics were also standardized using the formula:

$$\boldsymbol{hapFLKadj} = \frac{\boldsymbol{hapFLK\_mean(hapFLK)}}{\boldsymbol{Sd(hapFLK)}} \qquad (2.3)$$

The calculation of the raw *P-values* was based on the null distribution of empirical values (Fariello et al., 2013; Kijas, 2014). The *P-values* were plotted in a histogram to assess their distribution pattern and the cut-off value to determine significance was set at –Log10 (*P-value*) ≥ 3 (**Figure 2.3b**).

Using haplotype information, we performed the *Rsb* analysis implemented in *rehh* package (Gautier and Vitalis, 2012) of R. Haplotypes were estimated with SHAPEIT (Delaneau et al., 2014). To identify loci under selection, the *Rsb* values were log-transformed into P_{Rsb} (P_{Rsb} = −Log10 [1 − 2(Φ (*Rsb*) - 0, 5)]), where Φ(x) represents the Gaussian cumulative distribution

function (Gautier and Vitalis, 2012). Assuming that the *Rsb* values are normally distributed (under neutrality), the P_{Rsb} can be interpreted as -Log10 (*P*-value), where *P* is the two-sided *P*-value associated with the neutral hypothesis. For each comparison, SNPs that exhibited $P_{Rsb} \geq 3$ (*P*-value = 0.001) were taken to be under selection (de Simoni Gouveia et al., 2017). The *hapFLK* and *Rsb* analysis were also performed using window sizes of 200 Kb sliding along the genome by a distance of 60 Kb.

Gene annotation

Candidate regions that overlapped between F_{ST}, *hapFLK* and *Rsb* were identified and compared using the intersectBed function of Bed Tools software (Quinlan and Hall, 2010). Considering an average marker distance of between 60 and 200 Kb (Moioli et al., 2015) and the observed LD decay pattern (**Figure 2.2**), candidate regions under selection were identified by exploring the SNPs found up- and downstream, and within, the most significant windows. The Oar v3.1 Ovine reference genome assembly (Jiang et al., 2014) was used to annotate the candidate regions. Functional enrichment analysis was performed using the functional annotation tool in *DAVID* (Huang et al., 2008) using *Ovis aries* as the background species. Gene functions were determined using the NCBI (http://www. ncbi.nlm.nih.gov/gene/) and OMIM databases (http://www.ncbi. nlm.nih.gov/omim/) and a review of literature.

Results

Genetic Diversity and Population Structure

The average values of *Pn*, *He*, *Ho*, and *F*, as indicators of within-breed genetic diversity, are shown in **Appendix Table 5.2** and **Figure 2.4**. The lowest values of *Pn*, *He*, and *Ho* were observed in Bonga while the highest values were observed in Molale-Menz, Hammari and Kabashi, and Arabo, respectively.

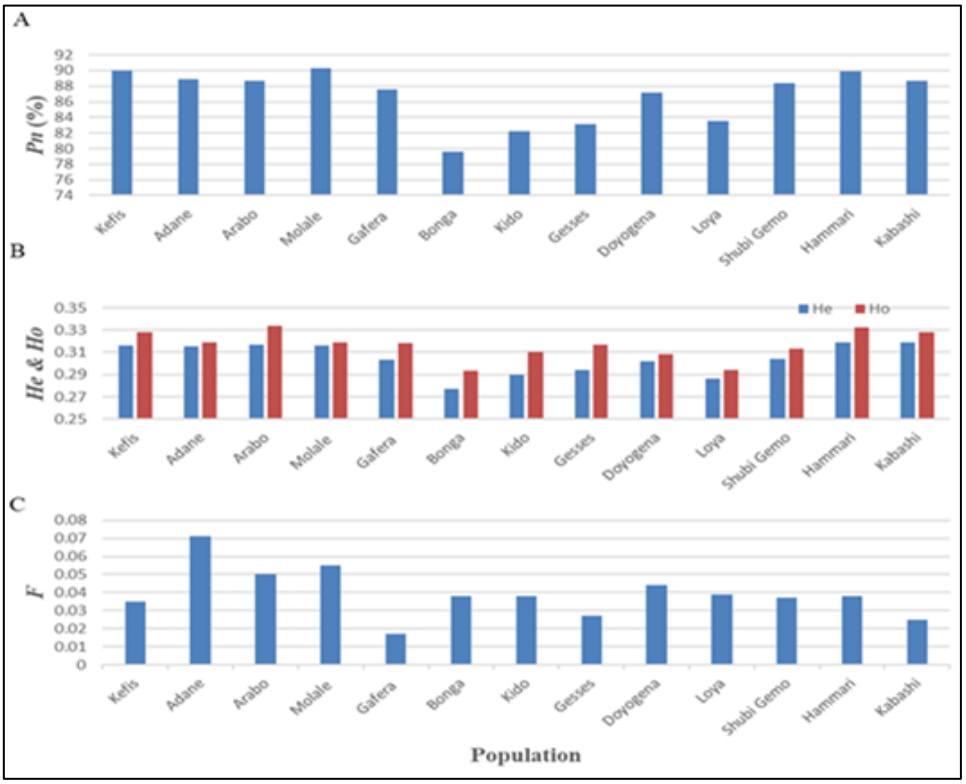


Figure 2.4 Distribution of the genetic diversity indices within each breed. (A) SNP displaying polymorphism (*Pn*), (B) expected heterozygosity (*He*); observed heterozygosity (*Ho*) and (C) the inbreeding coefficient (*F*)

The PCA plot incorporating the global populations and which was constructed using a sample size of five animals that were selected at random per population is shown in **Figure 2.5**. We used the uniform sample size of five animals since differences in sample sizes may influence clustering patterns on the PCA. The choice to use five samples per population was based on the smallest sample size of five individuals genotyped for Sidaoun and Berber breeds. In spite of the sample size rebalancing, the population cluster patterns did not differ from that observed when the PCA was performed using unequal sample sizes (**Appendix Figures 5.1, 5.2**). Generally, PC1 separates Ethiopian and South African fat-tail sheep, Sudanese thin-tail sheep, West African Djallonke and Algerian Sidaoun from the other sheep populations. Sheep from the Middle East and North Africa occur at the center of the PCA plot and, together with the

Cyprus fat-tail and Chinese sheep (which cluster close together) are separated by PC2 from African Dorper, Barbados Blackbelly and European sheep. The two populations of Ethiopian short fat-tail sheep diverge from each other; Gafera-Washera clusters near Ethiopian long fat-tail sheep while Molale-Menz clusters together with the Ethiopian fat-rump sheep. The West African Djallonke clusters close to the two South African breeds (Ronderib and Namaqua). Sidaoun and Berber (both from Algeria) cluster separate, while the Cyprus fat-tail clusters close to the Chinese sheep (**Figure 2.5**).

To obtain a clearer picture of the variation within the fat-tail sheep, we performed the PCA while excluding the thin-tail sheep (**Figure 2.6**). PC1 separates the Ethiopian fat-tails from their Middle East, North Africa, Mediterranean and Chinese counterparts. PC2 differentiates the South African breeds from the Ethiopian ones. Like in the global PCA, one Ethiopian short fat-tail sheep (Gafera-Washera) clusters with the Ethiopian long-fat tail sheep and the other (Molale-Menz) forms a cluster with the Ethiopian fat-rump sheep. Middle East sheep cluster together with the North African ones while the Mediterranean sheep unexpectedly cluster with the Chinese sheep despite the large geographic distance separating them.

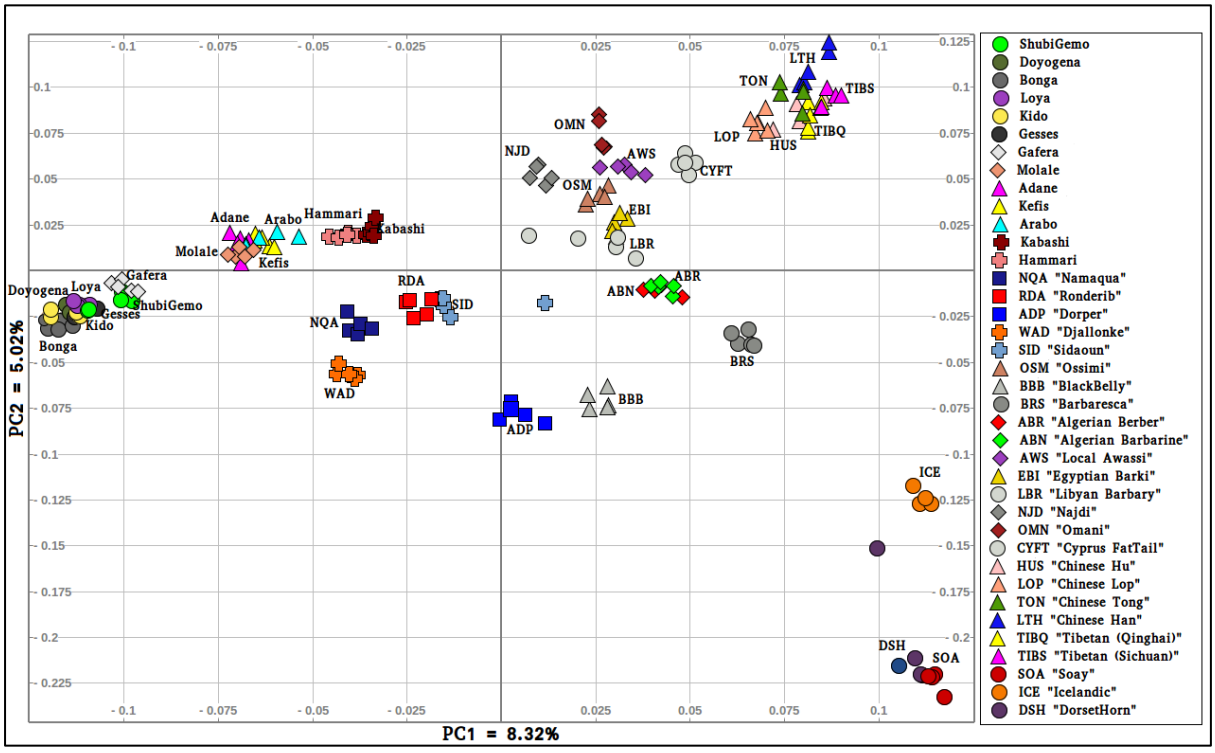


Figure 2.5 Genetic variation among the Ethiopian sheep populations in a global geographic context

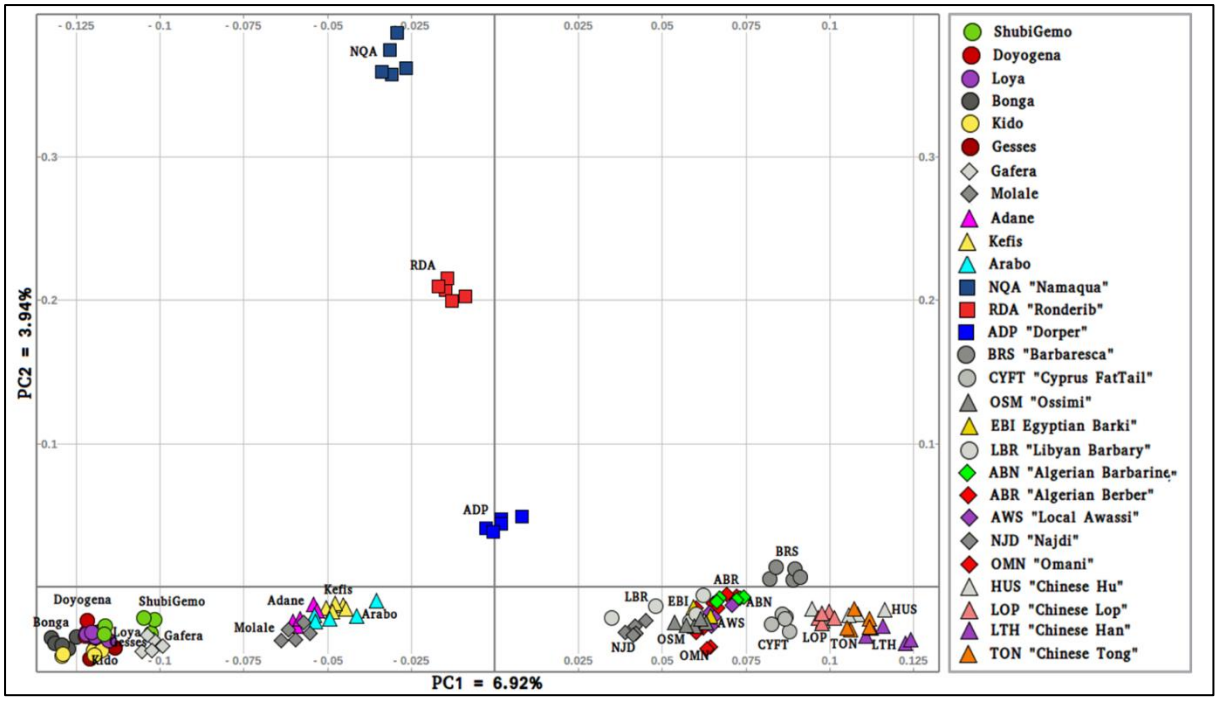


Figure 2.6 Distribution of genetic variation among the worldwide fat-tail sheep

To further illustrate the distribution of genetic variation among East African populations, we performed the PCA with only the Ethiopian and Sudanese thin-tail sheep (**Figure 2.7**). PC1 separates Ethiopian fat-rump, Molale-Menz (Ethiopian short-fat tail) and thin-tail sheep from the Ethiopian long fat-tail and Gafera-Washera (Ethiopian short-fat tail) sheep. Generally, PC1 separates the fat-rump sheep from the fat-tail ones derived from western and southern Ethiopia. PC2 reveals further separation of the Ethiopian sheep: (i) Molale-Menz, Adane and some Arabo individuals from Kefis population and the remaining Arabo individuals, and (ii) Gafera-Washera, Kido and Gesses from Doyogena, ShubiGemo, Bonga and Loya.

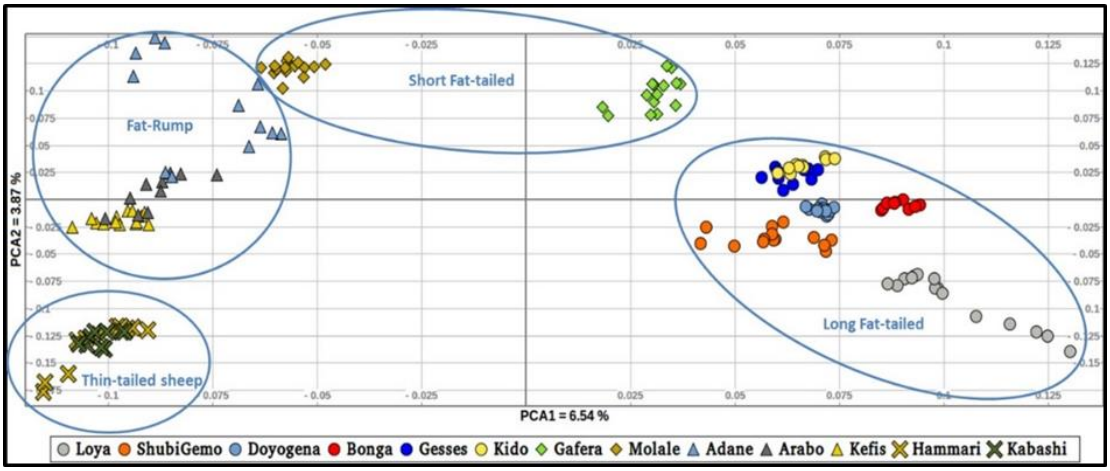


Figure 2.7 Distribution of the genetic variation among 144 individuals of 11 Ethiopian sheep populations (PC1 and PC2)

The admixture analysis of the global dataset separates the study populations following their geographic origins (**Figure 2.8**). The cross-validation (CV) error registered the lowest value at $K = 9$ suggesting this to be the most optimal number of clusters explaining the variation in this dataset (**Figure 2.9a**). Chinese sheep separate from the other populations at $K \geq 3$. Among African breeds, the South African ones (Namaqua, Dorper, Ronderib) and the West African Djallonke show a distinct but common genetic ancestry with the Ethiopian and Sudanese sheep for $3 \leq K \leq 6$.

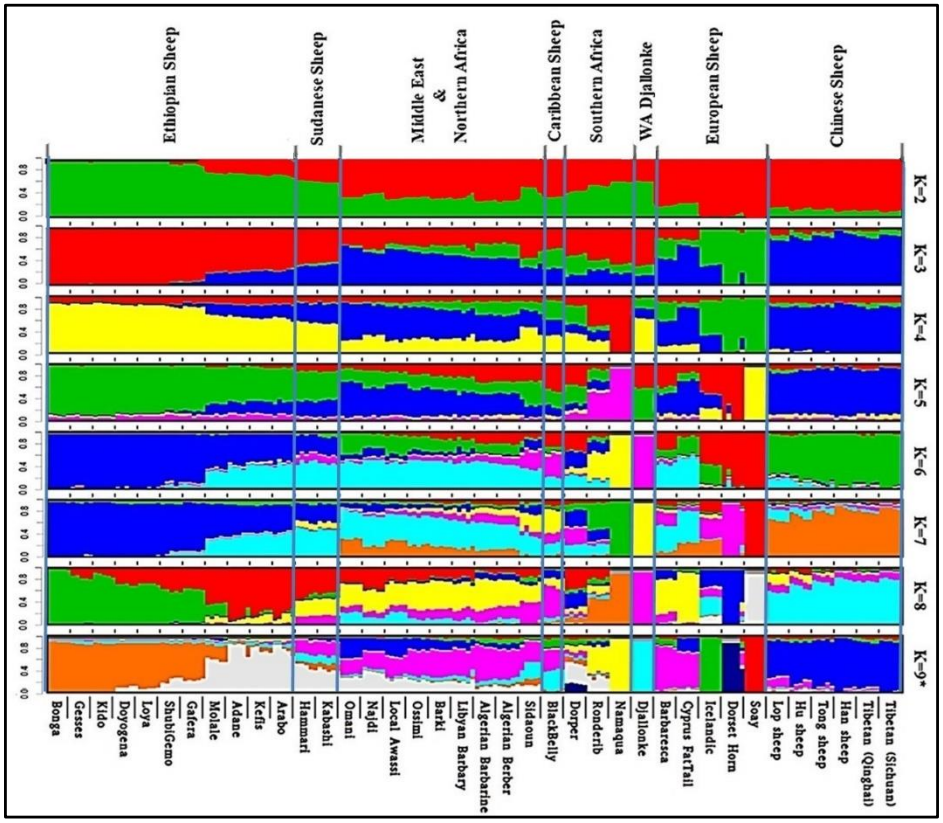


Figure 2.8 Admixture analysis of the studied populations in a global context ($K = 9$ had the lowest cross-validation error).

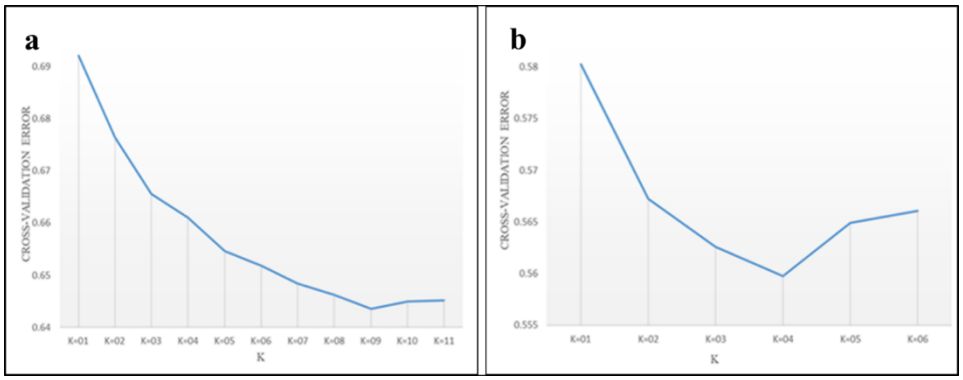


Figure 2.9 Cross-validation plot for admixture analysis of the studied populations (a) in a global and (b) national context

Two to six hypothetical ancestral clusters (*K*) were tested with Admixture on the East African dataset. The lowest CV error suggests *K* = 4 (**Figure 9.b**) as the optimal number of ancestral clusters present in Ethiopian and Sudanese thin-tail sheep. The proportion of each ancestral cluster (referred to as A, B, C, and D) in each population at *K* = 4 is shown in **Figure 2.10** and **Table 2**. They occur with the highest proportion (>90%) in Loya (cluster A), Bonga, Kido and Gesses (cluster B), Molale-Menz and a few individuals of Adane (cluster C) and in thin-tail sheep (cluster D). Clusters A, B, and C are observed in ShubiGemo and Doyogena; B and C in Gafera-Washera and Molale-Menz; B, C, and D in some individuals of Adane while Arabo and Kefis had C and D clusters. The analysis also shows that Gafera-Washera, Adane, Molale-Menz, Arabo, and Kefis share cluster C, while Hammari and Kabashi share the D cluster with Arabo and Kefis. ShubiGemo, Loya and Doyogena, all long fat-tail sheep from southern Ethiopia, share cluster A.

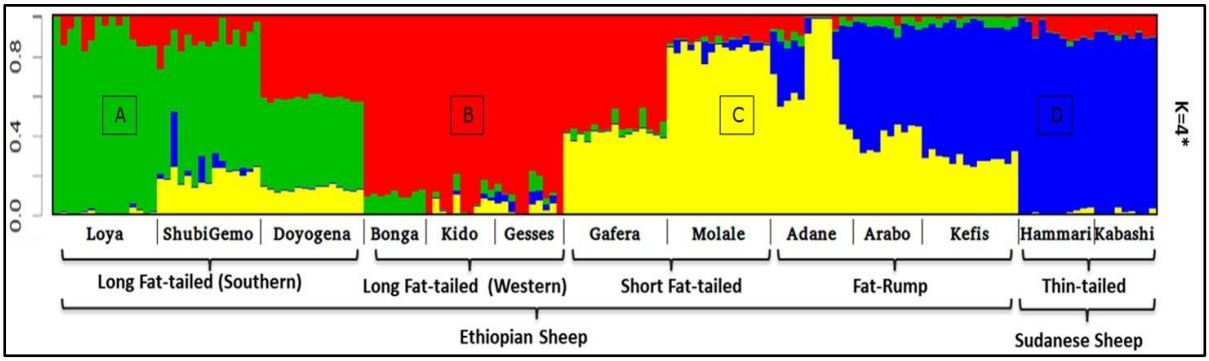


Figure 2.10 Admixture analysis involving Ethiopian indigenous sheep populations (*K* = 4 had the lowest cross-validation error). For brevity, the four genetic clusters are designated (A)–(D), respectively.

TreeMix revealed possibilities of gene-flow between East African sheep. The ***f*** index representing the fraction of the variance in the sample covariance matrix ($\widehat{\mathbf{W}}$) 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 15 possible migration vertices were computed. The first eight migration edges (gene flow) 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 0.51. We, therefore, chose *m* = 8 as the best predictive value for the migration model. Vectors from 9 to 15 resulted in only small incremental changes in the *f* value **(Figures 2.11a, b).**

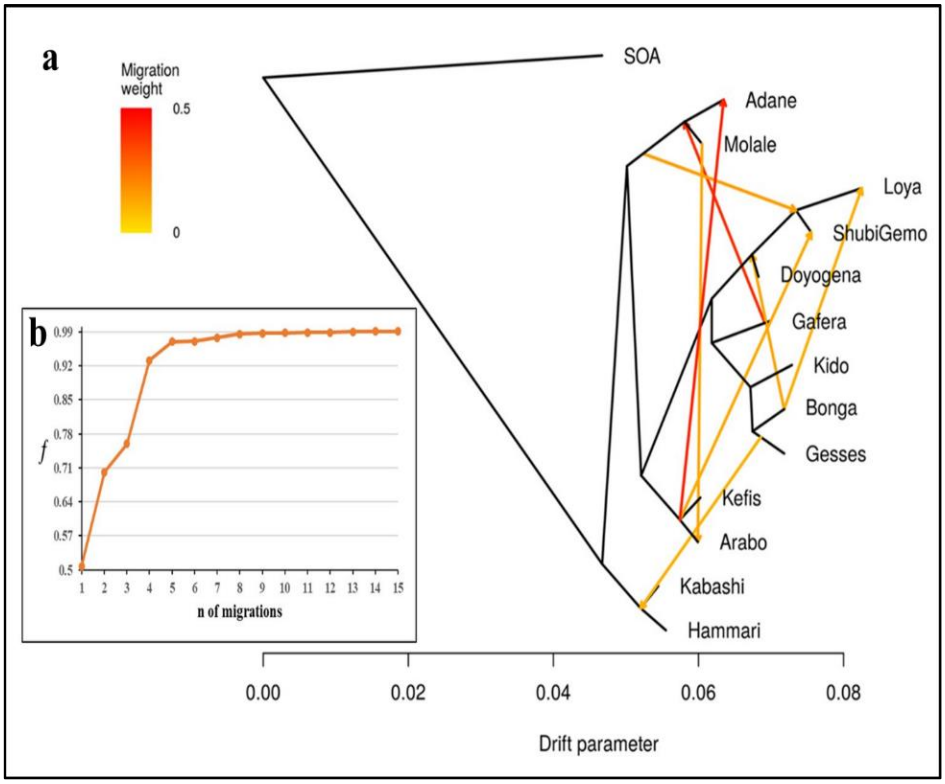


Figure 2.11 Tree-mix plot. (a) Phylogenetic network inferred by Tree-mix of the relationships between Ethiopian and Sudanese sheep populations. The first eight migration edges between populations are shown with arrows pointing in the direction toward the recipient group and colored according to the ancestry percentage received from the donor. (b) 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 modeled migration events.

The eight migration events were Loya and ShubiGemo (both long fat-tail); Arabo and Adane (both fat-rump); Gafera-Washera, Molale-Menz (both short fat-tail) and Adane (fat-rump); Molale-Menz (short fat-tail) and Adane (fat-rump) with ShubiGemo (long fat-tail); Bonga with ShubiGemo, Doyogena and Loya (all long fat-tail sheep); Molale-Menz (short fat-tail) and Arabo (fat-rump); ShubiGemo (long fat-tail) with Arabo (fat-rump) and Kefis (fat-rump); Gesses (long fat-tail) with Kabashi and Hammari (thin-tail).

Signatures of Selection

The PCA, Admixture and TreeMix **(Figures 2.7, 2.10, 2.11)** revealed three genetic groups in Ethiopian sheep *viz* fat-rump (E1), and long fat-tail from western (E2) and southern (E3) Ethiopia, respectively. The two short fat-tail sheep (Molale-Menz and Gafera-Washera) analyzed here were separated from each other **(Figure 2.7)** with Molale-Menz showing close genetic affinity to fat-rump sheep and Gafera-Washera appeared genetically distinct. The three groups are distinct from thin-tail (S) sheep **(Figure 2.7).** For selection signature analysis, we included Molale-Menz with the fat-rump sheep but excluded Gafera-Washera from the analysis

due to its low sample size. We selected, at random, 20 samples to represent each of the four genetic groups and performed the selection signature analysis. We contrasted the three groups of Ethiopian sheep (E1, E2, and E3) with the thin-tail sheep (S). The top windows which passed the significance threshold, for each method (*hapFLK* > 3, ZF_{ST} > 4, *Rsb* > 3) were used to define candidate regions under selection.

For E1*S comparison, the fat-rump sheep were differentiated from the thin-tail in 23 candidate regions that overlapped between at least two methods and which spanned 86 genes **(Figure 2.12, Table 2.3).**

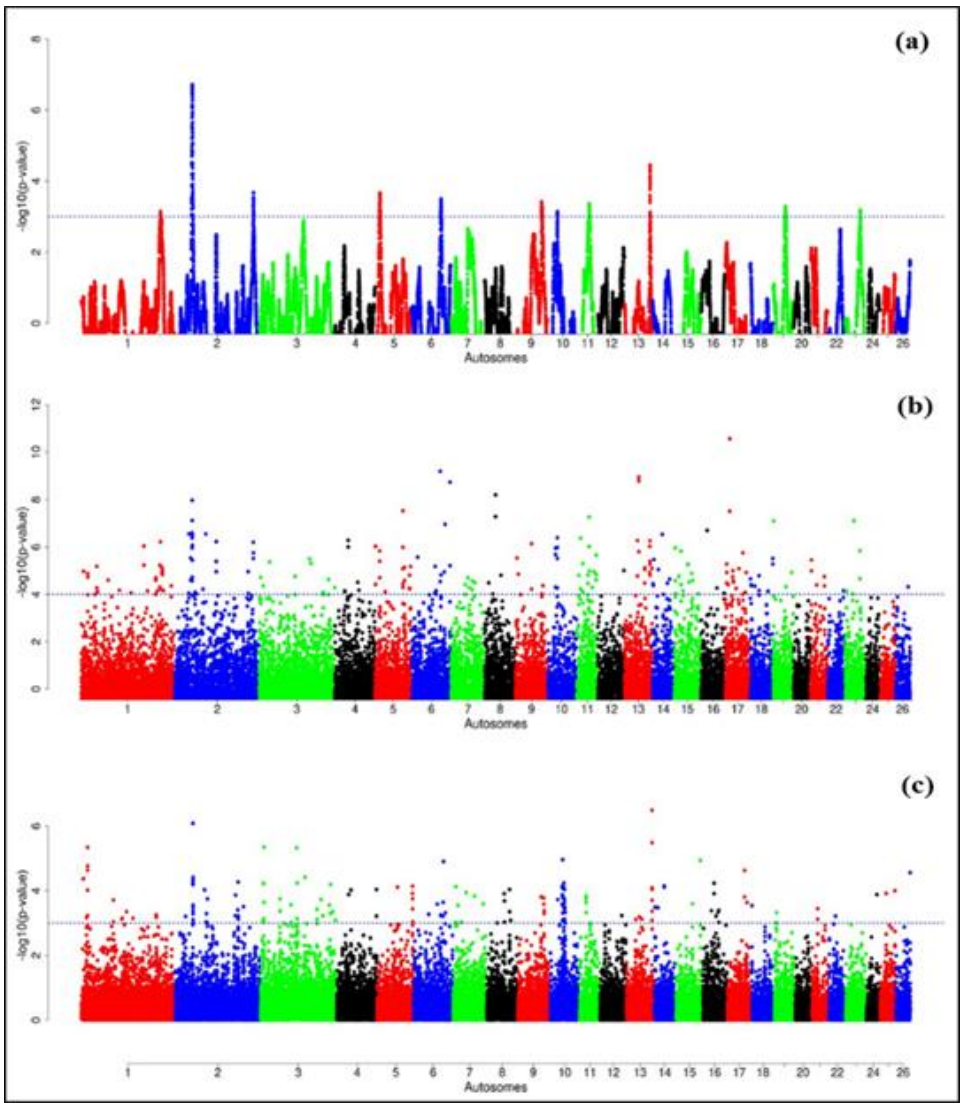


Figure 2.12 Manhattan plots of genome-wide autosomal *hapFLK* (a), ZF_{ST} (b) and *Rsb* (c) analyses of Ethiopian fat-rump (E1) *vs.* thin-tail (S) sheep.

Table 2.2 Candidate regions and genes identified to be under selection by a combination of at least two methods in the Ethiopian fat-rump *vs.* Sudanese thin-tail sheep.

Chr	Overlapping region	Gene location	Method	Candidate gene	Annotation
1	6600001-6800000	6767995-6787884	*Rsb*ZF$_{ST}$*	*SPP2*	secreted phosphoprotein 2
	19200001-19460000	19196066-19216520		*KIF2C*	kinesin family member 2C
		19227521-19230058		*RPS8*	ribosomal protein S8
		19233256-19236856		*BEST4*	bestrophin 4
		19251263-19255855		*PLK3*	polo like kinase 3
		19256190-19256846		*TCTEX1D4*	Tctex1 domain containing 4
		19258856-19263210		*BTBD19*	BTB domain containing 19
		19270475-19284541		*PTCH2*	patched 2
		19291159-19423029		*EIF2B3*	eukaryotic translation initiation factor 2B subunit gamma
		19425022-19433738		*HECTD3*	HECT domain E3 ubiquitin protein ligase 3
		19436350-19439595		*UROD*	uroporphyrinogen decarboxylase
		19442382-19608681		*ZSWIM5*	zinc finger SWIM-type containing 5
2	51660001-52220000	51686233-51755300	*hapFLK*ZF$_{ST}$*	*MELK*	maternal embryonic leucine zipper kinase
		51891433-51948694		*RNF38*	ring finger protein 38
		51989342-52042116		*GNE*	glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase
		52048202-52065307		*CLTA*	clathrin light chain A
		52087650-52089416		*CCIN*	Calicin
		52128947-52210749		*RECK*	reversion inducing cysteine rich protein with kazal motifs
	52020001-53180000	52411021-52417200	*hapFLK*Rsb*Z*	*FAM221B*	family with sequence similarity 221 member B
		52421111-52423389	*F$_{ST}$*	*HINT2*	histidine triad nucleotide binding protein 2
		52423298-52426475		*SPAG8*	sperm associated antigen 8
		52423842-52445175		*NPR2*	natriuretic peptide receptor 2
		52480200-52481163		*MSMP*	microseminoprotein, prostate associated
		52480334-52485038		*RGP1*	RGP1 homolog, RAB6A GEF complex partner 1
		52485320-52495944		*GBA2*	glucosylceramidase beta 2
		52496387-52500153		*CREB3*	cAMP responsive element binding protein 3
		52506528-52531560		*TLN1*	talin 1
		52537459-52544952		*TPM2*	tropomyosin 2
		52546134-52551851		*CA9*	carbonic anhydrase 9

		52560703-52563910		*ARHGEF39*	Rho guanine nucleotide exchange factor 39
		52564548-52567161		*CCDC107*	coiled-coil domain containing 107
		52572730-52573775		*SIT1*	signaling threshold regulating transmembrane adaptor 1
		52594675-52607206		*CD72*	CD72 molecule
		52603605-52607846		*TESK1*	testis-specific kinase 1
		52616756-52618641		*FAM166B*	family with sequence similarity 166 member B
		52619243-52632387		*RUSC2*	RUN and SH3 domain containing 2
		52817902-53036532	*hapFLK*Rsb*	*UNC13B*	unc-13 homolog B
		53056098-53059144		*FAM214B*	family with sequence similarity 214 member B
		53061224-53067598		*STOML2*	stomatin like 2
		53070391-53079125		*PIGO*	phosphatidylinositol glycan anchor biosynthesis class O
		53079030-53084363		*FANCG*	Fanconi anemia complementation group G
		53089776-53099744		*VCP*	valosin containing protein
		53138867-53146827		*DNAJB5*	DnaJ heat shock protein family (Hsp40) member B5
		53159612-53165463		*PHF24*	PHD finger protein 24
	232620001-232940000	232749221-233048136	*hapFLK*ZF_{ST}*	*DIS3L2*	DIS3 like 3'-5' exoribonuclease 2
3	107100001-107240000	107108271-107174474	*Rsb*ZF_{ST}*	*TSPAN8*	tetraspanin 8
	205800001-206000000	205801838-205853818		*A2ML1*	alpha-2-macroglobulin like 1
		205889722-205909753		*RIMKLB*	ribosomal modification protein rimK like family member B
		205954117-205968927		*MFAP5*	microfibril associated protein 5
		205985865-205999107		*AICDA*	activation induced cytidine deaminase
5	13620001-13940000	13733596-13879145	*hapFLK*ZF_{ST}*	*INSR*	insulin receptor
6	70200001-70520000	70189729-70234612	*Rsb*ZF_{ST}*	*KIT*	KIT proto-oncogene receptor tyrosine kinase
	87180001-87560000	87097877-87386270	*hapFLK*Rsb*	*ADAMTS3*	ADAM metallopeptidase with thrombospondin type 1 motif 3
7	63420001-63620000	63450344-63456226	*Rsb*ZF_{ST}*	*BMP4*	bone morphogenetic protein 4
9	76740001-77300000	76741376-76818820	*hapFLK*Rsb*	*SPAG1*	sperm associated antigen 1
		76826125-76827336		*POLR2K*	RNA polymerase II subunit K
		76838577-76849876		*FBXO43*	F-box protein 43
		76870092-77006581		*RGS22*	regulator of G protein signaling 22
		77057424-77839842		*VPS13B*	vacuolar protein sorting 13 homolog B
	78000001-78380000	78104905-78377671	*hapFLK*ZF_{ST}* RRsb*	*STK3*	serine/threonine kinase 3

10	24240001-24500000	24289442-24435384	*Rsb*ZF$_{ST}$*	*TRPC4*	transient receptor potential cation channel subfamily C member 4
		24474862-24508794		*POSTN*	periostin
	29400001-29780000	29454677-29502617	*hapFLK*ZF$_{ST}$*	*RXFP2*	relaxin family peptide receptor 2
11	37140001-37400000	37140993-37148353		*SNF8*	SNF8, ESCRT-II complex subunit
		37146942-37164597		*UBE2Z*	ubiquitin conjugating enzyme E2 Z
		37173130-37175267		*ATP5MC1*	ATP synthase membrane subunit c locus 1
		37227823-37243185		*CALCOCO2*	calcium binding and coiled-coil domain 2
		37272426-37302473		*TTLL6*	tubulin tyrosine ligase like 6
		37337231-37338988		*HOXB13*	homeobox B13
	37920001-38120000	37924394-37928175		*SNX11*	sorting nexin 11
		37972076-37981743		*NFE2L1*	nuclear factor, erythroid 2 like 1
		37992980-38001708		*COPZ2*	coatomer protein complex subunit zeta 2
		38037788-38047808		*CDK5RAP3*	CDK5 regulatory subunit associated protein 3
		38063059-38063361		*PRR15L*	proline rich 15 like
		38069220-38075491		*PNPO*	pyridoxamine 5'-phosphate oxidase
		38082581-38118204		*SP2*	Sp2 transcription factor
13	38580001-38660000	38609366-38671551	*Rsb*ZF$_{ST}$*	*RIN2*	Ras and Rab interactor 2
	75120001-75680000	75066765-75328455	*hapFLK*ZF$_{ST}$**	*EYA2*	EYA transcriptional coactivator and phosphatase 2
		75666854-75730764	*Rsb*	*NCOA3*	nuclear receptor coactivator 3
		75726734-75771128	*hapFLK*Rsb*	*SULF2*	sulfatase 2
14	2220001-2360000	2251815-2262220	*Rsb*ZF$_{ST}$*	*GABARAPL2*	GABA type A receptor associated protein like 2
		2276128-2300712		*ADAT1*	adenosine deaminase, tRNA specific 1
		2302972-2319972		*KARS*	lysyl-tRNA synthetase
	28860001-29000000	28747069-29125550		*CDH8*	cadherin 8
15	72540001-72620000	72556058-72606253	*Rsb*ZF$_{ST}$*	*ALX4*	ALX homeobox 4
17	51780001-51800000	51771124-51788976	*Rsb*ZF$_{ST}$*	*RILPL2*	Rab interacting lysosomal protein like 2

Similarly, a total of 65 genes were present across 18 candidate regions that overlapped between at least two approaches in the E2*S (western Ethiopia long fat-tail verses thin-tail) comparison **(Figure 2.13, Table 2.4).** Furthermore, 10 genes that seemed to be highly selected were identified by *Rsb* in three candidate regions on Oar8, Oar14, and Oar18, respectively **(Figure 2.13, Table 2.4).**

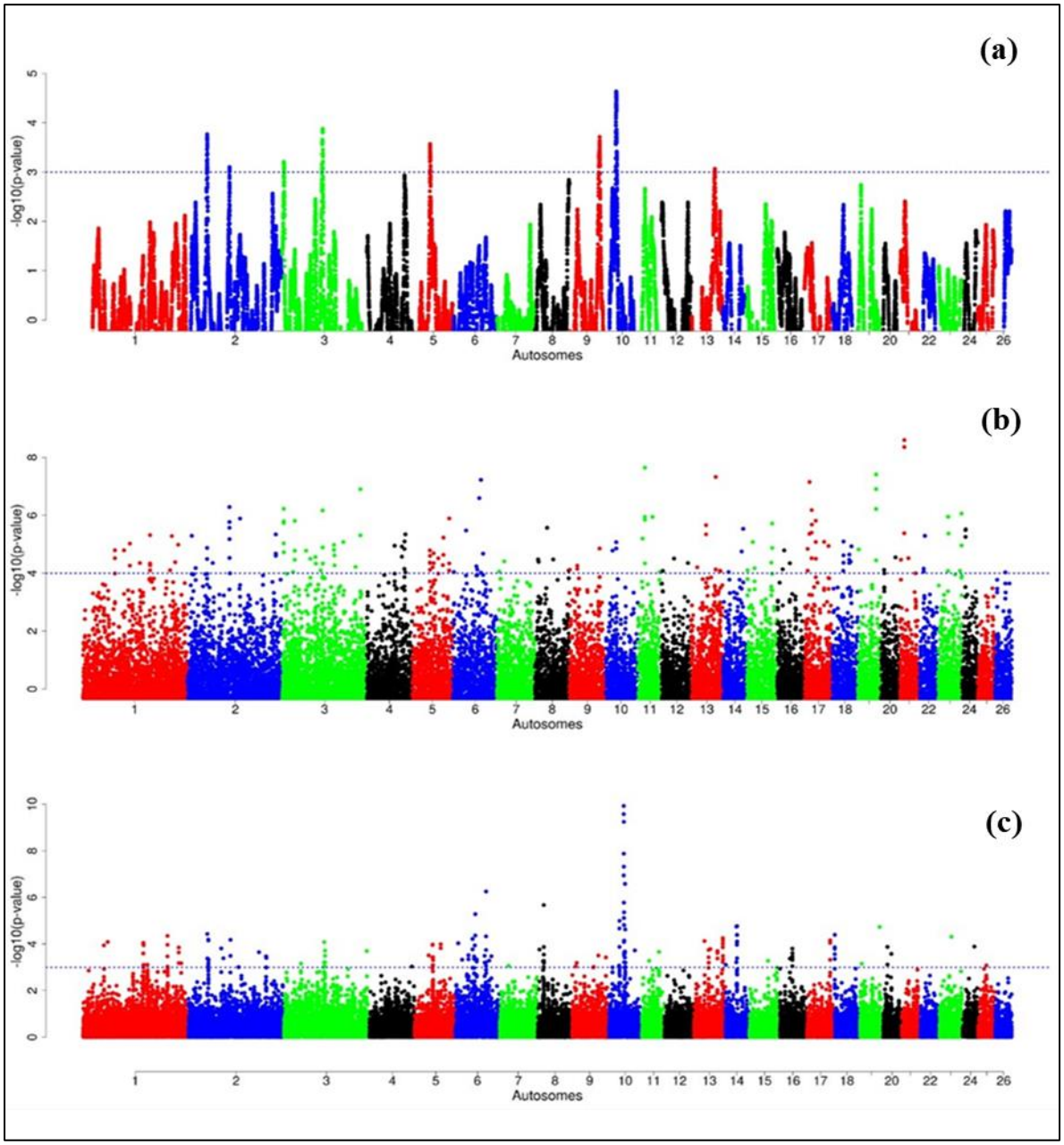


Figure 2.13 Manhattan plots of genome-wide autosomal *hapFLK* (a), ZF_{ST} (b) and *Rsb* (c) analyses of western Ethiopian long fat-tail sheep (E2) *vs.* thin-tail (S) sheep.

Table 2.3 Candidate regions and genes identified to be under selection by a combination of at least two methods in the Ethiopian western long fat-tail *vs.* Sudanese thin-tail sheep.

Chr	Overlapping region	Gene location	Method	Candidate gene	Annotation
2	51960001-52880000	51989342-52042116	*hapFLK*Rsb*	*GNE*	glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase
		52048202-52065307		*CLTA*	clathrin light chain A
		52087650-52089416		*CCIN*	Calicin
		52128947-52210749		*RECK*	reversion inducing cysteine rich protein with kazal motifs
		52411021-52417200		*FAM221B*	family with sequence similarity 221 member B
		52421111-52423389		*HINT2*	histidine triad nucleotide binding protein 2
		52423298-52426475		*SPAG8*	sperm associated antigen 8
		52423842-52445175		*NPR2*	natriuretic peptide receptor 2
		52480200-52481163		*MSMP*	microseminoprotein, prostate associated
		52480334-52485038		*RGP1*	homolog, RAB6A GEF complex partner 1
	51720001-51980000	51686233-51755300	*hapFLK*ZF$_{ST}$*	*MELK*	maternal embryonic leucine zipper kinase
		51891433-51948694		*RNF38*	ring finger protein 38
	110280001-110780000	110280423-110367262	*hapFLK*ZF$_{ST}$*Rsb*	*CLCN3*	chloride voltage-gated channel 3
		110404240-110525395		*NEK1*	NIMA related kinase 1
3	105360001-106220000	105840829-105932962	*hapFLK*Rsb*	*ANAPC1*	anaphase promoting complex subunit 1
		105945465-106063047		*MERTK*	MER proto-oncogene, tyrosine kinase
		106081347-106128188		*TMEM87B*	transmembrane protein 87B
		106141259-106197636		*FBLN7*	fibulin 7
	106860001-107240000	107108271-107174474	*hapFLK*ZF$_{ST}$*Rsb*	*TSPAN8*	tetraspanin 8
	4800001-5240000	5038319-5152424	*hapFLK*ZF$_{ST}$*	*RAPGEF1*	Rap guanine nucleotide exchange factor 1
		5207016-5299854		*UCK1*	uridine-cytidine kinase 1
		5212765-5239836		*POMT1*	protein O-mannosyltransferase 1

	107580001-107840000	107556327-107605339	*hapFLK*ZF$_{ST}$*	*ZFC3H1*	zinc finger C3H1-type containing
		107606256-107618590		*THAP2*	THAP domain containing 2
		107630616-107646736		*TMEM19*	transmembrane protein 19
		107781187-107834198		*TBC1D15*	TBC1 domain family member 15
5	46320001-46700000	46440670-46557263	*hapFLK*ZF$_{ST}$*Rsb*	*KLHL3*	kelch like family member 3
		46579802-46580796		*HNRNPA0*	heterogeneous nuclear ribonucleoprotein A0
	46740001-47120000	46741353-46780919		*PKD2L2*	polycystin 2 like 2, transient receptor potential cation channel
		46784304-46868016		*FAM13B*	family with sequence similarity 13 member B
		46910528-46915217		*WNT8A*	Wnt family member 8A
		46938902-46961998		*NME5*	NME/NM23 family member 5
		46972869-46993265		*BRD8*	bromodomain containing 8
		46994290-47001867		*KIF20A*	kinesin family member 20A
		47003373-47022642		*CDC23*	cell division cycle 23
		47062627-47080160		*GFRA3*	family receptor alpha 3
	47160001-47660000	47153096-47208649	*hapFLK*Rsb*	*KDM3B*	lysine demethylase 3B
		47209138-47212067		*REEP2*	receptor accessory protein 2
		47225278-47227722		*EGR1*	early growth response 1
		47292664-47309664		*HSPA9*	heat shock protein family A (Hsp70) member 9
		47473181-47642282		*CTNNA1*	catenin alpha 1
		47580182-47582084		*LRRTM2*	leucine rich repeat transmembrane neuronal 2
		47651357-47849675		*SIL1*	nucleotide exchange factor
	48060001-48140000	48060550-48066158		*SPATA24*	spermatogenesis associated 24
		48074043-48099657		*DNAJC18*	DnaJ heat shock protein family (Hsp40) member C18
		48118011-48122197		*SMIM33*	small integral membrane protein 33
		48123808-48127851		*TMEM173*	transmembrane protein 173
8	15780001-16700000	15790630-15823674	*Rsb*	*SERINC1*	serine incorporator 1
		15831623-15870470		*HSF2*	heat shock transcription factor 2
10	29700001-30320000	29893792-30043083	*hapFLK*Rsb*	*B3GLCT*	beta 3-glucosyltransferase
		30044800-30065505		*HSPH1*	heat shock protein family H (Hsp110) member 1

		30217152-30243100		*TEX26*	testis expressed 26
		30250695-30265933		*MEDAG*	mesenteric estrogen dependent adipogenesis
	29100001-29420000	28986741-29188660		*FRY*	FRY microtubule binding protein
	29280001-29540000	29454677-29502617	*hapFLK*ZF$_{ST}$*	*RXFP2*	relaxin family peptide receptor 2
13	61320001-61700000	61459737-61515972	*hapFLK*Rsb*	*DNMT3B*	DNA methyltransferase 3 beta
		61523883-61574930		*EFCAB8*	EF-hand calcium binding domain 8
		61581681-61607701		*SUN5*	Sad1 and UNC84 domain containing 5
		61611933-61633734		*BPIFB2*	BPI fold containing family B member 2
		61641482-61656002		*BPIFB6*	BPI fold containing family B member 6
		61665357-61680683		*BPIFB3*	BPI fold containing family B member 3
		61689117-61711550		*BPIFB4*	BPI fold containing family B member 4
	38580001-38660000	38609366-38671551	*Rsb*ZF$_{ST}$*	*RIN2*	Ras and Rab interactor 2
	38700001-38840000	38683625-38700031		*NAA20*	N(alpha)-acetyltransferase 20, NatB catalytic subunit
		38700000-38723332		*CRNKL1*	crooked neck pre-mRNA splicing factor 1
		38723037-38973982		*CFAP61*	cilia and flagella associated protein 61
14	1020001-1340000	1005889-1031106	*Rsb*	*COG4*	component of oligomeric golgi complex 4
		1032499-1045080		*FUK*	fucokinase
		1096949-1108688		*ST3GAL2*	ST3 beta-galactoside alpha-2,3-sialyltransferase 2
		1120895-1170694		*DDX19A*	DEAD-box helicase 19A
		1177773-1196812		*AARS*	alanyl-tRNA synthetase
		1265599-1291738		*PDPR*	pyruvate dehydrogenase phosphatase regulatory subunit, mitochondrial
		1316701-1359043		*GLG1*	golgi glycoprotein 1
16	33060001-33260000	33089170-33159243	*Rsb*ZF$_{ST}$*	*PLCXD3*	16 phosphatidylinositol specific phospholipase C X domain containing 3
18	1860001- 2420000	1810732-1994082	*Rsb*	*ATP10A*	ATPase phospholipid transporting 10A (putative)

Twelve overlapping candidate regions spanning 36 genes, were observed in the southern Ethiopian fat-tail verses thin-tail sheep (E3*S) **(Figure 2.14, Table 2.5**). There were also 16 genes found across one (Oar26, 3 genes), one (Oar3, 1 gene), and 12 (Oar2, 1 gene; Oar3, 9 genes; Oar10, 2 genes) candidate regions that were identified by *hapFLK*, ZF_{ST}, and *Rsb*, respectively **(Figure 2.14, Table 2.5)**. The overlapped genes between all comparisons are shown in **Figure 2.15**.

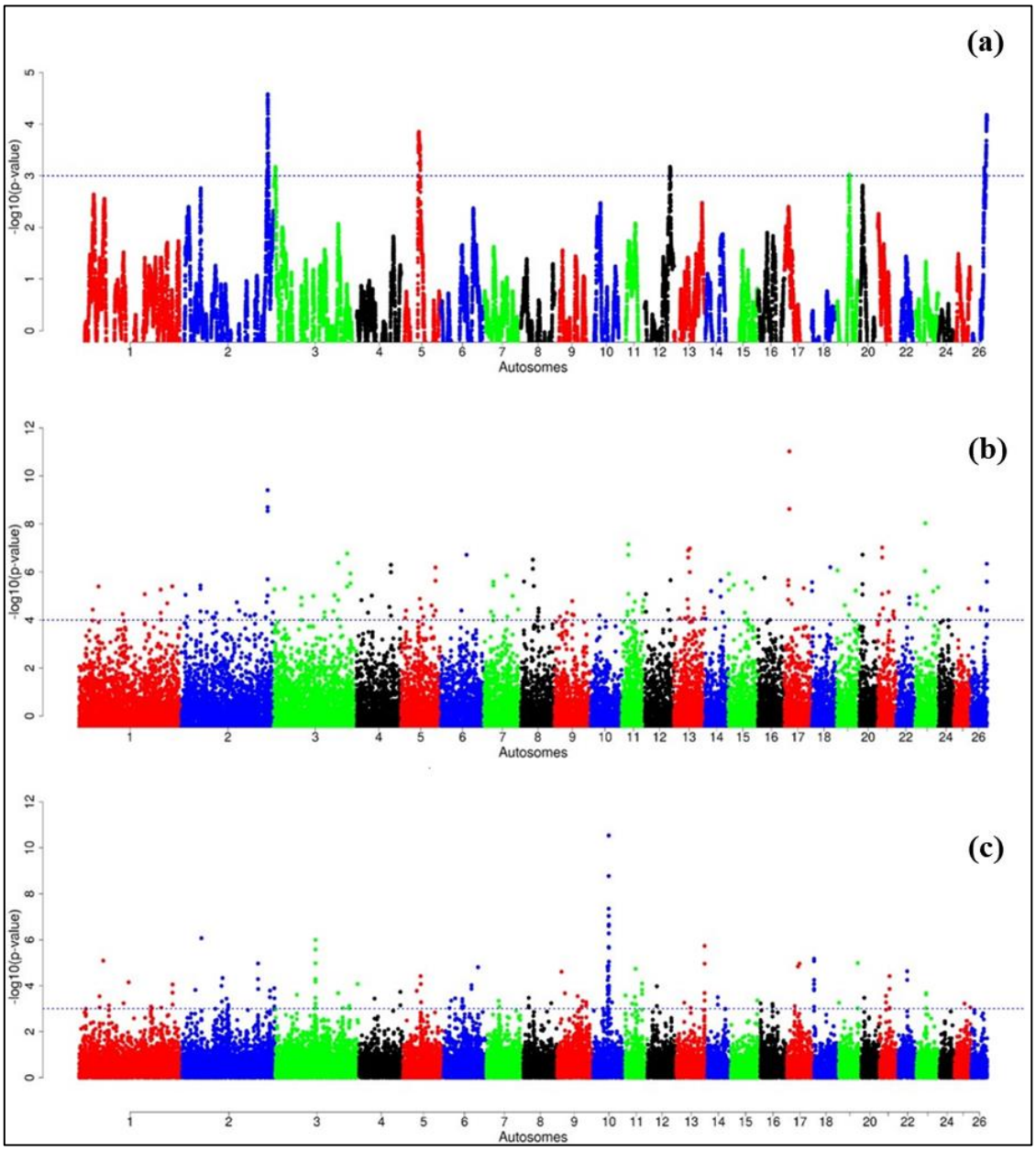


Figure 2.14 Manhattan plots of genome-wide autosomal *hapFLK* (a), ZF_{ST} (b) and Rsb (c) analyses of southern Ethiopian long fat-tail (E3) *vs*. thin-tail (S) sheep

Table 2.4 Candidate regions and genes identified to be under selection by a combination of at least two methods in the southern Ethiopia long fat-tail *vs.* thin-tail sheep.

Chr	Overlapping region	Gene location	Method	Candidate gene	Annotation
2	232500001-232940000	232511525-232515117	hapFLK*ZFST	PDE6D	phosphodiesterase 6D
		232572509-232589655		COPS7B	COP9 signalosome subunit 7B
		232749221-233048136		DIS3L2	DIS3 like 3'-5' exoribonuclease 2
	235131414-235231414	235135457-235145925	Rsb	FABP3	fatty acid binding protein 3
3	58380001-58700000	58404458-58464225	Rsb	RMND5A	required for meiotic nuclear division 5 homolog A
		58476359-58482584		D8CA	CD8a molecule
		58632941-58675138		SMYD1	SET and MYND domain containing 1
		58685544-58691134		FABP1	FABP1 fatty acid binding protein 1
	107880001-108560000	107854018-107953211		TPH2	tryptophan hydroxylase 2
		108235641-108685027		TRHDE	thyrotropin releasing hormone degrading enzyme
	181020001-181340000	181105243-181215802		SYT10	synaptotagmin 10
	220140001-220520000	220093324-220213264		ATXN10	ataxin 10
		220278340-220303444		WNT7B	Wnt family member 7B
3	106881919-107331750	107108271-107174474	ZFST	TSPAN8	tetraspanin 8
3	198240001-198380000	198318924-198332818	Rsb*ZFST	MGST1	microsomal glutathione S-transferase 1
5	46740001-47120000	46741353-46780919	hapFLK*Rsb	PKD2L2	polycystin 2 like 2, transient receptor potential cation channel
		46784304-46868016		FAM13B	family with sequence similarity 13 member B
		46910528-46915217		WNT8A	family member 8A
		46938902-46961998		NME5	NME/NM23 family member 5
		46972869-46993265		BRD8	bromodomain containing 8
		46994290-47001867		KIF20A	5 kinesin family member 20A
		47003373-47022642		CDC23	cell division cycle 23
		47062627-47080160		GFRA3	family receptor alpha 3
	47340001-47660000	47473181-47642282		CTNNA1	catenin alpha 1
		47580182-47582084		LRRTM2	leucine rich repeat transmembrane neuronal 2
		47651357-47849675		SIL1	nucleotide exchange factor
	48060001-48380000	48060550-48066158		SPATA24	spermatogenesis associated 24
		48074043-48099657		DNAJC18	DnaJ heat shock protein family

		48118011-48122197		SMIM33	small integral membrane protein 33
		48123808-48127851		TMEM173	transmembrane protein 173
		48371405-48408795		PSD2	pleckstrin and Sec7 domain containing 2
	51000001-51200000	51004236-51024531	hapFLK*ZFST	FGF1	fibroblast growth factor 1
		51190660-51671936		ARHGAP26	Rho GTPase activating protein 26
10	29160001-29480000	28986741-29188660	Rsb	FRY	FRY microtubule binding protein
		29454677-29502617		RXFP2	relaxin family peptide receptor 2
11	30240001-30380000	30273281-30323323	Rsb*ZFST	MAP2K4	mitogen-activated protein kinase kinase 4
12	69720001-69920000	69778803-69860630	hapFLK*ZFST	LPGAT1	lysophosphatidylglycerol acyltransferase 1
		69915816-69927483		NEK2	NIMA related kinase 2
13	38520001-38660000	38609366-38671551	Rsb*ZFST	RIN2	Ras and Rab interactor 2
18	1800001-2420000	1810732-1994082		ATP10A	ATPase phospholipid transporting 10A
19	51480001- 51680000	51478736-51488000	Rsb*ZFST	SPINK8	serine peptidase inhibitor, Kazal type 8
		51498015-51501665		NME6	NME/NM23 nucleoside diphosphate kinase 6
		51554521-51556734		CATHL3	BAC7.5 protein
		51566237-51569729		BAC5	5 kDa bactinecin precursor
20	9480001-9740000	9408372-9492510	Rsb*ZFST	PPARD	20 peroxisome proliferator activated receptor delta
		9523663-9535319		FANCE	Fanconi anemia complementation group E
		9551273-9570868		TEAD3	TEA domain transcription factor 3
		9574110-9588018		TULP1	tubby like protein 1
		9692040-9766541		FKBP5	FK506 binding protein 5
26	36480001-37520000	36642631-36727385	hapFLK	CSGALNACT1	Chondroitin sulfate N-acetylgalactosaminyltransferase 1
		36739622-36795389		SH2D4A	SH2 domain containing 4A
		37045901-37438136		PSD3	pleckstrin and Sec7 domain containing 3

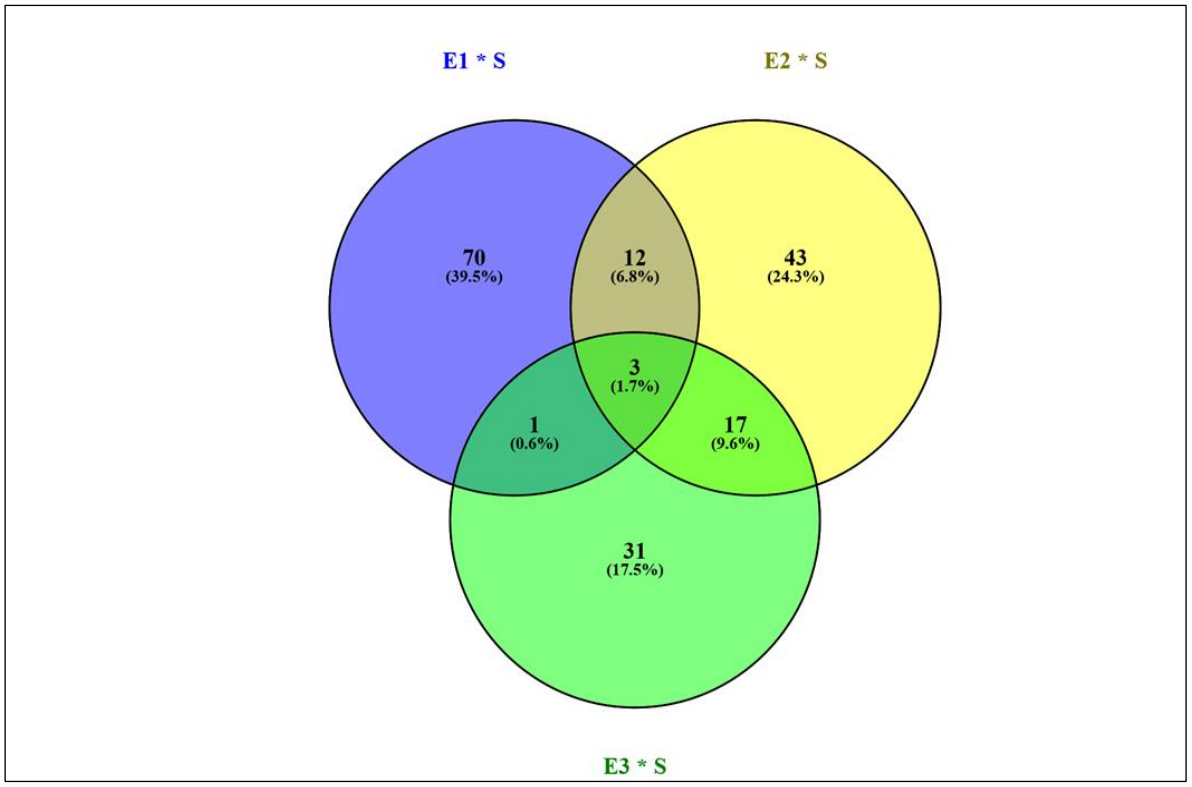


Figure 2.15 Venn diagram showing the distribution and number of genes shared between the three comparisons of sheep groups (E1*S, E2*S, E3*S) used in the analysis of selection signatures.

We performed gene ontology (GO) enrichment analysis for the candidate genes revealed in each pairwise comparison (**Table 2.6**). The five top most GO terms associated with the candidate genes from the E1*S comparison include embryonic skeletal system morphogenesis (GO:0009952, GO:0048704, GO:0030224, GO:0048706), response to cold (GO:0009409), innervation (GO:0060384), stem cell maintenance (GO:0019827) and positive regulation of cell adhesion (GO:0045785). The top GO terms associated with the E2*S candidate genes include cellular response to heat (GO:0034605), lipid binding (GO:0008289), magnesium ion binding (GO:0000287) and response to gamma radiation (GO:0000287). The GO terms for the genes from the E3*S comparison included skin development (GO:0043588), regulation of actin cytoskeleton reorganization (GO:2000249) and wound healing (GO:0042060).

Table 2.5 Enriched functional term clusters and their enrichment scores following DAVID analysis for genes identified in Ethiopian and Sudanese sheep

ID	Term	P-values	Associated genes	Comparison
GO:0009952	anterior/posterior pattern specification	0.0001	*HOXB3, HOXC6, HOXB4, HOXB1, HOXC8, HOXB2, HOXB7, HOXB5, HOXC4, HOXB6, HOXB9*	*Fat-rump vs Thin-tail*
GO:0048704	embryonic skeletal system morphogenesis	0.0010	*HOXB3, HOXB4, HOXB1, HOXB2, HOXB7, HOXB5, HOXB6*	*Fat-rump vs Thin-tail*
GO:0009409	response to cold	0.0040	*CDH8, ADRB3, THRA, TRPM8, PLAC8*	*Fat-rump vs This-tail*
GO:0030224	monocyte differentiation	0.0045	*BMP4, MED1*	*Fat-rump vs Thin-tail*
GO:0048706	embryonic skeletal system development	0.0096	*HOXC6 SULF2, WNT11, HOXB9*	*Fat-rump vs thin-tail*
GO:0060384	Innervation	0.0149	*SULF2, RNF165, LRIG2, UNC13B*	*Fat-rump vs thin-tail*
GO:0019827	stem cell population maintenance	0.0161	*MED28, NODAL, DIS3L2, MED24, LEO1, FZD7*	*Fat-rump vs thin-tail*
GO:0045785	positive regulation of cell adhesion	0.0209	*VAV3, ERBB2, ITGAV, ANGPT1, SKAP1*	*Fat-rump vs thin-tail*
GO:0034605	cellular response to heat	0.0033	*TFEC, CLPB, NF1, SLC52A3, MYOF*	*Western long fat-tail vs thin-tail*
GO:0008289	lipid binding	0.0051	*BPIFB1, BPIFA3, BPIFB2, BPIFB3, BPIFB4, BPIFA1, BPIFB6*	*Western long fat-tail vs thin-tail*
GO:0000287	magnesium ion binding	0.0069	*GSS, CIB4, EYA2, GTPBP10, SNCA, ATP10A, DIS3L2, ERN1, ITPK1, STK3, ADPRH*	*Western long fat-tail vs thin-tail*
GO:0010332	response to gamma radiation	0.0162	*BRCA2, TRIM13, PRKDC, PRKAA1*	*Western long fat-tail vs thin-tail*
GO:0043588	skin development	0.0027	*COL3A1, ITGA3, PTCH2, ARRDC3, COL5A2, DHCR24*	*Southern long fat-tail vs thin-tail*
GO:2000249	regulation of actin cytoskeleton reorganization	0.0072	*GMFG, SEMA3E, RAPGEF3, ARHGDIB*	*Southern long fat-tail vs thin-tail*
GO:0042060	wound healing	0.0232	*PPARD, COL3A1, NF1, GRHL3, PAK1*	*Southern long fat-tail vs thin-tail*

Discussion

In this study, we used the Ovine 50K SNP BeadChip generated genotype data to investigate the autosomal genetic diversity in Ethiopian indigenous sheep. Including populations from other regions of the world and the African continent allowed us to assess this diversity in a global geographic context. Our findings showed that the Ethiopian indigenous sheep are genetically differentiated from the other populations including other African fat-tail sheep (**Figures 2.5, 2. 6**). The finding that the Ethiopian fat-tail sheep are distinct from those found in North Africa, support the presence of at least two genetic groups of fat-tail sheep in the continent and two separate introduction events, one *via* the Northeast Africa and the Mediterranean Sea coastline, and the other *via* the Horn of Africa crossing through the straits of Babel-Mandeb, respectively. The distinct clustering of the thin-tail sheep suggests its independent introduction into the continent.

The fact that the South African Ronderib and Namaqua sheep occur on the same PC planar axis with Ethiopian sheep (**Figure 2.5**) may suggest, a common genetic heritage between the two rather than with the North African breeds. The movement of sheep southwards remains speculative; some linguistic evidence suggests a movement of Bantu speaking populations from West Africa to South Africa through central Africa and following a western route rather than the more traditionally postulated eastern routes from East to South Africa (Newman, 1995). In such context, a close clustering of the thin-tail West African sheep with some fat-tail southern African sheep breeds, such as the Namaqua from Namibia studied here is worth mentioning as it offers some possible insights. This, however, will require further investigation beyond the scope of this study.

Our results agree with previous findings that were arrived at using microsatellite loci (Muigai, 2003) and 50K SNP genotype data (Mwacharo et al., 2017). They are also in line with archaeological and anthropological evidence indicating the introduction first, of thin-tail sheep into the continent followed by fat-tail sheep, initially through the Sinai Peninsula and later the Horn of Africa region (Gifford-Gonzalez and Hanotte, 2011; Muigai and Hanotte, 2013).

Interestingly, the PCA results involving Ethiopian and Sudanese sheep separate the Ethiopian populations into three groups while ADMIXTURE revealed four genetic clusters in Ethiopian sheep irrespective of their geographic origins. TreeMix revealed extensive gene flow between populations of different geographic origins and tail-types. These results suggest, most likely,

current and historical intermixing of sheep following human socio-cultural and economic interactions. This appears to be a common feature in Ethiopia and most likely the Northeast and eastern Africa region as it was also observed in Ethiopian goats by Tarekegn et al. (2018). We propose here that the common D genetic background present in short fat-tail and fat-rump sheep may represent historical introgression of the thin-tail gene pool into short fat-tail and fat-rump genepool. This result calls for further investigation.

Our findings on the genetic relationships and differentiation between Ethiopian sheep populations agree with findings of previous studies, which were performed using either microsatellites (Gizaw, 2008) or 50K SNP genotype data (Edea et al., 2017) and which indicated a grouping of Ethiopian indigenous sheep populations based on their tail phenotypes. However, uniquely in our study, the long fat-tail populations were further subdivided into two secondary groups representing sheep populations from western and southern Ethiopia **(Figure 2.7)**. These two groups were also defined by different genetic backgrounds by ADMIXTURE **(Figure 2.10)** and they clustered separately in TreeMix **(Figure 2.11**). Although they are defined by the same tail phenotype, the two populations of Ethiopian short fat-tail sheep did not cluster together. Geographic isolation coupled, most likely, with adaptation to different eco-climates, as well as ethnic, cultural and religious practices and differences, that can impede gene flow, may have shaped this genetic sub-structuring (Madrigal et al., 2001; Gizaw et al., 2007).

In selection signature analysis, we contrasted groups of Ethiopian indigenous sheep that showed variation in the size of the fat-tail with thin-tail sheep. Our results identified several genes as potential candidates controlling tail morphotype and fat localization. Several genes occurred within candidate regions that overlapped between at least two of the three approaches used to detect signatures of selection (*hapFLK*, F_{ST}, *Rsb*). The F_{ST} approach detects signatures arising from an increase or decrease in allele frequency differentiation between populations/breeds, *hapFLK* detects the same but based on increase/decrease in haplotype frequency differentiation while accounting for hierarchical population structure (Kijas, 2014) while *Rsb* detects signatures associated with the LD patterns between loci across the genome (Oleksyk et al., 2010; de Simoni Gouveia et al., 2014).

Since these methods are based on different algorithms and assumptions, if common signatures are detected by at least two of the methods it suggests good reliability of the results while reducing the likelihood of interpreting false positives. They also detect signatures spanning

different periods; the F_{ST} and *hapFLK* detect signatures arising from long-term differential selection while *Rsb* detects ongoing signatures of selection including those that arise in the short to medium term (Oleksyk et al., 2010).

In the E1*S comparison, three genes associated with growth traits were present on the candidate region on Oar2, i.e., histidine triad nucleotide binding protein 2 (*HINT2*), sperm associated antigen 8 (*SPAG8*) and natriuretic peptide receptor 2 (*NPR2*). Previous studies reported these genes to be associated with birth and carcass weights, and fat depth, respectively, in cattle (Casas et al., 2000; McClure et al., 2010) and sheep (Moradi et al., 2012; Wei et al., 2015). We also identified two genes on Oar5 (*ANGPTL8, INSR*), which might be responsible for fat accumulation in adipose tissues. Angiopoietin-like 8 (*ANGPTL8*), when induced by insulin receptor (*INSR*), inhibits lipolysis and controls post-prandial fat storage in white adipose tissue and directs fatty acids to adipose tissue for storage during the fed state (Mysore et al., 2017).

The *ADAMTS3* (ADAM metallopeptidase with thrombospondin type 1 motif 3) gene was present in the region identified on Oar6. This gene is expressed in cartilage, where collagen II is a major component, as well as in embryonic bone and tendon, suggesting that it could be a major procollagen processing enzyme in musculoskeletal tissues (Dubail and Apte, 2015). The homeobox B13 (*HOXB13*) and ALX homeobox 4 (*ALX4*) were identified on the candidate region on Oar11 and Oar15, respectively. Mutations in the former result in overgrowth of the caudal spinal cord and tail vertebrae in mice (Economides et al., 2003), while the later is involved in the development of limbs and skeleton (Fariello et al., 2014).

Our enrichment analysis for the E1*S genes revealed a cluster of genes (*BMP4*, *MED1*) with functions that could be related to the tail formation **(Table 2.6)**. Bone Morphogenetic Protein 4 (*BMP4*) was revealed by *Rsb* and F_{ST} to be on a candidate region on Oar7 and it has been implicated in tail formation (Moioli et al., 2015). Peroxisome Proliferator Activated Receptor Gamma (*PPARG*) expression has been associated with back-fat thickness in sheep (Dervish et al., 2011). Ge et al. (2008) reported Mediator Complex Subunit 1 (*MED1*) is an essential protein for the optimal functioning of *PPARG.* Despite this association, our analysis did not reveal any signals spanning *PPARG,* but two of our methods (*Rsb* and F_{ST}) revealed a signature on Oar20 that spanned the *PPARD* gene, a paralogue to *PPARG.*

In the same comparison (E1*S), we identified a cluster of genes (*CDH8, ADRB3, THRA, TRPM8, PLAC8*) that are associated with the GO biological process, response to cold. This is

not surprising considering that three out of the four E1 populations are living at a high altitude and therefore in a relatively cold habitat. Indeed, Adreno receptor Beta 3 (*ADRB3*) plays a major role in energy metabolism and the regulation of lipolysis and homeostasis (Wu et al., 2012). It is also associated with birth weight, growth rate, carcass composition and survival in various breeds of sheep (Horrell et al., 2009). The ion channel *TRPM8* has been reported to play a major role in eliciting cold defense thermoregulation, metabolic and defense immune responses in humans (Kozyreva and Voronova, 2015).

Several other genes occurring in the E1*S candidate regions and which are associated with the GO term embryonic skeletal system development (GO:0048706) included *HOXC6, SULF2, WNT11,* and *HOXB9. WNT11* was identified by ZF_{ST} on Oar15 while *HOXC6* and *HOXB9* were revealed by *hapFLK* on Oar3 and Oar13, respectively. The *WNT* gene family and the *T* gene have been implicated in vertebral development in laboratory mice (Greco et al., 1996), and with the short-tail phenotype in sheep (Zhi et al., 2018). In addition, the roles of the *WNT* gene family in lipid metabolic processes in fat-tail sheep have also been reported (Kang et al., 2017). The *HOX* genes represent transcriptional regulatory proteins that control axial patterning in bilaterians (Garcia-Fernàndez, 2005), where the inactivation of one of the *HOX* genes often causes transformations in the identity of vertebral elements (Mallo et al., 2010). The *HOX* genes can control morphologies along the anteroposterior axis (Lewis, 1978). Furthermore, *HOXC11, HOXC12,* and *HOXC13* developmental genes were found to be expressed in the tail region indicating their possible associations with tail size and fat development in fat-tail sheep (Kang et al., 2017).

The candidate regions revealed by the E2*S comparison spanned 65 candidate genes. Three genes of the BPI fold Containing Family B (*BPIFB3, BPIFB4, and BPIFB6*) were present in a candidate region on Oar13. These, along with other paralogs (*BPIFB1, BPIFA3, BPIFB2, BPIFA1*), formed a cluster of functional genes related to the GO term, lipid binding functional process **(Table 2.6)**. In contrast to the E1*S comparison, the cluster of genes identified in the E2*S comparison were associated with the GO terms, Magnesium ion binding, response to gamma radiation and cellular response to heat. This suggests, most likely, the propensity of this group of sheep to adapt to the eco-climatic conditions prevailing in their home-tract. This is consistent with the humid highland and moist lowland conditions of the geographic area where the populations representing the E2 group (Bonga, Gesses, Kido) were sampled. Based on field experience, high fecundity and prolificacy is a common reproductive trait preferred by

farmers in the Bonga sheep. This may explain the occurrence of the *CIB4* and *PRKAA1* in a candidate region in the E2*S comparison. The *CIB4* gene was suggested to be linked, in some way, to high fecundity in the small Tail Han sheep (Yu et al., 2010) and *PRKAA1* is involved in ewe's follicular development (Foroughinia et al., 2017).

The third comparison (E3*S) resulted in 36 genes that occurred in candidate regions that were revealed by at least two methods used to detect selection signatures. Fatty acid binding proteins *FABP3 and FABP1* found on candidate regions on Oar2 and Oar3, respectively are the genes that relate most closely to fat deposition. *SREBF1* along with *PPARG* are the main transcription factors controlling lipogenesis in adipose tissue and skeletal muscle (Ferré and Foufelle, 2010), and are mainly regulated by fatty acid-binding proteins (*FABP*) (Lapsys et al., 2000). Recently, Bahnamiri et al. (2018) evaluated the effects of negative and positive energy balances on the expression pattern of these genes in fat-tail and thin-tail lambs. They observed differential transcriptional regulation of lipogenesis and lipolysis during periods of negative and positive energy balances in the two groups of lambs. In general, the cluster of genes identified in this comparison was significantly enriched for GO terms relating to skin development, wound healing and regulation of actin cytoskeleton reorganization **(Table 2.6)**.

The commonest genes between the three comparisons are *TSPAN8, RXFP2,* and *RIN2* (**Figure 2.15**). The *TSPAN8* (Tetraspanin 8) occurred in the candidate region on Oar3; it is among the genes that are reported to be associated with insulin release and sensitivity, and obesity in humans (Grarup et al., 2008), while the relaxin family peptide receptor 2 (*RXFP2*) has been associated with horn morphology (Johnston et al., 2011; Wiedemar and Drögemüller, 2015).

Twelve genes (*MELK, RNF38, GNE, CLTA, CCIN, RECK, HINT2, SPAG8, NPR2, FAM221B, MSMP, RGP1*) were common between E1*S and E2*S comparisons. On Oar2, three genes were identified within the overlapping candidate region, i.e., *CLTA* which is associated with prion protein deposition in sheep (Filali et al., 2014), *GNE* which is important for the metabolism of sialated oligosaccharides in bovine milk (Wickramasinghe et al., 2011) and *RECK* which encodes an inhibitor of angiogenesis, invasion and metastasis, DNA methylation, and increased mRNA in cell lines in humans (Su, 2012). Other genes (i.e., *HINT2, SPAG8,* and *NPR2*) are associated with fat deposition in sheep as herein discussed for each of the three comparisons.

Furthermore, one gene (*DIS3L2*) was in a candidate region that overlapped between the E1*S and E3*S comparisons. DIS3 like 3'-5' exoribonuclease 2 (*DIS3L2*) has also been identified,

among genes involved in cancer, cellular function and maintenance, and neurological disease, in a candidate region under selection in cattle (Gautier et al., 2009). In sheep, using F_{ST}, *iHS*, and *Rsb*, de Simoni Gouveia et al. (2017) indicated that *DIS3L2 is* among genes associated with height variation. In addition, *DIS3L2* has reportedly been associated with the Perlman syndrome, which is characterized by overweight in humans (Astuti et al., 2012).

Finally, seventeen genes (*PKD2L2, FAM13B, WNT8A, NME5, BRD8, KIF20A, CDC23, GFRA3, CTNNA1, LRRTM2, SIL1, SPATA24, DNAJC18, SMIM33, TMEM173, FRY, ATP10A*) were in candidate regions that overlapped between the E2*S and E3*S comparisons. Among these, DnaJ heat shock protein family (*HSP40*) member C18 (*DNAJC18*) and spermatogenesis associated 24 (*SPATA24*) on Oar5 were reported among genes involved in heat stress tolerance and male reproductive function, respectively, in East African Shorthorn Zebu cattle (Bahbahani et al., 2015).

Conclusion

Overall, our results revealed four distinct autosomal genomic backgrounds (A, B, C, D) in Ethiopian indigenous sheep. The genotypes of most of the individuals analyzed were made up of at least two genetic backgrounds which could be accounted for by some level of current or historical admixture between populations. Selection signature analysis identified several putative candidate regions spanning genes relating to skeletal structure and morphology, fat deposition and possibly adaptation to environmental selection pressures. Our results indicate that Ethiopian indigenous sheep may represent a valuable animal genetic model to understand genetic mechanisms associated with body fat metabolism and distribution in humans. This is especially important not only because fat deposits are a crucial component of adaptive physiology but also because excessive fat deposition in adipose tissue can result in obesity and overweight, and energy metabolism disorders.

References

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

Assefa, F., Tigistu, T., and Lambebo, A. (2015). Assessment of the production system, major constraints and opportunities of sheep production in Doyogena Woreda, Kembata Tambaro Zone, Southern Ethiopia. *J Biol. Agric. Healthc.* 5, 37–41.

Astuti, D., Morris, M. R., Cooper, W. N., Staals, R. H., Wake, N. C., and Fews, G. A. (2012). Germline mutations in DIS3L2 cause the Perlman syndrome of overgrowth and Wilms tumor susceptibility. *Nature Genet.* 44, 277–284.

Atti, N., Bocquier, F., and Khaldi, G. (2004). Performance of the fat-tailed Barbarine sheep in its environment: adaptive capacity to alternation of underfeeding and re-feeding periods. *Rev. Anim. Res.* 53, 165–176.

Bahbahani, H., Clifford, H., Wragg, D., Mbole-Kariuki, M. N., Van Tassell, C., Sonstegard, T., et al. (2015). Signatures of positive selection in East African Shorthorn Zebu: a genome-wide single nucleotide polymorphism analysis. *Sci. Rep.* 5:11729.

Bahnamiri, H. Z., Zali, A., Ganjkhanlou, M., Sadeghi, M., and Shahrbabak, H. M. (2018). Regulation of lipid metabolism in adipose depots of fat-tailed and thin-tailed lambs during negative and positive energy balances. *Gene* 641, 203–211.

Blench, R., and MacDonald, K. (eds.). (2006). *The Origins and Development of African Livestock: Archaeology, Genetics, Linguistics and Ethnography.* London; New York, NY: Routledge; Taylor & Francis Group.

Buchmann, R., and Hazelhurst, S. (2014). *Genesis Manual.* Johannesburg: University of the Witwatersrand. Available online at: http://www.bioinf.wits.ac.za/software/genesis/Genesis.pdf

Casas, E., Shackelford, S. D., Keele, J. W., Stone, R. T., Kappes, S. M., and Koohmaraie, M. (2000). Quantitative trait loci affecting growth and carcass composition of cattle segregating alternate forms of myostatin. *J Anim. Sci.* 78, 560–569.

de Simoni Gouveia, J. J., de, S., Silva, M.V.G.B. da., Paiva, S. R., and Oliveira, S. M. P., de. (2014). Identification of selection signatures in livestock species. *Genet. Mol. Biol.* 37, 330–342.

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

Delaneau, O., Marchini, J., McVean, G. A., Donnelly, P., Lunter, G., and Marchini, J. L. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Commun.* 5:3934.

Dervish, E., Serrano, C., Joy, M., Serrano, M., Rodellar, C., and Calvo, J. H. (2011). The effect of feeding system in the expression of genes related with fat metabolism in semitendinous muscle in sheep. *Meat Sci.* 89, 91–97.

Dubail, J., and Apte, S. S. (2015). Insights on ADAMTS proteases and ADAMTS-like proteins from mammalian genetics. *Matrix Biol*. 44, 24–37.

Economides, K. D., Zeltser, L., and Capecchi, M. R. (2003). *Hoxb13* mutations cause overgrowth of caudal spinal cord and tail vertebrae. *Dev. Biol.* 256, 317–330.

Edea, Z., Dessie, T., Dadi, H., Do, K. T., and Kim, K. S. (2017). Genetic diversity and population structure of Ethiopian sheep populations revealed by high-density SNP markers. *Front. Genet.* 8:218.

Epstein, H. (1971). *The Origin of Domestic Animals of Africa, Volume 2.* New York, NY: Africana Publication Corporation, 719.

Fariello, M. I., Boitard, S., Naya, H., San Cristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193, 929–941.

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

Ferré, P., and Foufelle, F. (2010). Hepatic steatosis: a role for *de novo* lipogenesis and the transcription factor SREBP-1c. *Diabetes Obes. Metab.* 12, 83-92.

Filali, H., Martín-Burriel, I., Harders, F., Varona, L., Hedman, C., and Mediano, D. R. (2014). Gene expression profiling of mesenteric lymph nodes from sheep with natural scrapie. *BMC Genom.* 15:59.

Foroughinia, G., Fazileh, A., and Eghbalsaied, S. (2017). Expression of genes involved in BMP and estrogen signaling and AMPK production can be important factors affecting total number of antral follicles in ewes. *Theriogenology* 91, 36-43.

Garcia-Fernàndez, J. (2005). Hox, ParaHox, ProtoHox: facts and guesses. *Heredity* 94, 145-152.

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.

Gautier, M., Flori, L., Riebler, A., Jaffrézic, F., Lalo,é, D., Gut, I., et al. (2009). A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genom.* 10:550.

Ge, K., Cho, Y. W., Guo, H., Hong, T. B., Guermah, M., and Ito, M. (2008). Alternative mechanisms by which mediator subunit MED1/TRAP220 regulates peroxisome proliferator-activated receptor γ-stimulated adipogenesis and target gene expression. *Mol. Cell. Biol.* 28, 1081–1091.

Gifford-Gonzalez, D., and Hanotte, O. (2011). Domesticating animals in Africa: Implications of genetic and archaeological findings. *J. World Prehistory* 24,1-23.

Gizaw, S. (2008). *Sheep Resources of Ethiopia: Genetic Diversity and Breeding Strategy.* A Ph.D. thesis, Wageningen University. 145.

Gizaw, S., Van Arendonk, J. A., Komen, H., Windig, J. J., and Hanotte, O. (2007). Population structure, genetic variation and morphological diversity in indigenous sheep of Ethiopia. *Anim. Genet*. 38, 621–628.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Resour*. 5, 184–186.

Grarup, N., Andersen, G., Krarup, N. T., Albrechtsen, A., Schmitz, O., Jørgensen, T., et al. (2008). Association testing of novel type 2 diabetes risk alleles in the JAZF1, CDC123/CAMK1D, TSPAN8, THADA, ADAMTS9, and NOTCH2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged Danes. *Diabetes* 57, 2534–2540.

Greco, T. L., Takada, S., Newhouse, M. M., McMahon, J. A., McMahon, A. P., and Camper, S. A. (1996). Analysis of the vestigial tail mutation demonstrates that Wnt-3a gene dosage regulates mouse axial development. *Genes Develop*. 10, 313–324.

Horrell, A., Forrest, R. H. J., Zhou, H., Fang, Q., and Hickford, J. G. H. (2009). Association of the ADRB3 gene with birth weight and growth rate to weaning in New Zealand Romney sheep. *Anim. Genet*. 40, 251–251.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protoc.* 4, 44–57.

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

Johnston, S. E., McEwan, J. C., Pickering, N. K., Kijas, J. W., Beraldi, D., Pilkington, J. G., et al. (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, 2555–2566.

Kang, D., Zhou, G., Zhou, S., Zeng, J., Wang, X., Jiang, Y., et al. (2017). Comparative transcriptome analysis reveals potentially novel roles of Homeobox genes in adipose deposition in fat-tailed sheep. *Sci. Rep.* 7:14491.

Kijas, J. W. (2014). Haplotype-based analysis of selective sweeps in sheep. *Genome* 57, 433-437.

Kozyreva, T. V., and Voronova, I. P. (2015). Involvement of neurogenomic regulation in maintenance of temperature homeostasis in the cold. *Russ. J. Genet. Appl. Res.* 5, 569–576.

Lapsys, N. M., Kriketos, A. D., Lim-Fraser, M., Poynten, A. M., Lowy, A., and Furler, S. M. (2000). Expression of genes involved in lipid metabolism correlate with peroxisome proliferator-activated receptor γ expression in human skeletal muscle. *J. Clin. Endocr. Metab*. 85, 4293–4297.

Lawal, R. A., Al-Atiyat, R. M., Aljumaah, R. S., Silva, P., Mwacharo, J. M., and Hanotte, O. (2018). Whole-genome resequencing of red junglefowl and indigenous village chicken reveal new insights on the genome dynamics of the species. *Front. Genet*. 9:264.

Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. *Nature* 276, 565-570.

Madrigal, L., Ware, B., Miller, R., Saenz, G., Chavez, M., and Dykes, D. (2001). Ethnicity, gene flow, and population subdivision in Limon, Costa Rica. *Am. J. Phys. Anthrop*. 114, 99–108.

Mallo, M., Wellik, D. M., and Deschamps, J. (2010). Hox genes and regional patterning of the vertebrate body plan. *Dev. Biol*. 344, 7-15.

Marshal, F. (2000). The origins and spread of domestic animals in East Africa. In "The Origins and Development of African Livestock: Archaeology, Genetics, Linguistics and Ethnography" (R. M. Blench and K. C. MacDonald, eds.), pp. 191-218. University College London Press, London.

McClure, M. C., Morsci, N. S., Schnabel, R. D., Kim, J. W., Yao, P., and Rolf, M. M. (2010). A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. *Anim. Genet*. 41, 597–607.

Moioli, B., Pilla, F., and Ciani, E. (2015). Signatures of selection identify loci associated with fat tail in sheep. *J. Anim. Sci.* 93, 4660–4669.

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

Muigai, A. W. T. (2003). *Characterization and Conservation of Indigenous Animal Genetic Resources: Genetic Diversity and Relationships of Fat-Tailed and Thin-Tailed Sheep of Africa*. Ph.D. thesis, Jomo Kenyatta, University of Agriculture and Technology Juja, Kenya.

Muigai, A. W. T., and Hanotte, O. (2013). The origin of African sheep: Archaeological and genetics perspectives. *Afr. Archaeol. Rev*. 30, 39–50.

Mwacharo, J. M., Kim, E.-S., Elbeltagy, A. R., Aboul-Naga, A. M., Rischkowsky, B. A., and Rothschild, M. F. (2017). Genomic footprints of dryland stress adaptation in Egyptian fat-tail sheep and their divergence from East Africa and western Asia cohorts. *Sci. Rep*. 7, 17647.

Mysore, R., Liebisch, G., Zhou, Y., Olkkonen, V. M., and Haridas, P. N. (2017). Angiopoietin-like 8 (Angptl8) controls adipocyte lipolysis and phospholipid composition. *Chem. Phys. Lipids* 207, 246–252.

Nejati-Javaremi, A., Izadi, F., Rahmati, G., and Moradi, M. (2007). Selection in fat-tailed sheep based on two traits of fat-tail and body weight versus single-trait total body weight. *Int. J Agri. Biol*. 9, 645–648.

Newman, J. L. (1995). *The Peopling of Africa: A Geographic Interpretation.* Dexter, MI: Yale University Press.

Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 185–205.

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*. 8, e1002967.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., and Bender, D. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559-575.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.

R Core Team (2012). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,* Vienna. Available online at: URL http:// www.R-project.org/

Ryder, M. L. (1983). *Sheep and Man*. London: Gerald Duckworth & Co. Ltd.

Su, L. J. (2012). *Diet, Epigenetics, and Cancer. In Cancer Epigenetics,* Totowa, NJ: Humana Press, 377–393.

Tarekegn, G. M., Tesfaye, K., Mwai, O. A., Djikeng, A., Dessie, T., Birungi, J., et al. (2018). Mitochondrial DNA variation reveals maternal origins and demographic dynamics of Ethiopian indigenous goats. *Ecol Evol.* 8, 1543–1553.

Wei, C., Wang, H., Liu, G., Wu, M., Cao, J., and Liu, Z. (2015). Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. *BMC Genom.* 16:194.

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

Wickramasinghe, S., Hua, S., Rincon, G., Islas-Trejo, A., German, J. B., Lebrilla, C. B., et al. (2011). Transcriptome profiling of bovine milk oligosaccharide metabolism genes using RNA-sequencing. *PLoS ONE* 6:e18895.

Wiedemar, N., and Drögemüller, C. (2015). A 1.8-kb insertion in the 3′-UTR of RXFP2 is associated with polledness in sheep. *Anim. Genet.* 46, 457–461.

Wu, J., Qiao, L., Liu, J., Yuan, Y., and Liu, W. (2012). SNP variation in ADRB3 gene reflects the breed difference of sheep populations. *Molecular Biol. Rep*. 39, 8395–8403.

Yu, Y., Zhang, Y., Song, X., Jin, M., Guan, Q., Zhang, Q., et al. (2010). Alternative splicing and tissue expression of CIB4 gene in sheep testis. *Anim. Repro. Sci.* 120, 1-9.

Yuan, Z., Liu, E., Liu, Z., Kijas, J. W., Zhu, C., Hu, S., et al. (2017). Selection signature analysis reveals genes associated with tail type in Chinese indigenous sheep. *Anim. Genet*. 48, 55-66.

Zhi, D., Da, L., Liu, M., Cheng, C., Zhang, Wang, Y., et al. (2018). Whole genome sequencing of Hulunbuir short-tailed sheep for identifying candidate genes related to the short-tail phenotype. *Gene Genom. Genet*. 8, 377-383.

Zhu, C., Fan, H., Yuan, Z., Hu, S., Ma, X., Xuan, J., et al. (2016). Genome-wide detection of CNVs in Chinese indigenous sheep with different types of tails using ovine high-density 600K SNP arrays. *Sci. Rep.* 6:27822.

Chapter 3

Autosomal genome-wide diversity, population structure and demographic dynamics of North and East African indigenous sheep

Abstract

The phenotypic variation within and among sheep populations is explained by differences in ancestral origins, migratory history and local natural and human-driven selection. Here, we generated and analyzed whole-genome data from 60 long fat-tailed, 32 short fat-tailed and 38 fat-rump sheep from Ethiopia and Libya (~30x coverage) as well as 20 thin-tailed sheep from Sudan (~10x coverage). Overall, 34.8 million high quality single nucleotide polymorphisms were identified and used to assess genomic diversity. The results from full genome sequence analysis are in agreement with those from **Chapter 2,** which used 50K Bead-Chip data generated. They are in line with the archeological history of African sheep in relation to their proposed entry points into, and migration events across, the continent. Diversity indices indicate moderate to high levels of genetic diversity, while differences in inbreeding values support possible bottleneck in Ethiopian and Sudanese sheep following possibly intensive selection for different tail types in the former. Population relationship and structure analysis clustered the study populations into four genetic groups representing Ethiopian fat-rump, long fat-tail from Ethiopia and Libya, respectively and thin-tail (SD) sheep from Sudan. The Libyan population displays a high level of genetic variation combined with reduced inbreeding values in light of the random mating breeding applied by their local owners. Further genome analysis studies including thin-tail sheep from Ethiopia and both thin and/or fat-tail sheep from West, North and South Africa will provide a more comprehensive view on the genome diversity of African sheep at the continental level.

Introduction

Since their domestication in the Near East approximately 10,000 years ago, sheep have adapted to a wide spectrum of geographic, ecosystems and management regimes due to their obedience behaviour, manageable size and better adaptability to variable biotic and abiotic stressors such as food shortage and extreme climatic conditions (Kijas et al., 2009). There are currently about 1400 recognized sheep breeds/populations across the world (Scherf, 2000). The phenotypic differences between these breeds/populations can be attributed to differences in their ancestral origins and migration history (Cañón et al., 2001), adaptation to local environments following selection pressure on specific loci in different ecological environments (Brehm et al., 2001), and cultural human preferences for specific phenotypes (Gizaw et al., 2007). Furthermore, human-directed breeding goals have led to the development of tailored sheep breeds and/or populations to produce specialised products, such as meat and wool.

The African continent has witnessed multiple entry points and migration events of domestic sheep. The first wave introduced hairy thin-tailed sheep followed by a second wave that introduced wooly thin-tailed sheep *via* the Isthmus of Suez (Epstein, 1954; Ryder, 1984; Marshal, 2000). Sheep from these two migratory events seem to be the most ancient of the African sheep and their phenotypic characters (e.g. convex nose, pendulous ears and exceptionally long legs) resemble ancient Egyptian sheep (Ryder, 1983; Wilson, 1991; Muigai and Hanotte, 2013). Presently, the thin-tailed sheep mainly inhabit the arid and semi-arid regions in North Africa from Morocco to Egypt, in Sudan and the sub-humid and humid regions of West Africa from Senegal to Nigeria (Muigai and Hanotte, 2013; Mwacharo et al., 2017). They likely reached East Africa following migration from the North (Marshal, 2000). The fat-tailed sheep comprised the third wave of arrival and dispersal of sheep in the African continent *via* two entry points; the Isthmus of Suez for North African fat-tailed sheep and the straits of Bab el Mandeb in the Horn of Africa for the fat-tail sheep found in the wider eastern Africa region (Ryder, 1984). The northern stream of fat-tail sheep expansion replaced the original hairy thin-tailed sheep in Egypt, Libya and Tunisia, but it did not diffuse to eastern and southern Africa regions, while those which entered through the straits of Bab el Mandeb migrated to southern Africa likely following the movement of pastoral communities (Epstein, 1954; Wilson, 2011).

The fat-tail sheep predominate in the deserts of North Africa, and in the highlands, semi-arid and arid environments of East and SouthAfrica (Mwacharo et al., 2017). The North African

fat-tail sheep have long pendulous fat tails. They are represented by the Libyan Barbary, Tunisian Barbarine and Egyptian Barki sheep. The East African fat-tail sheep, represented here by Ethiopian sheep, are classified based on their tail length into long fat-tailed types inhabiting mid to high-altitudes and the short fat-tail sheep which occur in very cool alpine high-altitude areas (Gizaw et al., 2007; Edea et al., 2017; Ahbara et al., 2019). The fat-rump sheep entered the Horn of Africa much later (Epstein, 1954). It likely originated from the Black Head Somali sheep (Wilson, 2011). This group of sheep now inhabit the arid low-altitude areas of Somalia, the eastern and northern parts of Ethiopia with neighbouring expansions in Eritrea, the Toposa area of Sudan and the northern part of Kenya (Gizaw et al., 2007; Wilson, 2011; Edea et al., 2017).

Several microsatellite-based genetic studies have assessed the genetic diversity of African sheep populations and attempted to reconstruct their genetic history (Gizaw et al., 2007; Gornas et al., 2011; Gaouar et al., 2014; Sassi-Zaidy et al., 2014; Elbeltagy et al., 2015; Gaouar et al., 2015). Recently, Edea et al. (2017) published the first genome-wide analysis of Ethiopian sheep using the High Density (~600K) ovine SNP Chip data, while we have reported in **Chapter 2**, genome-wide 50K SNP Chip data for different populations from the same region (Ahbara et al., 2019).

The traditional SNP based studies lack the power to identify causative phenotypic variants following incomplete LD between causal variants and genotyped SNPs (Yang et al., 2010; Das et al., 2015). Even with low or high density SNP genotype information, the few numbers of SNPs in LD with causative variants for complex traits makes it difficult to accurately predict genomic breeding values (Meuwissen and Goddard, 2010), a problem that is enhanced by the possible ascertainment bias of the screening tools. The advent of whole-genome sequence data has reduced these limitations while opening the doors towards more accurate genome selection approaches. Besides, whole-genome sequence information allows the reconstruction of full genome phylogenies including the assessment of broader evolutionary patterns and innovations (e.g. introgression from related species), population level genomic studies including the assessment of local adaptation or population divergence (Angeloni et al., 2012; Funk et al., 2012; Kardos et al., 2015; Yang et al., 2016), demographic history (Li and Durbin, 2011; Sheehan et al., 2013), and the identification of the genetic basis of phenotypic variation (Daetwyler et al., 2014). Finally yet importantly, besides the full genome identification of SNPs and InDels, genome sequencing allows the identification of copy number variations following

e.g. gene expansions (Dalloul et al., 2010; Telford and Copley, 2011; Prado-Martinez et al., 2013).

The latest sheep reference genome assembly is Oar_v3.1 (ISGC et al., 2010), derived from sequencing of a Texel sheep. As at 23rd December 2018, the Ensembl *Ovis aries* dbSNP database (https://www.ensembl.org/info/genome/variation/species/sources_documentation.html) release 150 contained a total of 61 million variants (including SNPs and indels).

Based on **Chapter 2** results, extending our analysis using the more robust whole-genome sequencing protocol could provide more comprehensive insights on the genome diversity and demographic patterns of sheep distribution around the two main historic entry points and dispersal of sheep into the continent. This could be supported by including sheep populations (such as Libyan Barbary) from northern Africa. Such a strategy could be valuable in identifying signatures of selection for adaptation and how such adaptations have shaped their genomes. Hence, identifying genes involved in adaptive evolution extends our knowledge of both the origins of biodiversity and the processes driving the evolutionary and phenotypic specialization in sheep.

Here, we report autosomal whole-genome sequence analysis of diversity, population structure and demographic dynamics of African indigenous sheep comprising 60 long fat-tailed, 32 short fat-tailed and 38 fat-rump sheep from Ethiopia and Libya (~30x coverage) as well as 20 thin-tailed sheep from Sudan (~10x coverage).

Materials and methods

DNA Samples and Extraction

The sampling strategy and the selected indigenous breeds of sheep from different geographic regions in Ethiopia, Sudan and Libya are presented in **Figure 3.1** and **Table 3.1.** Geographic positioning system (GPS) coordinates were recorded for all the populations. We used altitude to determine the agro-eco-climatic zones where the sheep were sampled. All efforts were made to include populations representing the different tail phenotypes found in the study areas. Twenty and 12 DNA samples from two thin-tail (Hammari and Kabashi) and pendulous fat-tail (Barbary) sheep were obtained from Sudan and Libya, respectively (see details in **Chapter 2**). Genomic DNA was extracted from 118 ear tissue samples, collected from 12 Ethiopian indigenous sheep breeds, using the NucleoSpin® Tissue Kit (www.mn-net.com) following the manufacturers protocol. All the 150 genomic DNA samples were sequenced using the Illumina HiSeq2000 at Edinburgh Genomics (University of Edinburgh - Roslin Institute).

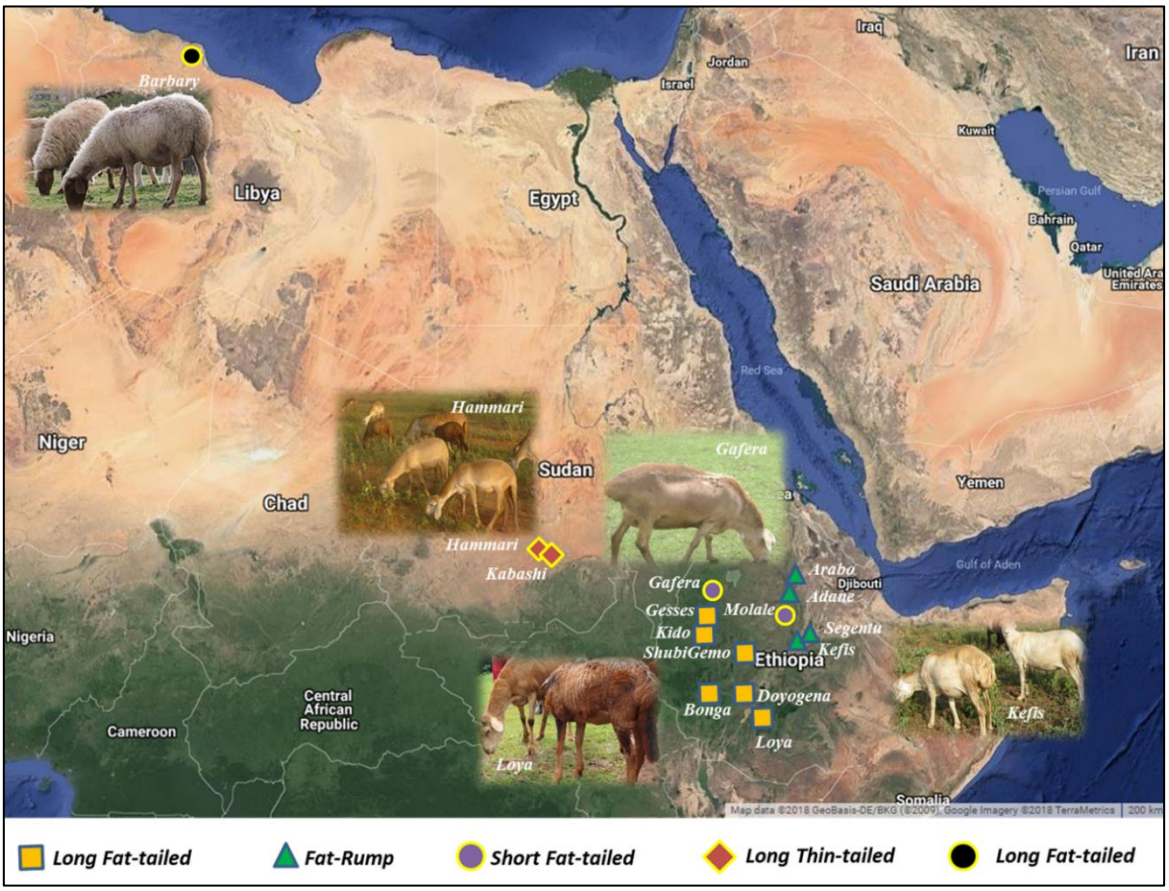


Figure 3.1 The sampling locations of the Ethiopian, Sudanese and Libyan sheep populations

Table 3.1 Description of the sheep populations sampled for this study

Origin	Population	Political Zone	Latitude (*N*)	Longitude (*E*)	Altitude (meter asl)	N		Tail Type	Agro-ecology
						Male	Female		
Ethiopia	Kefis	Zone 3	9°30’	40°10’	890	2	11	Fat-Rump	Arid lowland
	Segentu	Zone 3	9°27’	40°05’	740	3	3	Fat-Rump	Arid lowland
	Adane	South Wollo	11°14’	39°50’	2450	6	4	Fat-Rump	Cool highland
	Arabo	South Wollo	11°09’	39°54’	1500	2	7	Fat-Rump	Cool highland
	Gafera-Washera	Agew Awi	11°31’	36°54’	2500	1	9	Short, fat tail	Wet, warmer mid-highland
	Molale-Menz	North Shewa	10°70’	39°39’	3068	5	5	Short, fat tail	Sub-alpine
	Bonga	Keffa	7°16’	36°15’	1788	-	10	Long, fat tail	Humid mid-highland
	Gesses	Metekel	10°50’	36°14’	1300	1	9	Long, fat tail	Moist lowlands
	Kido	Metekel	10°71’	36°19’	1300	3	7	Long, fat tail	Moist lowland
	Doyogena	Kembata Tembara	7°21’	37°47’	2324	1	9	Long, fat tail	Cool, wet highland
	ShubiGemo	East Shewa	8°80’	38°51	1600	3	7	Long, fat tail	Cool, wet highland
	Loya	Sidama	6°29’	38°24’	1900	3	7	Long, fat tail	Cool, wet highland
Sudan	Hammari	North Kurdufan	13°09’	29°22’	620	3	7	Long, thin tail	Arid lowland
	Kabashi	North Kurdufan	13°09’	29°22’	620	5	5	Long, thin tail	Arid lowland
Libya	Barbary	Mid-costal Region	32°31’	15°09’	< 100	4	8	Pendulous fat tail	Arid lowland
						42	**108**		
		Total				**150**			

Library preparation and DNA sequencing

Ethiopian and Libyan sheep samples

Whole-genome sequences of 118 Ethiopian and 12 Libyan sheep were generated by Edinburgh Genomics as following:

Genomic DNA (gDNA) samples were evaluated for quantity and quality using an AATI Fragment Analyzer and the DNF-487 Standard Sensitivity Genomic DNA Analysis Kit. The AATI ProSize 2.0 software provides a quantification and a quality (integrity) score for each gDNA sample. Samples found to have a quality score < 7 and little or no high molecular weight values failed the sample quality threshold and replacement samples were requested. Based on the quantification results, gDNA samples were pre-normalised to fall within the acceptable range of the Illumina SeqLab TruSeq Nano library preparation using the Hamilton MicroLab STAR.

Next Generation Sequencing libraries were constructed using Illumina SeqLab specific TruSeq Nano High Throughput library preparation kit in conjunction with the Hamilton MicroLab STAR and Clarity LIMS X Edition. The gDNA samples were normalised to the concentration and volume required for the Illumina TruSeq Nano library preparation kit, then sheared to 450 bp mean insert sizes using a Covaris LE220 focused-ultrasonicator. The inserts were blunt-end ligated, A-tailed, size selected, and TruSeq adapters added and enriched using 8 cycles of PCR amplification.

Following library preparation, the libraries were evaluated for mean peak size and quantity using the Caliper GX Touch with a HT DNA 1k/12K/HI SENS LabChip and HT DNA HI SENS Reagent Kit. The libraries were normalised to 5 nM using the GX touch and the actual concentration was established using a Roche LightCycler 480 and a Kapa Illumina Library Quantification kit and Standards. The libraries were then normalised, denatured, and pooled in multiples of eight for clustering and sequencing using a Hamilton MicroLab STAR with Genologics Clarity LIMS X Edition. Libraries were clustered onto HiSeqX Flow cell v2.5 on cBot2s and the clustered flow cell were transferred to a HiSeq2000 for sequencing using a HiSeqX Ten Reagent kit v2.5.

Sudanese sheep

Twenty Sudanese thin-tailed sheep samples representing two populations (10 Hammari and 10 Kabashi) were sampled in North Kurdufan (**Figure 3.1**). About 10 ml of whole blood was collected from each animal into EDTA coated vacutainer tubes following standard procedures, under the supervision of a qualified Veterinarian. The DNA was extracted from whole blood using Qiagen DNeasy® extraction kit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol. DNA quality and quantity were evaluated using a Nano-Drop spectrophotometer (NanoDrop Technologies, USA) and gel electrophoresis. Whole-genome resequencing was performed on the Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA, USA).

Demultiplexing and trimming

Demultiplexing was performed using bcl2fastq (2.17.1.14), allowing 1 mismatch when assigning reads to barcodes. Adapters (Read1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA, Read2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT) were trimmed during demultiplexing. After trimming the two adapters and applying the demultiplexing process, two compressed FASTQ files 'fastq.gz' for each sample were obtained. The total coverage depth was ~30x for the 130 long and short fat-tailed sheep from Ethiopia and Libya and ~10x for the 20 thin-tailed sheep from Sudan.

Clean reads for the 150 samples were mapped to the *Ovis aries* (sheep) reference genome (Oar_v3.1; downloaded from Ensembl database release 95 using the Burrows-Wheeler Alignment tool (BWA) version bwa-0.7.16a (Li and Durbin, 2010). The alignment files generated in SAM format were converted to BAM format using SAMtools v.1.19 (Li et al., 2009).

Variant calling

We applied the Best Practices pre-processing Genome Analysis Toolkit (GATK) workflow from the Broad Institute (https://software.broadinstitute.org/gatk/best-practices) to perform variant discovery. The alignment files were sorted by coordinate and indexed using SAMTools. The Picard suite of tools v.1.119 (http://sourceforge.net/projects/picard) was used to mark duplicate reads. Base Quality Score Recalibration (BQSR) was performed using

BaseRecalibrator from the GATK v.3.7 (McKenna et al., 2010) with the "knownSites" set to the *Ovis aries* dbSNP from Ensembl release 88 (ftp://ftp.ensembl.org/pub/release-88/variation/vcf/ovis_aries/Ovis_aries.vcf.gz).

Variants were called using HaplotypCaller from GATK (with –ERC GVCF and -stand_call_conf set to 30) followed by GenotypeGVCFs to perform joint genotyping and generate VCF files containing SNPs and Indels for all the samples. Variant quality score recalibration was performed using VariantRecalibrator with the following parameters: -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR -an DP -an InbreedingCoeff and the resource parameters set as: "resource: eva, known=false, training=true, truth=true, prior=15.0 evaFile and resource: dbsnp, known=true, training=false, truth=false, prior=2.0 dbsnpFile, where "evaFile" and "dbsnpFile" refer to the set of high quality variants from the European Variation Archive (ftp://ftp.ebi.ac.uk/pub/databases/eva/PRJEB14685/eva_normalised_files/*.filtered_intersect.vcf.gz) and the dbSNP VCF file from Ensembl release 88 (ftp://ftp.ensembl.org/pub/release-88/variation/vcf/ovis_aries/Ovis_aries.vcf.gz) for *Ovis aries*".

Estimation of genetic diversity

For estimating genetic diversity, VCFtools v.0.1.15 (Danecek et al., 2011) was used to calculate inbreeding coefficients (*F*), observed (*Ho*) and expected (*He*) heterozygosity, nucleotide diversity (*Pi*) and runs of homozygosity (RoH) for each population. The *Pi* was estimated for each population using 100 kb non-overlapping sliding windows using "--window-pi", while the "--het" option in VCFtools was used to estimate the inbreeding coefficient, and the observed and expected heterozygosity for each population. Runs of homozygosity (RoHs), which are genomic stretches of a diploid genome that show identical alleles on both chromosomes, were estimated using BCFtools/RoH, the extension to the BCFtools software package. It detects regions of autozygosity for each population in sequencing data using a hidden Markov model (HMM). Briefly, the HMM is applied to genetic variation data (in VCF format) for the population containing the sample, with positions in the chain corresponding to segregating sites in the population and using either genotype calls or genotype likelihoods. The two hidden states represent extended homozygosity (H) and non-homozygosity (N) within the sample. RR represents genotypes for a homozygous site matching the reference, RA for a heterozygous site and AA for a homozygous alternate (non-reference) site. Thus, H tracts can

only include RR and AA sites, whereas N tracts can include sites of any genotype (Narasimhan et al., 2016).

We then estimated the genomic inbreeding coefficient (F_{RoH}) for each population using the following equation (McQuillan et al., 2008):

$$\boldsymbol{F_{ROH}} = \frac{\boldsymbol{L_{RoH}}}{\boldsymbol{L_{AUTO}}} \qquad (3.1)$$

Where L_{ROH} is the total length of *RoH* of each individual in the genome and L_{AUTO} is the length of the autosomal genome of sheep (set to ~2600 Mb (ISGC et al., 2010)).

Population Genetic Analyses

Genetic structure of the 15 sheep populations was inferred using principal component analysis (PCA) and estimation of the proportion of ancestry (admixture). All autosomal bi-allelic SNPs with a minimum minor allele frequency of 0.1 (MAF > 0.1) were selected using Vcftools v.0.1.15 and data converted into plink format (Purcell et al., 2007). For admixture analysis, a LD pruning using default options (50 SNPs step 5 SNPs, r2 0.5) was performed using Plink v.1.19.

Plink v1.19 was also used to perform PCA analysis and the first two eigenvectors were plotted. The ADMIXTURE v.1.3.0 program (Alexander et al., 2009) was used to implement the block relaxation algorithm and to infer population structure with Kinship (*K*) set from 2 to 6. As previously applied in **Chapter 2** (Ahbara et al., 2019), a 5-fold cross-validation procedure was used to determine the optimal number of ancestral genomes (*K*) and proportions of shared genome ancestry among the studied populations. PCA and Admixture plots were visualized using GENESIS (Buchmann and Hazelhurst, 2014).

We further evaluated the pattern of splits and gene flow among the studied populations, with the maximum likelihood tree-based approach implemented in TreeMix (Pickrell and Pritchard, 2012). The optimal number of migration events (m) was identified using the same approach employed in **Chapter 2** (Pickrell and Pritchard, 2012; Ahbara et al., 2019). The models were visualized using the R script provided in Treemix.

The LD parameter (r^2) was estimated using the --r^2 option in PLINK as described in previous studies (Liu et al., 2017; Berihulay et al., 2019). It denotes the squared correlation coefficient

between two alleles at a locus that differs from the expectation of independence and random sampling (VanLiere and Rosenberg, 2008). The estimation of r^2 was performed based on the equation (Barbato et al., 2015):

$$r^2_{X,Y} = \frac{[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad \textbf{(3.2)}$$

Where $\bar{X}$ and $\bar{Y}$ are the mean of genotype frequencies for two separate loci (X, Y) respectively, while X_i denotes the genotype of individual i at the first locus (X) and $\bar{Y}$ refers to the genotype of individual i at the second locus. The LD was estimated genome-wide across each of the 26 autosomes within each group of sheep based on the subgroups generated from the PCA analysis. We sorted and binned the values of LD (r^2) in 50 kb to 10 Mb inter-SNP distances. The LD decay was finally plotted using R software.

Based on the estimated values of linkage disequilibrium, an estimation of the effective population size (N_e) was performed for each chromosome and generation (up to 1000 generations ago) within each sheep group. N_e values were estimated following Sved, (1971) as:

$$N_e = \frac{1}{4c}\left(\frac{1}{E(r^2)} - 1\right) \quad \textbf{(3.3)}$$

Where N_e denotes the effective population size, c the genetic distance in centimorgans (considering that 1 *cm* equals 1 Mb), and $E(r^2)$ represents the expected LD (r^2) for a distance of c morgans. Time intervals indicating the generation number in the past (t) was estimated based on $t = 1/2c$ as per previous study (Hayes et al., 2003).

Results

Variant discovery and annotation

Following the alignment of all the generated sequences to the *Ovis aries* reference genome, we achieved a 30x average sequencing depth for Ethiopian and Libyan sheep, and a 10x depth coverage for Sudanese sheep. To minimize false positive variant calls, strict filtration criteria were applied on the variants detected by the variant calling process (see Materials and Methods section). The total number of SNPs detected in all sheep populations (Ethiopian, Sudanese and Libyan sheep) ranged between 25.4 and 37.5 million. The lowest number was detected in Segentu and the highest in Kefis followed by Libyan Barbary sheep. Nearly a similar number of SNPs (29.9 million) was observed in the Sudanese sheep populations (**Table 3.2**). A nearly similar trend was evident in InDel numbers with the highest (4.1 million) being observed in Barbary and Kefis and the lowest (3.4 million) in Hammari sheep populations (**Table 3.3**). Based on the comparison with the *Ovis aries* dbSNP, about 6% of the SNPs and InDels in each sheep population are not present in the SNP database suggesting they could be novel, which may indicate to high reliability of our annotation **(Table 3.2; Table 3.3)**.

Table 3.2 SNP statistics for each sheep population

Population	Total number of SNPs	Average per sample	Het/Hom	dbSNP (%)
Kefis	37,531,090	12,561,119	2.15	93.7
Segentu	25,436,133	12,472,311	1.65	93.5
Adane	26,713,625	12,574,438	2.13	93.5
Arabo	28,466,085	12,617,954	2.03	93.5
Gafera-Washera	34,128,747	12,398,645	1.98	93.7
Molale-Menz	34,547,334	12,401,271	2.10	93.7
Bonga	34,274,566	12,152,710	1.96	93.7
Gesses	34,076,190	12,400,542	1.87	93.7
Kido	33,711,605	12,334,147	1.89	93.6
Doyogena	35,184,652	12,256,756	2.07	93.8
ShubiGemo	33,804,134	12,376,058	1.98	93.6
Loya	33,822,572	12,172,853	1.98	93.6
Hammari	27,307,188	11,875,632	2.37	93.5
Kabashi	27,347,392	11,959,310	2.34	93.5
Barbary	36,190,249	12,695,368	2.31	93.7

Table 3.3 InDel statistics for each sheep population

Population	Total number of InDels	No of Insertions	No of Deletions	dbSNP (%)	Average per sample
Kefis	4,102,070	1,782,561	2,319,509	93.9	1,388,390
Segentu	4,097,280	1,779,570	2,317,710	93.9	1,374,066
Adane	4,099,244	1,780,776	2,318,468	93.9	1,384,721
Arabo	4,100,430	1,781,492	2,318,938	94.1	1,390,390
Gafera-Washera	4,100,616	1,781,596	2,319,020	94.0	1,367,159
Molale-Menz	4,100,955	1,781,867	2,319,088	94.0	1,365,234
Bonga	4,099,254	1,780,829	2,318,425	93.9	1,341,735
Gesses	4,099,990	1,781,255	2,318,735	93.9	1,366,817
Kido	4,099,819	1,781,183	2,318,636	93.9	1,358,046
Doyogena	4,100,349	1,781,507	2,318,842	94.0	1,350,591
ShubiGemo	4,100,213	1,781,428	2,318,785	93.9	1,362,645
Loya	4,099,535	1,780,952	2,318,583	93.9	1,340,999
Hammari	3,496,053	1,657,173	1,838,880	93.7	1,622,816
Kabashi	3,650,551	1,730,407	1,920,144	93.6	1,598,529
Barbary	4,103,108	1,783,169	2,319,939	94.0	1,403,320

SNP annotation of the 34,857,882 unique variants (data from the 150 sheep pooled together) is presented in **Table 3.4** and **Figure 3.2**. Around 36.8% and 55.9% occur in introns and intergenic regions, respectively, while exonic variants are represented by only 0.79% of the variants. These exonic variants are represented by 275,998 SNPs including 140,120 missense, 131,158 synonymous, 3,374 stop-gains, 681 stop-losses, 522 start-losses and 143 stop-retained variants.

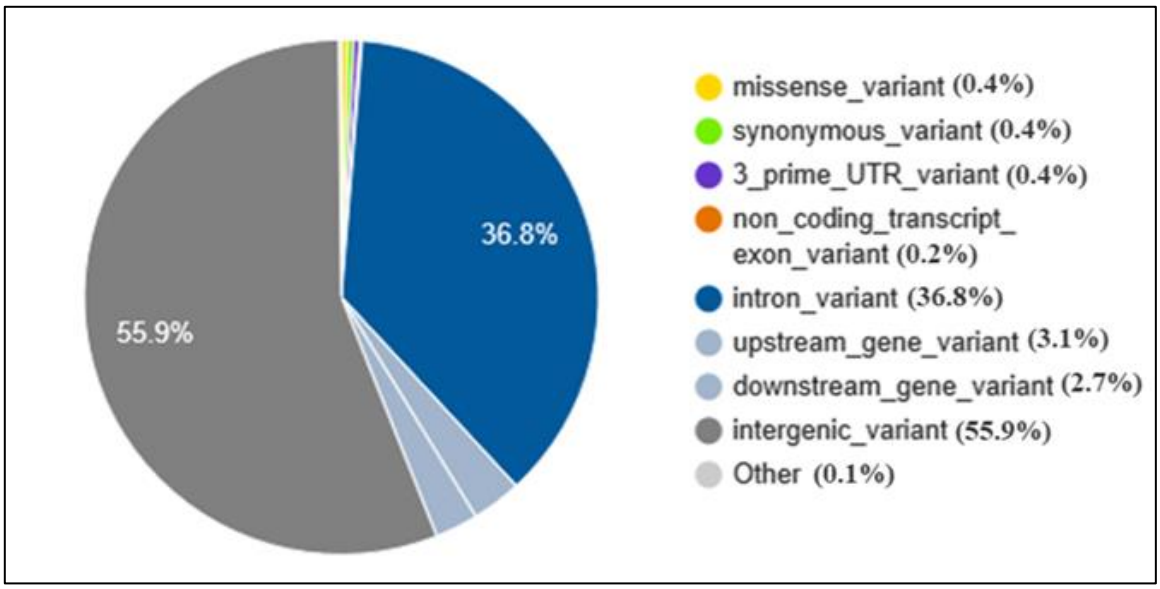


Figure 3.2 Distribution of SNPs following annotation categories (150 sheep samples)

Table 3.4 Number of SNPs in each category (150 sheep samples)

Category		Number of SNPs
Total		34,857,882
Upstream		1,093,873
Exonic	Missense	140,120
	Stop-gain	3,374
	Stop-loss	681
	Start-loss	522
	Stop retained	143
	Synonymous	131,158
Intronic		12,829,358
Coding sequence		147
5_prime UTR		22,392
3_prime UTR		125,545
Non coding transcript		24
Non coding transcript exon		56,561
Mature miRNA		247
Splicing		25,838
Splice donor		1,436
Incomplete terminal codon		9
Downstream		945,893
Intergenic		19,479,560

Genetic Diversity Estimates

The average estimates of *F, RoH, Pi, Ho* and *He*, indexes of within-breed genetic diversity, are shown in **Table 3.5**. The average inbreeding coefficients were generally low (less than 1%) with the exceptions of Hammari, Kabashi and Molale breeds with estimates of 8.3%, 7.3% and 1.1%, respectively, while the lowest level of inbreeding was observed in Libyan Barbary sheep (-0.006) (**Table 3.5**). These populations also recorded the highest (Hammari and Kabashi) and lowest (Barbary) *RoH* mean sizes with 0.034, 0.032 and 0.018 Mb, respectively (Table 3.5). The distribution of RoH within a range of 1.4 Mb for each breed are shown in **Appendix Figure 5.3.** While the Libyan Barbary sheep recorded the lowest (0.1771), the highest F_{RoH} was observed within Bonga (0.2797) and Loya (0.2745) sheep populations (**Appendix Figure 5.4).** The highest average nucleotide diversity was found in the Libyan Barbary (0.0032) and the lowest in Bonga and Loya (0.0027) (**Table 3.5**). The lowest values of *He* and *Ho* were recorded in Kido and the highest in Segentu and Adane, respectively.

Table 3.5 Estimates of genetic diversity parameters for each of the 15 populations analyzed.

Population	Inbreeding coefficient (*F*)	*RoHs Mean Size (Mb)*	F_{RoH} (*Mean*)	Nucleotide diversity (*Pi*)	Observed heterozygosity (H_o)	Expected heterozygosity (H_e)
Kefis	-0.020	0.0210	0.2056	0.0030	0.318	0.311
Segentu	-0.023	0.0221	0.2231	0.0029	0.377	0.368
Adane	-0.017	0.0239	0.2318	0.0030	0.357	0.351
Arabo	-0.005	0.0218	0.2096	0.0030	0.337	0.335
Gafera	-0.019	0.0239	0.2335	0.0029	0.335	0.329
Molale	0.011	0.0246	0.2377	0.0029	0.323	0.326
Bonga	-0.011	0.0262	0.2797	0.0027	0.337	0.333
Gesses	-0.035	0.0238	0.2360	0.0028	0.348	0.402
Kido	-0.025	0.0251	0.2495	0.0028	0.273	0.263
Doyogena	0.009	0.0254	0.2517	0.0028	0.326	0.396
ShubiGemo	-0.015	0.0239	0.2396	0.0028	0.332	0.327
Loya	-0.004	0.0258	0.2745	0.0027	0.335	0.334
Hammari	0.083	0.0342	0.2094	0.0029	0.295	0.321
Kabashi	0.073	0.0325	0.2042	0.0029	0.297	0.321
Barbary	-0.006	0.0182	0.1771	0.0032	0.303	0.301

Population Structure Analysis

We performed the PCA to illustrate the distribution of genetic variation among the African sheep populations (Figure 3.3). Around 17 million SNPs were retained for PCA analysis following the removal of SNPs with MAF < 0.1. Ethiopian sheep separated into two clusters (ET_G1 (fat rump) and ET_G2 (long fat tail)). The two Ethiopian short fat-tail sheep also cluster separately. While the Molale-Menz sheep cluster with ET_G1, the Gafera-Washera group with ET_G2. The Sudanese thin tail and the Libyan Barbary sheep also cluster separately and they are separated from the Ethiopian sheep.

Using about 2.4 million SNPs that are in linkage equilibrium (obtained after LD pruning of the 17 million SNPs used in PCA), admixture analysis was performed. It separates the populations following their geographic origins (**Figure 3.4a, b**). The cross-validation error registered the lowest value at $K = 3$ suggesting it as the optimal number of ancestral populations explaining the variation in the dataset **(Figure 3.4a,b**). At this level, the Segentu, Kefis, Arabo, Adane and Molale Ethiopian populations share predominantly one ancestry (ET_G1), while Bonga, Gesses, Kido Loya, Doyogena, ShubiGemo and Gafera Ethiopian populations share another ancestry (ET_G2). The Libyan (LD, Barbary) share a common third ancestry with the Sudanese (SD, Hammari, Kabashi) sheep breeds.

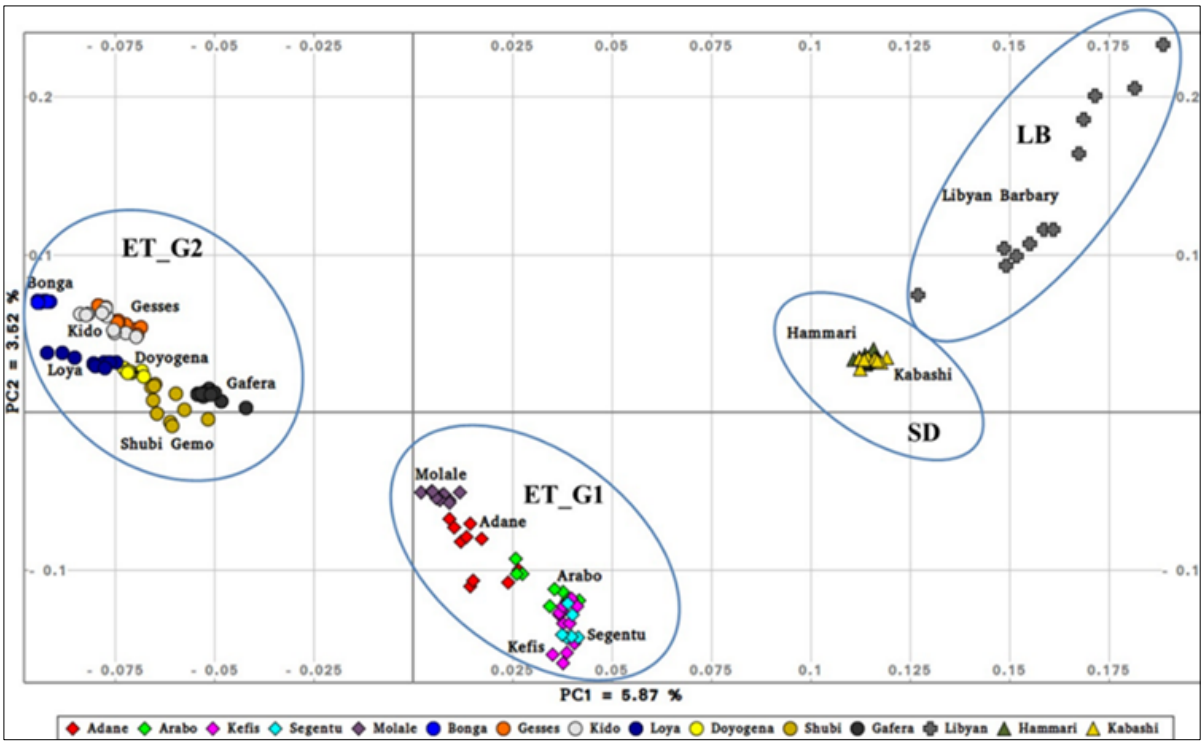


Figure 3.3 Distribution of genetic variation among the studied sheep populations (PC1 and PC2)

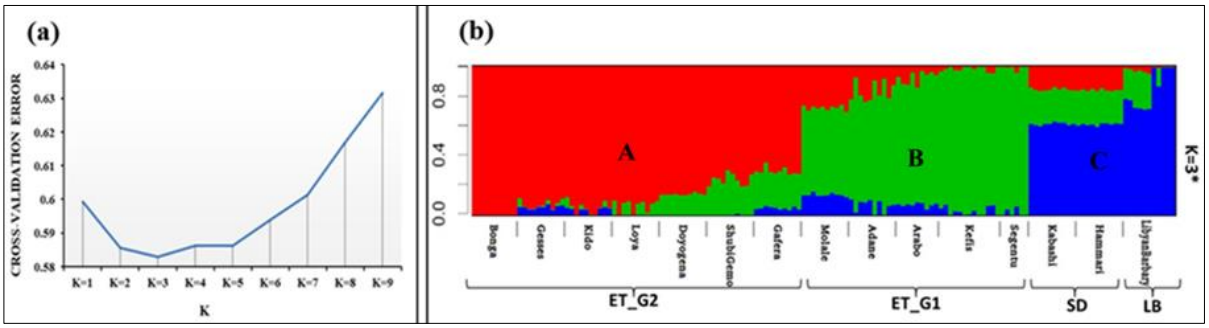


Figure 3.4 Admixture analysis of the studied populations (a) Cross-validation plot (b) Admixture plot

TreeMix reveals possibilities of gene-flow between the populations **(Figure 3.5)**. Following Pickrell and Pritchard (2012), the *f* index which represents the fraction of the variance in the sample covariance matrix ($\widehat{\mathbf{W}}$) that is accounted for by the model covariance matrix (**W**) was used to identify the information contribution of each migration vector that was added to the tree. Up to 15, equal to the number of studied populations, possible migration vertices were computed. The first seven migration edges (gene flow events) accounted for more than 90% of the total model significance explained by the *f* statistic, with the first migration edge having an *f* value of 0.91. We therefore chose m = 7 as the best predictive value for the migration model (Ahbara et al., 2019). Vectors from 8 to 15 resulted in only small incremental changes in the *f* value **(Figure 3.5a and 3.5b)**.

Based on the results of admixture analysis, population split was inferred by considering the Sudanese populations as the outgroup. The strongest migration events were observed between Ethiopian sheep populations (Adane, Molale, Arabo and Kefis) from the eastern part of the country and populations (Bonga, Loya and ShubiGemo) from the centre to the southern parts of the country. Furthermore, a gene flow links Bonga sheep with Gesses and Kido, at the West of the country. Similarly, Gafera population connects with Molale and Adane sheep, which have a similar type of tails. The later populations are also sharing a migration event with the Arabo fat-rump population. Finally, migrations events are supported between Segentu, Molale and the ShubiGemo long fat-tailed sheep population **(Figure 3.5a)**.

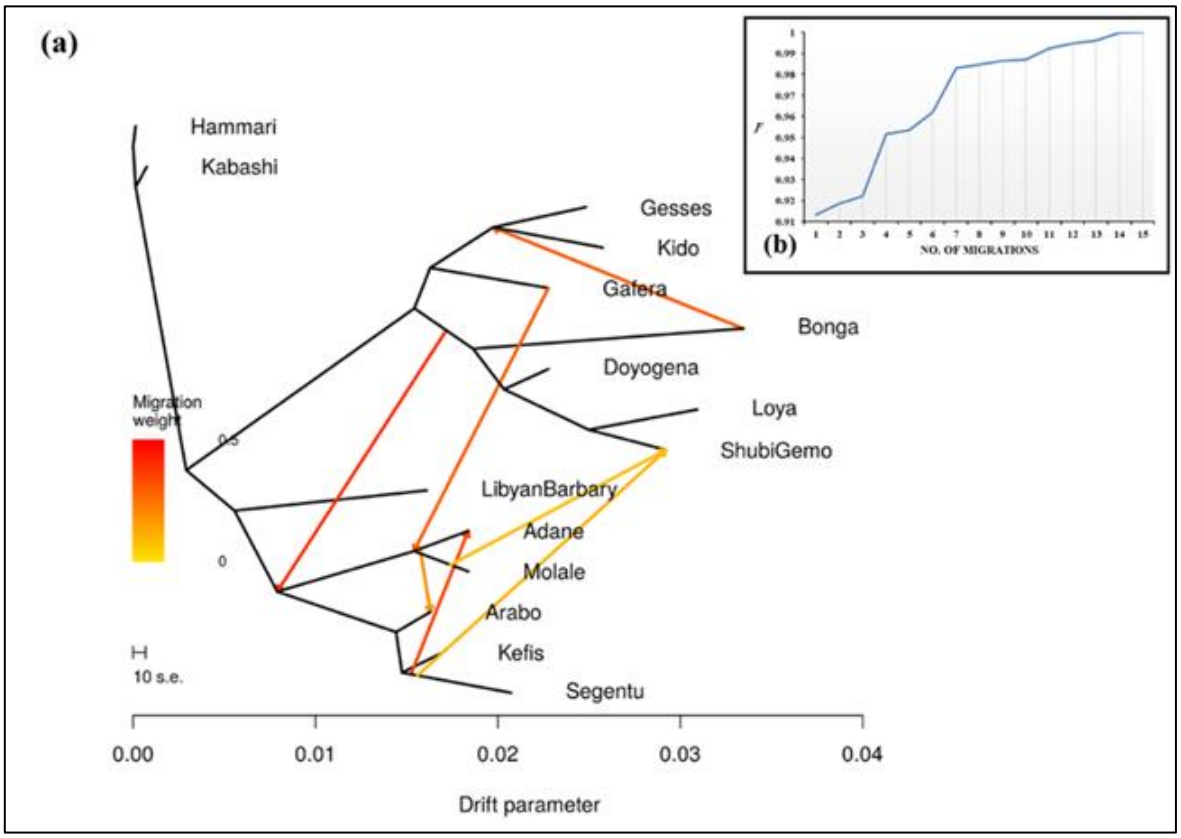


Figure 3.5 Tree-mix plot. (a) Phylogenetic network inferred by Tree-mix of the relationships between Ethiopian, Libyan and Sudanese sheep populations. The first seven migration edges between populations are shown with arrows pointing in the direction toward the recipient group and colored according to the percentage of ancestry received from the donor. (b) 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 modeled migration events.

Demographic History

The results of linkage disequilibrium (LD) analysis are shown in **Table 3.6**. The estimated average values of r^2 between adjacent variants for the 26 autosomes were the lowest for ET_G2

(0.063) and the highest for the LB sheep (0.130). The average values of r^2 decreased rapidly with genetic distance up to 100 kb, where it continued to decline but at a much slower space for all the sheep groups (**Figure 3.6**).

The inferred trends in *Ne* for the study populations are shown in **Figure 3.7 and Table 3.7**. In general, and except for LB, there is a general increase in *Ne* for all the population groups until 300 generations ago. This is followed by a gradual decline in *N*e till the present time. LB on the other hand, shows a gradual decline in *N*e until the present time.

Table 3.6 Summary of average linkage disequilibrium (r^2) values at different distances (Mb) across the 26 autosomes for the sheep groups.

	ET_G1		ET_G2		SD		LB	
CHR	r^2	**Distance (Mb)**	r^2	**Distance (Mb)**	r^2	**Distance (Mb)**	r^2	**Distance (Mb)**
1	0.064	0.224	0.064	0.224	0.087	0.224	0.133	0.224
2	0.067	0.230	0.065	0.227	0.092	0.227	0.132	0.224
3	0.064	0.235	0.062	0.233	0.089	0.238	0.126	0.234
4	0.065	0.212	0.062	0.212	0.087	0.211	0.129	0.212
5	0.062	0.211	0.060	0.211	0.085	0.210	0.130	0.212
6	0.070	0.197	0.068	0.209	0.092	0.198	0.129	0.203
7	0.059	0.215	0.0578	0.206	0.082	0.214	0.125	0.212
8	0.065	0.204	0.060	0.219	0.087	0.202	0.123	0.207
9	0.069	0.195	0.064	0.208	0.090	0.191	0.126	0.198
10	0.083	0.197	0.083	0.195	0.101	0.191	0.143	0.189
11	0.062	0.213	0.064	0.215	0.086	0.218	0.133	0.218
12	0.067	0.195	0.062	0.201	0.087	0.202	0.127	0.205
13	0.066	0.215	0.060	0.209	0.085	0.215	0.129	0.217
14	0.059	0.223	0.053	0.206	0.085	0.210	0.125	0.211
15	0.068	0.192	0.065	0.199	0.090	0.198	0.143	0.202
16	0.065	0.189	0.063	0.188	0.085	0.187	0.124	0.191
17	0.064	0.194	0.064	0.188	0.087	0.203	0.125	0.200
18	0.062	0.201	0.062	0.199	0.083	0.197	0.137	0.199
19	0.062	0.206	0.058	0.196	0.089	0.201	0.128	0.214
20	0.064	0.190	0.062	0.190	0.087	0.201	0.132	0.190
21	0.069	0.177	0.065	0.192	0.087	0.182	0.129	0.185
22	0.069	0.186	0.064	0.193	0.089	0.175	0.134	0.181
23	0.065	0.188	0.065	0.189	0.085	0.185	0.129	0.183
24	0.056	0.187	0.056	0.197	0.084	0.194	0.123	0.201
25	0.063	0.176	0.059	0.184	0.083	0.173	0.123	0.178
26	0.064	0.179	0.061	0.180	0.084	0.190	0.132	0.186
Mean	0.065	0.201	0.063	0.203	0.087	0.201	0.130	0.203

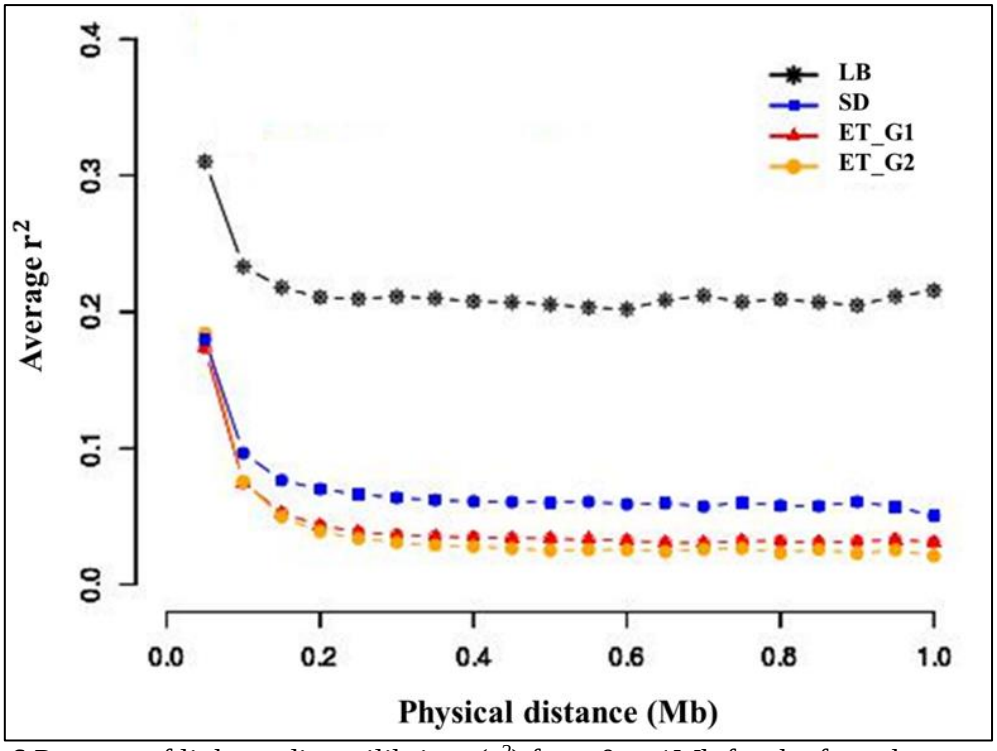


Figure 3.6 Patterns of linkage disequilibrium (r^2) from 0 to 1Mb for the four sheep groups

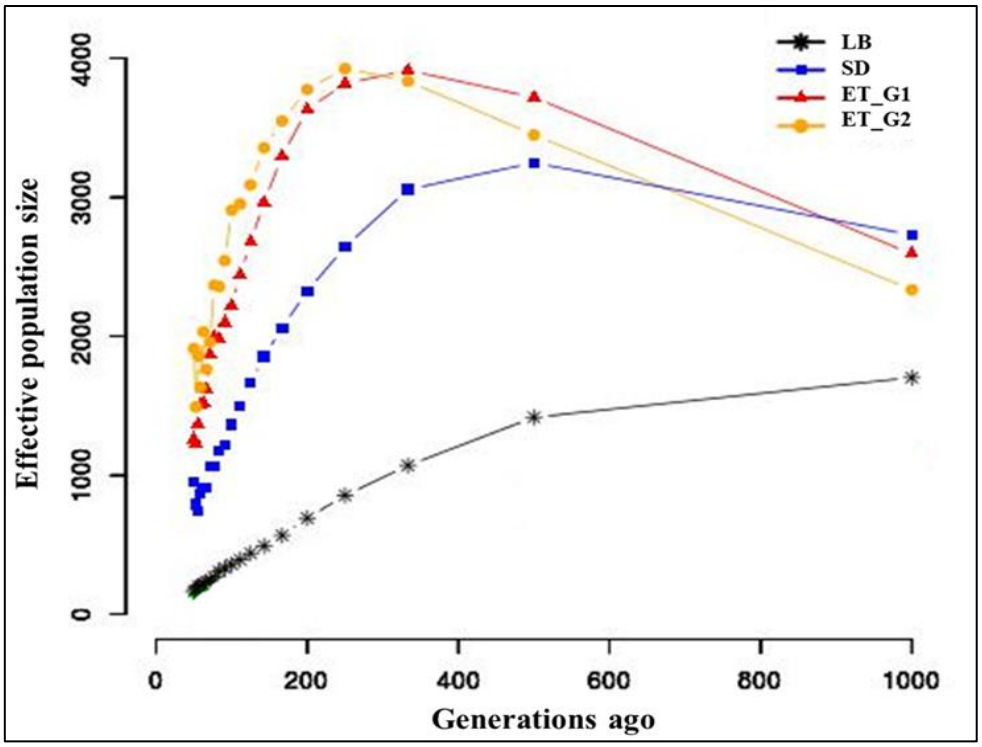


Figure 3.7 Average estimated effective population size (*Ne*) in the four sheep groups over the past 1000 generations

Table 3.7 The average autosomal effective population sizes over time for different sheep groups

Sheep Group	Generations ago				
	50	**200**	**350**	**500**	**1000**
ET_G1	1300	3550	3850	3750	2380
ET_G2	1500	3650	3850	3500	2680
SD	700	2300	3200	3242	2720
LB	350	700	1100	1395	1712

Discussion

In this chapter, we report whole-genome sequence sequences analysis of 150 African indigenous sheep from the North of the continent (Libyan Barbary, 12 samples), the Nile River Basin (Sudan, 20 samples) and Eastern Africa (Ethiopia, 118 samples) comparing the sequence reads to the sheep reference genome assembly Oar_v3.1. We obtained 34.85 million high-quality variants with an average of 93.62% of these SNPs validated in the sheep dbSNP database. Only 0.79% of the SNPs were located in an exonic region with the majority of the SNPs (55.9%) found in the intergenic regions (**Figure 3.3**). These ratios are consistent with the previous study on Chinese native sheep (Yang et al., 2016).

Among the sheep breeds analysed we observe a difference in variant number, likely to differences in sample numbers in each population. For example, the largest number of SNPs are found in Ethiopian Kefis (37.53 million SNPs) and Libyan Barbary (36.19 million SNPs) populations represented by 12 samples each, while the lowest number of SNPs (25.43 million SNPs) was found in Segentu sheep represented by only five individuals. The reduced number of variants in Kabashi (27.34 million SNPs) and Hammari (27.30 million SNPs) sheep from Sudan is possibly a consequence of the lower genome coverage for these two breeds; 10x against 30x coverage in Sudan and the other populations, respectively (**Table 3.2**).

As with the diversity estimates, the lowest inbreeding level (*F*) also translated by the highest runs of homozygosity (*RoHs and* F_{RoH}), nucleotide diversity (*Pi*) and high levels of observed ratios of heterozygous SNPs to homozygous non-reference SNPs were observed in the Libyan Barbary (**Table 3.2, Table 3.5 and Appendix Figure 5.3 and 5.4**). Comparatively, a high level of inbreeding as revealed by high ranges of RoH was observed in the Sudanese sheep and East African sheep represented by ET_G1 and ET_G2 groups of Ethiopian sheep indicating possibly past bottleneck or consanguineous mating. While the Barbary sheep recorded consistent estimation of both *F*, *RoH* and F_{RoH}, a contradicted estimation was observed within Hammari and Kabashi Sudanese sheep. The later populations recorded the highest *F* and *RoHs* size, while the highest F_{RoH} was observed within the Ethiopian Bonga and Loya sheep populations. This contradicted pattern, however, could be attributed to the different depth of genome sequencing coverage, which reflected on the fewer variants called in Hammari and Kabashi compared to the other populations (**Table 3.2 and 3.3**).

We identified four distinct groups of sheep by PC and admixture analysis (**Figure 3.3).** The first PC (5.87%) separated the Libyan Barbary (LB) from the sub-Saharan Africa sheep represented by the Sudanese (SD) and East African sheep from Ethiopia (ET_G1 and ET_G2) supporting the two distinct entry points of sheep into the continent, the Isthmus of Suez and the straits of Bab el Mandeb (Epstein, 1971; Muigai and Hanotte, 2013). Furthermore, the sub-grouping of Ethiopian sheep into fat-rump (ET_G1) and long fat-tail (ET_G2) sheep may indicate subsequent migration events for fat tail sheep *via* the same Horn of Africa routes. The later scenario is supported by PC2, which separates most of the long fat- and thin-tail sheep groups (ET_G2, SD and LB) from the fat-rump sheep group (ET_G1) (Epstein, 1954; Ryder, 1984; Marshal, 2000). The possible routes of introduction and spread of sheep pastoralism into Africa are presented in **Chapter 1** (**Figure 1.2**).

The sub-grouping of SD and LB sheep is indicative of gene flow between these two groups. These results are in agreement with the admixture analysis (**Figure 3.4a, b**). Three ancestral backgrounds are observed at the optimal number (*K*=3) of clusters. All of the three ancestries are represented in the Sudanese thin-tail sheep supporting this sheep as the most diverse and possibly ancient African sheep (Wilson, 1991; Muigai and Hanotte, 2013). Libyan (LB) and Sudanese (SD) sheep groups share the same ancestry supporting a shared route of introduction through the North-East side of the continent. Although they are all long fat-tail, the Ethiopian (ET_G2) and Libyan (LB) long fat-tail show different ancestral backgrounds, providing evidence of different introduction origins.

Corresponding to PCA results, admixture also clusters the Ethiopian sheep into two groups, ET_G1 and ET_G2, which support different routes of arrival and dispersal for the two groups. However, three populations (Doyogena, ShubiGemo, Gafera) from the ET_G2 group share 15 to 20% of their ancestry with ET_G1. In contrast, the ET-G1 group shares a lower proportion (xx%) of ancestry with ET_G2 represented by Molale, Adane and to some extent Arabo population. This pattern of intermixing could be attributed to the sheep movement guided by interactions between specific human ethnic communities (Gizaw et al., 2007). A similar pattern of admixture is observed between Ethiopian sheep (Molale, Gafera, Gesses, Kido, Adane and Arabo) and Sudanese sheep. This pattern is consistent with our admixture analysis in **Chapter 2** (Ahbara et al., 2019). Similarly, despite a lower number of samples, half of the individuals in the LB sheep group share around 30% of their ancestry with SD and ET_G1 groups. Along with PC sub-grouping, the low level of inbreeding and elevated nucleotide diversity in LB,

indicate possible gene flow between Libyan (Barbary) and Sudanese (Hammari and Kabashi) sheep populations.

The possible cause of such intermixing is the exchange of livestock between nomadic Arabs in the southern region of Libya with the Fulani tribes in the North-West border of Sudan (Mufarrih, 1991). Another possible reason could be the loss in monetary value of animal fat in recent decades due to the increased demand for low animal fat consumption to avoid health risks (Moradi et al., 2012; Moioli et al., 2015), which led Libyan sheep farmers to randomly cross their fat-tail sheep with the thin-tail types. This random crossing could erode the Barbary genepool, which represents around 95% of the total sheep population in Libya (Abdulkarim, 2015). Similar genetic dilution has been observed in two sheep breeds, Berber and Rembi, from the neighbouring, Algeria because of uncontrolled mating practices with the Ouled-Djellal breed (Gaouar et al., 2017). However, to the best of our knowledge, there are no genetic studies that have addressed, the current state of Libyan sheep diversity and, hence, further genome analysis is here crucial.

The maximum likelihood assessment considering the seven most optimal admixture events in our study populations has confirmed several aspects detected by admixture and PC analysis (**Figure 3.5**). Briefly, the short distance between LB and the proposed out-group (SD) supports, on one hand, the shared entry point across the northeast side of the continent, and on the other hand, the high level of diversity and relatively low inbreeding coefficient of Barbary Libyan sheep. The mid-position of the LB indicates different migratory waves. The absence of any migration events between LB and the two groups of Ethiopian sheep supports different origins of introduction and migration into the continent. The most genetically distant Ethiopian sheep populations from the outgroup were Bonga and Loya, which has the lowest (0.0027) nucleotide diversity (Table 3.5). The same level of diversity was observed with the 50K Bead-Chip data reported in **Chapter 2** (Ahbara et al., 2019).

The LD varies among sheep groups, genomic regions, and between pairs of markers (**Table 3.6**). The estimated average values of LD ranged between 0.063 and 0.130 which agrees with previous studies (Kijas et al., 2012; Liu et al., 2017). The observed differences in r^2 among sheep groups and chromosomes might be due to differences in genetic drift, recombination rates, and intensive selection, which results in a lower LD across the genome (Slatkin, 2008; Berihulay et al., 2019). A rapid decline in the average values of r^2 was observed when moving from 10 to 100 kb (**Figure 3.6**). This differs with previous studies that used 50K SNP genotype

data in sheep (Liu et al., 2017; Ahbara et al., 2019), goat (Visser et al., 2016) and cattle (Shin et al., 2013) which showed a decreasing trend from 10 to 200 kb. This could be attributed to the higher resolution of the whole genome coverage of our data.

Effective population size (*Ne*) is one of the most important parameters to assess the loss in genetic variation, determine inbreeding depression and calculate the power of genomic selection (Goddard, 2009). The Libyan sheep show decreasing *Ne* values, while an increase followed by a rapid decline is observed in the other sheep groups (**Figure 3.7** and **Table 3.7**). At around 1000 generations ago, SD sheep shows the highest Ne (~2720), while the lowest (~1712) is for the LB group, supporting again that the SD sheep most likely represents the most ancient of the African sheep studied here (Muigai and Hanotte, 2013) and/or a larger founder population. At 500 generations ago, while the *N*e value of LB decreased slightly, there was a minor but gradual increase in Ne for Sudanese (SD) and the Ethiopian sheep groups, respectively. The possible cause of this population increase could be optimal environmental conditions and resources such as climate, water, food and vegetation in the wider eastern Africa region. These optimal conditions could correspond to the African Humid Period (AHP) in the tropics and sub-Saharan Africa that within ~5.000 to 11.000 BP (Jolly et al., 1998; Skonieczny et al., 2015).

From 350 generations ago to the present time, LB and SD sheep show nearly a similar trend in the decline in *N*e, while that for Ethiopian sheep continued to increase up to 250 generations ago. The slight increase followed by a decline in the SD group could be linked to the presence of Sudanese sheep in the sub-Saharan Africa region at the border of the tropical optimal zone for population growth, where the drought period which subsequently affected the Sahara could have started earlier. From ~250 generations to present time, a sharp and sudden decrease in *N*e is observed for Ethiopian and Sudanese sheep, which possibly indicate a population bottleneck at around that period. The sudden and sharp decline may correspond to a similar incident of the onset of the termination of the humid period (AHP) that possibly occurred between 5 and 6 cal. ka BP, which was very abrupt and occurred within decades to centuries (Demenocal et al., 2000).

The confounding effects of admixture on LD (ALD) might, however, bias the estimates of Ne. Such confounding arises when two populations that have been separated for a long time rejoin. If allele frequencies at the same loci are markedly different, LD can arise between alleles at locus pairs on different chromosomes. ALD between unlinked loci quickly dissipates over a

couple of generations, but LD between linked loci decays more slowly and may be detectable 10 to 20 generations after the initial admixture (Rybicki et al., 2002).

Conclusion

In general, the results from this chapter are consistent with those obtained using the 50 K Bead-Chip data reported in **Chapter 2** and other studies that used whole-genome sequence data in non-African sheep supporting high reliability of our annotation and analysis. Our findings support the archaeological history of the proposed migration events and entry points of sheep into the African continent. The analyses indicate moderate to high levels of diversity, while differences in inbreeding indicate that a possible bottleneck may have occurred in the tropical zone followed by intensive selection, which may have resulted in different tail types in Ethiopian sheep. The Libyan sheep population show high level of genetic variation and low inbreeding values combined with reduced effective population size, which could be attributed to the random mating system followed by local owners. However, future studies including thin-tail sheep from Ethiopia and both thin and/or fat-tail sheep from West, North and South Africa could provide a comprehensive view on the genome diversity of African sheep.

References

Abdulkarim, A. (2015). Small ruminant contribution in meat production in Libya. *Egyptian Journal of Sheep and Goats Sciences*, 10(1), pp. 1-6.

Ahbara, A., Bahbahani, H., Almathen, F., Al Abri, M., Agoub, M. O., Abeba, A., et al. (2019). Genome-wide variation, candidate regions and genes associated with fat deposition and tail morphology in Ethiopian indigenous sheep. *Front. Genet.* 9, 699.

Angeloni, F., Wagemaker, N., Vergeer, P., and Ouborg, J. (2012). Genomic toolboxes for conservation biologists. *Evol. Appl.* 5, 130-143.

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, 109.

Berihulay, H., Islam, R., Jiang, L., and Ma, Y. (2019). Genome-Wide Linkage Disequilibrium and the Extent of Effective Population Sizes in Six Chinese Goat Populations Using a 50K Single Nucleotide Polymorphism Panel. *Animals* 9, 350.

Brehm, A., Khadem, M., Jesus, J., Andrade, P., and Vicente, L. (2001). Lack of congruence between morphometric evolution and genetic differentiation suggests a recent dispersal and local habitat adaptation of the Madeiran lizard Lacerta dugesii. *Genet. Sel. Evol.* 33, 671–685.

Buchmann, R., and Hazelhurst, S. (2014). *Genesis Manual.* Johannesburg: University of the Witwatersrand. Available online at: http://www.bioinf.wits.ac.za/software/genesis/Genesis.pdf.

Cañón, J., Alexandrino, P., Bessa, I., Carleos, C., Carretero, Y., Dunner, S., et al. (2001). Genetic diversity measures of local European beef cattle breeds for conservation purposes. *Genet. Sel. Evol.* 33, 311-332.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R. F., et al. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genet.* 46, 858-865.

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg, L. A., et al. (2010). Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS. Biol.* 8, e1000475.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156-2158.

Das, A., Panitz, F., Gregersen, V. R., Bendixen, C., and Holm, L.-E. (2015). Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. *BMC Genom.* 16, 1043.

Demenocal, P., Ortiz, J., Guilderson, T., Adkins, J., Sarnthein, M., Baker, L., et al. (2000). Abrupt onset and termination of the African Humid Period:: rapid climate responses to gradual insolation forcing. *Quat. Sci. Rev.* 19, 347-361.

Edea, Z., Dessie, T., Dadi, H., Do, K.-T., and Kim, K.-S. (2017). Genetic Diversity and Population Structure of Ethiopian Sheep Populations Revealed by High-Density SNP Markers. *Front. Genet.* 8, 218.

Elbeltagy, A. R., Aboul-Naga, A., Hassen, H., Rischkowsky, B., and Mwacharo, J. M. (2015). Genetic diversity and structure in Egyptian indigenous sheep populations mirror patterns of anthropological interactions. *Small. Rumenent. Res.* 132, 137-142.

Epstein, H. (1954). The fat-tailed sheep of East Africa. *The East African Agricultural Journal* 20, 109-117.

Epstein, H. (1971). "The origin of the domestic animals of Africa," Africana publishing corporation, New York.

Funk, W. C., McKay, J. K., Hohenlohe, P. A., and Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends. Ecol. Evol.* 27, 489-496.

Gaouar, S. B. S., Da Silva, A., Ciani, E., Kdidi, S., Aouissat, M., Dhimi, L., et al. (2015). Admixture and local breed marginalization threaten Algerian sheep diversity. *PLoS ONE* 10, e0122667.

Gaouar, S. B. S., Kdidi, S., Tabet Aouel, N., Aït-Yahia, R., Boushaba, N., Aouissat, M., et al. (2014). Genetic admixture of North-African ovine breeds as revealed by microsatellite loci. *Livestock Research for Rural Development* 26, 118.

Gaouar, S. B. S., Lafri, M., Djaout, A., El-Bouyahiaoui, R., Bouri, A., Bouchatal, A., et al. (2017). Genome-wide analysis highlights genetic dilution in Algerian sheep. *Heredity* 118, 293.

Gizaw, S., Van Arendonk, J. A., Komen, H., Windig, J., and Hanotte, O. (2007). Population structure, genetic variation and morphological diversity in indigenous sheep of Ethiopia. *Anim. Genet*. 38, 621-628.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245-257.

Gornas, N., Weimann, C., El Hussien, A., and Erhardt, G. (2011). Genetic characterization of local Sudanese sheep breeds using DNA markers. *Small Ruminant Res.* 95, 27-33.

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, 635-643.

ISGC, Archibald, A., Cockett, N., Dalrymple, B., Faraut, T., Kijas, J., et al. (2010). The sheep genome reference sequence: a work in progress. *Anim. Genet.* 41, 449-453.

Jolly, D., Prentice, I. C., Bonnefille, R., Ballouche, A., Bengo, M., Brenac, P., et al. (1998). Biome reconstruction from pollen and plant macrofossil data for Africa and the Arabian peninsula at 0 and 6000 years. *J. Biogeogr.* 25, 1007-1027.

Kardos, M., Luikart, G., Bunch, R., Dewey, S., Edwards, W., McWilliam, S., et al. (2015). Whole-genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. *Mol. Ecol.* 24, 5616-5632.

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., and San Cristobal, M. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 10, e1001258

Kijas, J. W., Townley, D., Dalrymple, B. P., Heaton, M. P., Maddox, J. F., McGrath, A., et al. (2009). A genome-wide survey of SNP variation reveals the genetic structure of sheep breeds. *PloS ONE* 4, e4668.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589-595.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078-2079.

Liu, S., He, S., Chen, L., Li, W., Di, J., and Liu, M. (2017). Estimates of linkage disequilibrium and effective population sizes in Chinese Merino (Xinjiang type) sheep by genome-wide SNPs. *Genes Genom.* 39, 733-745.

Marshal, F. (2000). The origins and spread of domestic animals in East Africa. In "The Origins and Development of African Livestock: Archaeology, Genetics, Linguistics and Ethnography" (R. M. Blench and K. C. MacDonald, eds.), pp. 191-218. University College London Press, London.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297-1303.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polašek, O., Tenesa, A., et al. (2008). Runs of Homozygosity in European Populations. *Am. J. Hum. Genet.* 83, 359–372.

Meuwissen, T., and Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185, 623-631.

Moioli, B., Pilla, F., and Ciani, E. (2015). Signatures of selection identify loci associated with fat tail in sheep. *J. Anim. Sci.* 93, 4660–4669.

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

Mufarrih, M. E. (1991). Sudan desert sheep: Their origin, ecology and production potential. *World Anim. Rev.* 66, 23-31.

Muigai, A. W. T., and Hanotte, O. (2013). Then origin of African sheep: Archaeological and genetics perspectives. *Afr. Archaeol. Rev*. 30, 39–50.

Mwacharo, J. M., Kim, E.-S., Elbeltagy, A. R., Aboul-Naga, A. M., Rischkowsky, B. A., and Rothschild, M. F. (2017). Genomic footprints of dryland stress adaptation in Egyptian fat-tail sheep and their divergence from East Africa and western Asia cohorts. *Sci. Rep.* 7, 17647.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32, 1749–1751.

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*. 8, e1002967.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., et al. (2013). Great ape genetic diversity and population history. *Nature* 499, 471–475.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., and Bender, D. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575.

Rybicki, B. A., Iyengar, S. K., Harris, T., Liptak, R., Elston, R. C., Maliarik, M. J., et al. (2002). Prospects of admixture linkage disequilibrium mapping in the African-American genome. *Cytometry: J. Int. Soc. Anal. Cyt.* 47, 63-65.

Ryder, M. L. (1983). *Sheep and Man*. London: Gerald Duckworth & Co. Ltd.

Ryder, M. L. (1984). Sheep. In "Evolution of Domestic Animals" (I. L. Mason, ed.), pp. 63-85. Longman, London.

Sassi-Zaidy, Y. B., Maretto, F., Charfi-Cheikrouha, F., and Cassandro, M. (2014). Genetic diversity, structure, and breed relationships in Tunisian sheep. *Small Ruminant Res.* 119, 52-56.

Scherf, B. D. (2000). "World watch list for domestic animal diversity," Food and Agriculture Organization (FAO), Rome, Italy.

Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* 194, 647-662

Shin, D.-H., Cho, K.-H., Park, K.-D., Lee, H.-J., and Kim, H. (2013). Accurate estimation of effective population size in the Korean dairy cattle based on linkage disequilibrium corrected by genomic relationship matrix. *Asian-Australas J. Anim. Sci.* 26, 1672–1679.

Skonieczny, C., Paillou, P., Bory, A., Bayon, G., Biscara, L., Crosta, X., et al. (2015). African humid periods triggered the reactivation of a large river system in Western Sahara. *Nature Commun.* 6, 8751.

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Rev. Genet.* 9, 477–485.

Sved, J. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* 2, 125-141.

Telford, M. J., and Copley, R. R. (2011). Improving animal phylogenies with genomic data. *Trends. Genet.* 27, 186-195.

VanLiere, J. M., and Rosenberg, N. A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. *Theor. Popul. Biol.* 74, 130-137.

Visser, C., Lashmar, S. F., Van Marle-Köster, E., Poli, M. A., and Allain, D. (2016). Genetic diversity and population structure in South African, French and Argentinian Angora goats from genome-wide SNP data. *PLoS ONE* 11, e0154353.

Wilson, R. T. (1991). "Small ruminant production and the small ruminant genetic resource in tropical Africa," Food & Agriculture Org., Rome, Italy.

Wilson, R. T. (2011). Populations and production of fat-tailed and fat-rumped sheep in the Horn of Africa. *Trop. Anim. Health Prod.* 43, 1419-1425.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* 42, 565–569.

Yang, J., Li, W.-R., Lv, F.-H., He, S.-G., Tian, S.-L., Peng, W.-F., et al. (2016). Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Mol. Biol. Evol.* 33, 2576-2592.

Chapter 4

Indigenous African sheep genomes reveal insights on their environmental adaptations and tail morphology

Abstract

With their relatively small body size, obedient behavior and resilience, domestic sheep often accompanied ancient human migrations and dispersion. Different environmental challenges, including on the African continent, have shaped their adaptive diversity including their tail phenotypes. Here, using whole-genome sequences, we aimed to identify candidate genome regions and genes associated with both environmental challenges and tail morphology variation in African sheep population representative of the different introductions of the species on the continent. Based on our population analysis (**Chapter 3**), our sheep population were initially clustered in four groups: Ethiopian fat-rump (ET_G1), Ethiopian long fat-tail (ET_G2), Libyan long fat-tail (LB) and Sudan thin-tail (SD) sheep. These groups represent sheep reared at sea level (< 50 m) (LB), inhabiting desert area (SD) (< 1000 m) or higher altitude (> 1300 m) (ET_G1 and ET_G2). Following morphological characterization of the caudal vertebrae, these populations were grouped into two groups: long-tailed (ET_G2 and SD) and short fat-tail (ET_G1 and LB) populations. Candidate regions associated with environmental challenges and tail morphology were identified with three selection scan indexes (*ZHp, ZFST* and *XP-EHH*), following the screening of 34 million autosomal SNPs. In Libyan sheep, we identified candidate genes putatively associated with salt-sensitivity hypertension, blood pressure and kidney function (*PLEKHA7*), horn size (*RXFP2*), tail phenotypes, nematode resistance (*PDGFRA*) and fat deposition (*PDS5B* and *VEGFA*). Besides, in Ethiopian sheep candidate genes associated with hypoxia responses (*NF1, EGLN1, BRIP1, VEGFA, TF, PLCG1, RELA, KIT* and *DSC2*) were identified in population living at high altitude. We detected also candidate genes associated with ear size (*MSRB3*), limb formation and body weight (*DIS3L2*) and genes associated with trypanotolerance (*CD84, LETM1, SLBP* and *PDE6B*). A tailored analysis for tail morphology only identified genes associated with fat tail deposition (*HINT2, SPAG8, NPR2, TUBB4A, DMXL2, CNTN4, EVI2B, OMG* and *ALOX12*) and length (*HOXB13*). Our results provide novel insights into sheep genomic adaptations to extreme environments and they illustrate the impact that environmental challenges may have had on the tail morphology of African sheep. They offer a valuable resource of new information for future research on livestock breeding in developing countries in response to climatic changes.

Introduction

Sheep is among the first domesticated species in the Fertile Crescent approximately 11,000 years ago (Zeder, 2008). Since then sheep dispersed worldwide across Europe, Africa and Asia alongside the spread of Neolithic culture and agricultural development (Taberlet et al., 2011). Since their domestication, sheep were selected for production traits desired by mankind such as meat, milk and wool (Ruiz-Larrañaga et al., 2018). With the natural adaptation to new environments, these selection processes have led to phenotypic variation resulting in over 1400 of now recognised sheep breeds (Kijas et al., 2012; FAO, 2015).

Domestication, artificial (human) and natural selection have markedly shaped the behavior, and appearance of most livestock populations (Megens et al., 2008). Uniquely, amongst our mammalian domesticates, some sheep have direct tails to reserve energy with the fat-tail phenotypes (short fat-tail, long fat-tail and fat rump) considered as a trait for sheep to cope with harsh environments such as drought seasons, extreme cold winters and food shortages (Ma et al., 2006; Nejati-Javaremi et al., 2007; Pourlis, 2011; Moradi et al., 2012; Lv et al., 2015). Such phenotype represents an important source of dietary fat for human societies, including in the Islamic world (Kashan et al., 2005) and among the pastoral society of the Silk Road. Thus, fat-tailed sheep are largely dispersed in the Middle East, North Africa as well as Central and East Asia where they represent the main type of sheep breeds (Almeida, 2011). These sheep breeds of different tail types provide an excellent animal model for studies associated with fat deposition and energy storage (Lv et al., 2014; Yang et al., 2016; Yuan et al., 2017), including human health related to lipid metabolism and obesity (Xiao et al., 2010; Kraja et al., 2012).

Several studies have been carried out to unravel how the genome of sheep populations has been shaped by environmental challenges. For instance, analysis of Chinese sheep populations living in contrasting environments, Yang et al. (2016) indicate rapid genomic adaptations to extreme environments (plateau, lowland, high-altitude, desert, highly humid and arid zone). Additionally, the analysis of Tibetan highland sheep has identified candidate genome-wide signatures of selection and genes (e.g. *EPAS1*), driving adaptation to high-altitude hypoxia (Wei et al., 2016). In addition, low-density SNP genotyping data has identified several genome regions in Egyptian fat-tail sheep enriched in genes of relevance to the adaptation to the dry-land environment (Mwacharo et al., 2017).

Several earlier studies using SNP Chip data have identified candidate functional genes and genome regions for fat deposition in a variety of fat-tail sheep breeds. For example, a genome-wide selection scan involving thin- and fat-tail sheep breeds identified three regions (on chromosomes 5, 7 and X) (Moradi et al., 2012). Furthermore, the *BMP2* and *VRTN* genes were identified as the stronger candidate genes accounting for the fat-tail phenotype in a genome-wide scan for selective signatures across two fat-tail and 13 thin-tail sheep breeds (Moioli et al., 2015). More recently, a genome-wide high-density SNP study on three Chinese native sheep breeds with various tail types (large-tailed Han, Altay and Tibetan sheep) detected several copy number variation regions harbouring functional genes associated with fat deposition, such as *PPARA, RXRA* and *KLF11* (Zhu et al., 2016). Yuan et al. (2017) also analysed selection signatures in fat-tailed and thin-tailed sheep and identified a 6.24 Mb candidate region and several candidate genes (i.e. *HOXA11, BMP2, PPP1CC, SP3, SP9, WDR92, PROKR1 and ETAA1*) that may affect fat-tail development.

In relation to the length of the tail, several candidate genes, including *HES7* (Bessho et al., 2001), *PAX1* (Wilm et al., 1998), *T* (Smith, 1999), *WNT3A* (Greco et al., 1996) and Hox genes (Mallo et al., 2010) have been associated with the short-tail phenotype in laboratory mice. In particular, mice homozygous for *Hoxb13* loss-of-function mutations show overgrowth in all major structures derived from the tail bud, including the developing secondary neural tube (SNT), the caudal spinal ganglia, and the caudal vertebrae (Economides et al., 2003). In sheep, our results identified many genes associated with tail morphology and *HOXB13* was among the strongest candidates (Ahbara et al., 2019). Furthermore, the *T* gene was identified as the strongest candidate gene for vertebral development in the Hulunbuir Short-Tailed Sheep (Zhi et al., 2018).

Sheep dispersed and became adapted to a wide range of agroecological conditions, especially those distributed on plateaus or in desert regions, which are particularly sensitive to climate change. Hence, indigenous sheep provide excellent model organisms to gain novel insights into genetic mechanisms underlying rapid adaptations to extreme environments.

In this chapter, we applied different within-and between-breed selection analysis including contrasting populations living in different environments to unravel the mechanisms of adaptation of African sheep to environmental challenges as well as the genetic control of fat deposition and tail morphology in the same breeds. Building-up on **Chapter 3** results, the sheep populations were categorised *a priori* based on their environments and tail morphology. To the

best of our knowledge, this is the first study to conduct a whole-genome sequence analysis targeting the adaptability of African sheep to different environments and aiming to elucidate genomic variations of their tail morphology.

Materials and methods

Grouping of Population

Based on the genetic clusters revealed by full genome PCA, admixture and Treemix analysis (**Chapter 3**), we initially separated the 15 Ethiopian, Sudanese and Libyan populations into four groups: ET_G1 (Kefis, Adane, Arabo, Segentu) and ET_G2 (Bonga, Kido, Gesses, Loya, ShubiGemo, Doyogena) Ethiopian fat-rump and long fat-tail sheep, respectively, thin-tail sheep (Hammari, Kabashi) from Sudan (SD) and pendulous fat-tail sheep (Barbary) from Libya (LB). One short fat-tail sheep (Molale) was included with the fat-rump (ET_G1) sheep, while the other population (Gafera) was clustered to ET_2 and included within this group of long fat-tail sheep.

For environmental adaptation analysis, based on sampling information reported in **Chapter 2** (**Figure 3.1 and Table 3.1**), we categorised the study breeds of sheep into different contrasting environments based on the altitude of the occupied region for each sheep group (**Table 4.1**).

Table 4.1 Classification of sheep groups and populations in contrasting environments based on altitude.

Group	N	Altitude category	Range (meter asl)	Populations
LB	12	Sea level	< 50	Libyan Barbary
SD	20	Low altitude	620	Hammari, Kabashi
ET_G2	70	High altitude	1300 - 2500	Gesses, Kido, ShubiGemo, Bonga, Loya, Doyogena, Gafera
ET_G1	19	Low altitude	740 - 890	Segentu, Kefis
	29	High altitude	1500 - 3068	Arabo, Adane, Molale

Our environmental classification considers ET_G2 as a group of sheep from high altitudes and humidity (average annual precipitation > 400 mm), SD as desert sheep (average annual precipitation < 10 mm) from low altitude, and LB as a sheep population from an arid (average annual precipitation < 400 mm) zone at sea level. The remaining group (ET_G1) includes sheep populations living at high or low-altitudes. This later group was not included in our genomic analysis in response to environmental challenges, but only for the tail phenotype analysis.

Measurement of tail phenotypes

We then measured the tail phenotypic characteristics of one sample from each group. Following the protocol used by Zhi et al. (2018), we carefully removed fat deposits and muscle tissue from the caudal vertebrae. We fixed the caudal vertebrae in ethanol for 5 days and cleared

them by immersing in 0.5% NaOH for 2 days. Residuals of adipose tissue were removed by immersing the caudal vertebrae in petrol for 3 days. Caudal vertebrae (CV) number, individual vertebrae length and the overall length of the tail skeleton were estimated and plotted. Based on these data and tail length related literature (Mason, 1991; Gizaw et al., 2007), we classified the study sheep groups into two tail phenotype categories: short-tailed sheep represented by ET_G1 and LB groups, and long-tailed sheep represented by ET_G2 and SD sheep.

Analysis of Signatures of Selection

Three approaches of detecting signatures of selection (*Hp*, F_{ST} and *XP-EHH*) were performed using 34.8 million annotated autosomal SNPs reported in **Chapter 3**. We used a sliding window approach to perform within-group heterozygosity (*Hp*) and among-group genetic differentiation (F_{ST}) analysis using VCFtools v.0.1.15 (Danecek et al., 2011; Li et al., 2014; Ahbara et al., 2019; Zhi et al., 2018).

Following Zhi et al. (2018) and based on the LD decay analysis reported in **Chapter 3 (Figure 3.2)**, a window size of 100 kb was selected. By implementing several analytical runs using different sliding-window distances of 10 kb, 20 kb, 50 kb, 75 kb and 100 kb, a 10 kb sliding-window distance was found to generate the best resolution of selection signals. For within-group analysis, we estimated the heterozygosity (*Hp*) across the genome using the following formula:

$$Hp = \frac{2 \sum nMaj \sum nMin}{(\sum nMaj \sum nMin)^2} \qquad (4.1)$$

Where ∑nMaj and ∑nMin are the sum of nMaj and nMin, respectively. We then transformed the *Hp* values into Z-scores using the formula:

$$ZHp = \frac{(Hp - \mu Hp)}{\sigma Hp} \qquad (4.2)$$

Where μHp is the average of the overall heterozygosity and σHp is the standard deviation for all windows within the group.

We measured genetic differentiation among the sheep groups using the fixation index (F_{ST}) as described in **Chapter 2** (Ahbara et al., 2019).

To better assess the direction of selection among, we adopted the cross-population statistic, *XP-EHH*, that contrasts extended haplotype homozygosity (*EHH*) profiles between populations (Sabeti et al., 2007). This haplotype-based method of detecting selection relies on estimating and contrasting the *iES* statistic (the pattern of the integrated *EHH* of the same allele) between sheep groups, as follows:

$$\ln (XP - EHH) = \ln \left(\frac{iES_{pop1}}{iES_{pop2}} \right) \quad (4.3)$$

Haplotype phasing was inferred simultaneously on all detected bi-allelic SNPs in the four study groups using BEAGLE (Browning and Browning, 2007). Assuming 100 Mb = 1 Morgan (Quinlan and Hall, 2010), phased haplotypes were converted to IMPUTE format using *--impute* function in VCFtools v.0.1.15. The generated haplotypes were used for *XP-EHH* estimation in pair-wise comparisons between the four groups using HAPIN (Maclean et al., 2015). To determine the empirical significance of the *XP-EHH* statistic, *XP-EHH* scores were normalized by subtracting the mean and dividing it by the standard deviation of all the scores. While the negative *XP-EHH* score indicates that selection occurred in the reference population, the top positive 0.1% of the empirical distribution were considered candidate regions for our targeted group in each comparison. Similar cut-off was used for ZF_{ST}, while the top negative 99 percentiles were used as significant cut-off values for *ZHp* analysis. The overlapping candidate regions among the different methods and comparisons were detected and merged using Bedtools v.2.25.0 (Quinlan and Hall, 2010).

To identify candidate genome regions associated with adaptation to different environments, high and low altitude comparisons, we contrasted the low altitude short-tailed groups (ET_G1 and LB) and long-tailed Sudanese (SD) against the high altitude long fat-tail sheep (ET_G2). The later in addition to a comparison contrasting SD sheep against all other sheep groups were considered identifying putative genes and genomic regions associated with the tail length and fat deposition in African sheep.

Functional annotation of candidate genes

Similar to the procedure followed in **Chapter 2**, genes overlapping genomic windows passing the proposed selection thresholds were retrieved based on the Oar-v3.1 Ovine reference genome assembly from the *Ensembl Genes 95 database* and the online *Ensembl BioMart* tool (http://www.ensembl.org/biomart) (Smedley et al., 2009). Functional enrichment analysis was

performed using the functional annotation tool in DAVID (Huang et al., 2008) using *Ovis aries* as the background species. Gene functions were determined using the NCBI (http://www.ncbi.nlm.nih.gov/gene/) and OMIM databases (http://www.ncbi.nlm.nih.gov/omim/) and review of the literature.

Haplotype structure analysis of candidate regions

We inferred haplotype structure within candidate regions of interest for evidence of distinct haplotypes that may be unique to a specific sheep group. Following **Chapter 2** procedure, we used *hapFLK* software (Fariello et al., 2013; Ahbara et al., 2018) and the scripts provided in the *hapFLK* webpage (https://forge-dga.jouy.inra.fr/projects/hapflk) for the estimation of haplotypes frequencies and visualization of haplotype clusters in the region of selected candidate genes (Fariello et al., 2013).

Results

Adaptations to high-altitude and arid environments

The PCA, admixture and TreeMix analysis reported in **Chapter 3** supports four distinct genetic groups: fat-rump (ET_G1), long fat-tail sheep from Ethiopia (ET_G2), long thin-tail from Sudan (SD) and Barbary pendulous fat-tail group from Libya (**Figure 3.3; Figure 3.4; Figure 3.5**). The two short fat-tail sheep (Molale-Menz and Gafera-Washera) were separated from each other with Molale-Menz showing close genetic affinity to the fat-rump (ET_G1) sheep and Gafera-Washera was genetically close to the Ethiopian long fat-tail (ET_G2) group. Environmentally, these groups of African sheep can be defined as sheep populations from high-altitude (ET_G2) and arid (LB and SD) zones as reported in Materials and Method (**Table 4.1**).

From this classification, we chose ET_ G2 to identify genome regions and genes associated with adaptation to high-altitudes in Ethiopia, while SD and LB were used to identify genome regions associated with low altitude. According to our results in **Chapter 3** (**Figure 3.6**), the fastest LD decay occurred in the first 100 kb of the genome in all the sheep groups. Therefore, 100 kb was taken as the optimal width of the window with 10 kb sliding step. Pooled heterozygosity (*Hp*) was evaluated for each sheep group using the proposed sliding window parameters. We then Z-transformed the *Hp* scores to standard normal distribution and the top 0.1% windows with negative values were regarded to be candidates for selection (**Figure 4.1**).

For each window, F_{ST} was calculated among sheep groups and then Z-transformed to display a standard normal distribution (**Figure 4.2**). The larger the ZF_{ST} values, the greater the genetic differentiation between the sheep groups. The *XP-EHH* LD based approach was implemented to determine the direction of selection among sheep groups. *XP-EHH* analysis was carried out between ET_G2 and ET_G1, LB and SD, respectively and *vice versa* (ET_G1, LB and SD vs ET_G2). The resulting *XP-EHH* scores were summarised for each 100 kb window, and the 0.1% top windows were considered significant (**Figure 4.3 and 4.4**).

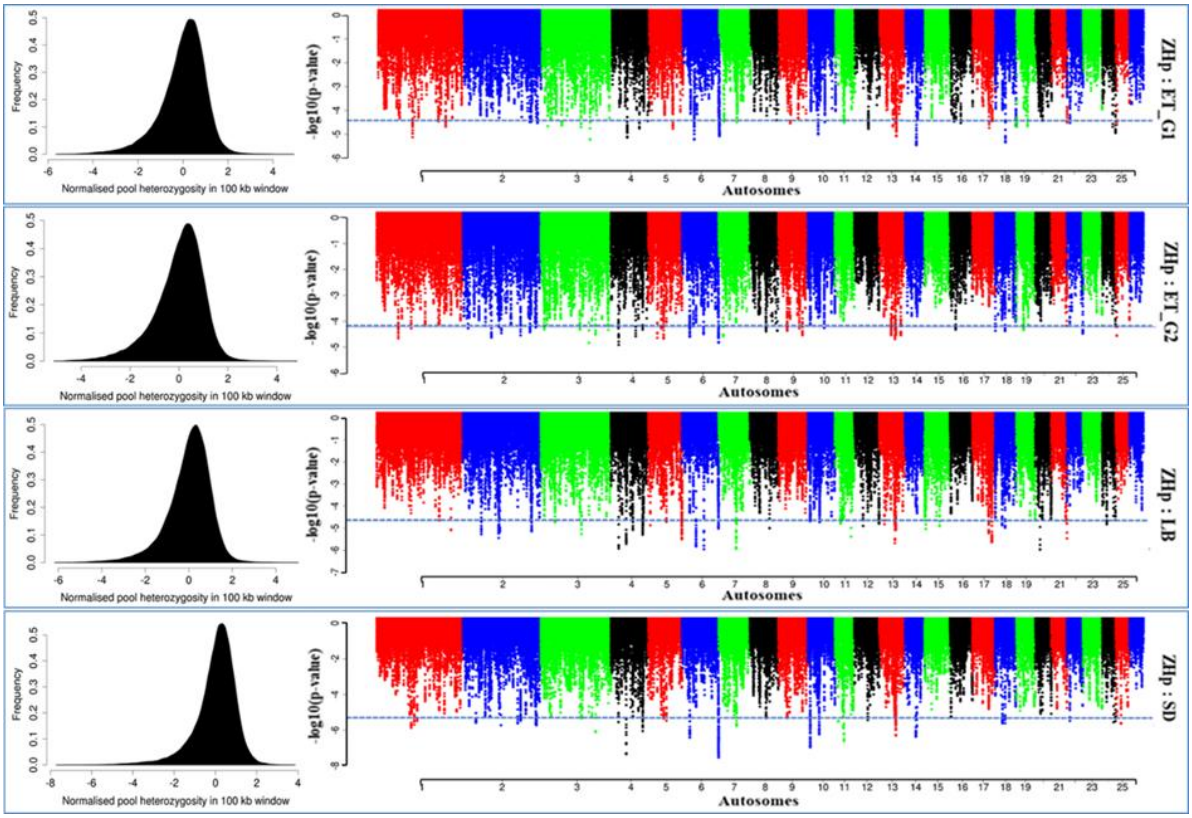


Figure 4.1, Negative end of the (*ZHp*) distribution plotted along chromosomes for each sheep group: Ethiopian fat-rump (ET_G1), Ethiopian long fat-tail (ET_G2), Libyan Barbary (LB) and Sudanese thin tail (SD) sheep.

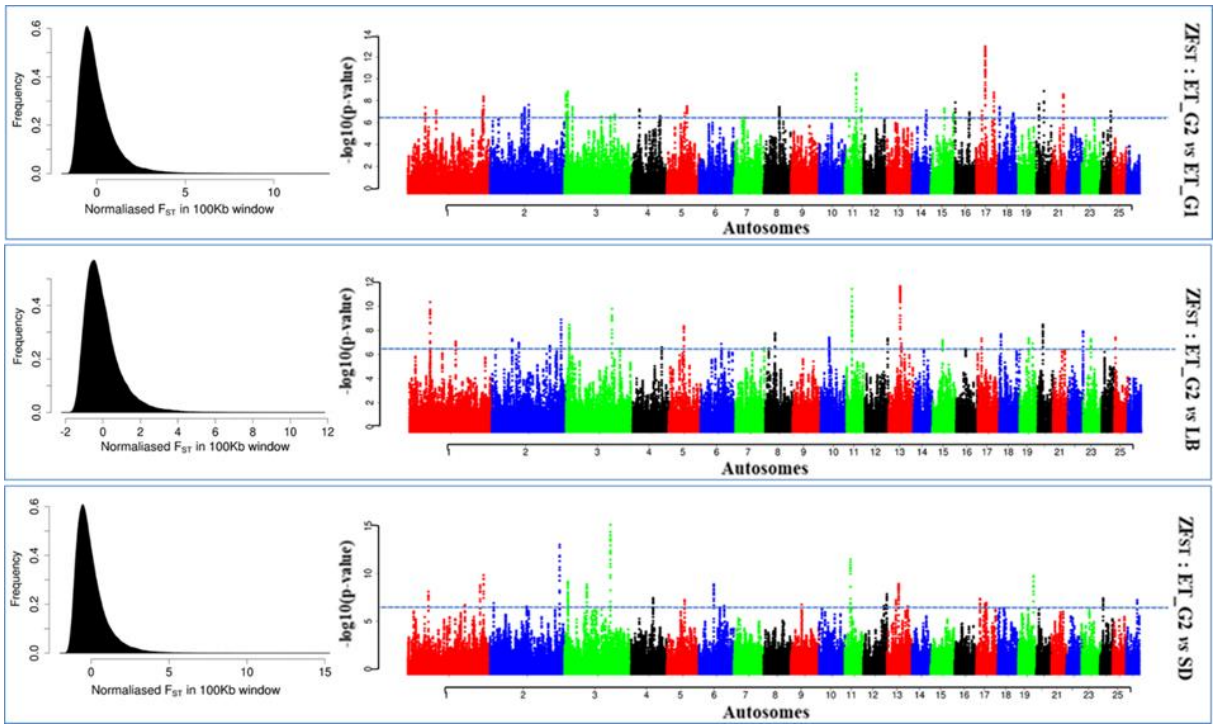


Figure 4.2. Manhattan plot of the genome-wide distribution of ZF_{ST} values following the comparison of Ethiopian long fat-tail (ET_G2) sheep with Ethiopian fat-rump (ET_G1), Libyan Barbary (LB) and thin-tail (SD) sheep group from Sudan.

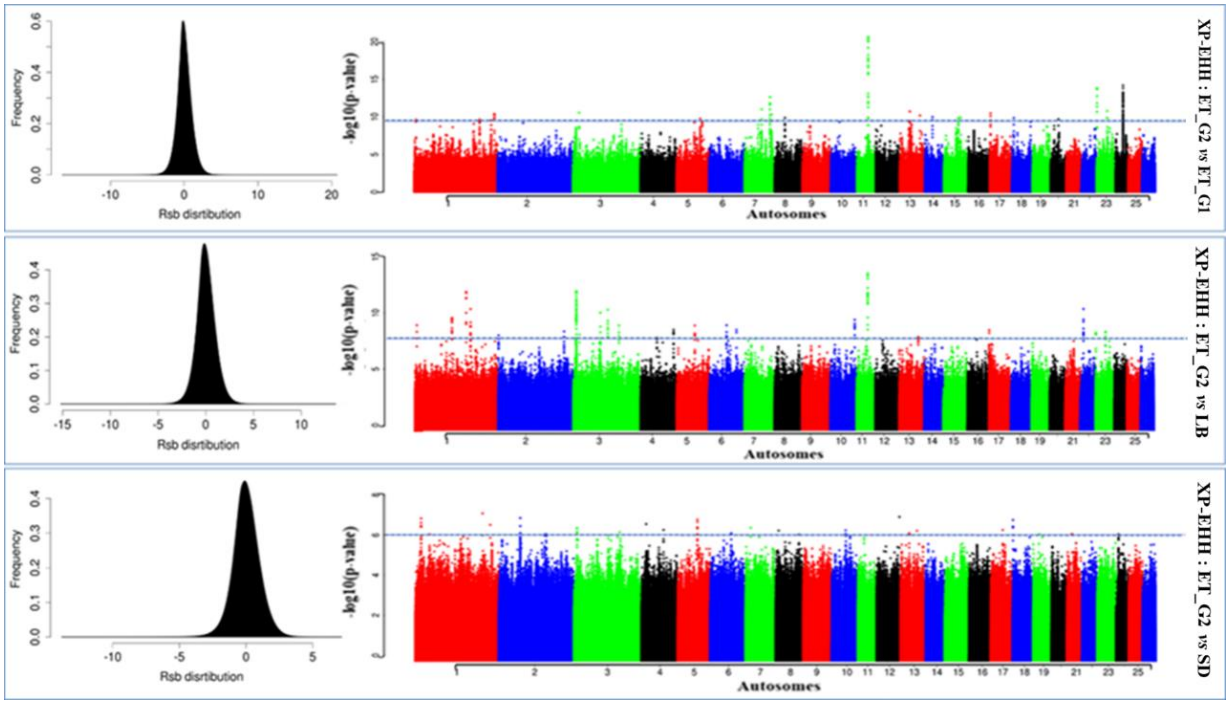


Figure 4.3. Manhattan plot of the genome-wide distribution of *XP-EHH* values following the comparison of Ethiopian long fat-tail (ET_G2) sheep with Ethiopian fat-rump (ET_G1), Libyan Barbary (LB) and thin-tail (SD) sheep group from Sudan.

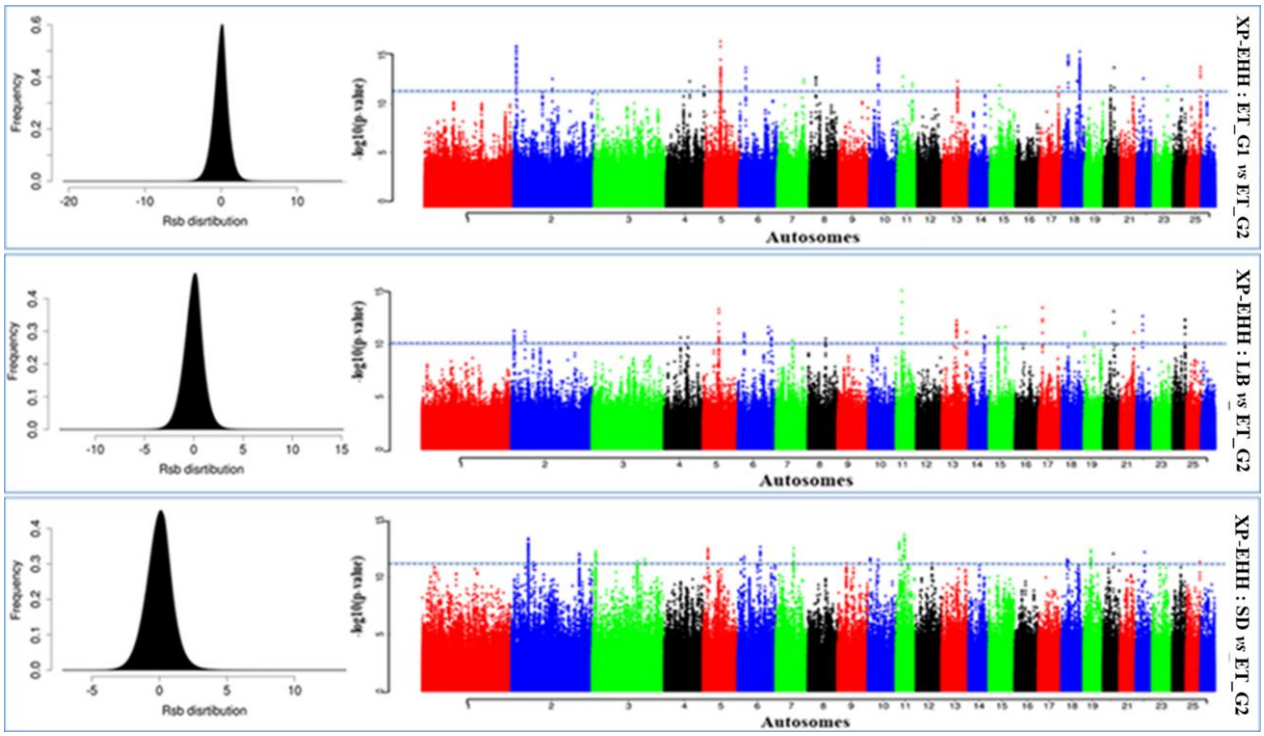


Figure 4.4 Manhattan plot of the genome-wide distribution of *XP-EHH* values following the comparison of Ethiopian fat-rump (ET_G1), Libyan Barbary (LB) and thin-tail (SD) sheep group from Sudan with Ethiopian long fat-tail (ET_G2) sheep.

From all these comparisons, 70 regions that overlapped between at least two methods and/or sheep group comparisons (**Appendix Table 5.3**) were further analysed. For better accuracy, results obtained from comparisons involving ET_G1 group (21 candidate regions) were excluded from the investigation of signatures associated with adaptation to environmental challenges. This exclusion was because the group included samples of sheep from contrasting environments. Following this exclusion, the remaining 49 regions across 21 autosomal chromosomes are expected to be associated with environmental challenges (**Table 4.2**). Of these, 11 were identified by the three methods or comparisons on chromosomes OAR1, OAR2 (2 candidate regions), OAR6 (2 candidate regions), OAR8, OAR11, OAR13 (2 candidate regions), OAR22 and OAR25; while 32 candidate regions were identified by two methods or comparisons. The remaining six putative selective regions were identified by only one analysis. It includes one region revealed by ZF_{ST} on OAR24, three regions identified by *XP-EHH* on OAR3 and one strong candidate region identified by *ZHp* on OAR13.

Annotation of these candidate regions identified 170 putative genes under positive selection, 81 of which only previously identified in **Chapter 2.** They included 18 candidate genes (*CD84, NF1, EVI2B, OMG, RIN2, NAA20, CRNKL1, CFAP61, PTMA, PDE6D, COPS7B, RF00091, DIS3L2, LNX1, CHIC2, GSX2, PDGFRA* and *TBX18*) found in significant regions identified simultaneously by the three approaches (*ZHp, ZF_{ST} and XP-EHH*).

Table 4.2 Candidate regions and genes associated with environmental challenges identified by a combination of at least two methods or comparisons in the Ethiopian long fat-tail *versus* Ethiopian fat-rump, Libyan Long fat-tail and Sudanese thin-tail sheep.

Chr	Region	Methods	Genes
1	68600001-68840000	ET_G2 **(*ZHp*)**, ET_G2 *vs* LB & SD **(*ZF$_{ST}$*)**	*BTBD8, RF00026, C1orf146, GLMN, RPAP2*
	250500001-253740000	ET_G2 vs ET_G1 & SD **(*ZF$_{ST}$*),** ET_G1 vs ET_G2 (***XP-EHH***)	*STAG1, PCCB, MSL2, PPP2R3A, EPHB1, RF00026, KY, CEP63, ANAPC13, AMOTL2, RYK, SLCO2A1, RAB6B, SRPRB, TF*
	5930001-6120000	ET_G2 *vs* ET_G1 & LB **(*XP-EHH*)**	*Close to SH3BP4*
	109920001-110020000	ET_G1, ET_G2 & SD **(*ZHp*)**, ET_G2 *vs* LB **(*ZF$_{ST}$*, *XP-EHH*)**	*CD84*
	154650001-154790000	LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*ZNF654, C3orf38*
2	68440001-68570000	LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*KANK1*
	90750001-90950000	ET_G2 (***ZHp***), LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*Span BMPR2 gene in cow*
	232440001-233000000	ET_G2 & SD **(*ZHp*)**, ET_G2 *vs* LB & SD **(*ZF$_{ST}$*, *XP-EHH*)**	*PTMA, PDE6D, COPS7B, RF00091, DIS3L2*
3	10920001-11320000	ET_G2 **(*ZHp*)**, ET_G2 *vs* ET_G1 & LB **(*ZF$_{ST}$*)**	*GOLGA1, ARPC5L, WDR38, RF00026, OLFML2A, MIR181B2, oar-mir-181a-2, NR6A1, NR5A1, ADGRD2*
	20980001-21170000	ET_G2 *vs* ET_G1 & LB (***XP-EHH***)	*ENSOARG00000016157*
	10630001-10820000	ET_G2 *vs* LB & SD (***XP-EHH***)	*GAPVD1, HSPA5, RABEPK, PPP6C, SCAI*
	12160001-12370000	SD vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*RF00402, CRB2*
	151770001-151960000	ET_G2 *vs* LB & SD (***XP-EHH***)	*Close to DYRK2*
	154100001-154390000	ET_G1 & LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***), SD vs ET_G2 (***ZF$_{ST}$***)	*MSRB3*
4	73080001-73320000	ET_G2 *vs* SD **(*ZF$_{ST}$*, *XP-EHH*)**	*Close to ZNF804B*
	96540001-96700000	ET_G1, LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*CHCHD3*
6	69010001-70090000	ET_G2, LB (***ZHp***), LB & SD vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*LNX1, CHIC2, GSX2, PDGFRA*
	69920001-70030000	LB **(*ZHp*)**, ET_G2 *vs* SD **(*ZF$_{ST}$*, *XP-EHH*)**	*Close to KIT, PDGFRA*
7	55870001-56140000	LB (***ZHp***), SD vs ET_G2 (***XP-EHH***)	*DMXL2, GLDN, CYP19*
	96270001-96390000	LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*Close to GALC*
8	51490001-51760000	ET_G2 & SD (***ZHp***), ET_G1 vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*Close to TBX18*
	32040001-32230000	LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*LIN28B*
10	28560001-28820000	LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*PDS5B*
	29390001-29510000	LB vs ET_G2 (***ZF$_{ST}$***, ***XP-EHH***)	*RXFP2*

11	10750001-10940000	SD (***ZHp***), SD vs ET_G2 (***XP-EHH***)	*INTS2, BRIP1*
	18190001-18520000	LB (***ZHp***), SD & LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*NF1, EVI2B, OMG*
	27010001-27180000	SD (***ZHp***), SD vs ET_G2 (***XP-EHH***)	*DNAH2, NAA38, KDM6B, TMEM88, CYB5D1, CHD3, RF00602, RNF227, KCNAB3, TRAPPC1, CNTROB*
	37200001-37430000	ET_G2 *vs* ET_G1 **(*ZF_{ST}, XP-EHH*)**, ET_G2 *vs* LB **(*XP-EHH*)**	*CALCOCO2, TTLL6, HOXB13, RF02133, RF02132, MIR196A1, HOXB9, HOXB7*
12	75850001-75990000	ET_G2 *vs* SD & LB **(*ZF_{ST}, XP-EHH*)**	*MIR181B1, oar-mir-181a-1*
13	42560001-42800000	ET_G1 (***ZHp***), LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*PAK4, NCCRP1, SYCN, LRFN1*
	38560001-38890000	ET_G2 (***ZHp***), LB & SD vs ET_G2 (***ZF_{ST}, XP-EHH***)	*RIN2, NAA20, CRNKL1, CFAP61*
	69370001-69520000	ET_G2 (***ZHp***)	*PLCG1*
15	35110001-35280000	LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*PLEKHA7, C11orf58*
16	31650001-31750000	LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*CCDC152*
17	29820001-30220000	ET_G1 & SD vs ET_G2 (***ZF_{ST}, XP-EHH***)	*INTU*
	13200001-13340000	LB & SD vs ET_G2 (***ZF_{ST}, XP-EHH***)	*RF02271, HHIP*
18	5540001-5710000	LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*ADAMTS17*
	6010001-6170000	LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*MEF2A*
19	23200001-23420000	ET_G2 (***ZHp***), SD vs ET_G2 (***XP-EHH***)	*CNTN4*
	30490001-30640000	LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*ENSOARG00000009693,* (*forkhead box P1*)
	43420001-43530000	LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*SLMAP*
20	17230001-17480000	LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*MAD2L1BP, RF00001, RSPH9, MRPS18A, VEGFA*
21	39200001-39360000	ET_G1 vs ET_G2 (***ZF_{ST}, XP-EHH***)	*DDB1, VWCE, TKFC, CYB561A3, TMEM138, TMEM216, CPSF7, SDHAF2, PPP1R32*
	43040001-43230000	LB vs ET_G2 (***XP-EHH***)	*LTBP3, SSSCA1, FAM89B, EHBP1L1, KCNK7, MAP3K11, PCNX3, SIPA1, RELA, KAT5, RNASEH2C, AP5B1, OVOL1*
22	50510001-50760000	ET_G2 (***ZHp***), LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*CYP2E1, ECHS1, FUOM, PAOX*
23	26220001-26380000	LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*DSC1, DSC2*
	1030001-1220000	ET_G2 *vs* ET_G1 & LB (***XP-EHH***)	*ENSOARG00000026145, Close to SALL3*
24	9840001-9940000	SD vs ET_G2 (***ZF_{ST}***)	*CLEC16A, PRM3*
25	4060001-4180000	ET_G1 & ET_G2 (***ZHp***), LB vs ET_G2 (***ZF_{ST}, XP-EHH***)	*GNPAT, EXOC8, SPRTN, EGLN1*

Following *DAVID* enrichment analysis, we identified the genes present in the 12 top most (*P*-value < 0.05) gene ontology (GO) and KEGG_PATHWAY terms and selected them as candidate genes associated with environmental challenges (**Table 4.3**). The most enriched GO terms are associated with response to oxygen level (GO:0070482) and hypoxia (GO:0070482) (GO:0001666).

The most significant KEGG_PATWAY was the HIF-1 signaling pathway (oas04066) with the identified genes by the enrichment analysis associated to high-altitude adaptation shown in **Figure 4.5**.

Table 4.3 Enriched functional term clusters and their enrichment scores following *DAVID* analysis for genes identified in association with environmental challenges.

Category	**Term**	***P*-values**	**Genes**
GO:0070482	response to oxygen levels	0.00020	*NF1, EGLN1, BRIP1, VEGFA*
GO:0001666	response to hypoxia	0.00052	*BRIP1, EGLN1, NF1, VEGFA*
oas04066	HIF-1 signaling pathway	0.00108	*TF, PLCG1, RELA, VEGFA, EGLN1*
oas04014	Ras signaling pathway	0.00347	*PLCG1, RELA, VEGFA, NF1, PDGFRA, KIT*
GO:0030325	adrenal gland development	0.00613	*NF1, PDGFRA, NR5A1*
oas05200	Pathways in cancer	0.00895	*PLCG1, RELA, VEGFA, PDGFRA, EGLN1, HHIP, KIT*
oas04015	Rap1 signaling pathway	0.01530	*PLCG1, SIPA1, VEGFA, PDGFRA, KIT*
GO:0008180	COP9 signalosome	0.01992	*PLCG1, COPS7B, HSPA5*
GO:1901098	positive regulation of autophagosome maturation	0.02539	*CLEC16A, CALCOCO2*
GO:0010587	miRNA catabolic process	0.02539	*DIS3L2, LIN28B*
GO:0060326	cell chemotaxis	0.04767	*PDGFRA, HOXB9, EPHB1*
GO:0001553	Luteinisation	0.05015	*PDGFRA, NR5A1*

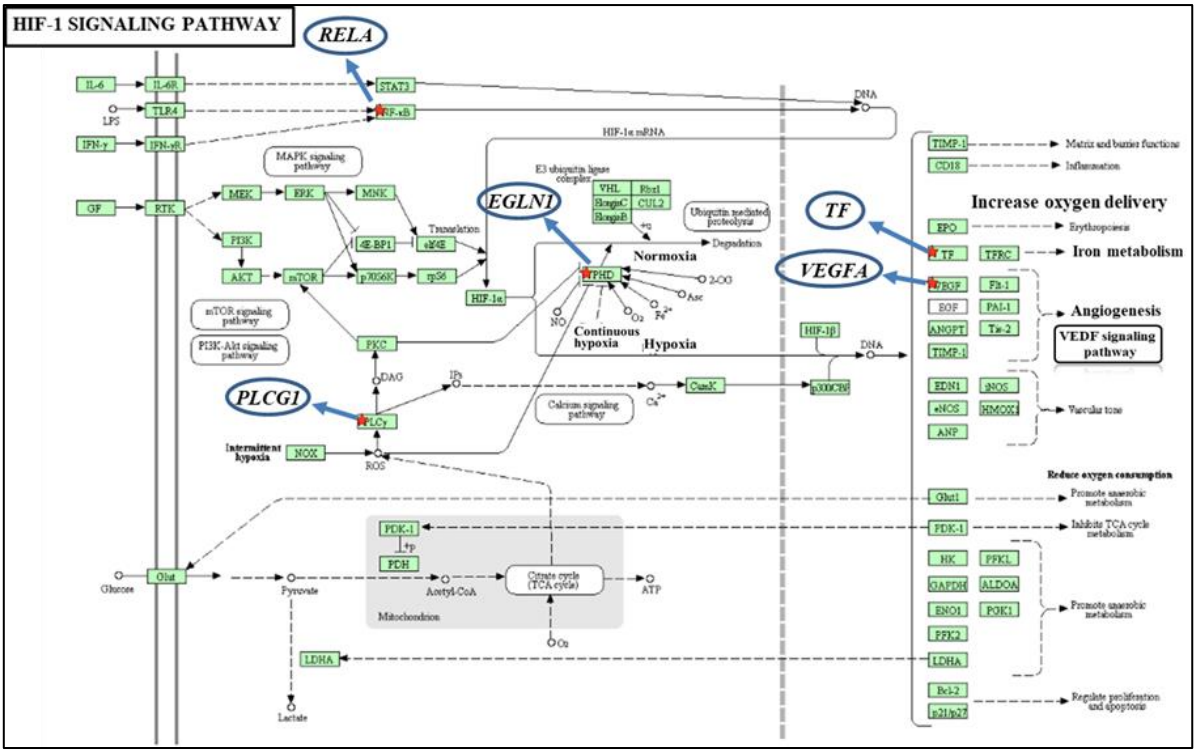


Figure 4.5. Plot of HIF-1 pathway showing the key role of the positively selected genes in high-altitude adaptation in African sheep.

The haplotype structures analysis was implemented for the strongest candidate regions **(Table 4.2)**. It includes: A candidate region spanning five genes (*PTMA, PDE6D, COPS7B, RF00091* and *DIS3L2*) on OAR2 was identified using all the three methodologies (**Table 4.2**). The haplotype structure of this region shows a shared haplotype approaching fixation but partly selected in Sudanese (SD) and Libyan sheep (LB); the other sheep group (ET_G2) having admixed haplotypes (**Figure 4.6**). This haplotype covers *DIS3L2* that most likely is linked to the formation and development of limbs under desert and arid environments.

On OAR3, the *MSRB3* (methionine sulfoxide reductase B3) region, was supported by ZF_{ST} and *XP-EHH* analysis. The haplotype diversity of this region shows a distinct and nearly fixed haplotype for the ET_G2 group, which is characterised by extremely short ears, whereas the other sheep groups (SD and LB) have a different admixture of haplotypes and their ear phenotypes are characterised by drooping and long earlobes (**Figure 4.7**).

The *NF1*, *EVI2* and *OMG* genes were identified on OAR11 in a candidate region supported by *ZHp* analysis in LB group and differentiating SD and LB from ET_G2 sheep group *via* ZF_{ST} and *XP-EHH* analysis. The spanned genes in this region are likely associated with fat deposition, particularly, in LB sheep inhabiting arid environments. The haplotype diversity of

this region shows a shared haplotype still under selection in SD but nearing complete fixation in the LB sheep group (**Figure 4.8**).

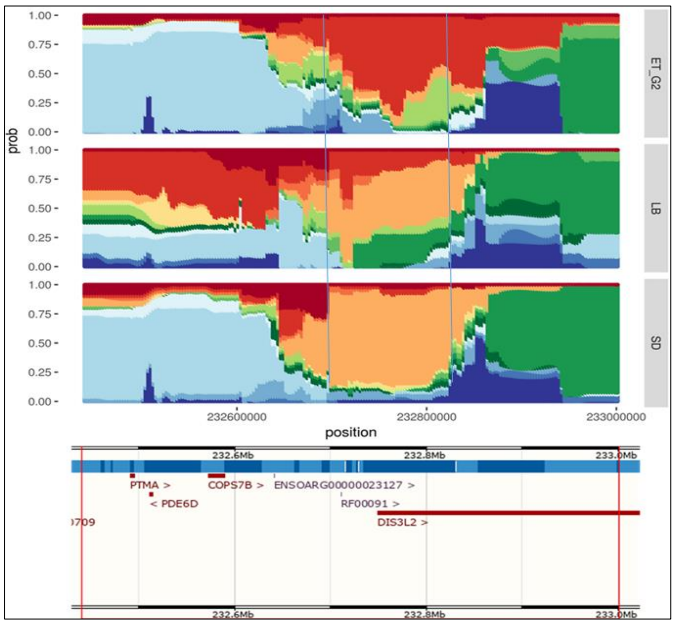


Figure 4.6. Haplotype structure of the *DIS3L2* candidate region.

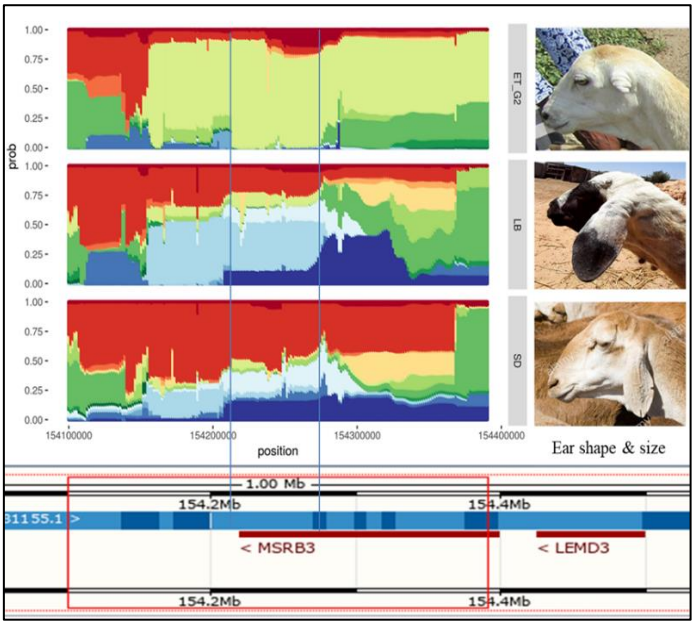


Figure 4.7. Haplotype structure of *MSRB3* candidate region likely associated to the size and shape of the ear.

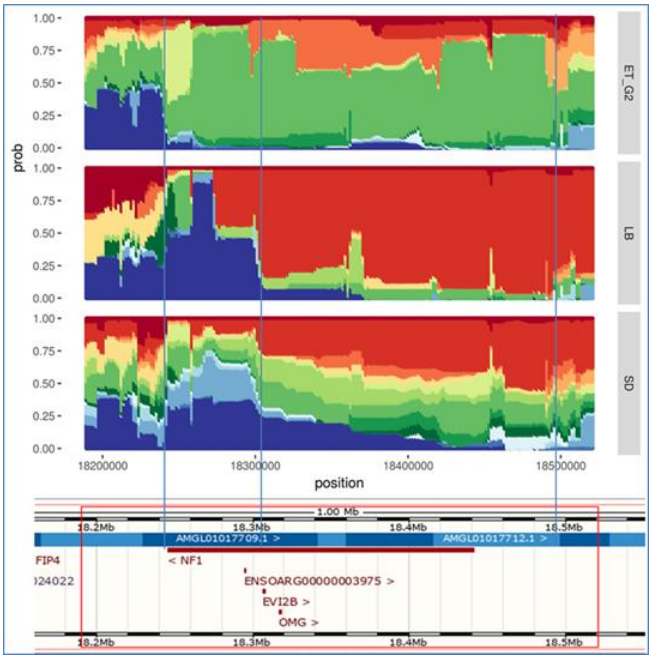


Figure 4.8. Haplotype structure of *NF1*, *EVI2* and *OMG* candidate region.

Another selection sweep region spanning 4 genes (*LNX1*, *CHIC2*, *GSX2* and *PDGFRA*) is also supported by the three analysis (**Table 4.2**). The haplotype diversity plot shows different backgrounds at the *PDGFRA* candidate gene region (**Appendix Figure 5.5**). The *PDGFRA* gene is associated with tail phenotype and involved in nematode resistance in LB sheep. Additionally, two candidate regions separated by 570 kb were detected on Oar10 using *ZFST and XP-EHH* analysis. Their haplotype diversity plots show distinct haplotypes in the LB sheep at the *PDS5B* and *RXFP2* regions, which are linked to fat deposition and horn size, respectively (**Figure 4.9**).

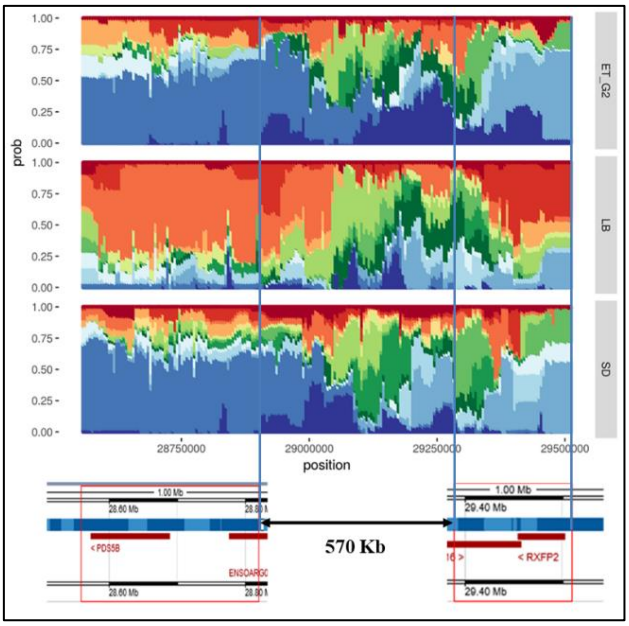


Figure 4.9. Haplotype structure of *PDS5B* and *RXFP2* candidate regions on OAR10.

Furthermore, among the candidate regions possibly under positive selection on OAR15 (ZF_{ST} and *XP-EHH* analysis), we found *PLEKHA7* with a distinct haplotype in LB sheep group (**Figure 4.10**). This gene plays a vital role in salt-sensitivity hypertension, blood pressure and kidney function.

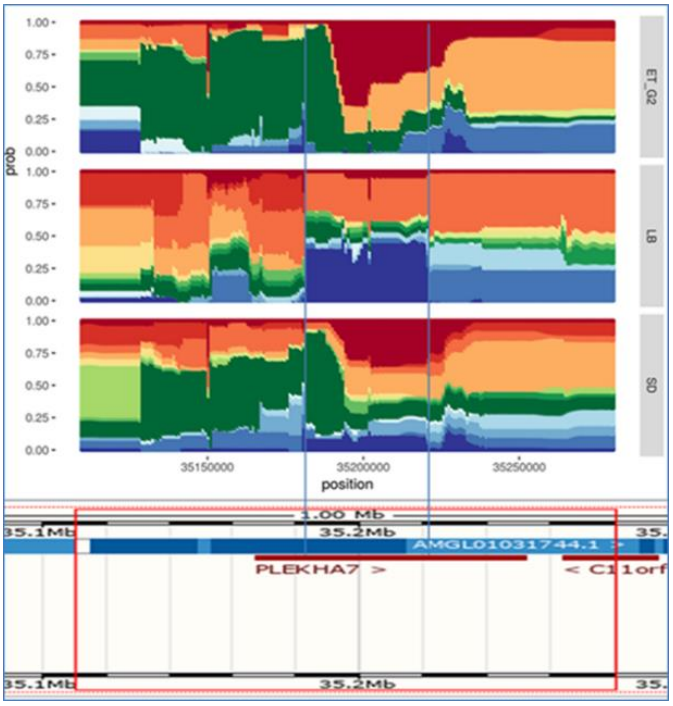


Figure 4.10. Haplotype structure of *PLEKHA7* candidate region on OAR15 within the contrasted groups of sheep

On OAR1, both ZF_{ST} and *XP-EHH* significant signals overlap at the candidate region spanning three candidate genes (*RAB6B*, *SRPRB* and *TF*). The haplotype diversity of this region supports a specific haplotype under selection at the region spanning *TF* gene, which is associated with oxygen consumption under high-altitudes (**Figure 4.11**).

Furthermore, a strong selection sweep spanning *VEGFA* gene was identified by ZF_{ST} and *XP-EHH* on OAR20. The haplotype diversity shows a shared haplotype near fixation in ET_G2 and a partly selected one in SD sheep group, while the LB group shows multiple haplotype at the *VEGFA* gene region (**Appendix Figure 5.6**). This gene has a complementary role to that played by *TF* related to oxygen delivery and transportation under hypoxic conditions. We also identified a strong peak on OAR25 harbouring 4 genes (*GNPAT*, *EXOC8*, *SPRTN* and *EGLN1*) supported by the three methodological analysis. The haplotype structure of this region shows a shared haplotype fixed in ET_G2 and near fixation in SD sheep around the *EGLN1* gene, which plays a key role in hypoxia responses under high-altitude environments. Its allele

frequencies are 1, 0.95 and 0.625 in ET_G2, SD and LB sheep groups, respectively (**Figure 4.12**).

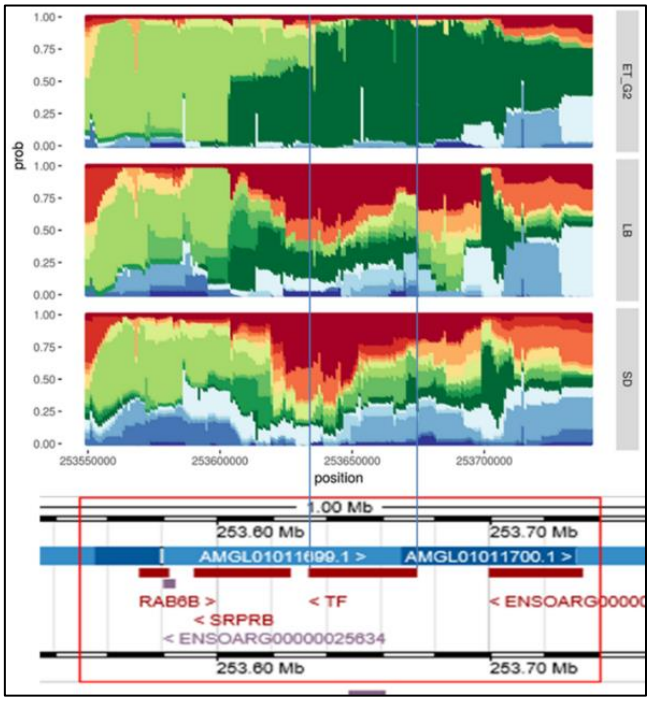


Figure 4.11. Haplotype structure of the *TF* candidate region on ORA1

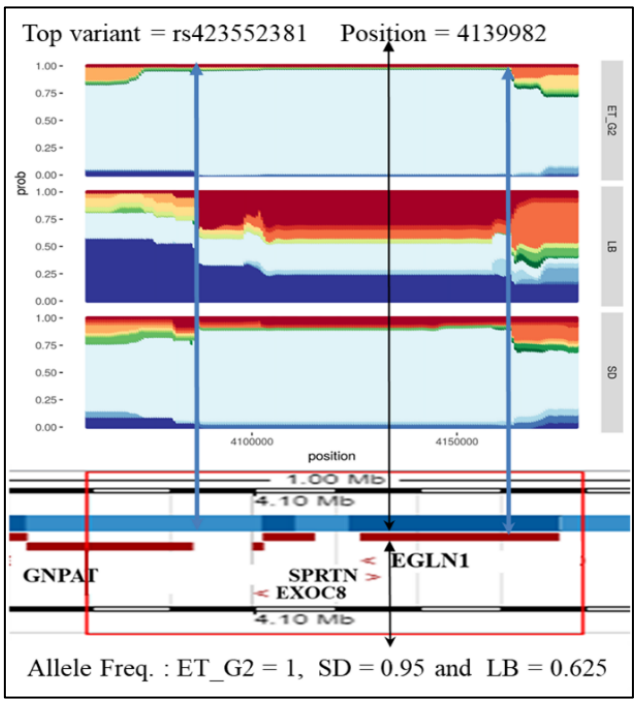


Figure 4.12. Haplotype structure of the *EGLN1* candidate region on ORA25. The position of the most significant variant is indicated above the arrow, and allele frequencies in the different groups are indicated below.

The last identified candidate region following the environmental adaptation analysis was detected on OAR23 using ZF_{ST} and *XP-EHH* with two haplotypes spanning *DSC1* and *DSC2* genes likely under selection in the ET_G2 sheep group (**Appendix Figure 5.7**). These genes are associated with hypoxia responses under humid high-altitude environments.

Tail morphology analysis

The phenotypic measurements of the fat-tailed sheep groups are shown in **Figure 4.13**. Extending to their hocks, while the Sudanese SD tail extends much lower, the ET_G1 sheep group represented by Menz and Kefis show extreme short tails. Libyan (LB) sheep and Ethiopian long fat-tail (ET_G2) show nearly similar tail lengths. The same number of 22 caudal vertebrae (CV) is observed in SD and LB groups, while ET_G1 and ET_G2 sheep have only 10 and 18 caudal vertebrae, respectively. The highest average of caudal vertebrae length (2.72 cm) was observed in SD followed by ET_G2 group (2.42 cm), whereas roughly similar CV lengths (2.24 and 2.20 cm) are observed in ET_G1 and LB sheep groups, respectively.

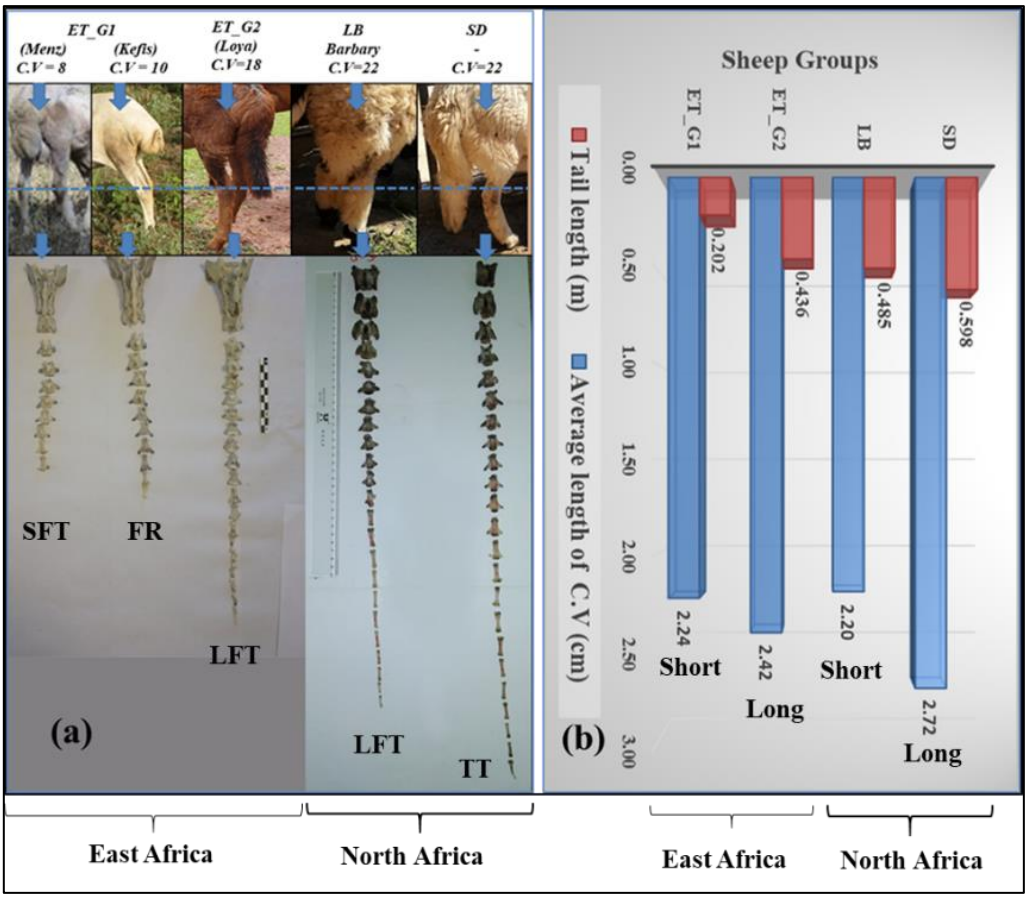


Figure 4.13. Tail phenotype measurements in different sheep groups: (a) visual length and caudal vertebrae (CV) account and shape. (b) Tail length (m) and average length (cm) of single vertebra. SFT = short fat-tail, FR = Fat-rump, LFT = Long fat-tail and TT = Thin tail.

Based on the phenotypic measurements, we categorised the fat-tailed sheep population into long-tailed sheep represented by SD (North African origin) and ET_G2 (East African origin) group, while ET_G1 (East African origin) and LB (North African origin) were considered as short-tailed sheep. To identify candidate regions and genes associated with tail length, we performed two sets of signatures of selection using ZF_{ST} and *XP-EHH*:

A. We contrasted the long fat-tail group with each of the short fat-tailed groups and the thin-tailed from Sudan to identify genome regions and genes associated with tail length (**Figures 4.14 A**). The comparisons between ET_G2 vs ET_G, LB and SD groups, respectively using ZF_{ST} and *XP-EHH* were previously performed for the study of environmental adaptation (**Figure 4.2 and 4.3**).
B. An additional comparison was performed by contrasting the long thin-tail sheep from Sudan (SD) with the other sheep groups (ET_G1, LB and ET_G2) to address here the genetic control of tail fat deposition (**Figures 4.14 B**).

These comparisons were further supported with *ZHp* analysis within each group (see **Figure 4.1**).

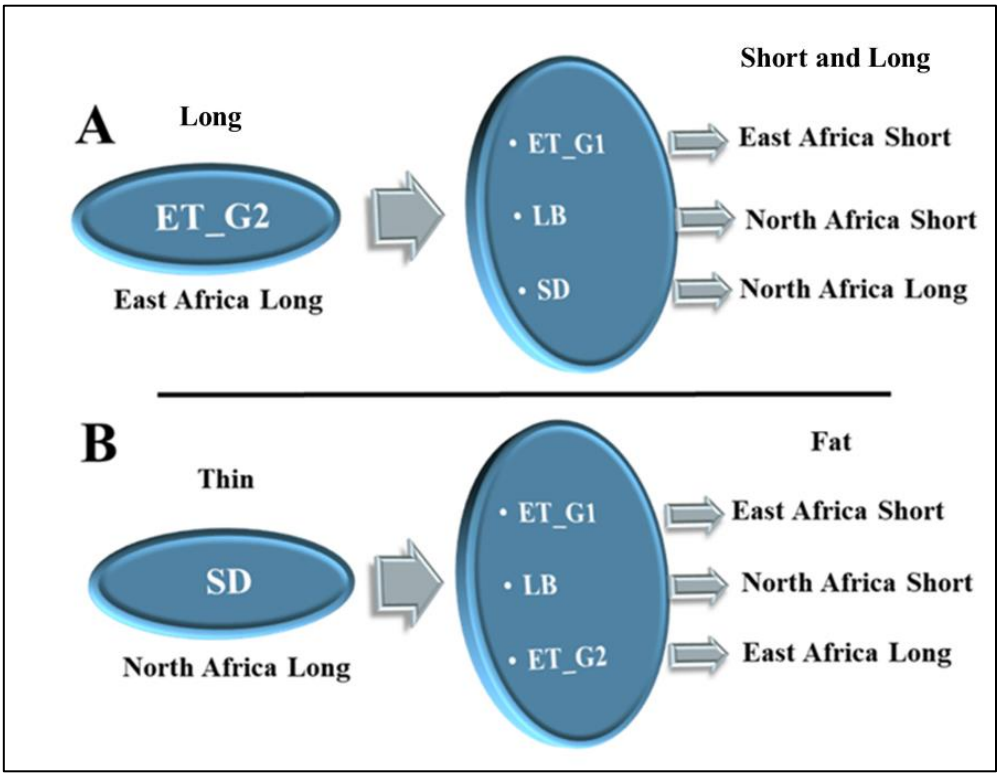


Figure 4.14. The proposed comparisons of signature of selection analysis to identify genes associated with tail morphology (length and fat): (A) long fat-tail sheep group (ET_G2) from East Africa against the other short fat-tail groups (ET_G1 and LB) from different origins and the North African long thin tail (SD). (B) long thin tail sheep group from North Africa (Sudan) against the other fat-tail groups, short fat-tailed (ET_G1 and LB) and long fat-tail (ET_G2).

In total, 54 candidate regions were identified by selection metrics across 20 of the 26 autosomal chromosomes of the sheep genome (**Table 4.4**). These candidates overlapped between at least two methodologies in the proposed comparisons (A and B). Of these, 4 were identified by three methodologies (*ZHp, ZF_{ST}* and *XP-EHH*). They are onOAR6 (two regions), OAR7 (one region) and OAR11 (one region). Seventeen candidate regions were identified by two methods and/or comparisons. The remaining 33 putative regions were identified by only one analysis but in more than one comparison. These include five regions identified using *ZF_{ST}* on OAR1, OAR2 (two regions), OAR3 and OAR4, while 17 selection sweeps were revealed by only *XP-EHH* in different comparisons across 11 autosomes.

Gene annotation of these candidate regions identified 183 candidate genes under positive selection, 20 (*FAM221B, HINT2, SPAG8, NPR2, MSMP, RGP1, GBA2, CREB3, TLN1, PDE6D, COPS7B, DIS3L2, CALCOCO2, TTLL6, HOXB13, RIN2, NAA20, CRNKL1, CFAP61, KIT*) of which were already reported in **Chapter 2.** Twenty-three candidate genes (*KIT, PDGFRA, NELFA, NSD2, LETM1, FGFR3, TMEM129, SLBP, GAK, CPLX1, UVSSA, MAEA, SLC49A3, PDE6B, PIGG, RF00427, DMXL2, GLDN, CYP19, NF1, EVI2B, and OMG*) were identified using the three methodologies (*ZHp, ZF_{ST} and XP-EHH*).

The strongest candidate region observed in the first comparison (Figure 4.14A) is on OAR11 (**Figure 4.2 and 4.3**). This genome region identified by both *ZF_{ST}* and *XP-EHH* spans 8 genes (*CALCOCO2, TTLL6, HOXB13, RF02133, RF02132, MIR196A1, HOXB9, and HOXB7*) (**Table 4.4 and Figure 4.15**). Amongst these, a strong candidate gene controlling tail length is *HOXB13,* the haplotype structure around this gene is presented at **Figure 4.16.** The candidate selected haplotype show a frequency of 0.843 and 0.750 in ET_G2 and LB sheep groups, at the *HOXB13* gene region.

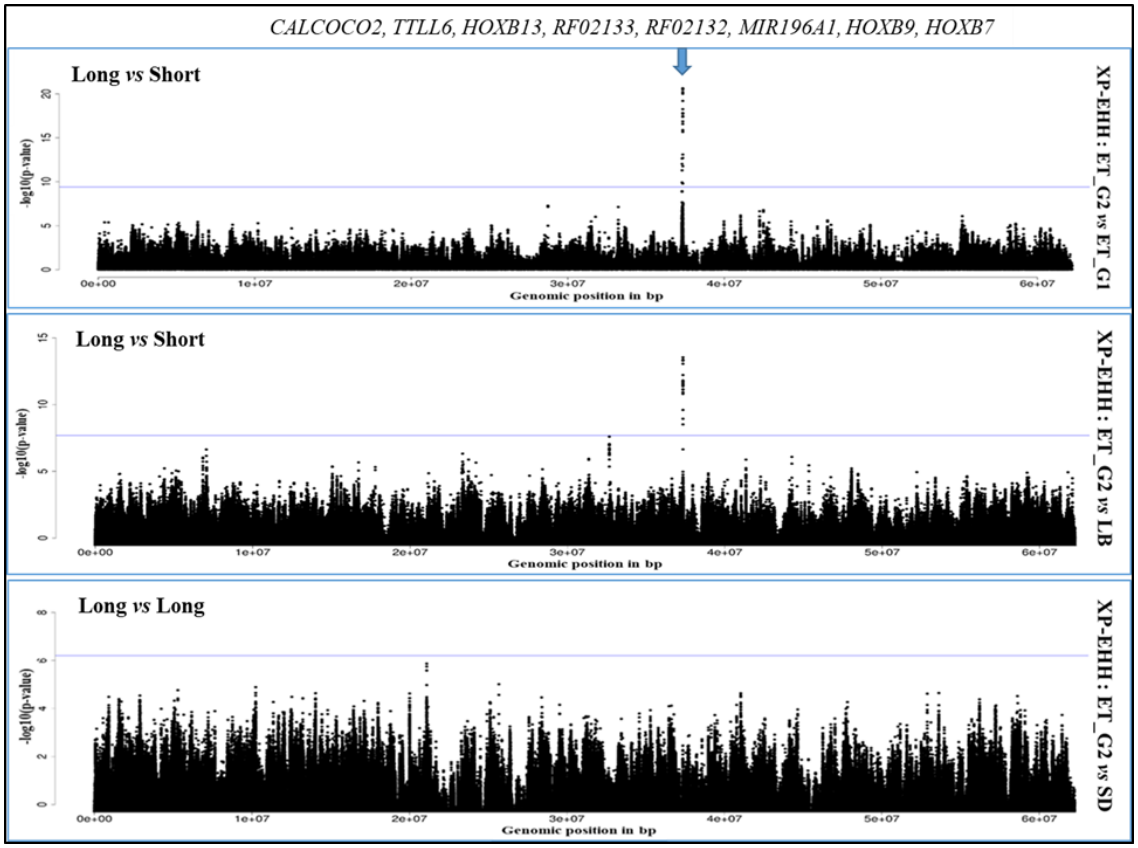


Figure 4.15. Candidate signature of selection for tail length on OAR11. XP-EHH analysis for comparison A (Figure 14 A)

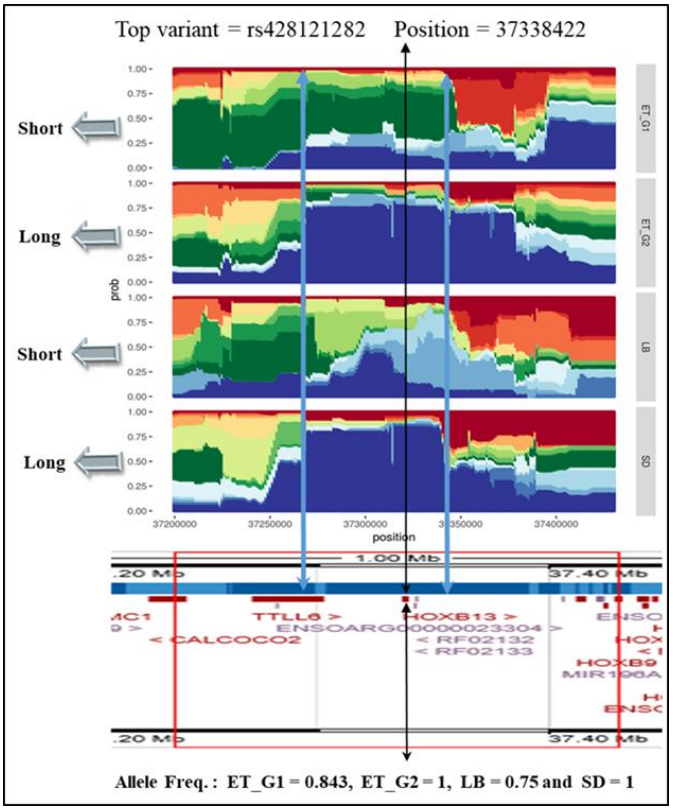


Figure 4.16. The haplotype structure of HOXB13 candidate region

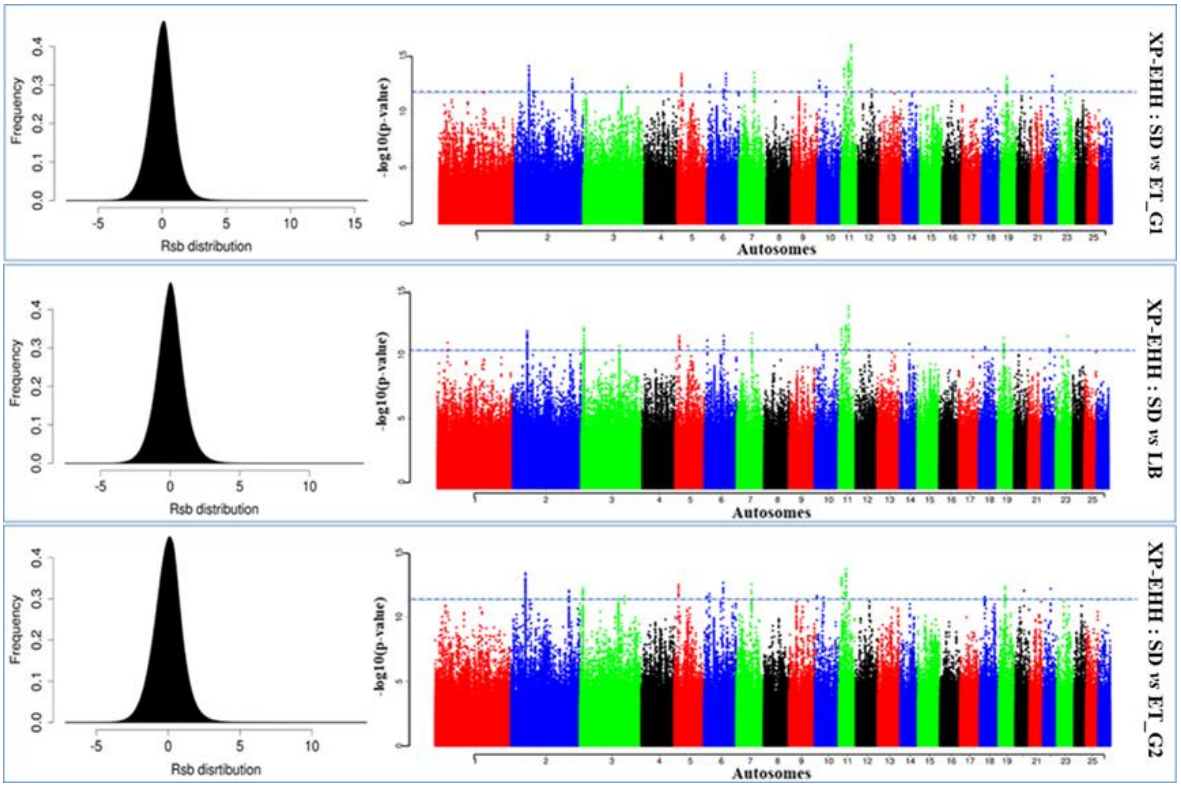


Figure 4.17 Manhattan plot of the genome-wide distribution of XP-EHH values following the comparison of thin-tail sheep group from Sudan (SD) with Ethiopian fat-rump (ET_G1), Libyan Barbary (LB) and Ethiopian long fat-tail (ET_G2) sheep.

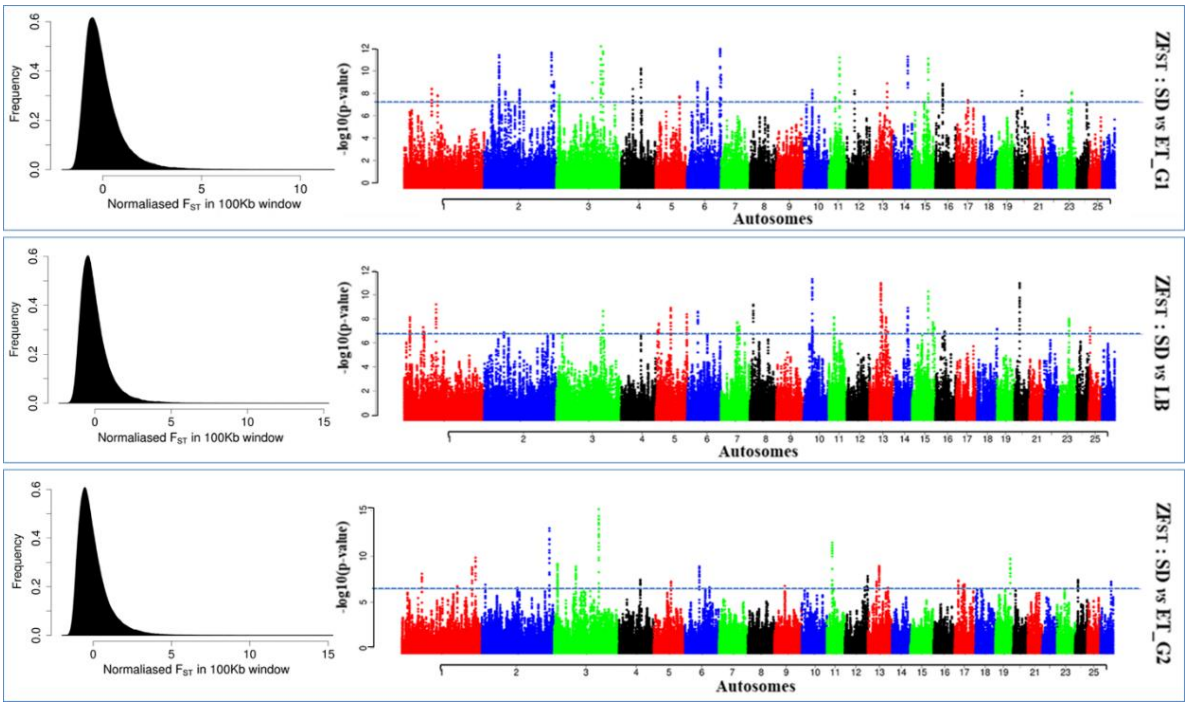


Figure 4.18 Manhattan plot of the genome-wide distribution of ZFST values following the comparison of thin-tail sheep group from Sudan (SD) with Ethiopian fat-rump (ET_G1), Libyan Barbary (LB) and Ethiopian long fat-tail (ET_G2) sheep.

Table 4.4. Candidate regions and genes associated with tail morphology (fat deposition) identified by a combination of at least two methods/comparisons in the Sudanese thin-tail *vs* Ethiopian fat-rump, Ethiopian long fat-tail and Libyan Long fat-tail sheep.

Chr.	Region	Comparison/Method	Genes
1	68620001-68860000	ET_G2 (***ZHp***), SD *vs* ET_G2 (***ZF$_{ST}$***)	*BTBD8, RF00026, C1orf146, GLMN, RPAP2*
	253550001-253740000	ET_G2 vs ET_G1 & SD (***ZF$_{ST}$***)	*RAB6B, SRPRB, TF*
	5930001-6120000	ET_G2 *vs* ET_G1 & LB **(*XP-EHH*)**	*Close to SH3BP4*
	109920001-110020000	ET_G1, ET_G2 & SD **(*ZHp*)**, ET_G2 *vs* LB (***ZF$_{ST}$***)	*CD84*
2	210250001-210440000	SD *vs* ET_G1 & ET_G2 (***XP-EHH***)	*MAP2, ENSOARG00000019133*
	90750001-90950000	ET_G2 **(*ZHp*)**, ET_G2 *vs* LB (***ZF$_{ST}$***)	*Span BMPR2 gene in cow*
	122520001-122620000	SD vs ET_G1 & ET_G2 (***ZF$_{ST}$***)	*FSIP2*
	240720001-240890000	SD vs ET_G1 & LB (***ZF$_{ST}$***)	*RUNX3*
	232480001-233000000	SD, ET_G2 (***ZHp***), SD *vs* ET_G2&LB (***ZF$_{ST}$***)	*PTMA, PDE6D, COPS7B, RF00091, DIS3L2*
	52290001-52530000	SD *vs* ET_G1 (***ZF$_{ST}$***), SD *vs* ET_G1, ET_G2 & LB (***XP-EHH***)	*TMEM8B, FAM221B, HINT2, SPAG8, NPR2, MSMP, RGP1, GBA2, CREB3, TLN1*
	67930001-68120000	SD *vs* ET_G1 & ET_G2 (***XP-EHH***)	*DOCK8*
	73560001-73750000	SD *vs* ET_G1 (***ZF$_{ST}$, XP-EHH***)	*KIAA2026*
3	10920001-11320000	ET_G2 **(*ZHp*)**, ET_G2 *vs* ET_G1 & LB (***ZF$_{ST}$***)	*GOLGA1, ARPC5L, WDR38, RF00026, OLFML2A, MIR181B2, oar-mir-181a-2, NR6A1, NR5A1, ADGRD2*
	20980001-21170000	ET_G2 *vs* ET_G1 & LB (***XP-EHH***)	*ENSOARG00000016157*
	10630001-10820000	ET_G2 *vs* LB & SD (***XP-EHH***)	*GAPVD1, HSPA5, RABEPK, PPP6C, SCAI*
	151770001-151960000	ET_G2 *vs* LB & SD (***XP-EHH***)	*Close to DYRK2*
	154100001-154390000	ET_G2 vs ET_G1 & LB (***ZF$_{ST}$***)	*MSRB3*
	124800001-124910000	ET_G1 & ET_G2 (***ZHp***), SD *vs* LB (***ZF$_{ST}$***)	*Close to KITLG gene*
	10630001-10850000	SD *vs* ET_G1&LB (***ZF$_{ST}$***), SD *vs* LB (***XP-EHH***)	*GAPVD1, HSPA5, RABEPK, PPP6C, SCAI*
	12160001-12410000	SD *vs* ET_G2 (***ZF$_{ST}$***), SD *vs* ET_G2&LB (***XP-EHH***)	*CRB2, RF00402*
	163330001-163520000	SD *vs* ET_G1 & ET_G2 (***XP-EHH***)	*GDF11, ITGA7, METTL7B, MMP19, ORMDL2, PYM1, RDH5, SARNP*
	140340001-140530000	SD *vs* ET_G2 & LB (***XP-EHH***)	*SCAF11, ARID2*
4	69930001-70070000	LB (***ZHp***), SD *vs* LB (***ZF$_{ST}$***)	*Close to HNRNPA2B1*
	73080001-73320000	ET_G2 *vs* SD **(*ZF$_{ST}$, XP-EHH*)**	*Close to ZNF804B*
	96560001-96700000	ET_G2 vs ET_G1 & LB (***ZF$_{ST}$***)	*CHCHD3*

5	107140001-107440000	LB (***ZHp***), SD *vs* LB (***ZF$_{ST}$***)	*WDR36*
	15510001-15700000	SD *vs* ET_G1, ET_G2 & LB (***XP-EHH***)	*ALKBH7, CLPP, CRB3, DENND1C, GTF2F1, PSPN, SLC25A23, SLC25A41, TUBB4A*
6	69060001-69160000	ET_G2 **(*ZHp*)**, ET_G2 *vs* LB (***ZF$_{ST}$***)	*LNX1*
	36000001-36430000	ET_G2 (***ZHp***), SD *vs* ET_G1 (***ZF$_{ST}$***)	*RF02114, NAP1L5, HERC5, HERC6, RF00026, PPM1K*
	10070001-10260000	SD *vs* ET_G1, ET_G2 & LB (***XP-EHH***)	-
	69830001-70100000	LB (***ZHp***), SD *vs* ET_G1, ET_G2&LB (***ZF$_{ST}$***, ***XP-EHH***)	*Close to KIT, PDGFRA*
	115960001-117040000	ET_G1, EG2 & SD (***ZHp***), SD *vs* ET_G1 (***ZF$_{ST}$***, ***XP-EHH***)	*NELFA, NSD2, LETM1, FGFR3, TMEM129, SLBP, GAK, CPLX1, UVSSA, MAEA, SLC49A3, PDE6B, PIGG, RF00427*
7	55860001-56300000	LB (***ZHp***), SD *vs* LB (***ZF$_{ST}$***), SD *vs* ET_G1, ET_G2&LB (***XP-EHH***)	*DMXL2, GLDN, CYP19*
8	51490001-51760000	ET_G2 & SD **(*ZHp*)**, ET_G2 *vs* ET_G1 (***ZF$_{ST}$***)	*TBX18*
9	28360001-28620000	ET_G2, SD (***ZHp***), SD *vs* ET_G1 (***XP-EHH***)	*MTSS1, NDUFB9, TATDN1, RNF139, TMEM65*
10	7260001-7650000	LB, SD (***ZHp***), SD *vs* ET_G1, ET_G2&LB (***XP-EHH***)	*ENSOARG00000006641, ENSOARG00000006632*
	31930001-32120000	SD *vs* ET_G1, ET_G2 (***XP-EHH***)	*FLT1, PAN3*
11	10750001-10940000	SD (***ZHp***), SD *vs* ET_G1, ET_G2&LB (***XP-EHH***)	*INTS2, BRIP1*
	18190001-18490000	LB (***ZHp***), SD *vs* ET_G2 (***ZF$_{ST}$***), SD *vs* ET_G1, ET_G2 &LB (***XP-EHH***)	*NF1, EVI2B, OMG*
	26350001-26540000	SD *vs* ET_G1, ET_G2 & LB (***XP-EHH***)	*ALOX12, ASGR1, ASGR2, BCL6B, CLEC10A, MIR195, MIR497, RF00026, RF00408, RNASEK, SLC16A11, SLC16A13*
	37240001-37430000	SD *vs* ET_G1 & LB (***ZF$_{ST}$***, ***XP-EHH***), ET_G2 *vs* ET_G1 (***ZF$_{ST}$***), ET_G2 *vs* ET_G1 & LB **(*XP-EHH*)**	*CALCOCO2, TTLL6, HOXB13, RF02133, RF02132, MIR196A1, HOXB9, HOXB7*
	27040001-27180000	ET_G2, SD (***ZHp***), SD *vs* ET_G1, ET_G2 & LB (***XP-EHH***)	*DNAH2, NAA38, KDM6B, TMEM88, CYB5D1, CHD3, RF00602, RNF227, KCNAB3, TRAPPC1, CNTROB*
12	75850001-75990000	ET_G2 *vs* SD & LB **(*ZF$_{ST}$*, *XP-EHH*)**	*MIR181B1, oar-mir-181a-1*
	49560001-49700000	SD *vs* ET_G1 & LB (***ZF$_{ST}$***, ***XP-EHH***)	*PERM1, PLEKHN1, KLHL17, NOC2L, SAMD11, ZBTB17*

13	38600001-38890000	ET_G2 (***ZHp***), SD *vs* ET_G2 (***ZF$_{ST}$***) , ET_G2 *vs* SD & LB (***ZF$_{ST}$***)	*RIN2, NAA20, CRNKL1, CFAP61*
14	34380001-34670000	ET_G1, SD (***ZHp***), SD *vs* LB (***XP-EHH***)	*HSD11B2, ATP6V0D1, AGRP, RIPOR1, CTCF, CARMIL2, ACD, PARD6A, ENKD1, C16orf86, GFOD2, RANBP10, TSNAXIP1*
18	23370001-23560000	SD *vs* ET_G1, ET_G2 & LB (***XP-EHH***)	*SAXO2, EFL1*
19	23180001-23420000	ET_G2 (***ZHp***), SD *vs* ET_G1, ET_G2 & LB (***XP-EHH***)	*CNTN4*
22	50510001-50760000	ET_G2 (***ZHp***), ET_G2 *vs* LB (***ZF$_{ST}$***)	*CYP2E1, ECHS1, FUOM, PAOX*
	27380001-27570000	SD *vs* ET_G1 & ET_G2 (***XP-EHH***)	*ENSOARG00000026748*
23	43340001-43530000	SD *vs* ET_G1 & LB (***XP-EHH***)	*CEP76, PSMG2, RF00410, PTPN2*
	1030001-1220000	ET_G2 *vs* ET_G1 & LB (***XP-EHH***)	*ENSOARG00000026145, Close to SALL3*
24	24800001-24990000	ET_G2 *vs* ET_G1 (***XP-EHH***)	*NSMCE1, IL4R, IL21R*
25	4060001-4180000	ET_G1 (***ZHp***), ET_G2 *vs* ET_G1 & LB (***ZF$_{ST}$***)	*GNPAT, EXOC8, SPRTN, EGLN1*

Compared to the first set of comparisons (**Figure 4.14A)**, in addition to the reported candidate region that is spanning *HOXB13*, the second contrasting procedure (**Figure 4.14, B)** resulted in several new candidate regions under positive selection (**Table 4.4**, **Figures 4.17 and 4.18**). Among which, candidate region on OAR2 included 10 genes (*TMEM8B, FAM221B, HINT2, SPAG8, NPR2, MSMP, RGP1, GBA2, CREB3,* and *TLN1*). It is detected following by ZF_{ST} and *XP-EHH*. In addition, a strong selection sweep spanning 9 genes (*ALKBH7, CLPP, CRB3, DENND1C, GTF2F1, PSPN, SLC25A23, SLC25A41, TUBB4A*) differentiates SD thin-tailed sheep from the three other groups on OAR5. Moreover, the three methodologies identified selection sweep on OAR7 covering three candidate genes (*DMXL2, GLDN, CYP19*). One candidate gene (*CNTN4*) was identified *via ZHp and XP-EHH* within a candidate region on OAR19. Additionally, five candidate regions under selection were identified on OAR11 (**Figure 4.19**).

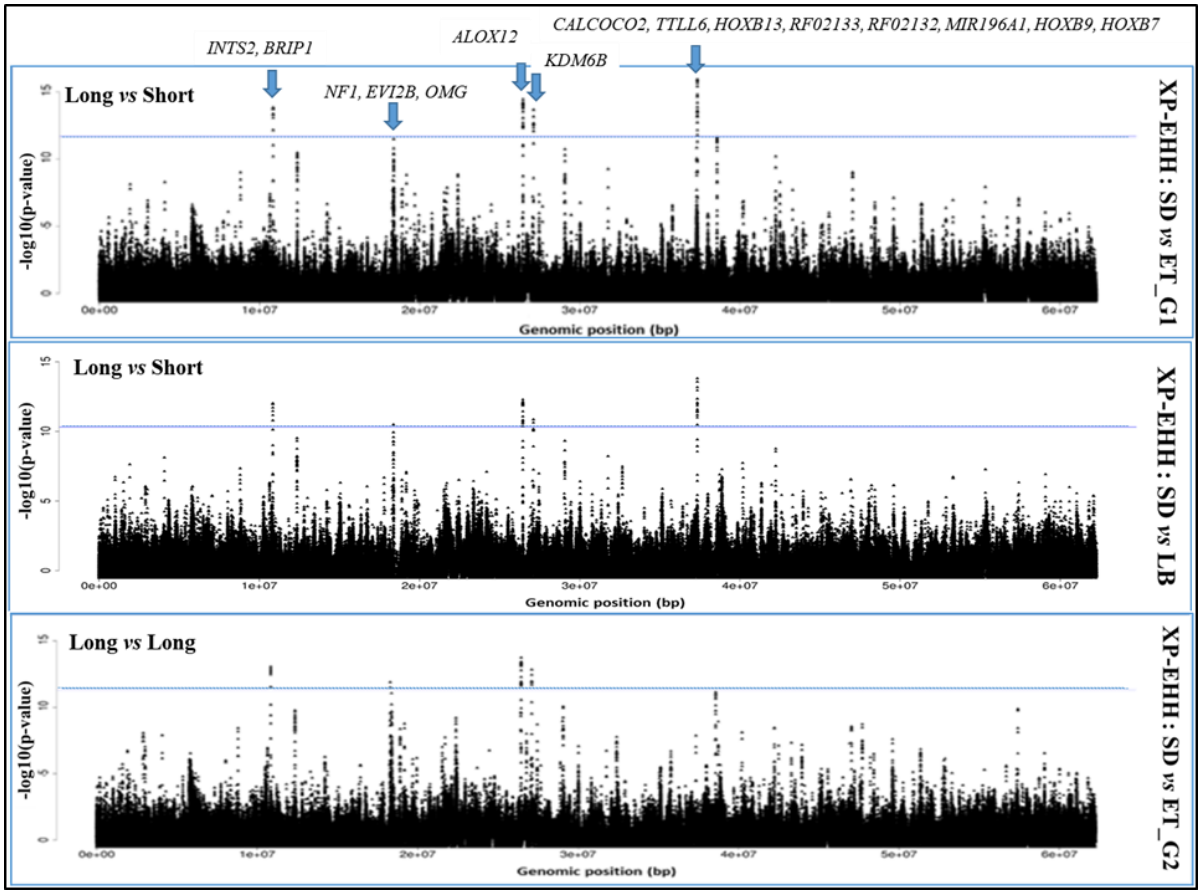


Figure 4.19. Signature of selection (*XP-EHH*) analysis on OAR11 showing the candidate selected region (comparison **B, Figure 4.14**). It includes the candidate regions, including the candidate region with the *HOXB13* gene, previously identified in connection to the length of the tail as well new candidate regions including genes related to fat metabolism (e.g. *ALOX12, NF1*, *EVI2B* and *OMG*).

Among these, starting from the beginning of the chromosome 11, the first selection sweep, supported by *ZHp* and *XP-EHH*, spanned two genes (*INTS2, BRIP1*). It is followed by a candidate region including *NF1, EVI2B* and *OMG* genes supported by the three approaches (**Figure 4.19).** A candidate region harbouring 12 genes (*ALOX12, ASGR1, ASGR2, BCL6B, CLEC10A, MIR195, MIR497, RF00026, RF00408, RNASEK, SLC16A11 and SLC16A13*) (**Figure 4.19)** with *ALOX12* known to be linked with obesity in humans. A 570 kb candidate region, including 11 genes (*DNAH2, NAA38, KDM6B, TMEM88, CYB5D1, CHD3, RF00602, RNF227, KCNAB3, TRAPPC1,* and *CNTROB*), with *KDM6B known to* play a vital role in regulating the expression of *HOX* gene family.

Discussion

In this chapter, we aim to identify genomic regions and candidate genes associated with both environmental challenges and tail morphology by contrasting the genomes of different sheep populations from East Africa (Ethiopia), Sudan and North Africa (Libya). Building on the previous results of **Chapter 2** (**Figure 3.1 and Table 3.1**) and the full genome sequences analysis of SNPs polymorphisms performed here, the populations were clustered in four groups: ET_G1 (almost Ethiopian fat-rump), ET_G2 (Ethiopian long fat-tail), SD (Sudanese thin-tail) and LB (Libyan long fat-tail). Interestingly, such grouping reflects to some extent the separation of the population based on their environments and tail phenotypes (**Table 4.1 and Figure 4.13**).

Three methodologies (*ZHp*, *ZFST* and *XP-EHH*) to identify candidate selection sweep regions/genes under positive selection associated with environmental challenges and tail morphology were used. For the environmental challenges, we considered the ET_G2 group to represent the sheep from the high altitude humid area (Gizaw et al., 2007), with sheep occupying an area at sea level (LB) considered as the most contrasting group overall in the term of altitude, temperature and annual precipitation (arid, average annual precipitation < 400 mm) (Zeleriakoval et al., 2012) with, however, the Sudanese sheep living in a desert environment (average annual precipitation < 10 mm) (El-Hag et al., 2001). The ET_G1 group of Ethiopian sheep occupies different agro-ecosystems including high and low zones as well as humid and arid ones. Therefore, this later group was only included for the morphology analysis and not the environmental ones.

Following these distinctions, the genome profile for each group was scanned using first the *ZHp* analysis (**Figure 4.1**). We then performed ZF_{ST} and *XP-EHH* analysis contrasting the genomes of these groups in pair-wise comparisons (**Figure 4.2, 4.3 and 4.4**). The environmental analysis identified 49 putative regions spanning 170 candidate genes possibly under positive selection (**Table 4.2**). Among these genes, 81 genes were previously identified in signatures of selection signals reported in **Chapter 2** (Ahbara et al., 2019). Interestingly, 19 genes (*DIS3L2, NF1, EVI2B, OMG, HOXB9, HOXB7, TTLL6, HOXB13, CHD3, KCNAB3, TRAPPC1, CNTROB, BMPR2, GLDN, CALCOCO2, CRB2, LIN28B, INTS2, BRIP1*) coincident with candidate genes revealed under selection using whole-genome sequencing in Chinese native sheep living in desert and arid environments (Yang et al., 2016). Among which, four genes (*DIS3L2, NF1, EVI2B, OMG*) were present in candidate selected regions for both

dessert and arid sheep populations, whereas 10 genes (*HOXB9, HOXB7, TTLL6, HOXB13, CHD3, KCNAB3, TRAPPC1, CNTROB, BMPR2* and *GLDN*) are associated with desert sheep and five genes (*CALCOCO2, CRB2, LIN28B, INTS2, BRIP1*) with the arid environment populations.

The *DIS3L2* (DIS3 like 3'-5' exoribonuclease 2) gene on OAR2 is identified within a strong candidate region and it is supported by the three methodological analysis (**Table 4.2**), both SD and LB groups are sharing a similar haplotype at the region of the gene (**Figure 4.6)**. This ribonuclease is involved in the regulation microRNA interaction and it is likely conserved in eukaryotic species (Malecki et al., 2013). This gene has been also identified as a strong candidate gene under selection in **Chapter 2** (Ahbara et al., 2019). Furthermore, it was identified among genes associated with animal height variation in sheep by de Simoni Gouveia et al., (2017). Also, *DIS3L2* has been associated with different phenotypes e.g. the Perlman syndrome in human, which is characterized by overweight individuals (Astuti et al., 2012) and the disruption of wing imaginal discs during *Drosophila* development (Towler et al., 2015). It might also be associated in humans with the hypoxic effects of living at high altitude, and/or reflects might reflect a genetic adaptation to such stress (Frisancho and Baker, 1970). In sheep living in desert and arid ecosystems walking long distances is required because of grazing shortage, these sheep have developed longer legs that may be reflected on animal stature (Mufarrih, 1991).

Additionally, we identified two genes (*HSPA5, MSRB3*) in candidate selective sweeps on OAR3. The heat shock protein family A (Hsp70) member 5 (*HSPA5*) was identified by *XP-EHH* in a candidate region differentiating ET_G2 that occupying high-altitude zone from the other contrasting SD and LB sheep groups inhabiting desert and arid environments, respectively. In general, expression of the Hsp70 gene family has been reported to be positively correlated with variation in thermotolerance in different organism (He et al., 2013; Nagayach et al., 2017), and the genes is believed to play a major role in cellular heat stress responses (Pennarossa et al., 2012). The second gene (*MSRB3*) was supported by both the ZF_{ST} and *XP-EHH* analysis. This gene (methionine sulfoxide reductase B3) has been reported to be associated with ear morphology in sheep (Wei et al., 2015), dogs (Boyko et al., 2010; Vaysse et al., 2011) and pig (Chen et al., 2018). It is interesting to note that a polytypic (morphologically variable) species will be expected to show shorter extremities of limbs and ears in the predominant cold area with those inhabiting warmer climates having longer

extremities and ears (Allen, 1877). Interestingly, this is consistent with the haplotype diversity and ear phenotypes observed in the different sheep groups at the selective sweep of *MSRB3* gene (**Figure 4.7**).

Candidate genes (*NF1, EVI2B,* and *OMG*) were identified on OAR11 by the three methodologies in both the desert and arid sheep ecotypes (**Table 4.2**). These genes are associated with tail type and fat deposition in sheep (Moioli et al., 2015; Wei et al., 2015; Yuan et al., 2017; Ahbara et al., 2019). These results are highly supported by the shared haplotypes between LB and SD sheep groups in the region spanning these genes (**Figure 4.8**).

In addition, a strong candidate region spanning four genes (*LNX1, CHIC2, GSX2* and *PDGFRA*) was identified by three signature of selection indexes. Among which, the platelet derived growth factor receptor alpha (*PDGFRA*) was previously identified under selection in Libyan Barbary sheep using low-density SNP data and it was reported to be associated with tail phenotype and involved in nematode infection resistance (Mastrangelo et al., 2018). These previous results are clearly supported by our haplotype structure analysis where the LB sheep group show a distinct haplotype around the gene within the candidate region (**Appendix Figure 5.5**). This gene is also included in two gene ontology terms, luteinisation (GO:0001553) and cell chemotaxis (GO:0060326) (**Table 4.3**).

Libyan Barbary (LB) sheep also differentiated themselves from the ET_G2 group in two candidate regions spanning *PDS5B* and *RXFP2* on Oar10 (ZFst and XP-EHH). The *PDS5B* (PDS5 cohesin associated factor B) has been identified among sheep tail associated genes (Yuan et al., 2017), while the relaxin family peptide receptor 2 (*RXFP2*) is highly associated with horn size in domestic ungulates (Kardos et al., 2015; Ahbara et al., 2019). Interestingly, the haplotype diversity plot shows a distinct haplotype near fixation, which differentiates the LB sheep from the other sheep groups at the *PDS5B* gene location (**Figure 4.9**).

As for the *RXFP2* gene, according to Hoefs (2000), the estimated index of heat-exchange capacity, which is the relationship between the core surface of sheep horns (estimated from the horn core length and circumference data) and animal mass, shows an evolutionary trend of variation of the horn core size as a response to ambient temperature. This trend is independent of a parallel trend of increased horn size for the benefit of enhancing reproductive success (Johnston et al., 2011). In the recent decades, based on the author's observation, sheep owners in Libya build their assessment of the chosen rams for mating on the horn size and tail fatness:

the larger horns and tail the more chance to be selected as a matting ram. From their point of view, rams of larger horns and tails are the most probable producers of strong and healthy progenies, and so, male-male competitions are widely practised in order to choose the ram for mating based on these criteria (**Appendix Figure 5.8**).

Interestingly, both *PDS5B* and *RXFP2* selection signal are within a close distance from each other (570 Kb) (**Figure 4.9**). This pattern of selection sweeps, along with the reported literature and the mentioned above human selection criteria for sheep matting support parallel evolutionary and human-mediated selection in both horn size (fighting adaptation) and fat deposition in tails (adaptation to heat stress) in LB sheep group.

In a similar way, among the other candidate selected regions which differentiate the LB sheep group we do find a region on OAR15 (ZF_{ST} and *XP-EHH* analysis) spanning the pleckstrin homology domain containing A7 (*PLEKHA7*) gene. This gene was reported previously as a candidate gene which attenuates salt-sensitivity hypertension and renal disease in rats (Mattson et al., 2008; Endres et al., 2014), associated with blood pressure and kidney function in humans (Taal et al., 2012) and blood pressure and/or hypertension in Chinese people from Ningde City in Fujian province (Lin et al., 2011).

Pastures in Libya represent 86.36% of the total agriculture area (Bazza et al., 2018). Salinisation is the main type of soil degradation (30 %) (Abagandura et al., 2017). Furthermore, salinity is the main concern in the regard to water quality following intrusion of seawater in the coastal lands, where salinity levels records have shown an increase from 150 ppm in 1950 to over 5000 ppm in 1990 (FAO, 2016). This increasing trend of salinity coincides with the reported increase of drought days along the Libyan coastline (Busby et al., 2010), which along with the increased evaporation may directly lead to a reduction in the runoff and groundwater levels (Bindra et al., 2013). All these factors are likely a consequence of global warming.. Interestingly, at the genome level, this increased water and soil salinity are supported by the ongoing selection indicated by *XP-EHH* analysis and haplotype structure at the region that spans *PLEKHA7* in the Libyan sheep (LB) population (**Figure 4.10**).

The function of *PLEKHA7* as a candidate gene for salinity adaptation is also supported by our gene ontology (GO) analysis (**Table 4.3**). Among the highly enriched GO terms, we identified a cluster of genes (*NF1, PDGFRA, NR5A1*) that are associated with the adrenal gland development (GO:0030325). The adrenal gland plays a major role in the homeostatic

adaptations in an integrated mode with other endocrines glands (Harvey et al., 1984). The overexpression of *NF1* (neurofibromin 1) results in lower transcription of *HSD11B2* with a possible association with salt sensitivity (Alikhani-Koupaei et al., 2007) and with vasculopathy, hypertension and congenital heart defects in humans (Friedman et al., 2002). *NR5A1* has been reported to play a crucial role in the aldosterone synthase gene (*CYP11B2*) expression (Takeda et al., 2018). Aldosterone plays a key role in salt-sensitive hypertension development in human (Fujita, 2008), and greater synthesis of aldosterone was observed in Dahl salt-sensitive (SS) rats compared with normotensive Sprague–Dawley (SD) rats (Gomez-Sanchez et al., 2010).

Common candidate selected genes (*DIS3L2*, *NF1, EVI2B* and *OMG*) in Libyan and Sudanese sheep (*ZHp* and *XP-EHH*) analysis **(Table 4.2)** and haplotype structure results (**Figure 4.6 and 4.8**) coincident with population distribution, topography and the predominant climatic conditions across both arid and desert environmental challenges. In the case of Libyan Barbary sheep, they represent 95 % of the total sheep reared with unlimited stock exchange between the different climatic zones (more specifically the low-altitude) across the country (Abdulkarim, 2015). Therefore, the population occupies different climatic conditions from humid in the western and eastern mountains, to arid conditions in the mid-costal region with gradual expansion across the southern area towards the completely desert conditions (Zeleriakoval et al., 2012). These signatures of selection signals may be related to the adaptabilities of Libyan sheep to both arid and desert conditions.

At the other hand, among the candidate genes identified specifically in the desert Sudanese sheep group, the *TTLL6*, *HOXB13*, *RF02133*, *RF02132*, *MIR196A1*, *HOXB9* and *HOXB7* genes have been also associated with tail formation (Economides et al., 2003; Mallo et al., 2010; Ahbara et al., 2019). Interestingly, these genes are found within a candidate region positioned upstream the fat deposition selective sweep (*NF1*, *EVI2B* and *OMG*) on the same autosomal chromosome (OAR11). These results support co-evolution between tail morphology and fat deposition. The *KCNAB3* (potassium voltage-gated channel subfamily A regulatory beta subunit 3) is among the desert related genes identified on OAR11. Its expression has been reported to be down-regulated expression in rats under chronic mild stress (Terenina et al., 2019) and among ion channel associated genes that highly expressed in ventral respiratory column in rats (González-Castillo et al., 2018). The *CNTROB* (centrobin, centriole duplication and spindle assembly protein) gene has been also identified on the same autosomal region, this

gene has been associated with skeletal alterations and limb phenotype in rats (Liška et al., 2013).

Our analysis has also identified nine genes (*NF1, EGLN1, BRIP1, VEGFA, TF, PLCG1, RELA, KIT* and *DSC2*) possibly associated with the high-altitude challenge (**Table 4.2**). Among which, three genes (*NF1, KIT* and *DSC2*) have been reported as candidate genes responsible for hypoxia adaptation in sheep (Wei et al., 2016; Yang et al., 2016). The serotransferrin gene (*TF*) was identified by ZF_{ST} analysis in a candidate region which differentiate ET_G2 from the other sheep groups on OAR11. The *TF* gene was identified among the highly expressed genes in the acute phase response of Jersey and Holstein cattle exposed to high-altitude hypoxia (Kong et al., 2019a; Kong et al., 2019b). Haplotype structure supports selection at this gene, with a high frequency distinct haplotype in the ET_G2 group (**Figure 4.11**). The neurofibromin 1 (*NF1*) gene has been reported to have an angiogenesis exaggeration effect indicated by increased neovascularization in both the retina and cornea in the *NF1* heterozygous mice in response to hypoxia compared to the wild-type (Wu et al., 2006). However, it seems that the *NF1* gene acts differently based on the inhabited environment. It is reflected on the haplotype structure observed here as both LB and ET_G2 have selected haplotypes indicating selective sweep but distinct one **(Figure 4.8)**.

The BRCA1 interacting protein C-terminal helicase 1 (*BRIP1*) has been reported to have a protective effect on rat cardiomyocytes against hypoxia/reoxygenation, reactive oxygen species production and apoptosis (Shen et al., 2019). A strong selection sweep was identified by the ZF_{ST} and *XP-EHH* analysis spanning the *VEGFA* hypoxia related gene on OAR20. This gene has been suggested to be associated with cerebrovascular adaptation to chronic hypoxia in sheep (Castillo-Melendez et al., 2015) and protects pregnant sheep from hypoxic effects by increasing blood flow in the uterine arteries (Mehta et al., 2011). Surprisingly, the haplotype structure plot shows an identical shared haplotype near fixation and partly selected spanning *VEGFA* gene in ET_G2 and SD sheep groups, respectively, against an admixed haplotype in LB group (**Appendix Figure 5.6**).

In addition, we identified a strong selection sweep peak on OAR25 including the *EGLN1* gene, which function is also most likely associated with response to hypoxia challenge. It has been shown that positively selected haplotypes of *EGLN1* (Egl-9 Family Hypoxia Inducible Factor 1) were significantly associated with the low haemoglobin content of Tibetan people, which is a unique characteristic of this extreme high-altitude population (Simonson et al., 2010). In

Tibetan sheep, the potential high-altitude hypoxia adaptation mechanisms include the role of *EPAS1* (Endothelial PAS Domain Protein 1) (Wei et al., 2016). Both *EGLN1*and *EPAS1* candidates are key genes found upstream the hypoxia inducible factor (*HIF*) pathway with functional mutations at *EGLN1* present in Tibetan people (Simonson et al., 2010; Xiang et al., 2013; Lorenzo et al., 2014) and at *EPAS1* in both Tibetan sheep and people (Xu et al., 2014; Wei et al., 2016). However, surprisingly at this region, following our haplotype structure analysis we found that the SD sheep group, which inhabits low-altitude areas share an identical haplotype with populations occupying high-altitude zones (ET_G2), while sheep inhabiting sea level area (LB) have a different set of admixed haplotypes (**Figure 4.12**). A result further supported by the estimated allele frequency for each group. The highest allele frequencies in ET_G2 and SD sheep are located within the *EGLN1* region. It requires further investigation noting that it might be consistent with the reported seasonal mobility of nomadic flocks of desert sheep across the sub-Saharan between low and high altitude areas (El-Hag et al., 2001).

Furthermore, the desmocollin 1 and 2 genes (*DSC1, DSC2*) are present within the selection peak on OAR23 (ZF_{ST} and *XP-EHH* analysis). They have been reported to regulate fungal response to hypoxia in *Histoplasma capsulatum* (Dickerhoof, 2014) and *Aspergillus fumigatus* (Willger et al., 2012; Bat-Ochir et al., 2016). These fungi species are spread widely in nature where they are typically found in moist soil and decaying organic matter, such as compost heaps, where they play an essential role in carbon and nitrogen recycling (Fang and Latgé, 2018). Interestingly, the haplotype structure of the candidate region of *DSC1* and *DSC2* shows a homogeneous haplotype in ET_G2, while several haplotypes are present in the other sheep groups (**Supplementary Figure 5.7**). These suggest that the ET_G2 group most likely has adapted to the humid environment of the Ethiopian highlands, while the SD sheep group seems to be more adapted to the dry and desert highlands environments.

The reported genes related to high altitude adaptation were highly enriched according to the gene ontology analysis (**Table 4.3**). The GO showed that four of the examined genes (*NF1, EGLN1, BRIP1, VEGFA*) are related to the hypoxia response to oxygen levels (GO:0070482) and general response to hypoxia (GO:0001666). Moreover, we found 5 genes among the candidate genes (*TF, PLCG1, RELA, VEGFA,* and *EGLN1*) that have a key role function in *HIF-1* signaling pathway, associated with oxygen consumption as a response to hypoxia in high-altitude (**Figure 4.5**). Under hypoxic conditions, the elevated expression of *TF* and

VEGFA genes increase oxygen delivery by stimulating iron metabolism and angiogenesis, respectively, according to the molecular positive feedback received from *ELNG1* gene.

To identify candidate regions associated with tail morphology, we integrated the phenotypic measurements of the tails with the genetic variation clusters following the population genetic analysis from **Chapter 3** (**Figure 4.13**). Archaeological records indicate two origins for the African sheep: the North of Africa with two migration waves represented here by the thin tail sheep from Sudan (SD) and the long fat-tail from Libya (LB), while sheep introduced into the Eastern part of the continent, possibly also in two migratory waves, are represented here by the long fat-tail (ET_G2) and fat-rump (ET_G1) groups. This is in agreement with the comparative visual length at hock level (**Figure 4.13)**. While the thin tail of SD group (TT) expands below the hock level, the long fat-tail (LFT) sheep from North (LB) and East (ET_G2) Africa have nearly visual tail length at the hock level while it is much shorter in the fat-rump (FR) sheep from East Africa. Interestingly, while the estimated number (22) of caudal vertebrae (C.V) was similar in the North Africa sheep (SD and LB), we observe a longer tail in the SD group compared to LB sheep group. Also, we observe a longer tail in the ET_G2 long fat-tail sheep (LFT) compared to the East African fat-rump sheep group (number of vertebrae and visual observation).

Building on these observations, we estimated the average length of each single vertebra within each group. Subsequently, based on these estimations, we observe nearly similar average vertebra length for the LB and ET_G1groups, 2.20 cm and 2.24 cm, respectively. At the opposite, the SD and ET_G2 sheep groups show longer vertebrae compared to the LB groups despite that the later have a similar number of vertebrae compared to ET_G2 sheep group (**Figure 4.13**). Then, to identify candidate regions associated with tail morphology (length and fat amount), we combine SD and ET_G2 group as representing long-tailed sheep from North and East origin, respectively, with the groups LB and ET_G1 representing the short-tailed sheep from the same regions. Thereafter, we perform selection scans within each group using *ZHp* analysis followed by runs of signature of selection using *FST* and *XP-EHH* (**Figure 4.14**). Following this approach, we were able to differentiate selective sweep signals for tail length (**Figure 4.14, A**) from tail fat deposition (**Figure 4.14, B**) with selective signature signal for tail length obtained only in comparison **A** and for both tail length and fat deposition in comparison **B**.

Fifty-four candidate regions spanning 183 candidate genes were commonly detected between different selection scan methodologies and/or groups comparisons **(Table 4.4).** Almost all these candidate genes were identified in comparison **B** (**Figure 4.12, B)**. Among fat deposition related candidates (**Table 4.4**, **Figures 4.15 and 4.16)**, 10 genes are found within a strong selection sweep on OAR2, three of which (*HINT2, SPAG8, NPR2*) associated previously with fat deposition using low-density SNP Chip data (see details in **Chapter 2**). The *TUBB4A* (tubulin beta 4A class Iva) gene was identified on OAR5 and it has been reported among the genes that may relate to fat metabolism and the accumulation of lipids in human cells (Guo et al., 2017; Heischmann et al., 2017). In addition, *Dmx like 2* (*DMXL2*) is found within a strong candidate region on OAR7. Transcriptome profile analysis indicates that this gene in among genes associated with fat deposition and meat quality in sheep (Wang et al., 2014). The *CNTN4* (contactin 4) genes on OAR19 has been associated with obesity, lipids and blood pressure traits according to association studies (Kraja et al., 2012) and analysis of DNA methylation of adipose tissue in human (Benton et al., 2015).

Moreover, on OAR11, we identified five candidate regions under selection with both *FST* and *XP-EHH* analysis. One candidate region was detected by in both **A** and **B** comparisons, while the remaining ones were identified in **B** comparison only (**Figures 4.14, 4.17 and 4.18)**. They include the *NF1, EVI2B* and *OMG* genes, which were reported among the candidate genes that play an important role in fat tail formation in Chinese sheep (Yuan et al., 2017) and *ALOX12* (arachidonate 12-lipoxygenase, 12S type) that contributes to the variation in the obesity phenotypes in young Chinese men (Xiao et al., 2010). These were identified on two subsequent candidate regions in the second comparison (comparison B). Also of interest is *KDM6B* (lysine demethylase 6B), which play a crucial role in the regulation of the expression of the *HOX* gene family (McLaughlin-Drubin et al., 2011; Ye et al., 2012; Bhatlekar et al., 2018). This family of genes has an exclusive ability to regulate morphologies along the anteroposterior (AP) axis in *Drosophila* (Lewis, 2004). In mice, the controlled "spatio-temporally" expression of *HOX* genes are important regulators of embryonic development (Mallo et al., 2010).

Interestingly, three genes (*HOXB7, HOXB9, and HOXB13*) belonging to this family are present in a signature of selection region on OAR11. The haplotype structure of this candidate region shows an identical haplotype near complete fixation shared between the two long-tailed ET_G2 and SD sheep groups originating from the East and North Africa, respectively (**Figure 4.19**).

The nearly fixed haplotype covering *HOXB13* (homeobox B13) supports causative mutation in this gene leading to the phenotypic variation in the studied population's tail length. These results are in well agreement with a previous study in mice that indicate that mutations at this gene are responsible for the overgrowth of the caudal spinal cord and the tail vertebrae (Economides et al., 2003). In addition to *HOXB13*, we identified two genes (*NR6A1* and *TBX18*) possibly associated with tail vertebrae growth and/or number. *NR6A1* (nuclear receptor subfamily 6 group A member 1) is present within a candidate selected region in OAR3. In Licha Black and possibly Laiwu pigs pig, it is a strong candidate gene associated with increased vertebral numbers (Yang et al., 2009). It has also been associated with increased vertebrae number and carcass traits in Kazakh sheep (Zhang et al., 2017). The *TBX18* (T-Box 18) gene was identified close to a candidate region under selection on OAR8 *via* ZF_{ST} in addition to *ZHp* analysis in both ET_G2 and SD long-tailed sheep groups **(Table 4.4)**. This gene is highly expressed in the dorsal and ventral regions of the somites in zebrafish (Begemann et al., 2002). Its expression is required to maintain craniocaudal sclerotome polarity at the somite stage of the embryonic development (Scaal, 2016). The whole-genome sequencing of Hulunbuir Short-Tailed Sheep has identified a functional mutation in the *T* gene responsible for the extreme short-tail phenotype of the breed (Zhi et al., 2018). These signatures of selection need to be functionally validated especially the ones relating to *HOXB13*.

Considering all our results, it may be argued that interaction between genes regions, whether on the same chromosome or located on different autosomal chromosomes driven by the environmental condition are responsible for the diversity of sheep tail phenotypes. For example, on OAR11 the expression of *KDM6B* within a candidate selected region close to other candidate regions associated with fat accumulation may regulate the expression of *HOXB13* (**Figure 4.18)**.

Furthermore, in both our environmental adaptation and tail morphology results, we identified a strong candidate region on OAR1 supported by *ZHp* scans in the ET_G1, ET_G2 and SD sheep groups. This candidate region includes *CD84* (CD84 molecule), a candidate gene for trypanosomiasis resistance in mice (Yan et al., 2004; Martin and Tarleton, 2005). Similarly, a selection sweep spanning 14 genes *(NELFA, NSD2, LETM1, FGFR3, TMEM129, SLBP, GAK, CPLX1, UVSSA, MAEA, SLC49A3, PDE6B, PIGG,* and *RF00427*) on OAR6 includes three candidate genes (*LETM1, SLBP,* and *PDE6B)* also most probably associated with trypanotolerance. This region is supported by *ZHp* analysis in the ET_G1, ET_G2 and SD

sheep groups, tropical sheep population which may have been exposed to trypanosomiasis challenges at the opposite of the North African Libyan (sheep). Among these trypanotolerant candidate genes *LETM1* (leucine zipper and EF-hand containing transmembrane protein 1), also called Trypanosome Letm1 Protein, plays an essential function in the infection stages of *Trypanosoma brucei* maintaining the parasite mitochondrial integrity (Nowikovsky et al., 2012; Hashimi et al., 2013; Nowikovsky and Bernardi, 2014). Further investigations are required to confirm the significance of these results in relation to trypanosomiasis tolerance/resistance.

It is worth pointing that some of the selection signatures found within the candidate clusters could be endogenous, but an adaptive introgression may cause others (especially the tail morphology related ones). In addition, additional selection signatures might be identified by analysing the dataset breed by breed and among breeds within genomic clusters.

Conclusion

Overall, our results support possible interaction between environmental challenges and tail morphology in sheep occupying diverse environments. From the environmental adaptation perspective, our studied populations illustrated different adaptive responses towards environmental challenges (arid, desert and high-altitude). While in the Libyan sheep (LB) we found candidate signatures of selection with genes of function related adaptation towards both the arid and desert conditions including increased levels of salinity in water and vegetation, we have identified in the Ethiopian long fat-tail sheep (ET_G2) selected genome region with genes linked to survivability under the humid high-altitude ranges. In the thin tail sheep from Sudan (SD), the presence in the genome of both selected genes links to the adaptation to high-altitude and desert environments could be related to the movement of these sheep populations across contrasting environments in the sub-Saharan region. Selection for fat tail deposition and horn size in Libyan sheep support both environmental and human-mediated selection. Moreover, our whole-genome sequences analysis support interaction between several genomic regions for the tail phenotype in sheep. Many candidate genes possibly under selection were identified by our analysis, among which, the salt-sensitivity related gene (*PLEKHA7*) in Libyan sheep, the hypoxia associated gene (*EGLN1*) in Ethiopian sheep and the *HOXB13* most likely are playing major rules in the unique adaptations of African sheep. To the best of our knowledge, we are the first reporting these candidate genes in sheep. Future direction in our work will include the search of the causative mutation at these genes.

References

Abagandura, O., Gandura, Park, D., White, D., and Bridges, W. (2017). Modelling Soil Degradation in Libya. *J. Nature Sci. Res.* 7, 30-40

Abdulkarim, A. (2015). Small ruminant contribution in meat production in Libya. *Egyptian Journal of Sheep and Goats Sciences*, 10, 1-6.

Ahbara, A., Bahbahani, H., Almathen, F., Al Abri, M., Agoub, M. O., Abeba, A., et al. (2019). Genome-wide variation, candidate regions and genes associated with fat deposition and tail morphology in Ethiopian indigenous sheep. *Front. Genet.* 9, 699.

Alikhani-Koupaei, R., Fouladkou, F., Fustier, P., Cenni, B., Sharma, A. M., Deter, H.-C., et al. (2007). Identification of polymorphisms in the human 11beta-hydroxysteroid dehydrogenase type 2 gene promoter: functional characterization and relevance for salt sensitivity. *FASEB J.* 21, 3618-3628.

Allen, J. A. (1877). The influence of physical conditions in the genesis of species. *Radic. Rev.* 1, 108-140.

Almeida, A. M. (2011). “By endurance we conquer”: fat-tailed sheep in the twenty-first century. *Trop. Anim. Health. Prod.* 43, 1233-1235.

Bat-Ochir, C., Kwak, J. Y., Koh, S. K., Jeon, M. H., Chung, D., Lee, Y. W., et al. (2016). The signal peptide peptidase SppA is involved in sterol regulatory element-binding protein cleavage and hypoxia adaptation in *Aspergillus nidulans*. *Mol. Microbiol.* 100, 635-655.

Bazza, M., Kay, M., and Knutson, C. (2018). "Drought characteristics and management in North Africa and the Near East." Water for Food, Dauthery Global Institute, University of Nebraska, Rome, Italy.

Begemann, G., Gibert, Y., Meyer, A., and Ingham, P. W. (2002). Cloning of zebrafish T-box genes tbx15 and tbx18 and their expression during embryonic development. *Mech. Dev.* 114, 137-141.

Benton, M. C., Johnstone, A., Eccles, D., Harmon, B., Hayes, M. T., Lea, R. A., et al. (2015). An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. *Genome Biol.* 16, 8.

Bessho, Y., Sakata, R., Komatsu, S., Shiota, K., Yamada, S., and Kageyama, R. (2001). Dynamic expression and essential functions of Hes7 in somite segmentation. *Genes Dev.* 15, 2642-2647.

Bhatlekar, S., Fields, J. Z., and Boman, B. M. (2018). Role of HOX genes in stem cell differentiation and cancer. *Stem Cells Int.* 2018, 1–15.

Bindra, S., Abulifa, S., Hamid, A., Al Reiani, H., and Abdalla, H. K. (2013). Assessment of impacts on ground water resources in Libya and vulnerability to climate change. *Scientific Bulletin of the "Petru Maior" University of Targu Mures* 10, 63-69.

Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., and Lohmueller, K. E. (2010). A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol.* 8, e1000451.

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084-1097.

Busby, J., White, K., and Smith, T. G. (2010). Mapping climate change and security in North Africa. *The German Marshall Fund of the United States: Climate & Energy Policy Paper Series.*

Castillo-Melendez, M., Yawno, T., Allison, B. J., Jenkin, G., Wallace, E. M., and Miller, S. L. (2015). Cerebrovascular adaptations to chronic hypoxia in the growth restricted lamb. *Int. J. Dev. Neurosci.* 45, 55-65.

Chen, C., Liu, C., Xiong, X., Fang, S., Yang, H., Zhang, Z., et al. (2018). Copy number variation in the MSRB3 gene enlarges porcine ear size through a mechanism involving miR-584-5p. *Genet. Sel. Evol.* 50, 72.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156-2158.

Dickerhoof, E. (2014). Analysis of Hypoxia Response Mechanisms in *Histoplasma capsulatum*. Honors Research Thesis. The Ohio State University.

Economides, K. D., Zeltser, L., and Capecchi, M. R. (2003). *Hoxb13* mutations cause overgrowth of caudal spinal cord and tail vertebrae. *Dev. Biol.* 256, 317–330.

El-Hag, F., Fadlalla, B., and Mukhtar, H. (2001). Some production characteristics of Sudan desert sheep under range conditions in North Kordofan, Sudan. *Trop. Anim. Health Prod.* 33, 229-239.

Endres, B. T., Priestley, J. R., Palygin, O., Flister, M. J., Hoffman, M. J., Weinberg, B. D., et al. (2014). Mutation of Plekha7 attenuates salt-sensitive hypertension in the rat. *Proc. Natl. Acad. Sci. USA* 111, 12817-12822.

Fang, W., and Latgé, J.-P. (2018). Microbe Profile: Aspergillus fumigatus: a saprotrophic and opportunistic fungal pathogen. *Microbiology* 164, 1009–1011.

FAO (2015). The second report on the state of the world's animal genetic resources for food and agriculture. In "FAO Commission on Genetic Resources for Food and Agriculture Assessments" (B. D. S. D. Pilling, ed.), pp. 606, Rome.

FAO (2016). AQUASTAT website: "Regional report - Libya" Food and Agriculture Organization of the United Nations (FAO).

Fariello, M. I., Boitard, S., Naya, H., San Cristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193, 929-941.

Friedman, J. M., Arbiser, J., Epstein, J. A., Gutmann, D. H., Huot, S. J., Lin, A. E., et al. (2002). Cardiovascular disease in neurofibromatosis 1: Report of the NF1 Cardiovascular Task Force. *Genet. Med.* 4, 105–111.

Frisancho, A. R., and Baker, P. T. (1970). Altitude and growth: a study of the patterns of physical growth of a high altitude Peru*via* n Quechua population. *Am. J. Phys. Anthropol.* 32, 279-292.

Fujita, T. (2008). Aldosterone in salt-sensitive hypertension and metabolic syndrome. *J. Mol. Med.* 86, 729–734.

Gizaw, S., Van Arendonk, J. A., Komen, H., Windig, J., and Hanotte, O. (2007). Population structure, genetic variation and morphological diversity in indigenous sheep of Ethiopia. *Anim. Genet*. 38, 621-628.

Gomez-Sanchez, E. P., Gomez-Sanchez, C. M., Plonczynski, M., and Gomez-Sanchez, C. E. (2010). Aldosterone synthesis in the brain contributes to Dahl salt-sensitive rat hypertension. *J. Exp. Physiol.* 95, 120-130.

González-Castillo, C., Muñoz-Ortiz, E., Guzmán-Brambila, C., Rojas-Mayorquín, A. E., Beltran-Parrazal, L., Ortuño-Sahagún, D., et al. (2018). Differential Expression of Ion Channels in Adult and Neonatal Rat Ventral Respiratory Column. *J. Mol. Neurosci.* 64, 51-61.

Greco, T. L., Takada, S., Newhouse, M. M., McMahon, J. A., McMahon, A. P., and Camper, S. A. (1996). Analysis of the vestigial tail mutation demonstrates that Wnt-3a gene dosage regulates mouse axial development. *Genes Develop*. 10, 313-324.

Guo, N., Zhang, N., Yan, L., Cao, X., Lv, F., Wang, J., et al. (2017). Down-regulation of single-stranded DNA-binding protein 1 expression induced by HCMV infection promotes lipid accumulation in cells. *Braz. J. Med. Biol. Res.* 50, e6389,

Harvey, S., Phillips, J., Rees, A., and Hall, T. (1984). Stress and adrenal function. *J. Exp. Zool.* 232, 633-645.

Hashimi, H., McDonald, L., Stříbrná, E., and Lukeš, J. (2013). Trypanosome Letm1 protein is essential for mitochondrial potassium homeostasis. *J. Biol. Chem.* 288, 26914-26925.

He, Y., Luo, M., Yi, M., Sheng, Y., Cheng, Y., Zhou, R., et al. (2013). Identification of a testis-enriched heat shock protein and fourteen members of hsp70 family in the swamp eel. *PLoS ONE* 8, e65269.

Heischmann, S., Dzieciatkowska, M., Hansen, K., Leibfritz, D., and Christians, U. (2017). The Immunosuppressant Mycophenolic Acid Alters Nucleotide and Lipid Metabolism in an Intestinal Cell Model. *Sci. Rep.* 7, 45088.

Hoefs, M. (2000). The thermoregulatory potential of Ovis horn cores. *Can. J. Zool.* 78, 1419-1426.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protoc.* 4, 44-57.

Johnston, S. E., McEwan, J. C., Pickering, N. K., Kijas, J. W., Beraldi, D., Pilkington, J. G., et al. (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, 2555–2566.

Kardos, M., Luikart, G., Bunch, R., Dewey, S., Edwards, W., McWilliam, S., et al. (2015). Whole-genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. *Mol. Ecol.* 24, 5616-5632.

Kashan, N., Azar, G. M., Afzalzadeh, A., and Salehi, A. (2005). Growth performance and carcass quality of fattening lambs from fat-tailed and tailed sheep breeds. *Small Ruminant Res.* 60, 267-271.

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M., et al. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 10, e1001258.

Kong, Z., Zhou, C., Chen, L., Ren, A., Zhang, D., Basang, Z., et al. (2019a). Multi-Omics Analysis Reveals Up-Regulation of APR Signaling, LXR/RXR and FXR/RXR Activation Pathways in Holstein Dairy Cows Exposed to High-Altitude Hypoxia. *Animals* 9, 406.

Kong, Z., Zhou, C., Li, B., Jiao, J., Chen, L., Ren, A., et al. (2019b). Integrative plasma proteomic and microRNA analysis of Jersey cattle in response to high-altitude hypoxia. *J. Dairy Sci.* 102, 4606-4618.

Kraja, A. T., Lawson, H. A., Arnett, D. K., Borecki, I. B., Broeckel, U., de las Fuentes, L., et al. (2012). Obesity–insulin targeted genes in the 3p26-25 region in human studies and LG/J and SM/J mice. *Metabolism* 61, 1129-1141.

Lewis, E. B. (2004). A Gene Complex Controlling Segmentation in Drosophila. In "Genes, Development and Cancer: The Life and Work of Edward B. Lewis" (H. D. Lipshitz, ed.), pp. 205-217. Springer US, Boston, MA.

Li, M., Tian, S., Yeung, C. K. L., Meng, X., Tang, Q., Niu, L., et al. (2014). Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Sci. Rep.* 4, 4678.

Lin, Y., Lai, X., Chen, B., Xu, Y., Huang, B., Chen, Z., et al. (2011). Genetic variations in CYP17A1, CACNB2 and PLEKHA7 are associated with blood pressure and/or hypertension in She ethnic minority of China. *Atherosclerosis* 219, 709-714.

Liška, F., Gosele, C., Popova, E., Chylíková, B., Křenová, D., Křen, V., et al. (2013). Overexpression of full-length centrobin rescues limb malformation but not male fertility of the hypodactylous (hd) rats. *PloS ONE* 8, e60859.

Lorenzo, F. R., Huff, C., Myllymäki, M., Olenchock, B., Swierczek, S., Tashi, T., et al. (2014). A genetic mechanism for Tibetan high-altitude adaptation. *Nature Genet.* 46, 951–956.

Lv, F.-H., Agha, S., Kantanen, J., Colli, L., Stucki, S., Kijas, J. W., et al. (2014). Adaptations to climate-mediated selective pressures in sheep. *Mol. Biol. Evol.* 31, 3324-3343.

Lv, F.-H., Peng, W.-F., Yang, J., Zhao, Y.-X., Li, W.-R., Liu, M.-J., et al. (2015). Mitogenomic meta-analysis identifies two phases of migration in the history of eastern Eurasian sheep. *Mol. Biol. Evol.* 32, 2515-2533.

Ma, Y. H., Rao, S. Q., Lu, S. J., Hou, G. Y., Guan, W. J., and Li, H. B. (2006). Phylogeography and origin of sheep breeds in Northern China. *Conserv. Genet.* 7, 117-127.

Maclean, C. A., Chue Hong, N. P., and Prendergast, J. G. (2015). hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Mol. Biol. Evol.* 32, 3027-3029.

Malecki, M., Viegas, S. C., Carneiro, T., Golik, P., Dressaire, C., Ferreira, M. G., et al. (2013). The exoribonuclease Dis3L2 defines a novel eukaryotic RNA degradation pathway. *EMBO J*. 32, 1842-1854.

Mallo, M., Wellik, D. M., and Deschamps, J. (2010). Hox genes and regional patterning of the vertebrate body plan. *Dev. Biol*. 344, 7–15.

Martin, D. L., and Tarleton, R. L. (2005). Antigen-specific T cells maintain an effector memory phenotype during persistent *Trypanosoma cruzi* infection. *J. Immunol.* 174, 1594-1601.

Mason, I. L. (1991). Classification and distribution of sheep breeds. In "Genetic resources of pig, sheep and goat." (K. Maijala, ed.), pp. 179-194. Elsevier, Amsterdam.

Mastrangelo, S., Moioli, B., Ahbara, A., Latairish, S., Portolano, B., Pilla, F., et al. (2018). Genome-wide scan of fat-tail sheep identifies signals of selection for fat deposition and adaptation. *Anim. Prod. Sci.* 59, 835-848.

Mattson, D. L., Dwinell, M. R., Greene, A. S., Kwitek, A. E., Roman, R. J., Jacob, H. J., et al. (2008). Chromosome substitution reveals the genetic basis of Dahl salt-sensitive hypertension and renal disease. *Am. J. Physiol. Renal. Physiol.* 295, F837-F842.

McLaughlin-Drubin, M. E., Crum, C. P., and Münger, K. (2011). Human papillomavirus E7 oncoprotein induces KDM6A and KDM6B histone demethylase expression and causes epigenetic reprogramming. *Proc. Natl. Acad. Sci. USA* 108, 2130-2135.

Megens, H.-J., Crooijmans, R. P., San Cristobal, M., Hui, X., Li, N., and Groenen, M. A. (2008). Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genet. Sel. Evol.* 40, 103–128.

Mehta, V., Abi-Nader, K. N., Peebles, D. M., Benjamin, E., Wigley, V., Torondel, B., et al. (2011). Long-term increase in uterine blood flow is achieved by local overexpression of VEGF-A165 in the uterine arteries of pregnant sheep. *Gene Ther.* 19, 925-935.

Moioli, B., Pilla, F., and Ciani, E. (2015). Signatures of selection identify loci associated with fat tail in sheep. *J. Anim. Sci.* 93, 4660–4669.

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

Mufarrih, M. E. (1991). Sudan desert sheep: Their origin, ecology and production potential. *World Anim. Rev.* 66, 23-31.

Mwacharo, J. M., Kim, E.-S., Elbeltagy, A. R., Aboul-Naga, A. M., Rischkowsky, B. A., and Rothschild, M. F. (2017). Genomic footprints of dryland stress adaptation in Egyptian fat-tail sheep and their divergence from East Africa and western Asia cohorts. *Sci. Rep.* 7, 17647.

Nagayach, R., Gupta, U., and Prakash, A. (2017). Expression profiling of hsp70 gene during different seasons in goats (Capra hircus) under sub-tropical humid climatic conditions. *Small Ruminant Res.* 147, 41-47.

Nejati-Javaremi, A., Izadi, F., Rahmati, G., and Moradi, M. (2007). Selection in fat-tailed sheep based on two traits of fat-tail and body weight versus single-trait total body weight. *Int. J Agri. Biol*. 9, 645–648.

Nowikovsky, K., Pozzan, T., Rizzuto, R., Scorrano, L., Bernardi, P. (2012). The pathophysiology of LETM1. *J. Gen. Physiol.* 139:445–54

Nowikovsky, K., and Bernardi, P. (2014). LETM1 in mitochondrial cation transport. *Front. Physiol.* 5, 83.

Pennarossa, G., Maffei, S., Rahman, M. M., Berruti, G., Brevini, T. A., and Gandolfi, F. (2012). Characterization of the constitutive pig ovary heat shock chaperone machinery and its response to acute thermal stress or to seasonal variations. *Biol. Reprod.* 87, 119, 1-9.

Pourlis, A. F. (2011). A review of morphological characteristics relating to the production and reproduction of fat-tailed sheep breeds. *Trop. Anim. Health. Prod*. 43, 1267-1287.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.

Ruiz-Larrañaga, O., Langa, J., Rendo, F., Manzano, C., Iriondo, M., and Estonba, A. (2018). Genomic selection signatures in sheep from the Western Pyrenees. *Genet. Sel. Evol.* 50, 9.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918.

Scaal, M. (2016). Early development of the vertebral column. *Semin. Cell Dev. Biol*. 49, 83–91.

Shen, H., Yao, Z., Zhao, W., Zhang, Y., Yao, C., and Tong, C. (2019). miR-21 enhances the protective effect of loperamide on rat cardiomyocytes against hypoxia/reoxygenation, reactive oxygen species production and apoptosis *via* regulating Akap8 and Bard1 expression. *Exp. Ther. Med.* 17, 1312-1320.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., et al. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science* 329, 72-75.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., et al. (2009). BioMart–biological queries made easy. *BMC genomic.* 10, 22.

Smith, J. (1999). T-box genes: what they do and how they do it. *Trends. Genet.* 15, 154-158.

Taal, H. R., van den Hil, L. C., Hofman, A., van der Heijden, A. J., and Jaddoe, V. W. (2012). Genetic variants associated with adult blood pressure and kidney function do not affect fetal kidney volume. The Generation R Study. *Early Hum. Dev.* 88, 711-716.

Taberlet, P., Coissac, E., Pansu, J., and Pompanon, F. (2011). Conservation genetics of cattle, sheep, and goats. *C. R. Biol.* 334, 247-254.

Takeda, Y., Demura, M., Wang, F., Karashima, S., Yoneda, T., Kometani, M., et al. (2018). Epigenetic regulation of aldosterone synthase gene by sodium and angiotensin II. *J. Am. Heart Assoc.* 7, e008281.

Terenina, E. E., Mormede, P., Cavigelli, S., Zhao, W., Parks, C., Lu, L., et al. (2019). Genetic factors mediate the impact of chronic stress and subsequent response to novel acute stress. *Front. Neurosci.* 13, 438.

Towler, B. P., Jones, C. I., Viegas, S. C., Apura, P., Waldron, J. A., Smalley, S. K., et al. (2015). The 3'-5' exoribonuclease Dis3 regulates the expression of specific microRNAs in *Drosophila* wing imaginal discs. *RNA Biol.* 12, 728-741.

Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Rosengren Pielberg, G., and Sigurdsson, S. (2011). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.* 7, e1002316.

Wang, X., Zhou, G., Xu, X., Geng, R., Zhou, J., Yang, Y., et al. (2014). Transcriptome profile analysis of adipose tissues from fat and short-tailed sheep. *Gene* 549, 252-257.

Wei, C., Wang, H., Liu, G., Wu, M., Cao, J., and Liu, Z. (2015). Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. *BMC Genom.* 16:194.

Wei, C., Wang, H., Liu, G., Zhao, F., Kijas, J. W., Ma, Y., et al. (2016). Genome-wide analysis reveals adaptation to high altitudes in Tibetan sheep. *Sci. Rep.* 6, 26770.

Willger, S. D., Cornish, E. J., Chung, D., Fleming, B. A., Lehmann, M. M., Puttikamonkul, S., et al. (2012). Dsc orthologs are required for hypoxia adaptation, triazole drug responses, and fungal virulence in Aspergillus fumigatus. *Eukaryot. Cell.* 11, 1557-1567.

Wilm, B., Dahl, E., Peters, H., Balling, R., and Imai, K. (1998). Targeted disruption of Pax1 defines its null phenotype and proves haploinsufficiency. *Proc. Natl. Acad. Sci. USA* 95, 8692-8697.

Wu, M., Wallace, M. R., and Muir, D. (2006). Nf1 haploinsufficiency augments angiogenesis. *Oncogene* 25, 2297-2303.

Xiang, K., Peng, Y., Yang, Z., Zhang, X., Cui, C., Zhang, H., et al. (2013). Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol. Biol. Evol.* 30, 1889-1898.

Xiao, W. J., He, J. W., Zhang, H., Hu, W. W., Gu, J. M., Yue, H., et al. (2010). ALOX12 polymorphisms are associated with fat mass but not peak bone mineral density in Chinese nuclear families. *Int. J. Obes.* 35, 378–386.

Xu, X.-H., Huang, X.-W., Qun, L., Li, Y.-N., Wang, Y., Liu, C., et al. (2014). Two functional loci in the promoter of EPAS1 gene involved in high-altitude adaptation of Tibetans. *Sci. Rep.* 4, 7465.

Yan, Y., Wang, M., Lemon, W., and You, M. (2004). Single nucleotide polymorphism (SNP) analysis of mouse quantitative trait loci for identification of candidate genes. J. *Med. Genet.* 41, e111.

Yang, G., Ren, J., Zhang, Z., and Huang, L. (2009). Genetic evidence for the introgression of Western NR6A1 haplotype into Chinese Licha breed associated with increased vertebral number. *Anim. Genet.* 40, 247-250.

Yang, J., Li, W.-R., Lv, F.-H., He, S.-G., Tian, S.-L., Peng, W.-F., et al. (2016). Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Mol. Biol. Evol.* 33, 2576-2592.

Ye, L., Fan, Z., Yu, B., Chang, J., Al Hezaimi, K., Zhou, X., et al. (2012). Histone demethylases KDM4B and KDM6B promotes osteogenic differentiation of human MSCs. *Cell Stem Cell* 11, 50-61.

Yuan, Z., Liu, E., Liu, Z., Kijas, J. W., Zhu, C., Hu, S., et al. (2017). Selection signature analysis reveals genes associated with tail type in Chinese indigenous sheep. *Anim. Genet*. 48, 55-66.

Zeder, M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci. USA* 105, 11597-604.

Zeleriakoval, M., Purcz, P., and Hlavata, I. (2012). Comparison of precipitation trends in Libya and Slovakia. In "River Basin Management VII" (C. A. Brebbia, ed.), pp. 365. WIT Press, Southampton, UK.

Zhang, Z., Sun, Y., Du, W., He, S., Liu, M., and Tian, C. (2017). Effects of vertebral number variations on carcass traits and genotyping of Vertnin candidate gene in Kazakh sheep. *Asian-Australasia. J. Anim. Sci.* 30, 1234-1238.

Zhi, D., Da, L., Liu, M., Cheng, C., Zhang, Wang, Y., et al. (2018). Whole-genome sequencing of Hulunbuir short-tailed sheep for identifying candidate genes related to the short-tail phenotype. *Gene Genom. Genet*. 8, 377-383.

Zhu, C., Fan, H., Yuan, Z., Hu, S., Ma, X., Xuan, J., et al. (2016). Genome-wide detection of CNVs in Chinese indigenous sheep with different types of tails using ovine high-density 600K SNP arrays. *Sci. Rep.* 6:27822.

Chapter 5

General conclusions and future directions

General conclusions

The diversity of a livestock species is reflected by its morphological appearance, its production traits characteristic and its adaptability to the challenges of different environmental conditions. Hence, the phenotypic variation within and among sheep populations is explained by differences in ancestral origins, migration history and local natural and human-mediated selection. Because of their relatively small size, obedient behaviour and their high ability to cope with different ranges of environments, sheep were amongst the most suitable livestock species to accompany human migrations and movement, including towards the African continent, since Neolithic time. Today, sheep play a crucial role in the socio-economic development of African countries representing a valuable source of food for the growing populations of the African continent.

The possibility to detect regions and loci under selection at genome level will lead to a better understanding of the selection mechanisms, which have shaped the within and among breeds/populations sheep diversity. It is now possible, thanks to the availability of whole-genome sequences along with huge advances in statistical approaches, to detect selective sweep in the genome. Such development will help to understand the genetic control of productivity traits in relation to the local environments.

Up to now, African sheep were poorly studied at the genome-wide level. Therefore, this thesis aimed to provide a comprehensive view of the diversity, population structure and adaptation of African indigenous sheep at genome-wide level building-up on prior archaeological knowledge on the history of arrival and dispersion of African sheep into the continent. We, therefore, included populations of different tail types from Ethiopia (East African origin) and North Africa (Libyan long fat-tail and thin tail sheep populations from Sudan).

Our first objective was to assess the level of diversity of the studied populations within the country and at global levels using the Illumina Ovine 50 K SNP BeadChip assay. With the same dataset, we identified a first set of candidate genome regions putatively linked to a range of environmental adaptation. Here, our focus was mainly Ethiopia considering the large diversity of distinct agro-ecological zones found in the country as well as of human ethnic farmer diversity of ancient origin. Results from this dataset indicate two main genetic backgrounds and support two distinct genetic histories for the African fat-tail sheep, while within Ethiopian sheep; our results show that the short fat-tail sheep do not represent a

monophyletic group. Four genetic backgrounds are present in Ethiopian indigenous sheep but at different proportions among the fat-rump and the long fat-tail sheep from western and southern Ethiopia.

Surprisingly, the Ethiopian fat-rump sheep share a genetic background with Sudanese thin-tail sheep. However, more detailed studies are in demand to unravel factors surrounding this pattern of admixture. Our findings of this stage of analysis suggest that Ethiopian fat-tail sheep represent a uniquely admixed but distinct genepool that represents an important resource for understanding the genetic control of skeletal growth, fat metabolism and associated physiological processes.

The second goal of the study was to expand the knowledge obtained from the Illumina Ovine 50 K SNP BeadChip assay to the entire genome using whole-genome sequences information, paving the way to the identification of the causative mutation. The results obtained from these genome sequences analysis are in close agreement with those obtained using the 50K Bead-Chip data. Besides, the results of the genome-wide sequence are in line with the archaeological history of African sheep in relation to their proposed entry points into and migration events across the continent. However, further genome analysis studies including thin-tail sheep from Ethiopia and both thin and/or fat-tail sheep from West, North and South Africa will provide a more comprehensive view on the genomic diversity of African sheep at a continental level.

Our third objective using both 50K Bead-Chip and whole-genome sequences data was to identify population or group of population-specific candidate signatures of selection regions and genes by contrasting the genomes of sheep populations after genome-wide clustering of their diversity. Using this approach, we have been able to identify candidate regions associated with adaptation to environmental challenges and with productive traits (tail length and fat deposition). Also, it seems that they are some commonality of environmental challenges across sheep population occupying tropical regions, represented here by the Ethiopian and Sudanese sheep populations, with candidate signatures of selection regions linked to trypanosomiasis infection in both groups. We also identified population specific adaptability with a candidate selected region associated with adaptation to salinity in sheep reared at sea level in the northern part of the continent (Libya). Our findings also show signals of adaptation at the genome level in Ethiopian fat-tail sheep under humid high-altitude environments, with candidate regions under selection highly enriched in genes included in the hypoxia GO pathways. Among these, the *EGLN1* gene has been for the first time identified in sheep as being of relevance for

adaptation to hypoxia. Also, our findings suggest some compromise between sheep adaptability and tail morphology. In particular, while genome variation sub-divide sheep populations in tail-based groups, these subgroups show different adaptability to environmental challenges. For the tail morphology, we identified candidate genes associated with fat deposition and tail length, which together form a "genome network" of signature of selection controlling the outcome of the tail morphology. Among these, *HOXB13* is most likely the main gene controlling tail length in African sheep. Beside environmental adaptive consideration, fat deposition and tail length are of high importance in regards to human selection contributing to meat production and carcass traits. More specifically, while fat deposition is linked to meat quality, meat production traits are linked to the number of vertebrae and their length.

Overall, our results provide novel insights into African sheep diversity at genome level, their genome adaptations to extreme environments, and into the genetic control of sheep tail morphology. They provide a valuable resource of new information for future research on livestock breeding in developing countries in response to the global challenge of climate change.

Future directions

Although we were able to support the archaeological history of sheep dispersal at the genome level, future genome analysis studies including thin-tail sheep from Ethiopia and both thin and/or fat-tail sheep from West, North and South Africa will provide a more comprehensive view on the genomic diversity of African sheep at the continental level. In addition, our results support genepool erosion in some populations, Libyan Barbary sheep in particular, with the need to assess across all the country the genetic integrity of the breed in order to establish an effective plan to conserve its genetic identity. Moreover, functional genomic and association studies, supported with precise phenotypic measurements and large sample sizes, are required to identify the causative mutations responsible for the traits of interest. Thus, the signatures of selection identified here, especially those related to tail morphology, need to be functionally validated in order to detect the causal variants within genes spanned by these genomic regions.

To the best of our knowledge, this is the first study that has assessed the genome profile of African sheep using both SNP chips genotyping and whole-genome sequences information in the same populations. We believe that this study represents a milestone in support of breeding programs aiming to conserve and utilise African sheep genetic resources. In particular, our

selected candidate regions could be integrated into a marker-assisted selection breeding program aiming to improve sustainable productivity under different environments contributing to the challenge of food security in African developing countries.

Appendices

Appendix Table 5.1 Description of the worldwide breeds used in the study

Origin	Breed	N	Abbr.	Tail type	Coat type	Source*
Italy	Barbaresca	13	BRS	Fat-tailed	Wool	Mastrangelo et al., 2017
Libya	Libyan Barbary	24	LBR	Fat-tailed	Coarse-wooled	Mastrangelo et al., 2018
	Berber	6	ABR	Fat-tailed	Wool	
Algeria	Sidaoun	6	SID	Long-tailed	Hair	Gaouar et al., 2017
	Barbarine	5	ABN	Fat-tailed	Coarse-wooled	
Egypt	Ossimi	8	OSM	Fat-tailed	Coarse-wooled	Mastrangelo et al., 2018
	Barki	13	EBI	Fat-tailed	Coarse-wooled	
Israel	Local Awassi	13	AWS	Fat-tailed	Coarse-wooled	
UK	Dorset Horn	13	DSH	Thin-tailed	Fine wool	
UK	Soay	12	SOA	Short tail	Wool	
Iceland	Icelandic	13	ICE	Fluke-shaped tail	Fine wool	
Cyprus	Cyprus FatTail	14	CYFT	Fat-tailed	Coarse-wooled	Kijas et al., 2012
	African Dorper	13	ADP	Short fat-tail	Coarse-wooled	
South Africa	Namaqua Afrikaner	13	NQA	Long fat-tailed	Coarse-wooled	
	Ronderib Afrikaner	13	RDA	Long fat-tailed	Coarse-wooled	
Caribbean island	Barbados BlackBelly	13	BBB	Thin-tailed	Hair	
	Hu sheep	12	HUS	Short fat-tailed	Wool	
	Tong sheep	15	TON	Long fat-tailed	Wool	
China	Han sheep	15	LTH	Large-tailed	Wool	Yuan et al., 2017
	Lop sheep	13	LOP	Short fat-tailed	Wool	
	Tibetan (From Qinghai)	14	TIBQ	Thin-tailed	Wool	
	Tibetan (From Sichuan)	14	TIBS	Thin-tailed	Wool	
West Africa	Djallonke	10	WAD	Thin-tailed	Hair	Spangler et al., 2017
Saudi Arabia	Najdi	6	NJD	Fat-tailed	Coarse-wooled	Faisal Almathen and Fabio Pilla
Oman	Omani	10	OMN	Fat-tailed	Coarse-wooled	Mohammed Al Abri and Fabio Pilla
Total		301				

*Gaouar, S.B.S., Lafri, M., Djaout, A., El-Bouyahiaoui, R., Bouri, A., Bouchatal, A., Maftah, A., Ciani, E. and Da Silva, A.B., (2017). Genome-wide analysis highlights genetic dilution in Algerian sheep. *Heredity*, 118(3), p.293.

Kijas, J.W., Lenstra, J.A., Hayes, B., Boitard, S., Neto, L.R.P., San Cristobal, M., Servin, B., McCulloch, R., Whan, V., Gietzen, K. and Paiva, S., 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS biology*, *10*, e1001258.

Mastrangelo, S., Moioli, B., Ahbara, A., Latairish, S., Portolano, B., Pilla, F., et al. (2018). Genome-wide scan of fat-tail sheep identifies signals of selection for fat deposition and adaptation. *Anim. Prod. Sci.* 59, 835-848.

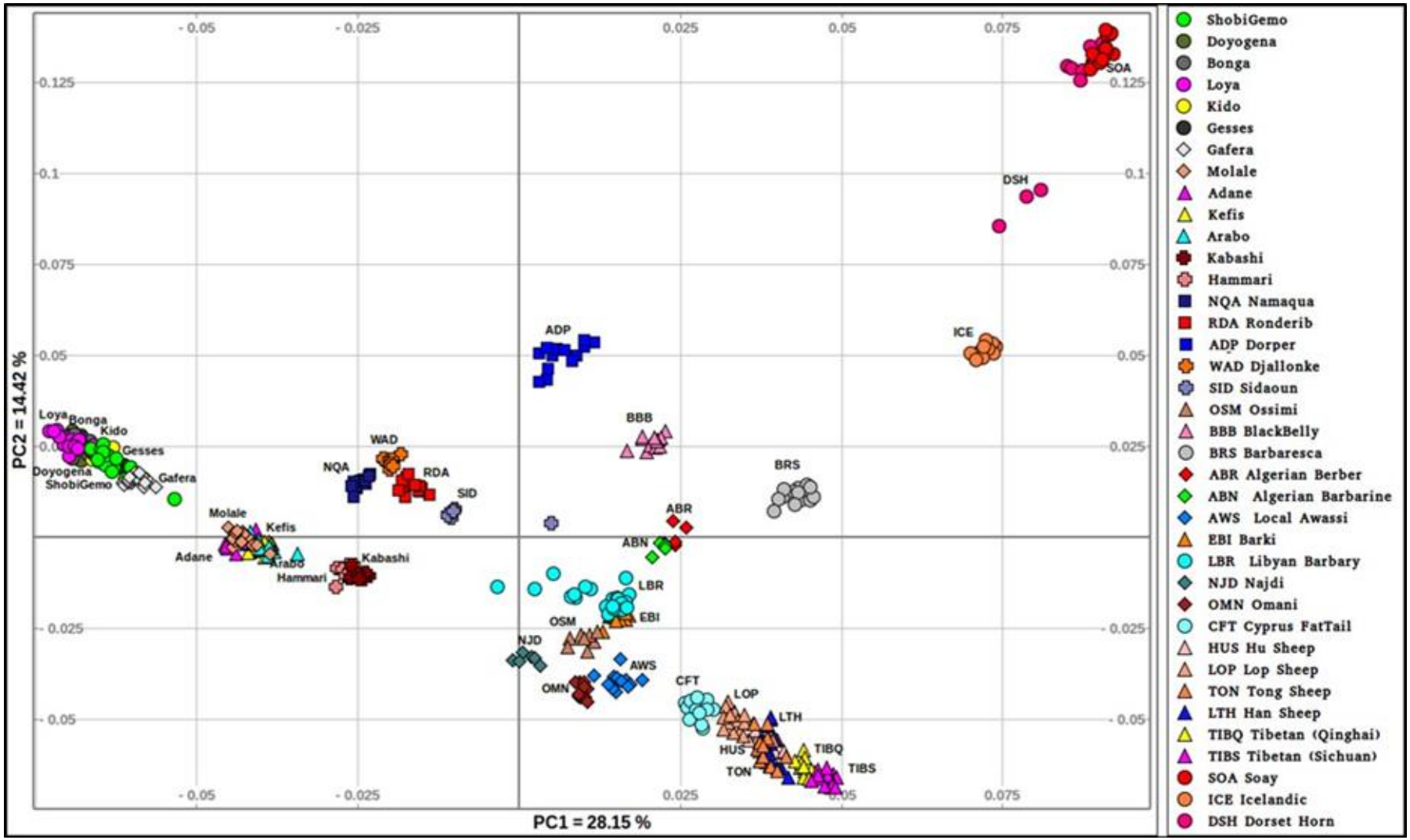
Mastrangelo, S., Portolano, B., Di Gerlando, R., Ciampolini, R., Tolone, M., Sardina, M. T. (2017). Genome-wide analysis in endangered populations: a case study in Barbaresca sheep. *Animal*. 11, 1107-1116.

Spangler, G.L., Rosen, B.D., Ilori, M.B., Hanotte, O., Kim, E.S., Sonstegard, T.S., Burke, J.M., Morgan, J.L., Notter, D.R. and Van Tassell, C.P. (2017). Whole genome structural analysis of Caribbean hair sheep reveals quantitative link to West African ancestry. *PloS One*, *12*(6), p.e0179021

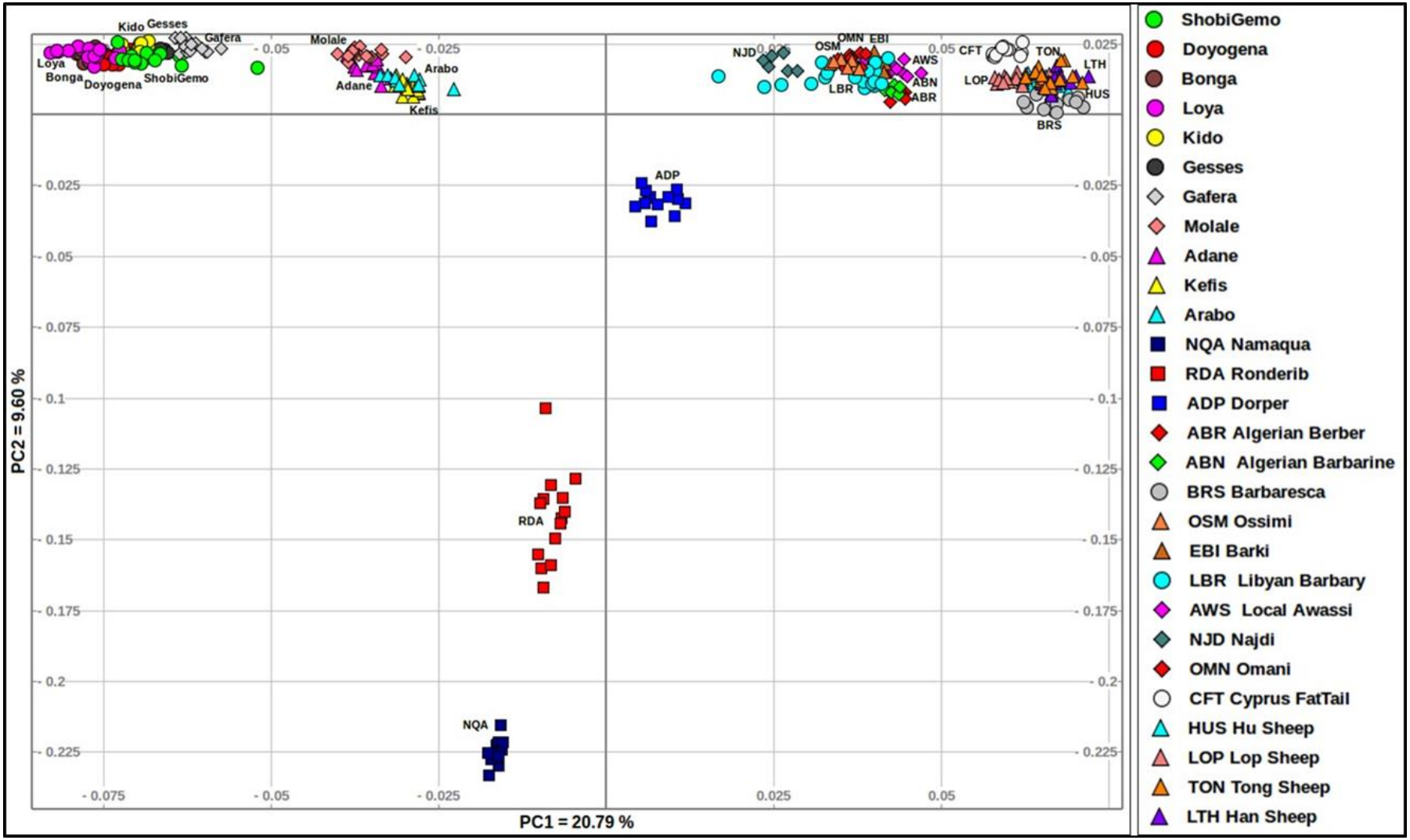
Yuan, Z., Liu, E., Liu, Z., Kijas, J.W., Zhu, C., Hu, S., et al. (2017). Selection signature analysis reveals genes associated with tail type in Chinese indigenous sheep. *Anim. Genet*. 48(1), 55-66.

Appendix Table 5.2 Measures of genetic diversity for each of the 13 populations analyzed.

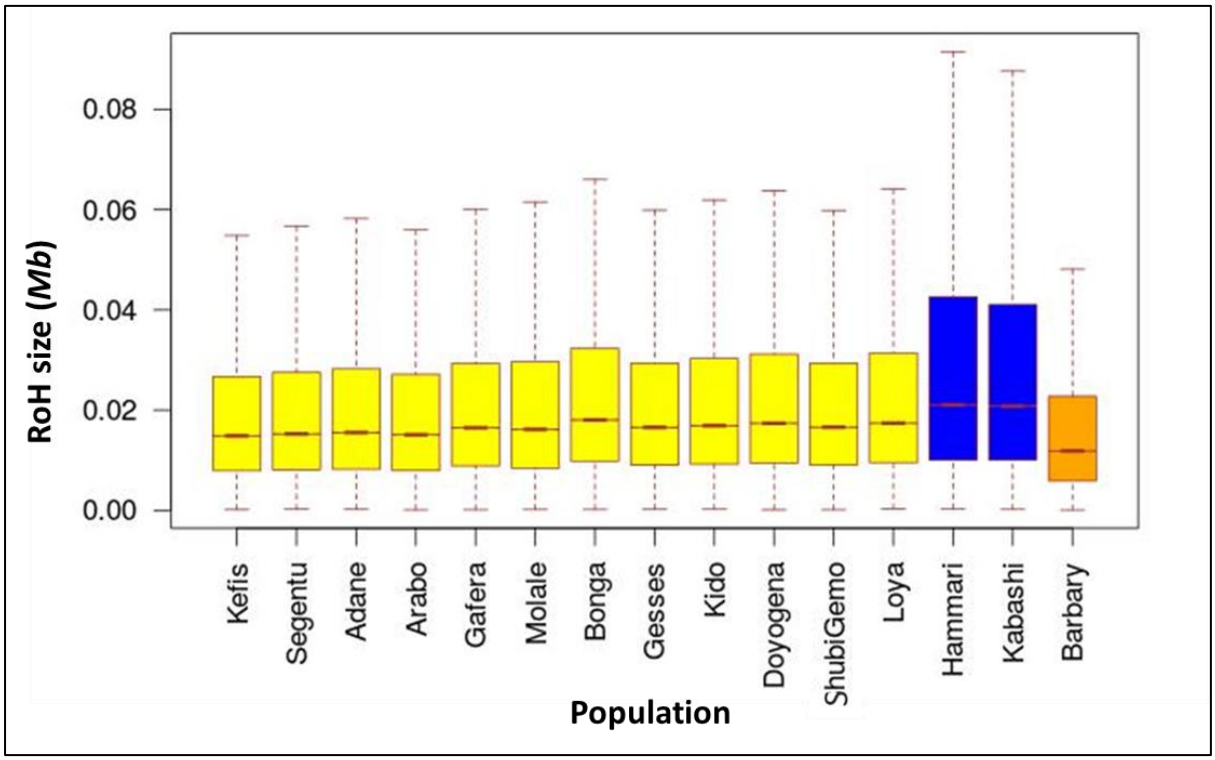
Breed	***N***	P_n **(%)**	H_e	H_o	***F***
Kefis	14	89.95	0.316	0.328	0.035
Adane	12	88.85	0.315	0.319	0.071
Arabo	10	88.69	0.317	0.334	0.050
Molale-Menz	15	90.29	0.316	0.319	0.055
Gafera-Washera	15	87.54	0.303	0.318	0.017
Bonga	9	79.59	0.277	0.293	0.038
Kido	10	82.18	0.290	0.310	0.038
Gesses	10	83.09	0.294	0.317	0.027
Doyogena	15	87.17	0.302	0.308	0.044
Loya	15	83.58	0.286	0.294	0.039
ShubiGemo	15	88.40	0.304	0.313	0.037
Hammari	11	89.93	0.319	0.332	0.038
Kabashi	9	88.64	0.319	0.328	0.025



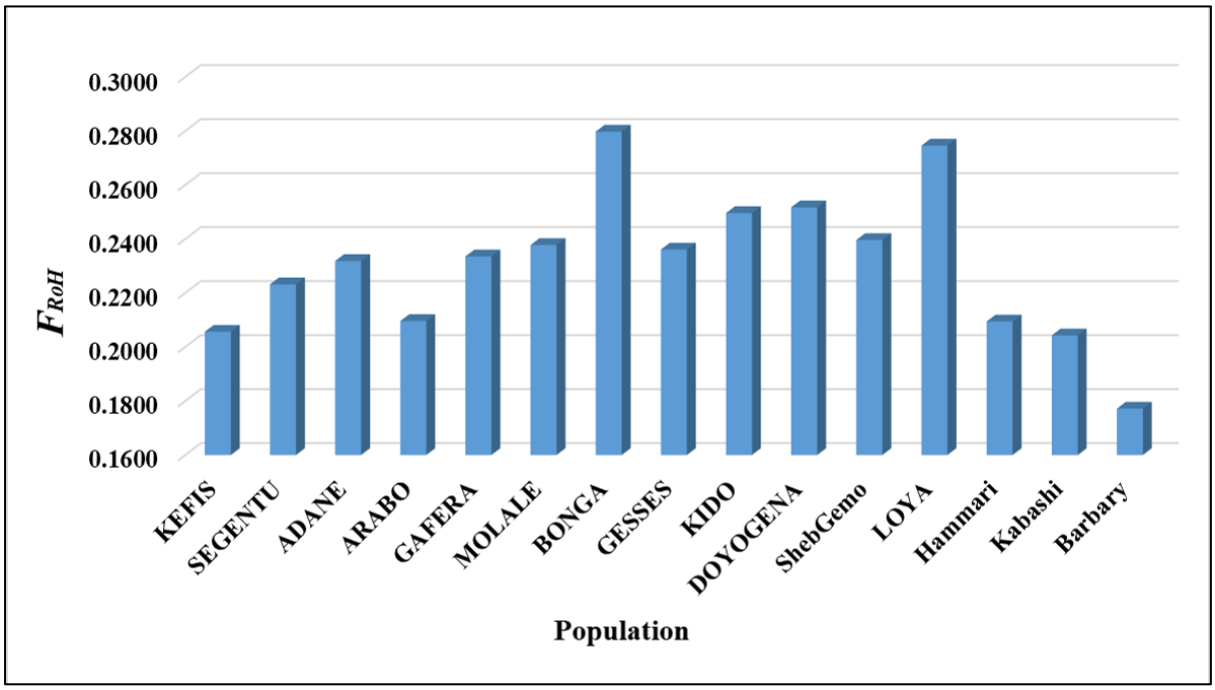
Appendix Figure 5.1 Genetic variation among the Ethiopian sheep populations in a global geographic context (all animals included for each population)



Appendix Figure 5.2 Distribution of the genetic variation among the worldwide fat-tail sheep (all animals included for each population)



Appendix Figure 5.3 Distribution of RoHs within range of 1.4 Mb for each sheep population



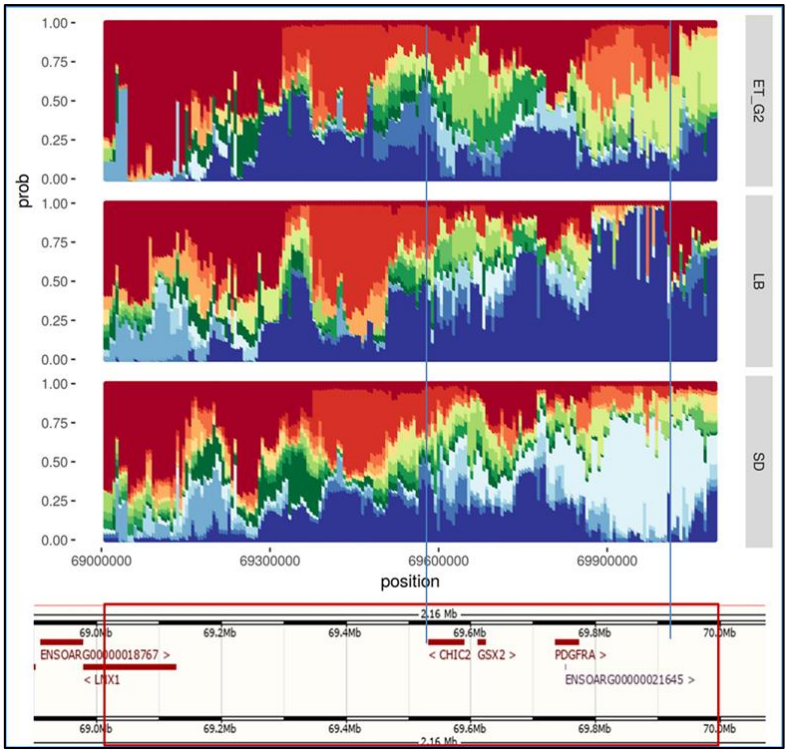
Appendix Figure 5.4 Distribution of the runs of homozygosity inbreeding coefficient (F_{ROH}) within each bree**d**

Appendix Table 5.3 Candidate regions and genes associated with environmental challenges identified by a combination of at least two methods/comparisons in the Ethiopian long fat-tail versus Ethiopian fat-rump, Libyan Long fat-tail and Sudanese thin-tail sheep.

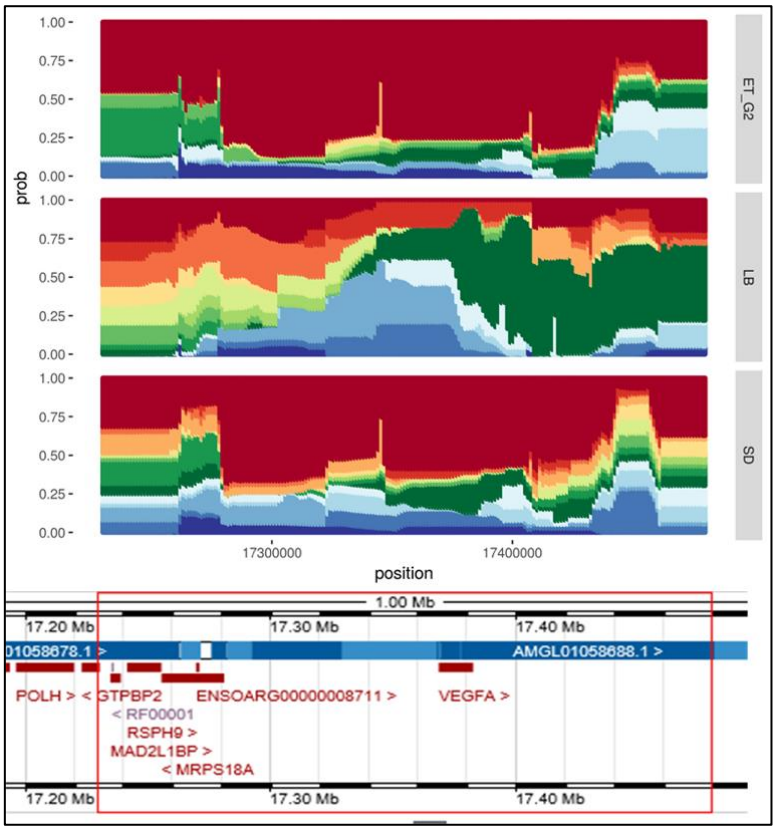
Chr	Region	Comparison/Method	Genes
1	68600001-68840000	ET_G2 **(*ZHp*)**, ET_G2 *vs* LB & SD **(F_{ST})**	*BTBD8, RF00026, C1orf146, GLMN, RPAP2*
	250500001-253740000	ET_G2 vs ET_G1 & SD **(F_{ST}),** ET_G1 vs ET_G2 (***XP-EHH***)	*STAG1, PCCB, MSL2, PPP2R3A, EPHB1, RF00026, KY, CEP63, ANAPC13, AMOTL2, RYK, SLCO2A1, RAB6B, SRPRB, TF*
	57960001-58090000	ET_G1 vs ET_G2 (F_{ST}, *XP-EHH*)	*ADGRL2*
	5930001-6120000	ET_G2 *vs* ET_G1 & LB **(*XP-EHH*)**	*Close to SH3BP4*
	109920001-110020000	ET_G1, ET_G2 & SD **(*ZHp*)**, ET_G2 *vs* LB **(F_{ST}, *XP-EHH*)**	*CD84*
	154650001-154790000	LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*ZNF654, C3orf38*
2	68440001-68570000	LB vs ET_G2 (F_{ST} , ***XP-EHH***)	*KANK1*
	115510001-115610000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*Spans MFF, TM4SF20 in cow*
	106240001-106390000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*GALNT7*
	129590001-129730000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*Upstream ZNF385B and downstream SESTD1*
	90750001-90950000	ET_G2 (***ZHp***), LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*Span BMPR2 gene in cow*
	232440001-233000000	ET_G2 & SD **(*ZHp*)**, ET_G2 *vs* LB & SD **(F_{ST} , *XP-EHH*)**	*PTMA, PDE6D, COPS7B, RF00091, DIS3L2*
3	4990001-5200000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*RAPGEF1, RF00026*
	10920001-11320000	ET_G2 **(*ZHp*)**, ET_G2 *vs* ET_G1 & LB **(F_{ST},)**	*GOLGA1, ARPC5L, WDR38, RF00026, OLFML2A, MIR181B2, oar-mir-181a-2, NR6A1, NR5A1, ADGRD2*
	20980001-21170000	ET_G2 *vs* ET_G1 & LB (***XP-EHH***)	*ENSOARG00000016157*
	10630001-10820000	ET_G2 *vs* LB & SD (***XP-EHH***)	*GAPVD1, HSPA5, RABEPK, PPP6C, SCAI*
	12160001-12370000	SD vs ET_G2 (F_{ST}, ***XP-EHH***)	*RF00402, CRB2*
	151770001-151960000	ET_G2 *vs* LB & SD (***XP-EHH***)	*Close to DYRK2*
	154100001-154390000	ET_G1 & LB vs ET_G2 (F_{ST}, ***XP-EHH***), SD vs ET_G2 (F_{ST})	*MSRB3*
	168750001-168900000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*FAM71C*
	26920001-27070000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*Close to OSR1*
4	27220001-27370000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*HDAC9*

	73080001-73320000	ET_G2 *vs* SD **(F_{ST}, *XP-EHH*)**	*Close to ZNF804B*
	96540001-96700000	ET_G1, LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*CHCHD3*
5	67200001-67380000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*Spans STAB2 in cow*
6	69010001-70090000	ET_G2, LB (***ZHp***), LB & SD vs ET_G2 (F_{ST}, ***XP-EHH***)	*LNX1, CHIC2, GSX2, PDGFRA*
	69920001-70030000	LB **(*ZHp*)**, ET_G2 *vs* SD **(F_{ST}, *XP-EHH*)**	*Close to KIT, PDGFRA*
7	55870001-56140000	LB (***ZHp***), SD vs ET_G2 (***XP-EHH***)	*DMXL2, GLDN, CYP19*
	96270001-96390000	LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*Close to GALC*
8	51490001-51760000	ET_G2 & SD (***ZHp***), ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*Close to TBX18*
	32040001-32230000	LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*LIN28B*
10	28560001-28820000	LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*PDS5B*
	29390001-29510000	LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*RXFP2*
11	10750001-10940000	SD (***ZHp***), SD vs ET_G2 (***XP-EHH***)	*INTS2, BRIP1*
	18190001-18520000	LB (***ZHp***), SD & LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*NF1, EVI2B, OMG*
	27010001-27180000	ET_G1, SD (***ZHp***), SD vs ET_G2 (***XP-EHH***)	*DNAH2, NAA38, KDM6B, TMEM88, CYB5D1, CHD3, RF00602, RNF227, KCNAB3, TRAPPC1, CNTROB*
	37200001-37430000	ET_G2 *vs* ET_G1 **(F_{ST})**, ET_G2 *vs* ET_G1 & LB **(*XP-EHH*)**	*CALCOCO2, TTLL6, HOXB13, RF02133, RF02132, MIR196A1, HOXB9, HOXB7*
	55080001-55210000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*RF00026, GRB2*
12	75850001-75990000	ET_G2 *vs* SD & LB **(F_{ST}, *XP-EHH*)**	*MIR181B1, oar-mir-181a-1*
13	42560001-42800000	ET_G1 (***ZHp***), LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*PAK4, NCCRP1, SYCN, LRFN1*
	38560001-38890000	ET_G2 (***ZHp***), LB & SD vs ET_G2 (F_{ST}, ***XP-EHH***)	*RIN2, NAA20, CRNKL1, CFAP61*
	69370001-69520000	ET_G2 (***ZHp***)	*PLCG1*
14	48070001-48190000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*PAK4, NCCRP1, SYCN, LRFN1*
15	35110001-35280000	LB vs ET_G2 (F_{ST}, ***XP-EHH***)	*PLEKHA7, C11orf58*
	45880001-46020000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*DNHD1, ARFIP2, TRIM3, HPX*
	75410001-75550000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*MYBPC3, SPI1, SLC39A13, PSMC3, RAPSN, CELF1*
	77710001-77820000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*SSRP1, P2RX3*
16	1300001-1430000	ET_G1 vs ET_G2 (F_{ST}, ***XP-EHH***)	*Close to SPDL1*

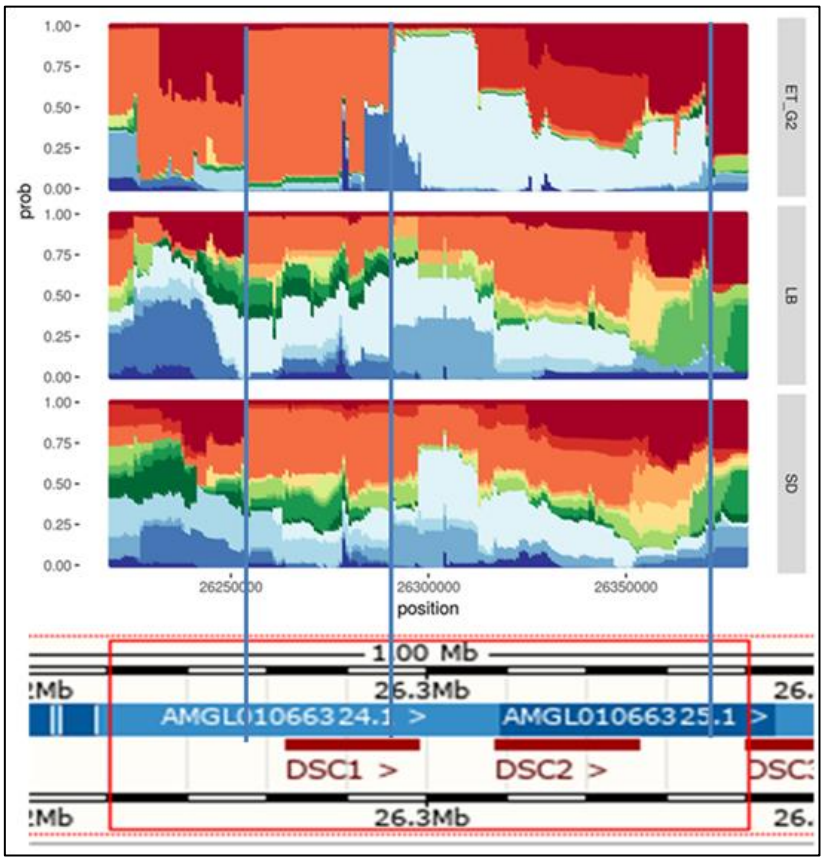
	31650001-31750000	LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*CCDC152*
	48890001-49010000	ET_G1 vs ET_G2 (ZFST, ***XP-EHH***)	*CDH10*
17	29820001-30220000	ET_G1 & SD vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*INTU*
	13200001-13340000	LB & SD vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*RF02271, HHIP*
	19260001-19370000	ET_G1 vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*SLC7A11*
18	5540001-5710000	LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*ADAMTS17*
	6010001-6170000	LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*MEF2A*
19	23200001-23420000	ET_G2 (***ZHp***), SD vs ET_G2 (***XP-EHH***)	*CNTN4*
	30490001-30640000	LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*ENSOARG00000009693*, (*forkhead box P1*)
	43420001-43530000	LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*SLMAP*
20	8730001-8860000	ET_G1 vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*C6orf106, RF00001*
	17230001-17480000	LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*MAD2L1BP, RF00001, RSPH9, MRPS18A, VEGFA*
	25440001-25630000	ET_G2 *vs* ET_G1 **(*F_{ST}*, *XP-EHH*)**	*DLA class II* (*histocompatibility antigen, DR-1 beta chain-lik*,), *SLA class II* (*histocompatibility antigen, DQ haplotype D alpha chain-like*)
21	39200001-39360000	ET_G1 vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*DDB1, VWCE, TKFC, CYB561A3, TMEM138, TMEM216, CPSF7, SDHAF2, PPP1R32*
	43040001-43230000	LB vs ET_G2 (***XP-EHH***)	*LTBP3, SSSCA1, FAM89B, EHBP1L1, KCNK7, MAP3K11, PCNX3, SIPA1, RELA, KAT5, RNASEH2C, AP5B1, OVOL1*
22	50510001-50760000	ET_G2 (***ZHp***), LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*CYP2E1, ECHS1, FUOM, PAOX*
23	26220001-26380000	LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*DSC1, DSC2*
	1030001-1220000	ET_G2 *vs* ET_G1 & LB (***XP-EHH***)	*ENSOARG00000026145, Close to SALL3*
24	24800001-24990000	ET_G2 *vs* ET_G1 (***XP-EHH***)	*NSMCE1, IL4R, IL21R*
	34640001-34770000	ET_G1 vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*CUX1, RF00001*
	9840001-9940000	SD vs ET_G2 (***F_{ST}***)	*CLEC16A, PRM3*
25	4060001-4180000	ET_G1 & ET_G2 (***ZHp***), LB vs ET_G2 (***F_{ST}***, ***XP-EHH***)	*GNPAT, EXOC8, SPRTN, EGLN1*



Appendix Figure 5.5 Haplotype structure at the *PDGFRA* candidate region in the contrasted groups of sheep groups.



Appendix Figure 5.6 Haplotype structure at the *VEGFA* candidate region on OAR20.



Appendix Figure 5.7 Haplotype structure at the *DSC1* and *DSC2* candidate region on OAR23.

Appendix Figure 5.8 Illustration of the phenotypic criteria (horns and tail) applied by the owners of sheep in Libya when selecting ram for reproduction.